PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH UNIVERSITY FERHAT ABBAS SETIF 1

FACULTY OF SCIENCES COMPUTER SCIENCE DEPARTMENT



END OF STUDIES THESIS WITH A VIEW TO OBTAINING A MASTER'S DEGREE IN COMPUTER SCIENCE

SPECIALTY: INGÉNIERIE DES SYSTÈMES D'INFORMATION ET DE L'IMAGE (F3I)

Theme:

Artificial Intelligence to unveil genetic characteristics in Cancer and rare disease

---

Presented by :
Guesmia Hibet-errahmane
Beddoud Ayat-errahmane

Suervised By :
zanbout imene

Promo: 2025 / 2026

Table of Contents

Table of Contents **1**
Acknowledgements **5**
Abstract **6**
Résumé **7**
ملخص **8**
Background and Problematic **9**
Motivation and Objectives **10**
General Introduction **5**
Chapter 1:Background **10**
Introduction 10
1. Definition 10
1.1 Biomarkers 10
1.2 Cancer 11
1.3 Rare Diseases 11
1.4 Omics Data 12
1.5 Precision Medicine 12
2. Biology of the Cell 13
2.1 Definition of a Cell 13
2.2 Types of Cells 13
2.3 Cell Structure 14
2.4 Cell Division 15
2.5 Cell Metabolism 15
3.Genetic Diseases 15
3.1 Definition 15
3.2 Types of Genetic Diseases 16
3.3 Genetic Mutations and Their Effects 17
3.4 Genetic Diseases and Cancer 17
3.5 Role of Modern Technologies in Studying Genetic Diseases 18
4.Bioinformatics 19
4.1 Definition 19
4.2 Importance of Bioinformatics 19
4.3 Major Areas of Bioinformatics 19
4.4 Bioinformatics Tools and Techniques 21
4.5 Applications of Bioinformatics: 21
4.6 Role of Bioinformatics 21
5. Sequencing Data 22
5.1 First-Generation Sequencing (Sanger Sequencing) 22
5.2 Next-Generation Sequencing (NGS) 23
5.3 Third-Generation Sequencing 23

5.4 Types of Omics Data Generated by Sequencing 23
6. Use of Artificial Intelligence 24
6.1 definition 24
6.1.1 Supervised Learning 24
6.1.2 Unsupervised Learning 24
6.1.3 Semi-Supervised and Self-Supervised Learning 25
6.2 Deep Learning for Omics Data Analysis 25
6.2.1 Autoencoders 25
6.2.2 Deep Neural Networks (DNNs) 25
6.2.3 Convolutional Neural Networks (CNNs) 25
6.2.4 Graph Neural Networks (GNNs) 25
6.2.5 Canonical Correlation Analysis-Based Autoencoders (DCCA) 26
7. Feature Selection and Biomarker Discovery 26
Conclusion 27
Chapter 2: Omic Biomarker Discovery 28
introduction 28
1. Scope and Objectives of the Review 28
2. Search Methodology and Paper Filtration 29
2.1 Search Strategy 29
3. Biomarker Discovery Using Omics Data 30
4. Machine Learning in Omic Biomarker Discovery 31
4.1 Gene Expression-Based Biomarker Discovery 31
4.2 miRNA Expression-Based Biomarker Discovery 31
4.3 Protein Expression (RPPA)-Based Biomarker Discovery 32
5 Machine Learning in Omic Data Analysis 32
5.1 Single-Level Omics Cancer Classification 32
5.2 Multi-Level Omics Integration for Biomarker Discovery 32
5.3 Explainability and Biological Interpretation 33
6. Machine Learning in Genetic Disease Research 34
6.1 ML for Variant Interpretation and Pathogenicity Prediction 34
6.2 Transfer Learning and Data Augmentation for Rare Diseases 35
6.3 Biomarker Discovery in Rare Genetic Diseases 35
7. General Discussion 35
Conclusion 35
Chapter 3: Proposed Methodology for Gene Expression-Based Biomarker Discovery 37
Introduction 37
1. Overview of the Proposed Model 37
2. Data Collection and Preparation 38
2.1 Data Source 38
2.2 Preprocessing 39

3. Bio-Inspired Feature Selection 40
3.1 Bat Algorithm (BA) 40
3.2 Differential Evolution (DE) 41
3.3 Genetic Algorithm (GA) 41
3.4 Grey Wolf Optimizer (GWO) 41
3.5 Particle Swarm Optimization (PSO) 41
3.6 Fitness Function 41
4. Pathway Integration 42
4.1 Motivation 42
4.2 Pathway Databases 42
4.3 Protein-Protein Interaction Network via STRING DB 43
4.4 Cancer Pathway Extraction 43
5. Intersection of Selected Genes and Pathway Genes 43
5.1 Rationale 43
5.2 Intersection Procedure 44
5.3 Union of Intersection Results 44
6. Classification Model 44
6.1 Model Selection 44
6.2 Experimental Setup 45
6.3 Evaluation Metrics 45
6.4 Explainable AI (XAI) and Biological Interpretation 45
7. Enrichment Analysis 46
7.1 Rationale 46
7.2 Enrichment Tools and Databases 46
7.3 Enrichment Procedure 47
Conclusion 47
Chapter 4: Experimental Results and Discussion 48
Introduction 48
1. Experimental Setup 48
2.Feature Selection Results 50
2.1 Variance Filtering 50
2.2 Genetic Algorithm (GA) 51
2.3 Particle Swarm Optimization (PSO) 52
2.4 Grey Wolf Optimizer (GWO) 53
2.5 Differential Evolution (DE) 54
2.6 Bat Algorithm (BA) 55
2.7 Comparative Summary 56
3.Pathway Integration Results 56
3.1 STRING Protein-Protein Interaction Network 56
3.2 Hub Protein Identification 57

3.3 Cancer Pathway Extraction 59
4. Intersection Results 59
4.1 Gene-Pathway Intersection 59
4.2 Union of Intersection Results 60
5. Classification Results 61
5.1 Overview 61
5.2 Classification Performance 61
6. Enrichment Analysis Results 63
6.1 Gene Ontology (GO) Enrichment 63
6.2 ShinyGO Pathway Enrichment 63
6.3 cBioPortal Genomic Alteration Analysis 63
6.4 STRING Enrichment Network 64
6.5 Selected Biomarkers 64
Conclusion 66
General Introduction 5
Chapter1:Background 10
Chapter 2: Omic Biomarker Discovery 28
Chapter 3: Proposed Methodology for Gene Expression-Based Biomarker Discovery 37
Chapter 4: Experimental Results and Discussion 48

Acknowledgements

We would like to express our profound gratitude to all those who have contributed to the successful completion of this master's thesis. This journey has been filled with challenges and milestones, and it would not have been possible without the unwavering support and encouragement of many individuals.

First and foremost, we extend our heartfelt appreciation to our supervisor, Dr. Zenbout Imene, for her exceptional guidance, patience, and invaluable mentorship throughout this research. Her expertise and commitment to our academic and professional development have been instrumental in shaping this work.

We are deeply thankful to our families, and most especially to our parents, whose unconditional love, endless sacrifices, and constant encouragement have been our greatest source of strength throughout this journey. This work is dedicated to them.

We also wish to express our sincere gratitude to our dearest friends, who stood by our side during every difficult moment, offering support, motivation, and positive energy when it was needed most.

Finally, we extend our acknowledgement to the faculty members of the F3I department at Université Ferhat Abbas Sétif 1, for their valuable teaching and support throughout our academic journey.
Beddoud Ayat-Errahmane & Guesmia Hibe-Errahmane.

Abstract

Gene expression analysis has emerged as a powerful approach for cancer biomarker discovery, enabling the identification of molecular signatures associated with tumour initiation, progression, and metastasis. However, the inherent high dimensionality of transcriptomic data, where the number of measured genes far exceeds the number of available patient samples, poses significant challenges for conventional machine learning approaches.

To address these challenges, this thesis proposes an integrated computational framework for gene expression-based biomarker discovery in cancer, applied to the TCGA Pan-Cancer dataset comprising 10,535 samples distributed across 33 cancer types. The proposed framework combines five bio-inspired metaheuristic feature selection algorithms the Bat Algorithm (BA), Differential Evolution (DE), Genetic Algorithm (GA), Grey Wolf Optimizer (GWO), and Particle Swarm Optimization (PSO) with a biological pathway integration strategy leveraging the KEGG, WikiPathways, and STRING databases. The biologically grounded gene subsets are evaluated through a comparative classification analysis using KNN and SVM classifiers, and validated through multi-platform enrichment analysis using GO, ShinyGO, cBioPortal, and STRING.

The experimental results demonstrate that all five algorithms outperformed the baseline classification accuracy of 42.80%, with GWO achieving the best performance by selecting only 371 genes while attaining a KNN accuracy of 74.6%. The pathway integration step reinforced the biological relevance of the identified gene panels, yielding a validated set of 23 candidate biomarkers grounded in well-established cancer signalling mechanisms.

Keywords: Gene expression analysis, Bio-inspired feature selection, Biological pathway integration, Cancer biomarker discovery, TCGA Pan-Cancer, Explainable AI, Enrichment analysis.

Résumé

L'analyse de l'expression génique est devenue une approche puissante pour la découverte de biomarqueurs du cancer, permettant l'identification de signatures moléculaires associées à l'initiation, la progression et la métastase tumorale. Cependant, la grande dimensionnalité inhérente aux données transcriptomiques, où le nombre de gènes mesurés dépasse largement le nombre d'échantillons de patients disponibles, pose des défis majeurs pour les approches conventionnelles d'apprentissage automatique.

Pour répondre à ces défis, ce mémoire propose un cadre computationnel intégré pour la découverte de biomarqueurs basée sur l'expression génique dans le cancer, appliqué à l'ensemble de données TCGA Pan-Cancer comprenant 10 535 échantillons répartis sur 33 types de cancers. Le cadre proposé combine cinq algorithmes métaheuristiques bio-inspirés de sélection de caractéristiques l'Algorithme des Chauves-souris (BA), l'Évolution Différentielle (DE), l'Algorithme Génétique (GA), l'Optimiseur par Meute de Loups Gris (GWO) et l'Optimisation par Essaim de Particules (PSO) avec une stratégie d'intégration de voies biologiques exploitant les bases de données KEGG, WikiPathways et STRING. Les sous-ensembles de gènes biologiquement fondés sont évalués par une analyse de classification comparative utilisant KNN et SVM, et validés par une analyse d'enrichissement multi-plateforme.

Les résultats expérimentaux démontrent que les cinq algorithmes ont surpassé la précision de classification de base de 42,80%, le GWO obtenant les meilleures performances en sélectionnant seulement 371 gènes tout en atteignant une précision KNN de 74,6%. L'étape d'intégration des voies biologiques a renforcé la pertinence biologique des panels de gènes identifiés, produisant un ensemble validé de 23 biomarqueurs candidats.

Mots-clés : Analyse d'expression génique, Sélection de caractéristiques bio-inspirée, Intégration de voies biologiques, Découverte de biomarqueurs du cancer, TCGA Pan-Cancer, IA explicable, Analyse d'enrichissement.

ملخص

أصبح تحليل التعبير الجيني نهجاً فعّالاً لاكتشاف المؤشرات الحيوية للسرطان، إذ يتيح تحديد البصمات الجزيئية المرتبطة ببدء الورم وتطوره وانتشاره. غير أن الأبعاد العالية المتأصلة في البيانات الجينومية، حيث يفوق عدد الجينات المقاسة عدد عينات المرضى المتاحة بشكل كبير، تُشكّل تحديات جوهرية أمام أساليب التعلم الآلي التقليدية، من بينها الإفراط في التخصيص والتكلفة الحسابية المرتفعة ومحدودية قابلية التفسير البيولوجي.

لمعالجة هذه التحديات، تقترح هذه المذكرة إطاراً حسابياً متكاملاً لاكتشاف المؤشرات الحيوية المبنية على التعبير الجيني التي تضم 10,535 عينة موزعة على 33 نوعاً من TCGA Pan-Cancer في السرطان، مطبّقاً على قاعدة بيانات السرطان. يجمع الإطار المقترح بين خمسة خوارزميات تحسين مستوحاة من الطبيعة — خوارزمية الخفافيش (BA)، والتطور التفاضلي (DE)، والخوارزمية الجينية (GA)، ومحسّن الذئب الرمادي (GWO)، وتحسين سرب الجسيمات (PSO) — واستراتيجية متكاملة لدمج المسارات البيولوجية من قواعد بيانات KEGG وWikiPathways. يتم تقييم المجموعات الجينية المختارة من خلال تحليل تصنيفي مقارن باستخدام مصنّفي SVM وKNN، والتحقق من صحتها عبر تحليل إثراء متعدد المنصات باستخدام STRING وcBioPortal وShinyGO وGO.

أثبتت النتائج التجريبية أن جميع الخوارزميات الخمس تجاوزت دقة التصنيف الأساسية البالغة 42.80%، إذ حقق محسّن الذئب الرمادي KNN مرحلة دمج المسارات بلغت 74.6%. كما عزّزت أفضل أداء باختيار 371 جيناً فقط مع تحقيق دقة عالية، مما أسفر عن مجموعة مُتحقق منها من 23 مؤشراً حيوياً مرشحاً مستنداً إلى آليات الإشارة السرطانية المعروفة.

الكلمات المفتاحية: تحليل التعبير الجيني، الاختيار البيو-مستوحى للخصائص، دمج المسارات البيولوجية، اكتشاف المؤشرات الحيوية للسرطان، TCGA Pan-Cancer، تحليل الإثراء، الذكاء الاصطناعي القابل للتفسير.