



شبكة المعلومات الجامعية
التوثيق الإلكتروني والميكرو فيلم

بسم الله الرحمن الرحيم



HANAA ALY



شبكة المعلومات الجامعية
التوثيق الإلكتروني والميكرو فيلم



شبكة المعلومات الجامعية التوثيق الإلكتروني والميكرو فيلم



HANAA ALY



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



HANAA ALY

Ain Shams University

Faculty of computer & Information sciences

Computer science department



Applying Computational Intelligence Techniques For Diagnoses of Breast Cancer

A thesis submitted as partial fulfillment of the requirements for the degree
of Master of Science in Computer and Information sciences

Submitted by

Muhammed Abd-elnaby Sadek Muhammed

Under supervision of

Prof. Dr. Mohamed Ismail Roushdy

Professor of Computer Science,

Faculty of Computer& Information sciences, Ain-shams University

Dean of faculty of Computers & Information Technology, Future University in Egypt

Dr. Marco Alfonse

Lecturer

Computer Science Department,

Faculty of Computer& Information sciences, Ain-shams University

Acknowledgements

بسم الله الرحمن الرحيم

"سُبْحَانَكَ لَا عِلْمَ لَنَا إِلَّا مَا عَلَّمْتَنَا إِنَّكَ أَنْتَ الْعَلِيمُ الْحَكِيمُ"

صدق الله العظيم

In the name of the Allah, the most Gracious, the most Merciful.

“Exalted are You; we have no knowledge except what You have taught us. Verily, the all-knower, the all-wise”

I would like to express my sincere gratitude to my supervisor, Prof. Dr. Mohamed Roushdy for their presence, patience, dedication, and their powerful help. I again thank him for their availability and readiness to read. Much thanks for their quick responses whenever I consulted. I am deeply grateful for their willingness to work with me.

Special thanks for Dr. Marco Alfonse for his guidance

I am also grateful to Dr. Ahmed Mohammed Alenany and Dr. Ahmed Mohamed Helmy for their guidance, encouragement along the road.

I grateful to my sister, Hager Abd-elnaby student of medicine for their effort to form biological background and understand the medical concepts.

I obliged to my master, my father Abd-elnaby Sadek, my brother Ahmed and all of my family for everything they have done for me.

Dedication

Dedicated to my father, the soul of my mother, Allah's grant who give hope and life to our souls, my beloved daughter, my lovely wife, my brother, and my sisters.

(I am always proud of all of you, I hope that I have done made you proud.)

Publications

- “Classification of Breast Cancer Using Microarray Gene Expression Data: A Survey”, **Muhammed Abd-Elnaby**, Marco Alfonse, Mohamed Roushdy, Journal of Biomedical Informatics, 117 (2021) 103764. (Indexed in Scopus).
- "A Hybrid Mutual Information-LASSO-Genetic Algorithm Selection Approach for Classifying Breast Cancer", **Muhammed Abd-elnaby**, Marco Alfonse, and Mohamed Roushdy, “Digital Transformation Technology”, Proceedings of ITAF 2020 (Internet of Things: Applications & Future 2020), Springer, Part of Lecture Notes in Networks and Systems Volume 224, pp. 547-560. (Indexed in Scopus)
- “A Hybrid Mutual Information-LASSO-Particle swarm optimization Selection Approach for Classifying Breast Cancer”, **Muhammed Abd-elnaby**, Marco alfonse, Mohamed Ismail Roushdy, International Journal of Intelligent Computing and Information Sciences, in press. (Accepted) / May 2021

Abstract

Cancer, in particular breast cancer, is considered one of the most common causes of death worldwide, according to the world health organization. For this reason, extensive research efforts have been done in the area of accurate and early diagnosis of cancer in order to increase the likelihood of cure. Among the available tools for diagnosing cancer, microarray technology is commonly used in biological and medical science to study gene expression in cells. When a healthy tissue becomes cancerous, gene expression levels change. Tissues can be classified by looking for changes in gene expression. Although the huge number of features or genes in the microarray data may seem advantageous, many of these features are irrelevant or redundant, resulting in the deterioration of classification accuracy. To overcome this challenge, feature selection techniques are a mandatory preprocessing step before the classification process. Two hybrid feature selection approaches are proposed and applied on five breast cancer datasets, namely, Van't veer, Chin, Chowdary, Gravier and West. First approach combined mutual information, Least Absolute Shrinkage and Selection Operator and genetic algorithm (MI-LASSO-GA). The second approach combined mutual information, Least Absolute Shrinkage and Selection Operator and particle swarm optimization (MI-LASSO-PSO). The following classifiers with 5-fold cross validation are used to assess the proposed approaches: Support Vector Machine (SVM), k-Nearest Neighbor (KNN), Random Forest (RF), Logistic Regression (LR), and extreme gradient boosting (XGBOOST). Statistical measures of accuracy, precision, recall and f1 metrics are used for performance analysis. MI-LASSO-PSO could produce subsets with fewer features than subsets produced by MI-LASSO-GA for all datasets except Chin. The average performance of MI-LASSO-GA on Van't veer and west is better than MI-LASSO-PSO of 96.4% and 100% respectively. On the other hand, the average of the performance of MI-LASSO-PSO on Gravier, Chowdary and Chin of 91.5%, 99.2% and 96.8% respectively. The proposed approaches outperformed state of art techniques for all datasets.

Keywords

Microarray, Gene expression, Breast cancer, Classification.

Contents

Chapter 1 Introduction	1
1.1 From DNA TO Gene Expression	1
1.2 From Normal Cell to Cancer	3
1.2.1 Normal cell.....	3
1.2.2 Mutation and cancer cell	4
1.3 Breast cancer	7
1.4 Microarray	8
1.5 Problem Statement	9
1.6 Objectives of the Study	9
1.7 Organization	9
Chapter 2 Background.....	10
2.1 Normalization.....	10
2.2 Feature Selection Methods.....	10
2.1.1 Filter approaches.....	12
2.1.2 Wrapper approach.....	14
2.1.1 Embedded approach.....	18
2.1.2 Hybrid approach.....	20
2.3 Machine learning.....	20
2.3.1 Supervised approach	21
2.3.2 Unsupervised approach.....	26
2.4 Evaluation	27
2.4.1 Confusion Matrix	27
2.4.2 Precision.....	28
2.4.3 Recall	28

2.4.4 F1-score.....	28
2.5 Validation Methods.....	28
2.5.1 Leave One Out Cross Validation (LOOCV).....	28
2.5.2 K-Fold Cross-Validation.....	29
Chapter 3 Literature Review	31
3.1 Related work	31
Chapter 4 Methodology	48
4.1 Dataset Description.....	48
4.2 First approach (MI-LASSO-GA):.....	49
4.3 Second approach (MI-LASSO-PSO):.....	52
4.4 Hyperparameters Tuning	54
4.4.1 Support Vector Machine.....	54
4.4.2 Random Forest.....	54
4.4.3 k-Nearest Neighbor.....	55
4.4.4 Logistic Regression.....	55
4.4.5 eXtreme Gradient Boosting.....	56
Chapter 5 Result and Discussion	59
5.1 Approach 1: MI-LASSO-GA.....	59
5.2 Approach 2: MI-LASSO-PSO	63
5.3 Comparison between the two proposed approaches	67
5.4 the performance of the proposed approach vs state of art on van't veer dataset	71
5.5 the performance of the proposed approach vs state of art on Chowdary dataset....	72
5.6 the performance of the proposed approach vs state of art on Gravier dataset	73
5.7 The performance of the proposed approach vs state of art on Chin dataset	74
5.8 The performance of the proposed approach vs state of art on west dataset.....	74

Chapter 6	 Conclusion and future work.....	75
References	77	

List of figures

Figure 1.1 Illustration of DNA [1].....	1
Figure 1.2 An illustration showing the process of transcription[2]	2
Figure 1.3 an illustration showing the process of translation[3].....	3
Figure 1.4 Levels of structural organization of the human body	4
Figure 1.5 proto-oncogene vs oncogene [5]	5
Figure 1.6 The function of tumor suppressor in normal state vs when mutated.....	6
Figure 1.7 Estimated number of new cases in 2020, worldwide (Female).....	7
Figure 1.8 Estimated number of new cases in 2020, Egypt (Female)	8
Figure 1.9 Illustration of curse of dimensionality.....	9
Figure 2.1 The pipeline of microarray analysis.	10
Figure 2.2 The taxonomy of the feature selection techniques	11
Figure 2.3 Wrapper approach	15
Figure 2.4 The taxonomy of wrapper methods.....	15
Figure 2.5 The taxonomy of embedded method	19
Figure 2.6 hybrid approach	20
Figure 2.7 Illustration of SVM.....	21
Figure 2.8 a) Original data in the input space b) Mapped data in the feature space [34]	22
Figure 2.9 Illustration of KNN, $k=3, 5$	25
Figure 2.10 Random Forest.....	26
Figure 2.11 Confusion matrix for two classes	27
Figure 2.12 LOOCV [40].....	29
Figure 2.13 Illustration of k-fold cross-validation [41]	30
Figure 4.1 flowchart of the proposed MI-LASSO-GA approach	51
Figure 4.2 flowchart of the proposed MI-LASSO-PSO approach.....	53
Figure 5.1 the performance of MI-LASSO-GA on van't veer dataset	61
Figure 5.2 the performance of MI-LASSO-GA on chin dataset.....	62
Figure 5.3 the performance of MI-LASSO-GA on west dataset	62
Figure 5.4 the performance of MI-LASSO-GA on chowdary dataset.....	62

Figure 5.5the performance of MI-LASSO-GA on Gravier dataset	63
Figure 5.6 the performance of MI-LASSO-PSO on van't veer dataset	66
Figure 5.7 the performance of MI-LASSO-PSO on chin dataset	66
Figure 5.8 the performance of MI-LASSO-PSO on chowdary dataset	66
Figure 5.9 the performance of MI-LASSO-PSO on gravier dataset.....	67
Figure 5.10the performance of MI-LASSO-PSO on west dataset.....	67
Figure 5.11 The performance of MI-LASSO-GA VS MI-LASSO-PSO for Van't-veer dataset ..	69
Figure 5.12the performance of MI-LASSO-GA VS MI-LASSO-PSO for chin dataset	69
Figure 5.13 the performance of MI-LASSO-GA VS MI-LASSO-PSO for Chowdary dataset.....	70
Figure 5.15 the performance of MI-LASSO-GA VS MI-LASSO-PSO for gravier dataset.....	70
Figure 5.14the performance of MI-LASSO-GA VS MI-LASSO-PSO for gravier dataset.....	70
Figure 5.16 Number of selected features by MI-LASSO-GA VS MI-LASSO-PSO	71
Figure 5.17 the performance of proposed approaches vs state of art on chowdary dataset.....	73
Figure 5.18 the performance of proposed approaches vs state of art on Gravier dataset	73
Figure 5.19 the performance of proposed approaches vs state of art on Chin dataset.....	74
Figure 5.20 the performance of proposed approaches vs state of art on west dataset	74

List of tables

Table 2.1 The pros and cons of feature selection techniques	12
Table 3.1 Different Methods for classifying breast cancer.....	40
Table 3.1(continued) Different Methods for classifying breast cancer	41
Table 3.1(continued) Different Methods for classifying breast cancer	42
Table 3.1(continued) Different Methods for classifying breast cancer	43
Table 3.1(continued) Different Methods for classifying breast cancer	44
Table 3.1(continued) Different Methods for classifying breast cancer	45
Table 3.2 Different Methods for classifying other cancer	46
Table 3.2 (continued) Different Methods for classifying other cancer.....	46
Table 4.1 The Description of Breast cancer datasets.....	49
Table 4.2 paramters of classifer used for MI-LASSO-GA approach	57
Table 4.3 parameters of classifiers used for MI-LASSO-PSO approach	58
Table 5.1 The performance of MI-LASSO-GA on Van't veer dataset	60
Table 5.2 The performance of MI-LASSO-GA on Chin dataset.....	60
Table 5.3 The performance of MI-LASSO-GA on West dataset	60
Table 5.4 The performance of MI-LASSO-GA on Chowdary dataset.....	60
Table 5.5 The performance of MI-LASSO-GA on Gravier dataset	61
Table 5.6 Number of selected features by MI-LASSO-GA	61
Table 5.7 The performance of MI-LASSO-PSO on Van't veer dataset.....	63
Table 5.8 The performance of MI-LASSO-PSO on Chin dataset	64
Table 5.9 The performance of MI-LASSO-PSO on Chowdary dataset	64
Table 5.10 The performance of MI-LASSO-PSO on Gravier dataset.....	65
Table 5.11 The performance of MI-LASSO-PSO on West dataset.....	65
Table 5.12 Number of selected features by MI-LASSO-PSO	65

Abbreviations

ABC	Artificial bee colony
AGA	Adaptive genetic algorithm
AGA	Adaptive genetic algorithm
AMB	approximated Markov blanket
BA	Bat algorithm
BDF	Binary dragonfly
CFS	Correlation-based Feature Selection
ClusterQGA	Cluster quantum Genetic Algorithm
CV	Cross validation
DNA	Deoxyribonucleic acid
DT	Decision tree
EGA	Entropy-based genetic algorithm
FCBF	Fast correlation-based filter
FN	False negative
FP	False positive
FPA	Flower pollination algorithm
FRFI-WSA	fuzzy-rough-set-based f-information & water swirl algorithm
FS	Feature selection