



شبكة المعلومات الجامعية
التوثيق الإلكتروني والميكرو فيلم

بسم الله الرحمن الرحيم



HANAA ALY



شبكة المعلومات الجامعية
التوثيق الإلكتروني والميكروفيلم



شبكة المعلومات الجامعية التوثيق الإلكتروني والميكروفيلم



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 تغييرات



يجب أن

تحفظ هذه الأقراص المدمجة بعيدا عن الغبار



HANAA ALY



شبكة المعلومات الجامعية
التوثيق الإلكتروني والميكروفيلم

بسم الله الرحمن الرحيم



HANAA ALY



شبكة المعلومات الجامعية
التوثيق الإلكتروني والميكروفيلم



شبكة المعلومات الجامعية التوثيق الإلكتروني والميكروفيلم



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



شبكة المعلومات الجامعية
التوثيق الإلكتروني والميكرو فيلم

بسم الله الرحمن الرحيم



HANAA ALY



شبكة المعلومات الجامعية
التوثيق الإلكتروني والميكروفيلم



شبكة المعلومات الجامعية التوثيق الإلكتروني والميكروفيلم



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



Information Systems Department
Faculty of Computer & Information Sciences
Ain Shams University

A Model Based on Genetic Data for Selection of the Modes of Management of Liver Cancer and Prediction of their Outcome

Thesis submitted as a partial fulfilment of the requirements for the degree of
Master of Science in Computer and Information Sciences

By

Esraa Hamdi Abbas Abdelaziz

B.Sc. in Bioinformatics, Demonstrator at Information Systems Department,
Faculty of Computer and Information Sciences, Ain Shams University.

Under Supervision of

Prof. Khaled El-Bahnasy

Professor
Information Systems Department
Faculty of Computer and Information Sciences
Ain Shams University, Egypt

Dr. Rasha Ismail

Associate Professor
Information Systems Department
Faculty of Computer and Information Sciences
Ain Shams University, Egypt

Dr. Abdelaziz Abdelhamid

Assistant Professor
Computer Science Department
Faculty of Computer and Information Sciences
Ain Shams University, Egypt

2020

Abstract

Liver Cancer is one of the main causes of cancer-related deaths worldwide. Due to the extreme heterogeneity of this disease, its prognosis and management are still not yet standardized. Different treatment modalities are available, however, the patient's response to each of them varies; Therefore, it's critical to establish a model to help physicians individualize the management of this aggressive tumor.

This work presents one of the first investigations into personalizing liver cancer treatment for patients with genotype 4 using their clinical and genetic data. The aim of this study is to build a model that can predict the best treatment for a liver cancer patient with genotype 4, which could mitigate the impact of the disease and overall improve the patient's prognosis.

Firstly, we proposed two pipelines, a Single-Model pipeline and a Multi-Model pipeline, for recommending the best liver cancer treatment for patients and therefore potentially improve their prognosis. The data of 1427 Egyptian patients with liver cancer was analyzed. We studied the performance of six regression methods in predicting a patient's survival months after undergoing a certain treatment. The best performing method which was Random Forest was used in building the models in the proposed pipelines. The performance of both pipelines was compared, the Single-Model pipeline proved to be more reliable than the Multi-Model pipeline with 80.6% successful treatment recommendations which is higher than the prediction accuracy achieved in existing work. Moreover, we also prove the crucial importance of

genetic data and their effect on patients' prognosis and response to treatment.

Secondly, we built a model to predict the HCC patients' survivability and extract important features that affect their prognosis. Three feature selection methods and seven data mining classification techniques were studied and compared. The model's features were selected using Wrapper Subset method and built using Random Forest. The model's predictive accuracy was 87.54%. Moreover, important genetic features such as p53 gene exon 6 and 9 mutations proved to have a significant impact on the patient's overall prognosis.

Acknowledgements

I wish to thank all the people whose assistance was a milestone in the completion of this work.

I want to express my sincerest gratitude to my supervisors: Prof. Khaled Elbhanasy and Dr Rasha Ismail. I'm very thankful for the hours they have spent revising my work. They supported and motivated me. Their advice and expert guidance were invaluable. Moreover, I wish to express my deepest gratitude to Prof. Sanaa Kamal for providing the raw data used in this work and for her medical experience that was essential to complete this work successfully.

I would like to convey my deep appreciation towards my family for being my rock. My mother was always an amazing source of motivation when the road got tough, and my father was continuously supportive and kept me going. I would like to thank my brother and two sisters for always believing in me when I desperately needed it.

A special thanks goes to my friend Samiha Abdelrahman for always being ready to help when I needed her.

Above all, I would like to thank Allah S.W.T, it's because of HIS blessing that I was able to finish this work successfully.

Table of Contents

ABSTRACT	2
ACKNOWLEDGEMENTS	4
TABLE OF CONTENTS	5
LIST OF FIGURES	8
LIST OF TABLES	9
CHAPTER 1 : INTRODUCTION	10
1.1 MOTIVATION	10
1.2 PROBLEM DEFINITION	11
1.3 OBJECTIVES	12
1.4 CONTRIBUTION	13
1.5 THESIS OUTLINE	14
CHAPTER 2 : BACKGROUND	16
2.1 PREPROCESSING TECHNIQUES	16
2.1.1 Data Discretization	16
2.1.2 Feature Selection	17
2.1.3 Feature Scaling	18
2.2 HYPERPARAMETER TUNING	18
2.2.1 Manual Tuning	19
2.2.2 Grid Search	19
2.3 REGRESSION METHODS	19
2.3.1 Gaussian Processes (GPs)	19
2.3.2 Linear Regression	20
2.3.3 Stochastic Gradient Descent Regression (SGD)	20

2.3.4 Random Forest.....	20
2.3.5 Decision Tree Regressor	21
2.3.6 Support Vector Machine (SVM) for regression	22
2.4 CLASSIFICATION METHODS.....	23
2.4.1 Artificial Neural Network (ANN)	23
2.4.2 Support Vector Machine (SVM) for classification.....	23
2.4.3 Decision Tree (J48)	24
2.4.4 RepTree	24
2.4.5 Random Forest.....	25
2.4.6 JRip.....	25
2.4.7 PART.....	25
2.5 BIOLOGICAL BACKGROUND	26
2.5.1 Medical Background	26
2.5.2 Genetics Background	28
CHAPTER 3 : RELATED WORK	34
3.1 RELATED WORK IN TREATMENT-DECISION MAKING	34
3.2 RELATED WORK IN THE APPLICATION OF DATA MINING TECHNIQUES IN THE RISK CLASSIFICATION OF PATIENTS.	38
CHAPTER 4: A REGRESSION-BASED MODEL FOR PREDICTING THE BEST MODE OF LIVER CANCER TREATMENT	43
4.1 PROPOSED PIPELINES.....	44
4.2 REGRESSION ANALYSIS	46
4.3 MATERIALS AND DATA PREPROCESSING	48
4.4 EXPERIMENTAL RESULTS.....	55
4.4.1 The best regression methods for the multi-model pipeline.....	55
4.4.2 The best regression method for the single-model pipeline.....	57
4.4.3 Building the pipelines	58
4.4.4 Choosing the best pipeline	60