

Title	Exploring independent omics data fusion for precision oncology: cancer genes, biomarkers, and machine learning
Authors	Younis, Haseeb
Publication date	2025
Original Citation	Younis, H. 2025. Exploring independent omics data fusion for precision oncology: cancer genes, biomarkers, and machine learning. PhD Thesis, University College Cork.
Type of publication	Doctoral thesis
Rights	© 2025, Haseeb Younis. - https://creativecommons.org/licenses/by/4.0/
Download date	2026-09-16 13:33:58
Item downloaded from	https://hdl.handle.net/10468/18421

Exploring Independent Omics Data Fusion for Precision Oncology: Cancer Genes, Biomarkers, and Machine Learning

Haseeb Younis

BSc, MSc

orcid.org/0000-0003-4093-3027

**Thesis submitted for the degree of
Doctor of Philosophy**



NATIONAL UNIVERSITY OF IRELAND, CORK

COLLEGE OF SCIENCE, ENGINEERING AND FOOD SCIENCE

SCHOOL OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY

October, 2025

Head of Department: Prof. Dirk Pesch

Supervisors: Dr. Rosane Minghim,
Dr. Adriana Franco Paes Leme

Research supported by Science Foundation Ireland

Contents

List of Figures	v
List of Tables	vii
Abstract	ix
Acknowledgements	xii
Acronyms	xiii
1 Introduction	1
1.1 Motivation	1
1.1.1 Research Questions	2
1.1.2 Research Objectives	2
1.2 Thesis Statement	3
1.3 Contributions	3
1.4 Publications	4
1.4.1 Peer-Reviewed Conference and Journal Papers	4
1.4.2 Preprints and Unpublished Manuscripts	5
1.5 Additional Academic Contributions	5
1.6 Thesis Structure	6
2 Background	8
2.1 Cancer Understanding and Analysis	8
2.1.1 Historical Overview of Cancer	8
2.1.2 Global Burden and the Pressing Challenge of Cancer	9
2.1.3 Limitations of Traditional Approaches in Cancer Diagnosis and Classification	10
2.1.4 The Promise of Independent-Omics Analysis in Unraveling Cancer Biology	11
2.1.5 Innovations Through Bioinformatics, Machine Learning, and Data Visualisation	13
2.2 Essential Background Concepts	15
2.2.1 Biology Related Concepts	15
2.2.2 Computer Science Related Concepts	17
3 Literature Review	19
3.1 Overview	19
3.1.1 Somatic Mutation Data in Cancer	20
3.1.2 Transcriptomic (RNA Expression) Data in Cancer	24
3.1.3 Proteomic Data in Cancer	31
3.1.4 Comparison of Omics Data Types: Strengths and Weaknesses	33
3.1.4.1 Strengths of DNA Somatic Mutation Data	33
3.1.4.2 Limitations of DNA Somatic Mutation Data	33
3.1.4.3 Strengths of Transcriptomic Data	34
3.1.4.4 Limitations of Transcriptomic Data	34
3.1.4.5 Strengths of Proteomic Data	35
3.1.4.6 Limitations of Proteomic Data	36

4 Somatic Mutation-Based Modelling for Cancer Classification and Subtype Prediction	37
4.1 Introduction	37
4.2 Methodology	38
4.2.1 Data Collection and Preprocessing	38
4.2.2 Desparsification Using SAMBAR	40
4.2.3 Deep Representation Learning and Clustering	42
4.2.4 Cluster Validation via Survival Analysis	43
4.2.4.1 Kaplan–Meier Estimator	43
4.2.4.2 Log-Rank Test	44
4.2.4.3 Cox Proportional Hazards Model	44
4.2.5 Classification of Molecular Subtypes	45
4.2.5.1 Model Architecture and Training	45
4.2.5.2 Explainability via LIME	47
4.2.5.3 Visualisation and Feature Importance Analysis	47
4.3 Results and Discussion	47
4.3.1 Experimental Setup	48
4.3.2 Dataset Description	48
4.3.3 Evaluation Measures	48
4.3.4 Clustering of desparsed data	49
4.3.5 Survival and Statistical Analysis	51
4.3.6 Robustness of our methodology	52
5 Transcriptomic Profiling and Deep Learning for Cancer Type and Subtype Classification	57
5.1 Introduction	57
5.2 Methodology	58
5.2.1 Data Preprocessing	58
5.2.2 Cancer Classification	60
5.2.2.1 SMOTE for Synthetic Data Generation	60
5.2.2.2 Model for Binary and Multiclass Cancer Subtype Classification	61
5.2.2.3 Early Stopping Criteria	63
5.2.3 Cancer Type Classification using RNA Data	64
5.2.3.1 PCA on RNA Dataset	64
5.2.3.2 Multiclass Cancer Type Classification Model with CNN	66
5.2.3.3 Optimisation and Class Balancing	68
5.2.4 Explainability and Visual Analysis	70
5.2.4.1 Explainable AI using LIME	70
5.2.4.2 Heatmap	72
5.2.4.3 Cluster Map	72
5.2.4.4 UMAP (Uniform Manifold Approximation and Projection)	73
5.2.4.5 Violin Plot	74
5.2.5 KEGG Pathway Enrichment Analysis	74

5.2.6	GO Enrichment Analysis	75
5.2.6.1	Equations for P-value Adjustment	75
5.3	Results and Discussion	76
5.3.1	Experimental Setup and Evaluation Measures	76
5.3.2	Dataset Description	76
5.3.3	Robustness of Our Methodology	77
5.3.3.1	Results for Microarray Datasets	89
6	Proteomic Complex Disruption Analysis for Biomarker Discovery in Oral Cancer	95
6.1	Introduction	95
6.2	Methodology	96
6.2.1	Data Collection and Preprocessing	96
6.2.2	Data Visualisation	97
6.2.3	Biomarker Ranking	98
6.2.3.1	ExtraTree Feature Selection	99
6.2.3.2	K-Best Feature Selection	100
6.2.3.3	Recursive Feature Elimination (RFE)	102
6.2.3.4	Boruta Feature Selection	103
6.2.4	Machine Learning Methods	104
6.2.4.1	k-Nearest Neighbours (k-NN)	104
6.2.4.2	Decision Tree Classifier	105
6.2.4.3	Random Forest Classifier	106
6.2.4.4	Adaboost Classifier	107
6.2.4.5	Gradient Boosting Classifier	108
6.2.4.6	Gaussian Naïve Bayes Classifier	109
6.2.4.7	Linear Discriminant Analysis	111
6.2.4.8	Quadratic Discriminant Analysis Classifier	112
6.2.5	Protein-Protein Interaction	113
6.3	Results and Discussion	115
6.3.1	Dataset analysis	115
6.3.2	Machine Learning for Biomarker ranking and importance	117
6.3.3	Analysis of Selected Biomarkers	119
6.3.4	Protein-Protein Interaction and Network Analysis	121
6.3.5	Validation of Protein Complexes	123
7	Conclusion and Future Directions	130
7.1	Conclusion	130
7.1.1	Somatic Mutation Analysis	130
7.1.2	Transcriptomic-Based Classification	131
7.1.3	Proteomic Complex-Based Biomarker Discovery	132
7.1.4	Addressing the Research Questions	132
7.1.5	Cross-Platform Impact	134
7.2	Ethical Considerations and Data Access	134
7.3	Recommendations for Future Work	135
7.4	Final Words	135

A Glossary of Terms	171
Glossary	171

List of Figures

4.1	Overview of CLARITY	39
4.2	Model Architecture for Cancer Multi-Class Classification	46
4.3	Representation of our desparsed data	50
4.4	Representation of clustered data by our Autoencoder + Spectral clustering method	51
4.5	Kaplan–Meier survival curves for patients grouped by molecular subtypes discovered through Autoencoder + Spectral clustering (for $k = 3$).	52
4.6	Accuracy and Loss Curves of our Model	54
4.7	Abundance Value Distribution of Top Hallmarks in BRCA	55
4.8	Cluster Map of Top Hallmarks	56
5.1	Overview of Our Transcriptomic Methodology	59
5.2	Model Architecture for Cancer Subtype Classification (Figure Generated using Netron [Roe25])	63
5.3	Model Architecture for Cancer Type Classification (Figure Generated using Netron [Roe25])	68
5.4	Results of Lung_GSE87340 Cancer Classification Model for Binary Class	79
5.5	Data Point Distribution of Tumour and Normal profiles	80
5.6	Abundance Value Distribution of Top Genes for Tissue GSE87340	81
5.7	Cluster Map of Top Genes for Tissue GSE87340. Columns correspond to genes, and rows correspond to samples. The colour bar indicates normalised expression values.	82
5.8	Results of Cancer Classification Model for Multi-Class Tasks	83
5.9	Data point distribution of Primary Tumour(Orange), Normal(Green), NormalLiver(Purple), Metastasis(Blue)	83
5.10	Abundance Value Distribution of Top Genes for Tissue SRR2089755	84
5.11	Cluster Map of Top Genes for Tissue SRR2089755. Columns correspond to genes, and rows correspond to samples. The colour bar indicates normalised expression values.	85
5.12	Results of Cancer Type Classification Model (8 Classes)	86
5.13	Data Point Distribution of Cancer Types and Normal	86
5.14	Abundance Value Distribution of Top Genes for 8 Classes	87
5.15	Cluster Map of Top Genes for 8 classes. Columns correspond to genes, and rows correspond to samples. The colour bar indicates normalised expression values.	88
5.16	Confusion Matrix for Cancer Type Classification	89
5.17	Comparison Between Proposed Method and Base Method	90
5.18	Gene Ontology (GO) Enrichment Analysis	91
5.19	Comparison of our Methodology with Base Study [GFD19] on CuMiDa.	92
5.20	Results of Our Methodology on CuMiDa	92

5.21 Data Point Distribution of Original Dataset and Model Separation on CuMiDa	93
6.1 Flow Diagram of the proposed proteomic analysis workflow . .	97
6.2 Correlation between Peptides	116
6.3 Correlation between Peptides Razors	116
6.4 Correlation between Proteins	117
6.5 Correlation between Proteins	117
6.6 Classification Accuracies with all Biomarkers	118
6.7 Classification Accuracies with top 50 Biomarkers	118
6.8 Classification Accuracies with top 100 Biomarkers	119
6.9 Classification Accuracies with top 200 Biomarkers	119
6.10 Common Proteins Within the Selected Sets	120
6.11 Distribution of Proteins Between Selected Sets	120
6.12 Distribution of Protein Abundance Across Control Cell Lines . .	121
6.13 Distribution of Protein Abundance Across Less Aggressive Cell Lines	122
6.14 Distribution of Protein Abundance Across Aggressive Cell Lines	122
6.15 Precision Curve for Control Complexes	123
6.16 Precision Curve for Less Aggressive Complexes	124
6.17 Precision Curve for Aggressive Complexes	125
6.18 Interaction network of EGFR and JUP in Less Aggressive Categories	126
6.19 Interaction network of CCT8 and CCT6A in Less Aggressive Categories	127
6.20 Interaction network of RPLs and RPSs in Aggressive Categories	127
6.21 Interaction network of GAPDH and PKM in Aggressive Categories	128

List of Tables

3.1	Comparison of Existing Cancer Classification Studies Using Genomic Data	23
3.2	Comparative Analysis of Relevant Studies	30
4.1	Comparison of Cluster Memberships for Different Values of k Using Autoencoder + Spectral Clustering vs. k-means	50
4.2	Comparison of Cox Proportional Hazards Model Results for Cluster-Based Survival Analysis	52
4.3	Performance Comparison of Classifiers with k-means and Autoencoder + Spectral Clustering (AES). Bold values indicate the better result for each classifier.	53
5.1	Overview of RNA-seq datasets used for Cancer Classification. . .	77
5.2	Classification performance metrics (Precision, Recall, F1-Score, and Accuracy) for each cancer type on the validation dataset. .	87
6.1	Gene Classification into Control, Less Aggressive, and Aggressive Categories	129

I, Haseeb Younis, certify that the work I am submitting is my own and has not been submitted for another degree, either at University College Cork or elsewhere. All external references and sources are clearly acknowledged and identified within the contents. I have read and understood the regulations of University College Cork concerning plagiarism and intellectual property.

Haseeb Younis

Abstract

Cancer is a multifaceted and heterogeneous disease that continues to pose a major global health challenge. Despite technological advances, accurate classification of cancer types and subtypes, discovery of reliable molecular biomarkers remain difficult due to the biological complexity and diversity of patients and tumour types. With the emergence of high-throughput technologies, large-scale omics datasets have become available, capturing distinct molecular layers of cancer, such as genomic mutations, gene expression, and protein interactions. While many studies attempt to integrate multiple omics, this work emphasises the importance of understanding the strengths, limitations, and biological markers within each modality independently. Such focused analysis enables the design of data-specific algorithms with clearer interpretation, helping reveal complementary insights from each omics layer.

This thesis presents novel computational approaches to the investigation of three independent omics data types for cancer classification and biomarker discovery: somatic mutation data (genomics), RNA Sequencing (RNA-seq) and microarray-based gene expression data (transcriptomics), and Mass Spectrometry (MS)-based protein co-fractionation data (proteomics). Each data type is investigated separately using tailored machine learning and deep learning algorithms, allowing precise modelling and interpretation of the biological and computational challenges unique to each dataset.

With regards to genomics, a pipeline termed “CLARITY” was proposed, to desparsify somatic mutation data by summarising mutations into pathway scores, followed by clustering with autoencoders to assign cancer stage labels. Classical machine learning models, as well as newly designed domain-specific deep learning models, were employed to classify cancer stages. The results demonstrated high accuracy. Furthermore, survival analysis validated the biological relevance of the identified subtypes and potential driver genes.

For transcriptomic analysis, deep learning architectures, specifically Convolutional Neural Network (CNN) and transformer models, were developed for cancer subtype classification with a focus on RNA-seq and microarray datasets from multiple cancer types. Synthetic Minority Over-sampling Technique (SMOTE) was used to address class imbalance, while Local Interpretable Model-agnostic Explanations (LIME) helped interpret model outputs and identify gene-level biomarkers. The models achieved competitive performance across various

datasets and uncovered interpretable gene expression signatures with biological significance; this model was further validated through pathway enrichment analysis.

The proteomic section of this research involved MS-based co-fractionation data collected from normal and cancerous oral keratinocyte cell lines. A ranking pipeline using feature selection was implemented to select robust protein subsets. Machine learning classifiers were then trained to differentiate between normal, less aggressive, and highly aggressive cancer lines, preserving the important proteins. Protein–Protein Interaction Network (PPIN)s were constructed to identify disrupted complexes specific to aggressive phenotypes, offering insight into molecular mechanisms underlying oral squamous cell carcinoma progression.

Throughout this thesis, the challenges of high dimensionality, class imbalance, and data heterogeneity are addressed using domain-appropriate preprocessing, feature selection, and Explainable Artificial Intelligence (XAI) techniques. Rather than relying solely on performance metrics, a strong emphasis was placed on biological interpretability, making the findings valuable for both computational researchers and biomedical scientists.

Collectively, this thesis demonstrates that each omics modality offers a distinct and valuable lens for understanding cancer biology. Genomic data provides information about mutation-driven mechanisms and tumour staging; transcriptomic data enables subtype classification and functional gene identification; proteomic data reveals disruptions in protein abundance and complex formation. An in-depth examination of these layers independently establishes a strong basis for future work aiming to integrate them.

By providing validated models and feature sets, this thesis contributes to the development of interpretable, data-driven cancer classification frameworks and biomarker discovery pipelines. The methodologies and findings presented here lay the groundwork for more generalizable and biologically meaningful applications of machine learning in computational oncology, ultimately advancing efforts toward precision cancer diagnosis and therapy.

I would like to express my sincere gratitude to Rosane Minghim and Adriana Franco Paes Leme for their invaluable supervision and guidance throughout my PhD research. I am also thankful to Dirk Pesch, Luis Gómez de Membrillera, and James Duggan for their support through the ADVANCE CRT programme. My appreciation extends to my colleagues at UCC and CRT ADVANCE for their collaboration and encouragement during this journey. I am grateful to both academic and industry partners for providing essential data and insights that contributed to my research. A heartfelt thanks to Isabel Maria Ronan, whose unwavering support has made this journey truly meaningful. Finally, I am deeply thankful to my family and friends for their constant encouragement and belief in me.

Acknowledgements

This publication has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland under Grant number 18/CRT/6222.

Haseeb Younis
October 2025

List of Acronyms

- AdaBoost** Adaptive Boosting. 107, 108
- AI** Artificial Intelligence. 22, 26
- AICc** Corrected Akaike Information Criterion. 114
- ANN** Artificial Neural Network. 17, 27
- ANOVA** Analysis of Variance. 100, 101
- AP-MS** Affinity Purification–Mass Spectrometry. 31
- AUC** Area Under the Curve. 22, 23, 28
- BH** Benjamini-Hochberg. 75
- BIC** Bayesian Information Criterion. 114
- BRCA** Breast Cancer Gene. 22, 23, 28, 37–39, 48, 49, 51, 54
- cDNA** complementary DNA. 25, 26
- CF-MS** Co-Fractionation–Mass Spectrometry. 3, 31, 33, 130, 132
- CNN** Convolutional Neural Network. ix, 3, 17, 19, 27, 28, 61, 66, 77–79, 131
- Co-Fractionation** Co-Fractionation. 96
- CORUM** Comprehensive Resource of Mammalian protein complexes. 96, 123
- COSMIC** Catalogue Of Somatic Mutations In Cancer. 33, 38, 39, 48
- CuMiDa** Curated Microarray Database. v, vi, 16, 58, 89, 92, 93
- DM** DNA Methylation. 13
- DNA** Deoxyribonucleic Acid. 11–13, 15, 16, 20, 24, 33, 34
- DNN** Deep Neural Network. 56, 131
- dNTPs** Deoxynucleotide Triphosphates. 25
- Driver Gene** Driver Gene. 13
- DT** Decision Tree. 29, 99, 105–109
- EGFR** Epidermal Growth Factor Receptor. 20, 33, 125
- ER** Estrogen Receptor. 11, 21
- ExtraTrees** Extremely Randomized Trees. 99, 100
- GAN** Generative Adversarial Network. 27, 28
- GBC** Gradient Boosting Classifier. 108, 109

- GDC** Genomic Data Commons. 38
- GE** Gene Expression. 13, 14, 20, 24–29, 31, 34, 36, 56, 64, 130
- GEO** Gene Expression Omnibus. 22, 29, 58, 59
- GEP** Gene Expression Profiling. 24, 25
- GISTIC** Genomic Identification of Significant Targets in Cancer. 38, 39, 48
- GLOBOCAN** Global Cancer Observatory. 9
- GO** Gene Ontology. 41, 58, 75, 86
- HER2** Human Epidermal Growth Factor Receptor 2. 11, 21, 33
- HGP** Human Genome Project. 9, 15
- HMK** Human Oral Keratinocyte. 32, 95, 96, 98, 104–112, 115, 129, 132
- HSC-3** Human Squamous Carcinoma-3. 32, 95, 96, 98, 104–112, 115, 129, 132
- IHC** Immunohistochemistry. 21
- k-means** k-means Clustering. vii, 49, 50, 52, 53, 56
- KDE** Kernel Density Estimate. 74
- KEGG** Kyoto Encyclopedia of Genes and Genomes. 41, 48, 58, 74, 75, 84, 94, 131
- KNN** k-Nearest Neighbour. 29, 60, 73, 104, 105
- LC-MS/MS** Liquid Chromatography–Tandem Mass Spectrometry. 96
- LDA** Linear Discriminant Analysis. 111, 112
- LIME** Local Interpretable Model-agnostic Explanations. ix, 3, 18, 22, 38, 47, 54, 56, 70–72, 87, 131, 133
- lncRNA** Long Non-Coding RNA. 25
- LR** Linear Regression. 28, 29, 33, 95, 114
- Luminal A** Luminal A. 21, 24
- Luminal B** Luminal B. 21
- MIL** Multiple Instance Learning. 22, 23
- MKL** Multiple Kernel Learning. 22, 23
- MLP** Multilayer Perceptron. 29
- mRNA** Messenger Ribonucleic Acid. 20, 24, 25, 27, 31, 33–35

- MS** Mass Spectrometry. ix, x, 1, 3, 7, 31, 32, 36, 94, 132, 133
- MSE** Mean Squared Error. 42
- MSigDB** Molecular Signatures Database. 38, 48
- NGS** Next Generation Sequencing. 21, 24, 25
- PAAD** Pancreatic Adenocarcinoma. 29
- PCA** Principal Component Analysis. 18, 58, 60, 64–66
- PPI** Protein–Protein Interaction. 31, 32, 95–97, 113–115, 118, 121, 123
- PPIN** Protein–Protein Interaction Network. x, 4, 31, 35, 95, 96, 114, 121, 132, 133
- PR** Progesterone Receptor. 11, 21
- PRINCE** Predicting Interactomes from Co-Elution. 95, 114
- PSA** Prostate Specific Antigen. 10
- QDA** Quadratic Discriminant Analysis. 112
- ReLU** Rectified Linear Unit. 42, 45, 62, 68, 78
- RF** Random Forest. 18, 26, 28, 29, 99, 103, 106, 107, 114
- RFE** Recursive Feature Elimination. 102, 103
- RNA** Ribonucleic Acid. 12, 13, 16, 25, 26, 29, 59, 64, 65, 70, 72–74, 76, 130
- RNA-seq** RNA Sequencing. vii, ix, 1–3, 24–26, 29, 34, 35, 57–60, 64, 66, 70, 71, 74, 75, 77, 78, 82, 83, 131, 133, 135
- SAMBAR** Subtyping Agglomerated Mutations By Annotation Relations. 24, 40–42, 45
- SCC-9** Squamous Cell Carcinoma-9. 32, 95, 96, 98, 104–112, 115, 129, 132
- SHAP** SHapley Additive exPlanations. 18
- SMOTE** Synthetic Minority Over-sampling Technique. ix, 3, 18, 46, 58, 60, 133
- SNV** Single Nucleotide Variant. 33
- SVM** Support Vector Machine. 18, 26–29, 32, 33, 95, 102, 114
- TCGA** The Cancer Genome Atlas. 9, 12, 13, 16, 21, 22, 27–29, 33, 38, 48
- TP53** Tumor Protein 53. 9, 20
- UMAP** Uniform Manifold Approximation and Projection. 18, 43, 49, 73, 74

WES Whole Exome Sequencing. 33

WHO World Health Organization. 9

XAI Explainable Artificial Intelligence. x, 2, 3, 7, 18–20, 22, 37, 54, 56–58, 94, 131, 134

Chapter 1

Introduction

This chapter provides an overview of the research context, motivation of this work, research questions and objective of this thesis. A statement summarising our work done in this thesis, our contributions to the field of cancer classification and biomarker discovery. It also highlights our publication and academic contributions made during the degree. Finally, it presents a conceptual foundation for the methodologies and analyses developed in the subsequent chapters.

1.1 Motivation

Cancer is one of the leading causes of death worldwide, with millions of new cases and deaths reported annually [Wor25a]. Despite decades of computationally advanced research, accurate and early classification of cancer types and subtypes remains a significant challenge. This challenge is largely due to the complexity and heterogeneity of cancer at the molecular level. Advances in high-throughput technologies have enabled the generation of large-scale omics data, such as somatic mutation profiles, gene expression patterns, and proteomic measurements that capture different aspects of tumour biology. However, analysing this diverse data independently and effectively remains non-trivial [PABH⁺24].

This thesis focuses on utilising independent omics datasets: somatic mutations (genomic), RNA-seq and microarray-based gene expression (transcriptomic), and MS-based protein co-fractionation (proteomic) to identify meaningful patterns and biomarkers that can aid in the classification of cancer types and subtypes. Each modality provides a unique biological perspective: mutations in-

dictate underlying genomic changes; transcriptomics reveals gene activity; and proteomics captures functional protein interactions.

Machine learning, especially deep learning and XAI, offers powerful tools to extract insights from these high-dimensional datasets. However, developing reliable, generalizable models that work across data types and cancer conditions requires careful design, validation, and interpretation.

1.1.1 Research Questions

This thesis contributes to answering the following research question and sub-questions:

- **RQ1:** How can computational methods be improved to effectively classify cancers with state-of-the-art performance in parallel with discovering biomarkers from independent omics datasets?
 - **sub-RQ1:** How can sparse, high-dimensional somatic mutation data be transformed into biologically meaningful features for cancer subtype discovery and validated using statistical methods?
 - **sub-RQ2:** How can transcriptomic data from RNA-seq and microarray platforms be modelled using deep learning architectures to achieve robust and interpretable subtype classification, and validated using bioinformatics methods?
 - **sub-RQ3:** How can proteomic co-fractionation data be analysed better to identify protein-level biomarkers and disrupted complexes linked to cancer aggressiveness?

1.1.2 Research Objectives

Our research aims to:

- Explore the predictive value of each omics type independently.
- Develop tailored deep learning and machine learning models for the classification of different cancer types and subtypes.
- Identify interpretable biomarkers using XAI techniques.
- Demonstrate how insights from each modality contribute uniquely to cancer biology.

This work lays the groundwork for integrative, cross-domain analysis while preserving the depth of understanding achievable from each individual dataset.

1.2 Thesis Statement

This thesis presents a comprehensive investigation into the independent use of genomic, transcriptomic, and proteomic data for cancer classification and biomarker discovery. It develops and evaluates tailored machine learning and deep learning models for each omics modality, incorporates XAI for interpretability and identification of potential biomarkers. It also highlights the unique biological insights each dataset contributes to the identification of cancer types and subtypes. By analysing the strengths and limitations of each modality independently, this work establishes a foundation for informed, modality-specific research and sets the stage for future integrative frameworks in computational oncology.

1.3 Contributions

This thesis makes the following core contributions:

1. **Independent modelling of somatic mutation data:** We developed a method called “CLARITY” to desparsify mutation data using biological pathway scores, employing clustering and supervised learning to classify cancer stages. Our method also identified poor prognostic subtypes, validated by domain-specific statistical and survival analysis.
2. **Deep learning on transcriptomic data:** We designed CNN and transformer-based models to classify multiple cancer types and subtypes using microarray gene expression data. Class imbalance was addressed using SMOTE, and explainability was achieved using LIME. The models achieved near-perfect performance and revealed key gene biomarkers. Furthermore, these models can be used as pretrained networks for RNA-seq and microarray-based databases.
3. **Proteomic biomarker discovery using Co-Fractionation–Mass Spectrometry (CF–MS):** We used MS data to study protein abundance and complex formation in oral cancer cell lines. Feature selection methods were applied to reduce dimensionality, and classification was performed

using machine learning models. PPIN analysis identified key complex disruptions related to aggressiveness.

These contributions advance the potential of single-omics analysis combined with modern machine learning to advance cancer research and set the stage for future integrative approaches.

1.4 Publications

Publications and Academic Contributions

Following publications in leading Computer Science and Bioinformatics venues, including IEEE- and ACM-organised conferences, have emerged from this PhD research. The list of contributions is as follows:

1.4.1 Peer-Reviewed Conference and Journal Papers

- Haseeb Younis and Rosane Minghim, “Enhancing Cancer Classification from RNA Sequencing Data Using Deep Learning and Explainable AI,” *Machine Learning and Knowledge Extraction* 7, no. 4: 114. doi:10.3390/make7040114.
- Haseeb Younis and Rosane Minghim, “Gene-Level Feature Prioritization and Interpretability in Microarray Cancer Classification Using Deep Convolutional Architectures and LIME-Based Saliency,” accepted at the *20th International Conference on Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB)*, Milano, Italy, September 10–12, 2025.
- Haseeb Younis, Horiya Imane Brahmi, Jonathan Byrne, and Rosane Minghim, “UNCCER: Unified Network for Cancer Classification and Efficient Representation using Microarray Data,” in *Proceedings of the 18th International Conference on Open Source Systems and Technologies (ICOSST)*, 2024, pp. 1–6. doi:10.1109/ICOSST64562.2024.10871147
- Haseeb Younis, Muhammad Azeem, Isabel Ronan, and Rosane Minghim, “MANet: A Deep Learning Framework for Multi-Cancer Microarray Analysis, Classification, and Visualization,” in *Proceedings of the International Conference on Frontiers of Information Technology (FIT)*, 2024, pp. 1–6. doi:10.1109/FIT63703.2024.10838450

- Haseeb Younis and Rosane Minghim, “Understanding High Dimensional Spaces through Visual Means Employing Multidimensional Projections,” *International Journal on Engineering, Science and Technology*, vol. 4, no. 2, pp. 170–179, 2022.
- Paul Trust, Provia Kadusabe, Haseeb Younis, Rosane Minghim, Evangelos Milios, and Ahmed Zahran, “UNLPSat TextGraphs-16 Natural Language Premise Selection Task: Unsupervised Natural Language Premise Selection in Mathematical Text using Sentence-MPNet,” in *Proceedings of TextGraphs-16*, 2022.
- Provia Kadusabe, Haseeb Younis, Rosane Minghim, Evangelos Milios, Ahmed Zahran, et al., “SNLP at TextGraphs 2022 Shared Task: Unsupervised Natural Language Premise Selection in Mathematical Texts using Sentence-MPNet,” in *Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing*, 2022, pp. 119–123.

1.4.2 Preprints and Unpublished Manuscripts

- Haseeb Younis, Hafiz Abdul Rehman, and Rosane Minghim, “Uncovering Clinically Relevant Cancer Subtypes from Somatic Mutations via Deep Learning, Autoencoder-Driven Clustering and Explainable AI,” *Manuscript under review*, 2025.
- Haseeb Younis, Luciana Daniele Trino Albano, Daniela C. Granato, Carolina M. Carnielli, Fabio Patroni, Adriana F. Paes Leme, and Rosane Minghim, “Integrative Proteomic and Machine Learning Approaches Reveal Protein Biomarkers and Interaction Networks in Oral Cancer Progression,” *Manuscript under review*, 2025.
- Paul Trust, Haseeb Younis, and Rosane Minghim, “SuBIS: Submodular Subset Selection with Importance Sampling for Data-Efficient Machine Learning,” *Preprint on Research Square*, 2024. doi:10.21203/rs.3.rs-3771212/v1

1.5 Additional Academic Contributions

Our contributions, however, were not limited to conference and journal publications. Throughout the course of this PhD, I have actively engaged in academic, professional, and outreach activities that supported the broader goals of

research dissemination, community engagement, and personal development. These contributions are summarised below:

- **Poster Presentations:** Presented different posters related to “*Cancer Biomarker Study*” at multiple forums, including the SFI ADVANCE CRT Research Colloquium and the School of Computer Science, University College Cork, between 2022 and 2025.
- **Professional Certification:** Successfully completed Level-9 the *Postgraduate Certificate in Innovation, Commercialisation and Entrepreneurship*, enhancing my ability to translate research into real-world impact.
- **Industry Placement:** Completed a professional work placement at **Intel Ireland Ltd.**, where I was awarded **nine individual performance awards** in recognition of research, innovation, and contributions to team and project success.
- **Leadership and Representation:** Served as a *Cohort Ambassador* for the ADVANCE Centre for Research Training (CRT), representing the PhD cohort in institutional and public events.
- **Recognition and Awards:** Received the *Best Idea Award* in the Postgraduate Idea Generation Competition, in recognition of innovative thinking in cancer diagnostics.

1.6 Thesis Structure

The structure of this thesis is as follows:

- **Chapter 1** introduces the motivation, thesis statement, contributions, publications, and structure of the thesis.
- **Chapter 2** discusses a foundational context for this thesis. Initially, it presents a historical overview of cancer, discussing its global impact and the limitations associated with traditional diagnostic approaches, while emphasising how bioinformatics innovations have provided significant advances to the field. Subsequently, the chapter defines essential concepts from biology and computer science that underpin the methodologies and analyses presented throughout this thesis, facilitating comprehension of the interdisciplinary research conducted.

- **Chapter 3** provides essential background on cancer biology, omics technologies, and machine learning methods used in this research, making it accessible to readers from both biological and computational domains.
- **Chapter 4** presents the analysis of somatic mutation data, describing pre-processing, desparsification, feature engineering, clustering, validation analysis, and classification methodology as well as results in that domain.
- **Chapter 5** details the modelling of transcriptomic data using deep learning techniques, model interpretation using XAI, and biomarker identification.
- **Chapter 6** explores proteomic data from MS, showing how protein complex-level patterns are used for classification and discovery of cancer-related disruptions.
- **Chapter 7** concludes the thesis and provides recommendations for future research directions.

Having outlined the motivation, objectives, and structure of this thesis, the next step is to establish the necessary conceptual foundation. Chapter 2 provides this context by providing a background of cancer and introducing key biological and computational principles that underpin the analyses developed in later chapters.

Chapter 2

Background

This chapter is organised into two main parts: First, we give a historical overview of cancer, discussing its global burden, the limitations of traditional approaches to cancer diagnosis and bioinformatics innovation; Second, we define the basic concepts related to the fields of biology and computer science used in this thesis.

2.1 Cancer Understanding and Analysis

2.1.1 Historical Overview of Cancer

Cancer has burdened the human population for millennia. Ancient Egyptian records provide the earliest known descriptions of cancer; for example, the Edwin Smith Papyrus (1600 BC) describes tumours of the breast treated by cauterisation, noting that “there is no treatment” for the disease [Fag15]. However, by the eighteenth century, physicians had recognised the invasive nature of cancer. Around 1870, the English surgeon Campbell De Morgan observed that cancer can spread from the primary tumour through the lymphatic system to distant sites, an insight that foreshadowed the modern concept of metastasis [GSS02]. By the mid-19th century, Rudolf Virchow’s pathological studies established that cancers are made up of cells (dispelling ancient “black bile” theories) and that all cells (including cancer cells) arise from other cells. Soon after, in 1902, Theodor Boveri hypothesised that cancer begins with defects in chromosomes following abnormal cell divisions [Bov08].

A major turning point in cancer research came in 1971 with the signing of the

U.S. National Cancer Act, launching the so-called “War on Cancer” and massively expanding funding for cancer research. This era led to the identification of numerous cancer-causing genes (e.g. the 1970s discovery of the first oncogene, *src*, and the 1980s discovery of the *Tumor Protein 53 (TP53)* tumour suppressor). This period also heralded a new generation of targeted therapies. Chemotherapy was introduced in the 1940s–1950s and improved over subsequent decades, while the advent of hormonal therapies, combined modality treatments, and radiation therapy (pioneered after the discoveries of X-rays in 1895 and radium in 1898) greatly improved cancer control [KP04]. More recently, the 21st century has seen breakthroughs in immunotherapy (such as immune checkpoint inhibitors and CAR-T cell therapy) along with the completion of large-scale initiatives like the Human Genome Project (HGP) and The Cancer Genome Atlas (TCGA), which have catalogued genomic alterations in dozens of cancers. These landmark developments have transformed cancer from a black-box disease into one of the best-characterised diseases at the molecular level [DJR12].

2.1.2 Global Burden and the Pressing Challenge of Cancer

Despite this progress mentioned in 2.1.1, cancer remains one of the world’s most pressing health challenges. Cancer is now a leading cause of death worldwide, accounting for nearly 10 million deaths in 2020 alone (about one in six deaths globally) [Wor25a]. According to Global Cancer Observatory (GLOBOCAN) estimates, 19.3 million new cancer cases were diagnosed in 2020, and 10.0 million people died from cancer that year [SFS⁺21]. These figures are expected to rise sharply due to population growth and ageing. By 2040, the global cancer incidence is projected to reach 28.4 million cases per year (a 47% increase from 2020) [SFS⁺21]. The burden of cancer is not only measured in lives, but also in economic cost and societal impact. The World Health Organization (WHO) has estimated the total annual economic cost of cancer at around \$ 1.16 trillion (based on 2010 data), representing roughly 1.5% of the world’s GDP [NBS⁺22].

Several factors contribute to cancer’s status as a global crisis. First, the incidence of cancer has been rising as lifestyles change (increased exposure to risk factors like tobacco, alcohol, and sedentary behaviour) and as longevity increases (cancer is primarily a disease of ageing). Second, many cancers are still diagnosed at advanced stages, where cures are difficult, leading to

high mortality. Third, treatment costs for newer therapies are often exorbitant, straining healthcare systems and patients equally. The combination of high fatality, high treatment cost, and growing case numbers makes cancer a formidable challenge for public health. Indeed, even in high-income countries, cancer control remains difficult; while survival rates have improved for some cancers, others (such as pancreatic, liver, and lung cancers) still have poor prognoses [DST⁺24]. In lower-income regions, limited screening and treatment capacity mean that cancers are often detected late, resulting in worse outcomes. In summary, cancer's global impact on lives lost, disability, and economic burden now rivals that of cardiovascular disease and is set to increase in the coming decades. This reality highlights an urgent need for new approaches to cancer prevention, early detection, and treatment [Wor25b, Wor25a].

2.1.3 Limitations of Traditional Approaches in Cancer Diagnosis and Classification

One reason cancer continues to have such a detrimental impact is that our traditional methods of diagnosing and classifying tumours have inherent limitations. Historically, cancers have been classified mainly by their organ of origin and microscopic anatomy, also known as histopathology. While the histological classification of tumours (e.g. grading and staging) has been foundational for oncology, it often fails to capture the full biological heterogeneity of the disease. Patients with “the same” cancer by site and histology can have markedly different clinical outcomes and responses to therapy; these circumstances are largely due to underlying molecular differences that are not apparent using routine pathology. Likewise, conventional single-analyte biomarkers such as Prostate Specific Antigen (PSA) for prostate cancer or CA-125 for ovarian cancer have improved cancer detection and monitoring, but many have suboptimal specificity and sensitivity. For example, elevated levels of some blood biomarkers may occur in benign conditions, leading to false positives, while truly early-stage cancers often go undetected by available markers. The low diagnostic specificity of many traditional cancer biomarkers limits their clinical utility [PABH⁺24].

In the past two decades, high-throughput profiling “omics” technologies have enabled a more detailed molecular characterisation of tumours, examining genomes, epigenomes, transcriptomes, proteomes, and more. However, much of earlier work in cancer genomics and precision medicine relied on a single

data type (for instance, Deoxyribonucleic Acid (DNA) sequencing alone). Such single-omics approaches, while valuable, provide only a narrow window into tumour biology [MG21]. Cancer is a complex, multi-factorial disease involving changes at multiple molecular levels. Focusing on one level at a time (e.g. genes, transcripts, or proteins in isolation) can miss critical interactions between levels. The analysis of a single omics layer often fails to capture the interplay between genomic mutations, gene expression changes, protein activation, and metabolic alterations that collectively drive cancer progression [HZZ24]. In other words, single-omics data provide an incomplete picture, potentially leading to an incomplete or even misleading understanding of a tumour's behaviour. This limitation hinders our ability to accurately classify cancers into clinically relevant subtypes. For instance, in breast cancer, traditional subtypes defined by histology and a few receptors (Estrogen Receptor (ER), Progesterone Receptor (PR), Human Epidermal Growth Factor Receptor 2 (HER2)) do not fully explain the observed variation in outcomes; a patient's prognosis and optimal treatment might be better determined by a combination of genomic and transcriptomic features that cut across those classical subtypes [HZZ24].

Additionally, relying solely on clinical and histopathological features can limit early diagnosis. Subtle molecular changes signalling a high-risk lesion or an aggressive tumour clone may not manifest as observable changes until the cancer is advanced. This recognition has spurred the movement toward analysis of independently different omic strategies for cancer classification and biomarker discovery. Thus, new diagnostic approaches are needed to detect cancers at earlier, more treatable stages by using molecular information.

2.1.4 The Promise of Independent-Omics Analysis in Unravelling Cancer Biology

Independently, “omics” approaches have emerged as a powerful way to overcome the limitations of single-modal data. By simultaneously examining the genome, transcriptome, and proteome, often along with other layers such as epigenome or metabolome of cancers, researchers can obtain a far more comprehensive view of the biological processes driving tumour behaviour. Each omics layer provides a different perspective: for example, **genomic** DNA sequencing reveals the mutations and structural variants present; **transcriptomics** shows which genes are actively expressed; **proteomics** indicates the actual effector molecules and pathways active in the cell. These layers are

interdependent; a DNA mutation may or may not cause changes in Ribonucleic Acid (RNA) and protein; conversely, protein network activity might be influenced by epigenetic modifications, or microRNAs that are not evident from DNA sequence alone. Investigating omics data allows us to capture regulatory relationships and causal chains across the molecular hierarchy [SNC⁺19]. Data integration across different omics platforms “provides opportunities to understand causal relationships across multiple levels of cellular organisation”, so that no single data type can reveal relationships [MG21]. Together, these omics approaches map directly onto the central dogma of molecular biology DNA, RNA, and protein, highlighting how each level contributes to the flow of genetic information and ultimately to the regulation of life processes.

Studies have shown that omics analysis can stratify cancers into subgroups with greater clinical relevance than traditional methods. For example, TCGA [HZ18] projects applied multi-platform genomic, epigenomic, and proteomic profiling to thousands of tumours. It has also identified novel subtypes in cancers like breast, lung, and gastric cancer that correlate with prognosis or treatment response better than classical taxonomy. In general, combining multiple layers of data yields more robust and accurate tumour classifications and risk predictions [HZZ24]. One reason is that integrative analysis can detect convergent evidence of a disease mechanism. For example, a particular signalling pathway might be aberrantly activated in a tumour; this circumstance could be evidenced by a gene mutation in that pathway, overexpression of downstream genes, plus elevated phosphorylation of pathway proteins. Each piece of evidence might be important on its own, but together they paint a consistent picture describing this pathway as a driver in that tumour. Omics investigation effectively “connects the dots” in this way, increasing confidence in identified biomarkers and therapeutic targets. Various computational frameworks (e.g. network-based methods, matrix factorisation, machine learning models) have been developed to integrate omics datasets; they consistently demonstrate improved performance in patient stratification over single-omic analyses. As one study succinctly put it, “more is better” when it comes to omics data for understanding complex diseases [HZZ24].

These independent omics approaches have also been crucial in deciphering mechanisms of drug resistance and identifying potential combination therapies. For instance, if a cancer genome analysis indicates no actionable mutation, transcriptomic and proteomic data might still reveal dysregulated pathways “a stress-response pathway or immune evasion pathway”, that could be targeted

by drugs. In another scenario, a targeted therapy might fail due to adaptive changes at the protein level that would only be caught by proteomic profiling. By examining tumours at multiple molecular levels before and after treatment, researchers can pinpoint how cancers adapt, and thereby propose interventions to counteract resistance.

Importantly, the holistic insight gained from different omics can highlight novel biomarker molecules or signatures that indicate the presence, subtype, or aggressiveness of a cancer. Some modern biomarker signatures combine different data types, for example, a “Gene Expression (GE) plus DNA Methylation (DM) signature” that is more predictive than either alone. In sum, independent omics integration supplies a panoramic view of tumour biology, capturing the complex networks and feedback loops of cancer cells. This systems-level understanding is key to identifying the Driver Gene of malignancy amid the sea of molecular changes. By leveraging multiple omics data, scientists are beginning to unravel previously obscure aspects of cancer biology, from the influence of tumour microenvironment and immune interactions to the metabolic vulnerabilities of cancer cells. The promise of multiple omics is that it can reveal emergent patterns and biomarkers that single-dimensional analyses would miss, ultimately leading to more precise diagnostics and tailored therapies.

2.1.5 Innovations Through Bioinformatics, Machine Learning, and Data Visualisation

Extracting meaningful knowledge from omics data is a grand computational challenge. Each tumour sample can easily generate gigabytes of data across different omic platforms; cohorts like TCGA involve thousands of samples. Here, bioinformatics and data science are indispensable. Bioinformatics provides the tools and algorithms to process raw sequencing or spectrometry data into usable profiles (e.g. calling genetic mutations, quantifying GE, and identifying protein abundances). It also offers databases and knowledge bases that annotate these molecular features, for example, identifying which genes are oncogenes or which proteins belong to known pathways. Moreover, specialised bioinformatics methods, such as algorithms that construct network models linking DNA variants to RNA and protein changes, methods that combine data by matrix integration, or factor analysis to find common sources of variation, have been developed to integrate heterogeneous data. These techniques allow researchers to find signals in the noise and highlight candidate biomarkers that correlate

with clinical outcomes.

Machine learning approaches not only improve classification accuracy but can also generate new biological hypotheses. Techniques like feature importance scoring or layer-wise relevance propagation can highlight which combinations of genes or proteins were most influential in a model's prediction, pointing researchers to potential key drivers or biomarkers. Furthermore, machine learning models can help in dimension reduction (e.g. compressing tens of thousands of genes into a few composite variables) and in the imputation of missing data, both common needs in multi-omics integration. As computational power and algorithms have advanced, training such models on large cancer datasets has become feasible. These models can rapidly sift through multi-modal data and may discover signals like a subtle expression pattern that human analysts or simpler methods would not notice. In doing so, machine learning accelerates the pace of discovery in cancer biomarker research [WK23].

Data visualisation also plays a key role in the clinical adoption of complex biomarker models. Clinicians are more likely to trust and use multiple omics predictive models if their output can be presented in a clear, interpretable format. For instance, heatmaps or cluster dendrograms can show integrated clustering of samples across data types; network graphs can illustrate molecular interactions or regulatory relationships supported by omics evidence. Interactive visual analytics play an important role in understanding the inter- and intracellular behaviours, where clinicians can review a patient's multi-omic profile via dashboards that integrate genomics (mutations, copy-number changes), transcriptomics (GE levels), and proteomics (protein or phospho-protein levels), along with suggested treatments. In essence, advanced visualisation closes the loop between raw data and actionable knowledge, enabling the translation of these omics and machine learning findings into real-world decisions. Together, bioinformatics, machine learning, and visual analytics form a synergistic triad that is driving innovation in cancer biomarker discovery. Bioinformatics provides the data processing and integration backbone, machine learning provides the pattern-finding and predictive power, and visualisation provides interpretability and human insight.

2.2 Essential Background Concepts

In this section, we describe the basic concepts in a simple and straightforward manner, using the terminology and approaches employed in this thesis to clarify them for biology and computer science domain experts. The detailed biological concepts are given in the book named as *Genome 5* [Bro23].

2.2.1 Biology Related Concepts

A *gene* is a unit of DNA that carries instructions to build proteins, the molecules essential for cellular function. *Gene expression* is the process by which these DNA instructions are converted into functional proteins, determining the biological behaviour of cells. Studying gene expression patterns, especially differences between cancerous and normal cells, helps identify potential biomarkers and understand cancer mechanisms [Sta04].

The *genome* is the complete set of DNA in an organism, including all of its genes and non-coding sequences. It contains the instructions for building and maintaining the organism. In humans, the genome consists of approximately 3 billion DNA base pairs organised across 23 pairs of chromosomes. The HGP successfully mapped the entire human genome, enabling large-scale studies of genetic variations linked to diseases such as cancer [Bro23].

The *epigenome* consists of chemical modifications to DNA and histone proteins that regulate gene activity without altering the DNA sequence itself. These modifications, such as DNA methylation and histone acetylation, affect how tightly DNA is packaged and whether specific genes are turned on or off. Epigenetic changes can be influenced by environmental factors and are often involved in cancer development by silencing tumour suppressor genes or activating oncogenes [MRB05].

Mutations are changes in the DNA sequence that can affect how cells function. *Copy number changes* occur when sections of DNA are duplicated or deleted, altering the number of gene copies. *Protein abundance* refers to how much of a specific protein is present in a cell or tissue [GCM10].

Somatic mutations are genetic alterations that occur after birth, affecting the DNA of body cells, but are not inherited from parents. In cancer, these mutations accumulate and can alter cell function, potentially leading to tumour development. Identifying these mutations is crucial for understanding cancer

progression and personalising treatments [MC15].

Proteins perform essential biological functions within cells, such as signalling and metabolism. Often, proteins form groups called *protein complexes*, which work together to execute specific tasks. Changes in protein complexes, like their assembly or function, are associated with diseases, including cancer, and understanding these complexes can reveal new therapeutic targets [PRT⁺20].

A *biological pathway* is a series of interactions between molecules (like proteins and genes) that lead to specific biological functions or changes. Cancer often involves disruptions to these pathways, making pathway analysis crucial for identifying how normal cellular functions go awry during disease progression [SHZ⁺19].

Transcriptomics involves studying RNA molecules produced by genes, reflecting which genes are actively expressed at a given moment. Microarrays measure the expression levels of thousands of genes simultaneously, allowing researchers to identify gene patterns associated with specific cancer types or subtypes [ZFLB⁺14].

Proteomics studies the complete set of proteins expressed by cells. Mass spectrometry is a technique that precisely measures the abundance and interactions of proteins. This method helps reveal changes in protein expression and complex formation in cancer, providing insights into disease mechanisms and biomarker discovery [AM03].

Large-scale public databases like *TCGA* and *Curated Microarray Database (CuMiDa)* store extensive omics datasets across various cancers. These databases are crucial resources enabling comprehensive, comparative analyses and validation of computational methods in cancer research [DAAC20, PRZD⁺19].

Analysing biological data through *pathway and network-based approaches* provides insights into the collective behaviour of molecules, revealing how disruptions in networks or pathways contribute to cancer. This systems-level understanding is crucial for identifying potential therapeutic targets and biomarkers [ZLX⁺16, RHP⁺20].

A *biomarker* (short for biological marker) is a measurable indicator of a biological state, condition, or process. In the context of cancer, biomarkers can be molecules such as DNA mutations, RNA transcripts, proteins, or metabolites

that are associated with the presence, subtype, progression, or response to the treatment of the disease. Biomarkers play a critical role in modern medicine by supporting early diagnosis, predicting prognosis, guiding therapeutic decisions, and monitoring disease recurrence or treatment response [GMB⁺02, HH12].

Driver genes are genes where mutations promote cancer by helping cells grow or survive abnormally. *Passenger genes* are also mutated but do not affect cancer development. Identifying driver genes is important because they reveal how cancer starts and may suggest targets for treatment or diagnosis [PM15, HS07].

A *cancer type* refers to the organ or tissue where the cancer begins, like the lung, breast, or skin [HD17]. A *cancer subtype* describes specific molecular or cellular features within a cancer type, which can affect how the cancer behaves and responds to treatment [CBCB19].

2.2.2 Computer Science Related Concepts

Machine learning involves training algorithms to recognise patterns and make predictions from data without explicit programming. In cancer research, machine learning methods classify tumour types, predict outcomes, and identify biomarkers from complex biological datasets, significantly enhancing analytical accuracy and efficiency [ENM15, Zho21].

Deep learning is a subtype of machine learning involving Artificial Neural Network (ANN) with multiple layers that automatically extract patterns from complex data. Models like CNN and Autoencoders excel in discovering intricate features within omics datasets, facilitating accurate classification and biomarker discovery [LBH15].

A *CNN* is a deep learning model specifically effective at capturing local patterns in structured data, such as gene expression arrays. CNNs automatically learn relevant features from data, making them suitable for tasks like cancer subtype classification [LLY⁺21].

An *Autoencoder* is a neural network designed to compress (encode) data into a simpler representation, then reconstruct (decode) it. This process reduces data complexity, highlighting essential features and facilitating clearer data interpretation, especially useful in high-dimensional omics data [LPL23].

Transformers are advanced neural networks that effectively capture relationships within data through attention mechanisms. Initially developed for lan-

guage processing, transformers now assist in modelling complex interactions within biological datasets, enhancing prediction accuracy and interpretability [ENS24].

Feature selection methods (e.g. Boruta, SelectKBest, Recursive Feature Elimination) identify the most informative features in large datasets, eliminating irrelevant or redundant data. This elimination reduces computational complexity and improves model accuracy, crucial in high-dimensional biological data analyses [LCW⁺17].

SMOTE is a technique used to balance imbalanced datasets by creating synthetic examples of minority classes. This technique ensures that machine learning models correctly learn from all classes, improving their predictive reliability, especially in datasets where some cancer types are rare [BL13].

XAI techniques, such as SHapley Additive exPlanations (SHAP) and LIME, help interpret machine learning model predictions by clarifying the impact of individual data features on decisions. These insights enhance the transparency and trustworthiness of computational models in biomedical research [DDN⁺23].

Visualisation techniques like *Uniform Manifold Approximation and Projection (UMAP)* and *Principal Component Analysis (PCA)* [MR93] simplify high-dimensional data, making complex patterns visually interpretable. These methods are essential for verifying biological relevance and model robustness in omics studies [Liu20, RC22].

Classical machine learning algorithms, such as *Support Vector Machine (SVM)* and *Random Forest (RF)*, classify data based on defined mathematical principles. They have been widely used due to their simplicity, interpretability, and effectiveness in handling high-dimensional data in cancer studies [Zho21].

While Chapter 2 presented the essential disciplinary concepts, it did not yet position this work within the broader scientific landscape. Chapter 3 addresses this by critically reviewing existing research in cancer biology, omics technologies, and machine learning, highlighting both the opportunities and the methodological gaps that motivate the approaches adopted in this thesis.

Chapter 3

Literature Review

This chapter is organised as follows: First, we survey literature on independent omic data for cancer subtype classification and biomarker discovery. To conduct the literature review, a snowballing approach was used, which enables the identification of additional relevant studies through both backward and forward reference searches. Next, we examine deep learning applications (CNN-based and others) in cancer analysis and approaches for making these models interpretable via XAI. We then present a comparative overview of somatic mutation, proteomic (co-fractionation), and transcriptomic data modalities, including a formulated section of their respective advantages and limitations, along with the added value of their integration. Recent breakthroughs in identifying biomarkers through integrated omics and machine learning are highlighted.

3.1 Overview

Cancer is a complex disease driven by alterations at multiple molecular levels, including the genome, transcriptome, and proteome. It is not a single disease but a collection of hundreds of distinct types and subtypes, each characterised by diverse molecular alterations and clinical behaviours [CPLZ22]. This heterogeneity complicates diagnosis and treatment, as tumours of the same histological type can have vastly different genetic drivers and therapeutic responses. Effective biomarkers, measurable molecular indicators of cancer presence or subtype, are crucial for early detection, accurate classification, and personalised therapy. However, identifying robust biomarkers is difficult due to the complex, multi-factorial nature of cancer, which involves changes at the genomic, transcriptomic, and proteomic levels. Single-omics studies (focusing on one level of

data) often provide only a narrow view of this complexity. By examining multiple molecular layers together, researchers can uncover complex dysregulation patterns and interactions that single-omics analyses miss.

Recent advances in high-throughput *omics* technologies have enabled profiling tumours from multiple angles, including DNA mutations, Messenger Ribonucleic Acid (mRNA) expression, and protein abundances. Each data type offers unique insights: DNA somatic mutations can reveal the initiating genetic events; GE patterns reflect the functional state of the tumour; and proteomic data show the effector molecules and pathways at work. In this thesis, we leverage all three data types, including **genomic**, **transcriptomic**, and **proteomic**, to discover and classify cancer biomarkers. In the following sections, we introduce each omics data type and justify the importance of analysing them individually, before discussing how deep learning techniques enable the discovery of patterns in these high-dimensional datasets. We then highlight the rationale for integrating multiomics findings and emphasise the role of XAI and visualisation in ensuring that our models remain interpretable.

3.1.1 Somatic Mutation Data in Cancer

Genomic alterations are a hallmark of cancer, and profiling somatic mutations (DNA mutations acquired in tumours) has been fundamental to understanding and classifying cancers. Somatic mutation data, obtained from whole-genome or exome sequencing of tumours, pinpoint changes such as single-nucleotide variants, insertions/deletions, and structural variants. These mutations can drive oncogenesis (e.g. mutations in TP53 or EGFR) and often serve as biomarkers or therapeutic targets. Analysing mutation data in isolation is important because it directly identifies the genomic “ground truth” of a tumour, the irreversible DNA changes that may initiate cancer. Effective treatment and individualised therapy for cancer patients depend on early detection and correct diagnosis. Cancer can develop in different cell types and tissues. The leading causes of cancer are genetics, lifestyle factors, and the environment. It is essential to understand the role of genetics in the individualised manifestations of cancer.

Cancer detection and diagnosis rely heavily on morphological descriptors [VLP⁺18, ACQ⁺20, SK24, GHL⁺24]. For instance, a pathologist uses a microscope for specific morphological features that are linked to breast cancer when a sample of breast tissue is obtained through a biopsy or surgical

procedure [MON⁺00, All12, XZD⁺25]. These features include abnormal cell growth, changes in cell size or shape, and the presence of cancerous cells. Despite its continued importance as a diagnostic and screening tool for cancer, morphological examinations have certain drawbacks, including inter-observer variability, limitations in detecting subtle changes, and potential for misdiagnosis due to tumour heterogeneity [MNA⁺22, dNO20, GJW⁺21, OE22]. There are limitations to morphological characteristics in the diagnosis and detection of cancer [MMM⁺21].

The identification of cancer subtypes can help to classify patients into the same prognosis, clinical phenotype or treatment response group; for example, cancer classification can be performed based on histopathological examination of the absence or presence of IHC markers such as PR, ER, and HER2. The classification of breast cancer into four molecular subtypes is also based on these markers. These different subtypes respond to different hormone therapies and have different prognoses [HKDL⁺12]. Furthermore, these molecular classifications have discovered subclasses of PR-positive and ER-positive breast cancer types, such as Luminal A and Luminal B. To monitor the effective use of the monoclonal antibody trastuzumab in breast cancer treatment, the high expression level of the circulating tumour factor HER2 is used [GJKS10, SMAN10].

Recently, NGS techniques [RF09] have been used for cancer studies that include whole-exome sequencing [RTM14], whole-genome sequencing [NK10], whole-transcriptome sequencing [YK15], epigenomics [GT08] and targeted gene sequences [GSP⁺06]. NGS provides the base pair resolution of genomic and genetic information in cancer. With the help of these techniques, the molecular mechanism of cancer can be investigated. From all of these datasets, somatic mutation data is instrumental in discovering cancer mutations and genes [JWC⁺14, LSM⁺14, JPZ14]. Several large databases contain somatic mutation data. These databases include Wellcome Trust Sanger Institute's Cancer Genome Project [Doo06], TCGA [DBSP16], and International Cancer Genome Consortium (ICGC) [ZBA⁺19].

Computer technology is developing quickly; computational methods are becoming increasingly important in detecting cancer. Different studies are using various kinds of techniques and different modalities to improve the diagnosis and prognosis of cancer. Danaee et al. [DGH17] introduced a deep learning technique that employs a Stacked Denoising Autoencoder (SDAE) to find genes that may successfully distinguish between tumour and healthy cases of

breast cancer. Jia et al. [JCC⁺21] examined BRCA data from the GEO [CB16] and TCGA. The performance of the current models is not without severe challenges. Selecting the ideal feature representation and architecture, appropriate model parameters, and proper weight and bias values are crucial elements in enhancing performance [MX22, KM23, ZGT⁺22]. Furthermore, selecting the optimal regularisation parameters and learning rates can significantly impact the model's ability to generalise to new data.

Various other studies have focused on skin cancer classification using images and employing deep learning and transfer learning [YBA19, XWL⁺21, YJ22, AAA⁺22a], and XAI with breast and lung cancer data [KAU⁺24, MCA⁺24, WKB24]. XAI refers to a set of tools and approaches meant to make the judgments and predictions of advanced machine learning models more intelligible and interpretable to humans. Traditional AI models, especially deep learning models, typically function as “black boxes”, producing very accurate predictions without providing insight into how those predictions were formed. XAI seeks to overcome this gap by tracing the input that most influenced the outcome [DDN⁺23]. LIME [SREG⁺24] is a method in XAI to deliver human-readable explanations for predictions made by deep learning models. Some studies have focused on somatic mutation data for cancer classification; Zongzhen et al. [HZYZ21] propose an adapted Multiple Kernel Learning (MKL) method to classify breast cancer using somatic mutation data, achieving an Area Under the Curve (AUC) of 94.39. Quan et al. [LRC⁺22] proposed the oncogenic prioritisation artificial intelligence (OPAI) algorithm to classify cancer based on 12 evidence features and 23 functional features, and achieved a maximum of 85.4 AUC. Sanjaya et al. [SMK⁺23] presented an attention-based deep neural network for cancer type and subtype classification on the somatic mutation dataset and achieved a maximum score of 97.6% accuracy. Anaya et al. [ASMB24] proposed Multiple Instance Learning (MIL) for somatic mutation data and performed experiments on 24 cancer types; they achieved the overall highest accuracy of 98%, and for BRCA specifically, 55%. Similarly, Wang et al. [WTZQ25], proposed a deep learning based model, M-MNet, for the classification of the Pan-Cancer database using somatic mutation data and achieved 94.62% of accuracy. They also compared their methods with those of other pre-trained networks.

The comparison of all the above-mentioned studies is given in Table 3.1.

These methods are limited by the fact that somatic mutation data are very

Table 3.1: Comparison of Existing Cancer Classification Studies Using Genomic Data

Study	Year	Method	Classification	Data Type	Results
Zongzhen et al. [HZYZ21]	2021	MKL	BRCA	Somatic mutation	AUC = 94.39
Dikaos et al. [Dik22]	2022	SVM	32 Tumors	Somatic Mutations	Accuracy = 73%
Quan et al. [LRC ⁺ 22]	2022	OPAI	BRCA	Somatic mutation	AUC = 85.4
Sanjaya et al. [SMK ⁺ 23]	2023	MuAt	WGS	Somatic mutation	ACC = 97.6
Anaya et al. [ASMB24]	2024	MIL	24 cancer types	Somatic mutation	ACC = 98% and 54% for BRCA
Wang et al. [WTZQ25]	2025	MMNet	Pan-Cancer	Somatic mutation	ACC = 94.62%

sparse, which makes the classification of cancer subtype computationally expensive and challenging. Morvan et al. [LMZV17] proposed a network-based approach to desparsify somatic mutation data. They used the network from Pathway Commons by removing non-essential mutations and adding missing ones. Marieke et al. [KPS⁺18] presented a tool named Subtyping Agglomerated Mutations By Annotation Relations (SAMBAR) to desparsify somatic mutation data. This tool summarises data into pathway scores. Another limitation of using these studies is their reliance on deep learning algorithms; although these algorithms are highly precise, they do not typically output gene information, thereby losing important biomarker observations.

3.1.2 Transcriptomic (RNA Expression) Data in Cancer

GE profiling offers a dynamic view of tumour biology by measuring mRNA levels of thousands of genes simultaneously (via microarrays or RNA-seq). Tumours can be classified by their expression signatures, as different cancer types or subtypes often exhibit characteristic GE patterns reflecting their cell of origin and activated pathways. Analysing transcriptomic data independently is valuable because it captures regulatory and phenotypic changes that DNA mutations alone may not reveal. For example, molecular subtypes of breast cancer (e.g. basal-like vs. Luminal A) were originally defined by distinct mRNA expression profiles, leading to improved prognostication and treatment strategies. Expression biomarkers (such as gene signatures like Oncotype DX) are used clinically to predict outcomes and guide therapy in certain cancers [SDBLM18]. Furthermore, GEs are used to identify more specific breast cancer subtypes in triple negative breast cancer [LBC⁺11]. Numerous techniques are developed based on information on GE to precisely predict cancer [OMD⁺22, CKP⁺22, APPS22]. However, the main limitation of this method is its inability to assign samples to specific molecular subtypes consistently [WBRF10, OBA23]. The main reason for this inconsistency and subsequent misclassification is the dynamic nature of GE in individuals.

The emergence of NGS and improvements in microarray technology have made patient Gene Expression Profiling (GEP) more accessible. This accessibility has led to the creation of GE datasets related to different diseases. Personalised medicine has seen a dramatic change due to this move away from descriptive “morphological” classification methods and toward a more all-encompassing approach that takes immunohistochemistry biomarkers and

clinical features into account. GEP is already widely used in standard clinical practice [TDZ⁺22]. Cancer researchers have thoroughly studied GEP, and clinical oncologists are beginning to incorporate the results of these investigations into their routine operations. Additionally, mining GE-level data has been beneficial for the early diagnosis and treatment of various cancer types [Azi22]. Numerous techniques are developed based on information on GE to precisely forecast cancer [OMD⁺22, CKP⁺22, APPS22]. Using GE data, several machine learning, deep learning, and metaheuristic approaches have been developed and used to identify and categorise cancer.

Single and dual-channel microarrays are being increasingly replaced by RNA-seq as the primary method to assess GEP in biological investigations [GSB20]. Depending on the chosen NGS technology and the RNA to be examined, RNAs of interest are isolated from a biological sample, and an RNA fragment library is assembled. According to Hardlickova et al. [HTT17] and Berge et al. [VdBHS⁺19], this stage typically entails the isolation of the RNA molecule, complementary DNA (cDNA) production by reverse transcriptase reaction, random amplification by polymerase chain reaction, and inclusion of sequencing adapters. There are several sequencing platforms, and their choice must consider the analysis procedure and the experiment aims since they exhibit varying performance [LCC⁺12, JPP⁺21, FCS⁺16]. Currently, Illumina is state-of-the-art with its synthesis sequencing-based technology [HCMD21, BNR⁺25]. This method uses unique fluorescently tagged reversible terminator Deoxynucleotide Triphosphates (dNTPs), which are identifiable by the fluorescent colour they generate [GS11]. Each cycle of dNTPs contains one nucleotide. While mRNA is the most researched type of RNA, other functional varieties are just as important to comprehend the cell's regulatory apparatus, even if they do not code for proteins. RNA-seq stands out because it can select different RNAs, such as mRNA, Long Non-Coding RNA (lncRNA), and circular RNA, by varying RNA extraction and isolation protocols [RTSB18]. This ability enables the creation of unique protocols that go beyond simply formulating probes to identify individual RNA types [LL18].

Due to this flexibility, businesses have invested in developing the technology, and as a result, RNA-seq investigations may now be carried out at prices that are competitive with microarray tests [ZFLB⁺14]. Consequently, it is anticipated that RNA-seq would present additional data-handling issues due to its completely distinct character from microarray results [Zar15, GI23]. The first challenge is the rigorous preprocessing necessary for precise analysis of RNA-

seq. RNA-seq addresses read abundance as opposed to microarrays, which quantify the expression of a specific RNA in a biological sample through the appropriate hybridisation of cDNA with its probe [EB00, BGE01, Bla03]. It would be completely incorrect only to use the raw GE data as a result. However, in order to use RNA-seq, one must analyse the quality of the sample, eliminate low-quality bases, eliminate experimental artefacts, remove any remaining ribosomal RNA, estimate transcript-level abundance, and normalise RNA-seq read counts. Therefore, before implementing any AI methods, rigorous data processing must be performed. The second challenge lies in developing an algorithm capable of precisely examining such data. In general, sample quality analysis, background correction, and normalisation must be properly applied to a machine learning approach when working with microarray data sets [FCGD19].

With respect to the file input, RNA-seq raw data are given as non-negative elements and integer-valued counts rather than log-intensities like microarrays do [Zar15]. Therefore, it becomes essential to provide an input matrix that has undergone all preprocessing stages before the analysis stage provided by machine learning. Furthermore, the “curse of dimensionality”, which is linked to data sets with a large number of features but few samples, is very common in biological datasets [VF05]. High dimensionality often results in the overfitting of the model, which impairs its generalisation and increases computation time. Therefore, the development of more accurate algorithms is closely tied to the availability of high-quality data.

A detailed overview of recent cancer research efforts that use GE data from different cancer kinds was provided by Khalsan et al. [KMAS⁺22]. They covered a number of applications of machine learning in cancer research, including the use of RNA-seq and microarray data, cancer prediction and classification. Yuan et al. [YLZ20] used GE data to use several machine learning techniques for the identification of lung cancer. Wang et al. [WLZ⁺18] presented a unique computational approach to diagnose breast cancer by combining SVM, dagging [MKM21], rough set-based rule learning, RF, and Monte Carlo Feature Selection [DRIE⁺08]. Alshareef et al. [AAA⁺22b] suggested a deep learning model for prostate cancer detection in conjunction with an AI-based feature selection approach (AIFSDL-PCD) utilising GE data.

In recent years, notable advancements in cancer prediction have been witnessed using deep learning and machine learning techniques based on GE data. To address the above-mentioned challenges, this work seeks to develop a state-of-

the-art framework for the use of CNN for classifying GE data by identifying an accurate prediction model and employing metaheuristic techniques to optimise the CNN model. Optimisation algorithms that repeatedly explore a vast search space and refine potential solutions are known as metaheuristic algorithms [ABAFS18]. They can provide near-optimal solutions in a reasonable period of time for NP-hard problems, which are computationally intractable issues that cannot be solved with exact methods [RSO⁺21, TBLA20, MD20]. Meanwhile, large-scale bioinformatics optimisation problems have been found to be effectively solved by metaheuristic optimisation techniques. Since many of these issues fall under the NP-hard category, academics have mostly relied on metaheuristic approaches to solve them. Large-scale sample problems can be solved effectively with the use of metaheuristic techniques, which also minimise the need for computer resources. Although a variety of optimisation techniques exist, metaheuristic algorithms are valuable for addressing such problems because of their adaptability in producing superior solutions within very short processing times [CBJP23]. The challenges of high dimensionality, complex variable correlations, and noisy data specific to GE data can all be solved with the use of metaheuristic models. Additionally, by utilising strategies like randomisation and simulated annealing to break out from local optima, metaheuristic models can manage noisy and non-linear data [STRC20].

An ANN-based metaheuristic approach for classifying skin diseases has been presented by Chakraborty et al. [CMC⁺17]. In MotieGhader et al. [MMSAMN20], mRNA and micro-RNA expression data were utilized to identify breast cancer using the combination of an SVM classifier and metaheuristic techniques such as genetic algorithms [OM01], world competitive contest [MSM16], particle swarm optimization [ZWWC11], cuckoo optimization [GYA13], imperialist competitive algorithm [KK17], learning automata [LÖJ17], heat transfer optimization algorithm [PS15], ant colony optimization [DBS06], forest optimization algorithm [GFD14], discrete symbiotic organisms search [EP19], and league championship algorithm [Kas14]. As a result, the suggested algorithm chose 186 mRNAs out of 9,692 and 116 miRNAs out of 489, which produced results with accuracy levels above 90% for the miRNAs dataset and 100% for the mRNA dataset. A Generative Adversarial Network (GAN) model on cancer genomic data was proposed by Wei et al. [WLH⁺22]. They employed twelve distinct GE data sets from the TCGA. Additionally, they employed a reconstruction loss to improve stability when training the model. Their suggested model obtained a 92.6% accuracy rate.

A two-stage gene selection model was presented by Deng et al. [DLDW22] for the categorisation of cancer in microarray datasets. Their method coupled gradient boosting (XGBoost) with a multi-objective optimisation genetic algorithm (XGBoost-MOGA). In the first step, a set of genes most pertinent to the class is left behind after irrelevant genes are efficiently ranked using XGBoost-based feature selection. In the second step, multi-objective optimisation with XGBoost-MOGA is used to identify a subset of optimal genes from the collection of the most relevant genes. A comparative study was conducted between the suggested method and other cutting-edge feature selection techniques, utilising two commonly used learning classifiers on 14 publicly available microarray datasets. The outcomes showed that XGBoost-MOGA performed better than earlier techniques in terms of recall, accuracy, F-score, and precision. Elbashir et al. [EEMS19] proposed a CNN architecture for the classification of GE data from PAN-Cancer Atlas and achieved 98.76% accuracy. They transformed their data into 2D images before passing it to the model, and this transformation loses the information about the biological structure and the relationships among genes.

By integrating the Barnacles Mating Optimiser [SMS⁺18] algorithm with an SVM, also known as BMO-SVM, Houssein et al. [HAH⁺21] managed to select genes that contribute to the prediction of cancer from GE datasets with the best accuracy based on microarray data. They assessed the suggested model using four benchmark microarray datasets: leukemia1, lymphoma, a small-round-blue-cell tumour (SRBCT), and leukemia2. According to their findings, the suggested BMO-SVM technique outperformed other well-known techniques, including GA, Artificial Bee Colony [Kar10], and Tunicate Swarm Algorithm [KASD20]. An approach for gene selection called the Improved Whale Optimisation approach [CSSS22] was proposed by Devi et al. [DP23]. A multi-objective fitness function that strikes a balance between feature selection and error rate minimisation was employed in the suggested approach. The findings demonstrate that the proposed method, which used the Gradient Boost Classifier to obtain a limited selection of genes needed for the BRCA classification, accomplished 97.7% accuracy. Similarly, Mohamed et al. [MEFD⁺23] proposed a CNN architecture combined with EOSA to classify BRCA data from TCGA. This architecture outperformed all the techniques mentioned above. Wang et al. [WHH⁺24] used cfRNA data for biomarker identification and cancer classification; using RF and Linear Regression (LR), their solution achieved 90.5% AUC.

A GAN model on cancer genomic data was proposed by Wei et al. [WLH⁺22].

They employed twelve distinct GE data sets from the TCGA. Additionally, they employed reconstruction loss to enhance stability during model training. Their suggested model obtained a 92.6% accuracy rate. Additionally, Jagadeeswararao et al. [JS24] used multiple machine learning algorithms, including Decision Tree (DT) [Qui96], RF [LWZ12], k-Nearest Neighbour (KNN) [LL14], Multilayer Perceptron (MLP) [TM18], SVM [PS20], and LR [LaV08] to classify Pancreatic Adenocarcinoma (PAAD) using GE data from TCGA and achieved 96% accuracy. Similarly, Alharbi et al. [AVZ⁺25] proposed a graph neural network-based architecture termed LASSO-MOGAT for classifying 31 cancer types and normal tissue, comprising messenger RNA, microRNA, and DNA methylation data. They integrated these datasets into multi-omics data and achieved an overall performance of 95.9%.

One very important aspect of RNA-seq data is reliability and uniformity, as previously mentioned. In order to tackle that, Feltes et al. [FPD21] proposed a dataset “BARRA:CuRDa”, which is made up of 17 carefully selected RNA-seq datasets for humans taken from the GEO and subjected to stringent filtering standards. Each dataset was subjected separately to analyses of sample quality, low-quality base removal, ribosomal RNA removal, experimental process artefacts, and transcript level abundance. The result is a curated RNA-seq data set that works as a benchmark for human cancer classification experiments. The same work also used multiple machine learning algorithms, including DT, RF, KNN, MLP, and SVM to classify the cancer profiles. The comparison of all the above-mentioned studies is given in Table 3.2.

The main limitations of these studies revolve around their use of TCGA Data, and therefore, the comparison between conditions may be impacted by the fact that raw “fastq” normal samples are not publicly available in TCGA for all institutions; the lack of public availability of raw “fastq” at TCGA hinders the ability to do high-quality custom analysis and statistical treatments. Of critical importance, most studies use distinct algorithms, which perform differently for different data sets, suggesting that the results are hard to reproduce for new data sets of the same kind. Additionally, studies that use deep learning algorithms, although highly precise, do not usually output gene information, losing important biomarker information. A method that can handle different types of cancers without changing the methodology and parameters every time is much needed. Last but not least, most of these studies are focused on cancer subtype classification, not cancer type classification.

Table 3.2: Comparative Analysis of Relevant Studies

Year	Study	Method	Result	Optimized	DSA	SD	MCT	SA
2017	Danaee et al. [DGH17]	SDAE	98.29% Acc	✗	✓	✓	✗	✗
2018	Wang et al. [WLZ ⁺ 18]	PDX	MCC (0.777), 92.9% Acc	✗	✗	✗	✗	✗
2019	Elbashir et al. [EEMS19]	CNN	98.76% Acc	✗	✓	✓	✗	✓
2020	Yuan et al. [YLZ20]	SVM, RF	100% Acc	✗	✗	✗	✗	✗
2020	MotieGhader et al. [MMSAMN20]	Metaheuristic + SVM	90% Acc	✓	✗	✗	✗	✗
2021	Jia et al. [JCC ⁺ 21]	WGCNA	97.36% Acc	✗	✓	✗	✗	✓
2021	Houssein et al. [HAH ⁺ 21]	BMO-SVM	99.36% Acc	✓	✗	✗	✓	✗
2021	Feltes et al. [FPD21]	RF, SVM, KNN, DT, MLP	100% Acc	✓	✗	✓	✓	✗
2022	Alshareef et al. [AAA ⁺ 22b]	IFSDL-PCD	97.19% Acc	✗	✓	✗	✗	✓
2022	Wei et al. [WLH ⁺ 22]	GANs	92.6% Acc	✗	✓	✓	✓	✓
2022	Deng et al. [DLDW22]	XGBoost-MOGA	56.67% Acc	✓	✗	✗	✓	✓
2023	Devi et al. [DP23]	IWOA	97.7% Acc	✓	✗	✗	✗	✓
2023	Mohamed et al. [MEFD ⁺ 23]	CNN	98.3% Acc	✓	✓	✓	✗	✓
2024	Jagadeeswararao et al. [JS24]	RF, SVM, KNN, LR	96% Acc	✗	✗	✓	✗	✗
2024	Wang et al. [WHH ⁺ 24]	SVM, LR	90.5% AUC	✗	✗	✓	✓	✗
2025	Alharbi et al. [AVZ ⁺ 25]	LASSO-MOGAT	95.9% Acc	✓	✓	✓	✓	✗

^a The table compares various cancer classification models, their results, and whether they incorporate specific criteria such as optimization, domain-specific algorithms (DSA), sufficient data (SD), multiple cancer types (MCT), or single algorithms (SA).

3.1.3 Proteomic Data in Cancer

Proteins are the functional effectors in cells, and proteomic changes often directly influence phenotype and therapeutic response. MS-based proteomics has advanced significantly, enabling the quantification of thousands of proteins (and their post-translational modifications) in tumour samples. Analysing proteomic data independently is essential because protein levels and activities can diverge from mRNA levels due to translational control, protein degradation or modification [ZCC⁺22]. In other words, transcriptomic data provides an incomplete picture of cellular state, and proteomics can uncover biomarkers that would be missed if one looked only at mRNA. For example, phosphorylation of signalling proteins or the presence of protein isoforms can be critical cancer biomarkers that cannot be detected via GE analysis alone. Protein complexes, composed of multiple proteins, are pivotal in a range of biological functions, including maintaining cellular balance, growth, and division [HHN⁺12]. An example is the “26S proteasome”, comprising 31 subunits, which is critical for regulating the cell cycle and apoptosis through the degradation of aged or defective proteins [WD18]. Unravelling the structure and roles of these complexes is key to understanding cell biology. Despite significant research [MT15, WLL20, HTB⁺15], identifying the full range of human proteins in these complexes across the entire proteome is still challenging. Various experimental methods are used to detect protein complexes from Protein–Protein Interaction (PPI). Techniques like yeast two-hybrid assays facilitate the identification of PPIs and complexes by using bait and prey fusion proteins that bring together different domains of a transcription factor [PMDLB19, RVH⁺05].

Advanced experimental methods like Affinity Purification–Mass Spectrometry (AP–MS) [HHP⁺15, HBP⁺17] and CF–MS [HHN⁺12, LST⁺18] have been instrumental in the large-scale characterisation of PPIs. AP–MS involves isolating target proteins that interact with a bait protein attached to a matrix [DMBM17], while CF–MS entails fractionating cellular lysates through various biochemical methods to identify co-eluting protein complexes. These methods have significantly expanded our knowledge of PPIN in humans and other organisms. Recent large-scale studies, such as BioPlex [HTB⁺15], AP–MS, [HHP⁺15], and CF–MS [WBP⁺15], have enhanced our understanding of human PPI networks. However, these studies have shown a limited overlap in their findings [DLH⁺17], possibly due to different experimental techniques capturing various interaction types [SS19]. To address this, Drew et al. [DLH⁺17] amal-

gamated these datasets using an SVM classifier, creating a comprehensive map of human protein complexes. Additionally, the availability of numerous PPI datasets in public databases like BioGRID [ORC⁺21], BioPlex [SHHG18], and STRING [SGN⁺21] has enabled researchers to integrate and analyse these data through computational methods.

The integration of biological information and computing power has changed our understanding of difficult diseases like cancer in the discipline of bioinformatics. Proteomics, the comprehensive examination of proteins, is important to cellular function and the genetic code's executors; it is at the front of this transition. In particular, when it is applied to cancer research, it offers a precise map of the functions and interactions of proteins, which is vital for finding out aberrant pathways that lead to the development of cancer. Since PPIs are the foundation of all biological functions, researching them is very crucial [MKZ⁺22]. Apoptosis, cell proliferation, and other processes essential to the initiation and spread of cancer are frequently dysregulated in cancer due to abnormalities in PPI networks [SMS⁺14]. With an emphasis on oral cancer, a complicated illness with variable levels of aggressiveness, bioinformatics offers the resources required to analyse the complexity present in its proteome landscape. Oral cancer cell lines provide a unique opportunity to investigate the role that variations in protein expression and interaction play in the progression of cancer. These cell lines can be classified into three types: normal control (Human Oral Keratinocyte (HMK)), less aggressive (Squamous Cell Carcinoma-9 (SCC-9)), and more aggressive (Human Squamous Carcinoma-3 (HSC-3)).

High-throughput proteomics techniques are vital to the search for such knowledge. With the development of techniques such as MS, protein profiling has never before been possible in such depth and with such breadth [CCZ22]. Nevertheless, significant computational analyses are required due to the complexity of the resultant data. This requirement is where bioinformatics comes into play, using statistical techniques and algorithms to sort through the massive amounts of proteomic data and turn it into biologically meaningful insights. One crucial stage in this process that comes to light is feature selection [CS14]. The proteome data is reduced to digestible, informative subsets using methods like Boruta [KJR10], SelectKBest [SH23], Recursive Feature Elimination [GFBG06], and ExtraTrees [ASF⁺22]. In our work, the approach strikes a balance between computational tractability and data depth by choosing 50, 100, and 200 protein attributes, ensuring that the most informative proteins are pursued for additional examination. These feature sets are then subjected to classification meth-

ods implemented in machine learning frameworks such as scikit-learn [KK16]. The cell lines are categorised into distinct groups using a variety of algorithms, including ensemble approaches, LR, and SVM.

3.1.4 Comparison of Omics Data Types: Strengths and Weaknesses

Different omics data types each contribute uniquely to cancer analysis. We performed a detailed comparison of somatic mutation (genomic DNA) data, transcriptomic (mRNA expression) data, and proteomic (with an emphasis on CF-MS) data, outlining their individual strengths and limitations.

3.1.4.1 Strengths of DNA Somatic Mutation Data

- **Direct identification of driver alterations:** Mutations (Single Nucleotide Variant (SNV)s, indels, copy-number changes) in oncogenes or tumour suppressors can pinpoint causal drivers of cancer. Many targeted therapies are designed for specific mutations (e.g. EGFR mutations in lung cancer, HER2 amplification in breast cancer) [CPLZ22].
- **Stable markers:** DNA mutations are essentially permanent *logs* of events; they do not vary with transient conditions. This lack of variation makes them reliable lineage markers and useful for tracking tumour evolution (e.g. in liquid biopsies) [Ray22].
- **Rich knowledge base:** There are extensive databases of cancer mutations (TCGA, Catalogue Of Somatic Mutations In Cancer (COSMIC), etc.) linking mutations to clinical outcomes, which aids interpretation [CPLZ22].

3.1.4.2 Limitations of DNA Somatic Mutation Data

- **Sparse and high-dimensional:** The mutation matrix is mostly zeros (most genes are not mutated in a given tumour). Important mutations can be rare events, making statistical detection hard without large cohorts. Also, WES focuses on coding regions and might miss non-coding regulatory mutations [TBN⁺15].
- **Functional uncertainty:** Not all mutations impact protein function; distinguishing driver vs passenger mutations is challenging. A mutation's

presence does not indicate if it actually alters pathways; it might have no effect if the gene is not expressed or the protein is compensated for by others [Bro02].

- **No dynamic information:** DNA data alone does not tell us which pathways are active. It is a static blueprint; it must be combined with expression or proteomic data to see the downstream consequences [QZY⁺21].
- **Limited by tumour heterogeneity:** Minor subclonal mutations can be diluted in bulk sequencing signals [HL22].

3.1.4.3 Strengths of Transcriptomic Data

- **Reflects active biological processes:** mRNA levels indicate which genes are up- or down-regulated, thus revealing pathway activation, cell proliferation rates, etc. Many cancer subtypes were originally defined by distinct GE signatures [CSCW21].
- **Quantitative and integrative:** Continuous expression values allow clustering of samples and correlating with phenotypes. Differential expression analysis can identify candidate biomarkers (e.g. an overexpressed oncogene in a subtype).
- **Mature technology and data abundance:** Thousands of tumour expression profiles exist (from microarrays and RNA-seq). Methods for normalisation and analysis are well-established, and cross-study meta-analyses are feasible [CMF⁺21].
- **Capture of downstream effects:** Expression can reflect the impact of DNA mutations or epigenetic changes, providing a functional readout (e.g. a KRAS mutation might be inferred by the activation of MAPK pathway genes even if the mutation is not observed directly) [MJ22].

3.1.4.4 Limitations of Transcriptomic Data

- **Indirect measure:** mRNA abundance does not always correlate with protein levels or activity due to post-transcriptional regulation. A gene may be highly expressed but heavily regulated at the protein level (or vice versa). Thus, relying solely on mRNA can be misleading for functional status [FdPP21].

- **Context sensitivity:** Expression profiles can change with micro-environment, drug exposure, or even sample handling. This plasticity means biomarkers from mRNA might be less stable across conditions [MM04].
- **Normalisation and batch effects:** Especially with microarrays, technical biases can be strong. Batch effects or differences in platforms can create false signals, requiring careful normalisation [Alt05].
- **Bulk tissue averaging:** Typical expression data from bulk tumours mixes signals from different cell types (cancer cells, stroma, immune cells) [SLAN20]. A high expression could be due to immune infiltrates rather than the cancer cells, complicating interpretation (single-cell RNA-seq can address this, but is more recent and not the focus here).

3.1.4.5 Strengths of Proteomic Data

- **Proteins are direct effectors:** Protein levels and modifications determine cell phenotype. Measuring proteins (the targets of most drugs) is often more predictive of phenotype than measuring mRNA alone. Proteomic changes can reveal activated signalling (phosphoproteins) and metabolic states [NLZJ23].
- **Complexes and interactions:** Co-fractionation data adds another layer by indicating which proteins likely interact or co-complex. This data can identify PPIN and complexes that are perturbed in cancer. For instance, if multiple members of a protein complex co-elute and show altered abundance in tumours, that complex may be dysregulated [BBBH⁺23].
- **Post-transcriptional insights:** Proteomics can capture regulation missed by transcriptomics (e.g., protein degradation, alternative splicing leading to different protein isoforms, etc.). It can validate whether mRNA changes led to protein changes or not [VM12].
- **Emerging large-scale datasets:** Efforts like CPTAC have produced large proteomic datasets matched with genomics. Also, human interactome mapping (hu.MAP, BioPlex, etc.) gives reference networks to interpret co-fractionation results [EOT⁺15].

3.1.4.6 Limitations of Proteomic Data

- **Coverage and missing data:** MS may not detect all proteins, especially low-abundance or membrane proteins. Co-fractionation profiles can be noisy, and not all co-eluting proteins truly interact (some co-elution is coincidental). Data can have missing values if a protein isn't identified in some fractions or samples [LAS07].
- **Quantitative complexity:** Protein quantification is semi-quantitative in many MS experiments. Technical variability can be high, and dynamic range issues mean changes of very high or low-abundance proteins are hard to measure [BMD21].
- **Interpretation requires domain knowledge:** Co-fractionation data is complex; one must often use computational algorithms to infer interaction networks from it. Distinguishing biologically meaningful co-fractionation patterns from artefacts can be challenging [ZRM⁺23].
- **Lower throughput than genomics:** Proteomic experiments (especially interactome mapping) are resource-intensive. Sample sizes are often smaller than GE studies. This circumstance can limit statistical power for biomarker discovery unless combined with other data [SWGZ21].

These limitations highlight the need for investigation of all independent omics approaches that can capture the complexity of biological systems more comprehensively. The background and literature review together frame the interdisciplinary space in which this thesis is situated. Building on this foundation, Chapter 4 begins the analytical contributions of the thesis by focusing on somatic mutation data. This chapter demonstrates how data processing, feature engineering, clustering, and classification can reveal biologically and clinically meaningful subgroups.

Chapter 4

Somatic Mutation-Based Modelling for Cancer Classification and Subtype Prediction

4.1 Introduction

This chapter extends the concepts introduced in our earlier study [YTM22], where visualisation techniques and clustering evaluation measures were examined, and applies them to the analysis of somatic mutation data. In this chapter, we deal with the limitations of comparative classification studies by proposing a pipeline termed “CLARITY” to identify and classify cancer stages. First, we curated somatic mutation datasets using gold-standard cancerous gene datasets and pathway score summarisation to desparsify and filter somatic mutation data. Second, we integrated autoencoders with a spectral clustering algorithm to categorise cancer profiles into different stages based on their mutation scores in pathways. To validate our clustering algorithm, we embedded our generated cluster into clinical information and performed Kaplan–Meier survival estimation [DWC16] and Cox proportional hazards modelling [BAB05]. Then, we designed a deep learning and XAI-based lightweight model to classify cancer into different stages (i.e. initial, middle, late). With the use of XAI, we could preserve hallmark information, which impacts the classification result. These hallmarks are candidates for biomarkers. We demonstrate our methodology on BRCA and achieved 99.47% accuracy, outperforming state-of-the-art results. Additionally, we employed traditional machine learning algorithms to demonstrate the ro-

bustness of our methodology, which shows independence from complex classification algorithms. The technical details of our methods are given in Section 4.2, while detailed results, our findings, results comparison, and visual analyses are provided in Section 4.3.

4.2 Methodology

This section describes the technical details of our methodology, including methods, equations, and illustrations. The summarised flow of our approach is shown in Figure 4.1. The diagram represents the curation of somatic mutation data from TCGA and COSMIC, as well as cancerous genes from GISTIC and COSMIC. Then, the curated data was de-sparsified using MSigDB Pathways Data and transformed into cancer hallmarks. Self-designed autoencoders integrated with a spectral clustering algorithm were used to cluster the hallmarks into cancer molecular subtypes. Next, the clustered data was passed to our developed deep neural network for the classification and LIME for the interpretability. The generated clustered data was also validated using TCGA Firehose Clinical Data, as well as survival and Cox Hazards analysis. Furthermore, the visual analysis was done to demonstrate the robustness of our designed pipeline. The gold standard cancer genes information was taken from the Genomic Identification of Significant Targets in Cancer (GISTIC) [MSH⁺11] and COSMIC [TBJ⁺19] databases. To disperse the mutation data, we used biological pathways from the Molecular Signatures Database (MSigDB) [LSP⁺11]. The technical details of these steps are given below.

4.2.1 Data Collection and Preprocessing

We obtained clinical data for BRCA from TCGA via the Genomic Data Commons (GDC) data portal [JFGS17], as well as from the Broad Institute's Firehose repository [VLL⁺22]. Clinical annotations include patient demographics, tumour stages, overall survival, and treatment details. The technical details of these steps are given below.

Mutation data for BRCA was retrieved using the TCGAbiolinks R package [CSO⁺16], which provides programmatic access to TCGA data through the GDC API. The data were downloaded in Mutation Annotation Format (.maf). We applied barcode filters to isolate data exclusively from primary tumour samples (`sample_type == 'Primary Tumour'`). Each cancer type con-

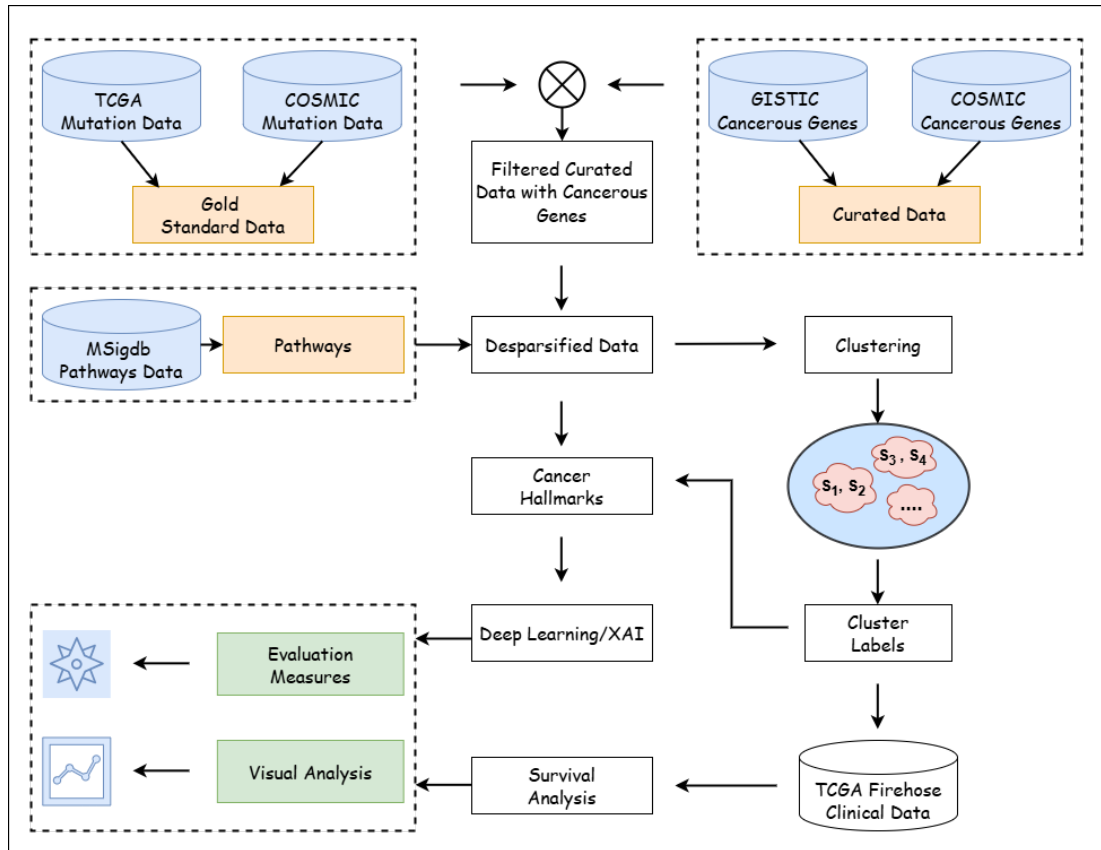


Figure 4.1: Overview of CLARITY

tained four mutation files corresponding to different variant callers: Mutect, SomaticSniper, MuSE, and VarScan. For consistency, we selected the Mutect mutation call set across all cancer types. After removing silent (synonymous) mutations from the dataset, we retained 1398 BRCA samples for further analysis.

Gold-Standard Gene List Curation To define a high-confidence list of cancer-associated genes, we downloaded curated cancer gene lists from two sources:

- COSMIC, which provides experimentally validated cancer driver genes.
- GISTIC, which predicts driver genes based on significant somatic copy-number alterations across cancer types.

Genes from both sources were merged to form a non-redundant gold-standard gene list, resulting in 18,230 unique cancer-associated genes. The original mutation dataset was then filtered using this curated gene list to retain only relevant gene alterations.

Mutation Matrix Construction The filtered mutation dataset was converted into a binary matrix $\mathbf{M} \in \{0, 1\}^{n \times m}$, where n denotes the number of samples (patients) and m denotes the number of curated genes. Each entry $M_{ij} = 1$ indicates the presence of a non-silent mutation in a gene j for patient i ; otherwise, $M_{ij} = 0$. Equations 4.1–4.6, used for de-sparsifying somatic mutation data, are adapted from the method described by Kuijjer et al. [KPS⁺18].

Sample Scoring and Mutation Rate Calculation We computed a normalised *sample mutation score* P_{ij} for each patient i with respect to gene j as in Equation 4.1.

$$P_{ij} = \frac{M_{ij}}{G_j} \quad (4.1)$$

Where G_j represents the genomic length of a gene j in base pairs, retrieved from the ensemble annotations.

The total mutation burden for a patient i was computed as in Equation 4.2.

$$S_i = \sum_{j=1}^m P_{ij} \quad (4.2)$$

Subsequently, the normalised mutation rate R_i for each patient i was defined as in Equation 4.3.

$$R_i = \frac{S_i}{\sum_{i=1}^n S_i} \quad (4.3)$$

This mutation rate quantifies the relative burden of mutations in a sample, normalised across the cohort, thereby enabling comparative analysis across patients with varying mutation loads.

4.2.2 Desparsification Using SAMBAR

Somatic mutation data is characteristically sparse, with the majority of patients having few mutations across a large gene set. To mitigate this sparsity and enhance interpretability, we employed the SAMBAR algorithm [KPS⁺18], which aggregates gene-level mutations into biologically meaningful gene sets.

SAMBAR performs desparsification by transforming the binary mutation matrix

$\mathbf{M} \in \{0, 1\}^{n \times m}$ into a more compact and informative feature matrix $\mathbf{F} \in \mathbb{R}^{n \times k}$, where $k \ll m$, by leveraging prior biological knowledge such as Gene Ontology (GO) [ABB⁺00], Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [KG00], or curated biological processes.

Let $\mathcal{G} = \{g_1, g_2, \dots, g_k\}$ denote the set of k annotated gene sets, and let each gene set g_l be a subset of genes from the curated list. For each patient i and gene set g_l , the aggregated mutation score A_{il} is defined as in Equation 4.4.

$$A_{il} = \sum_{j \in g_l} \frac{M_{ij}}{G_j} \quad (4.4)$$

where M_{ij} is the binary mutation status for a gene j in sample i , and G_j is the length of gene j in base pairs, as before.

The result is a matrix $\mathbf{A} = [A_{il}] \in \mathbb{R}^{n \times k}$, where each row corresponds to a patient and each column to a gene set, capturing cumulative functional mutation scores.

To normalise across gene sets with differing sizes and lengths, SAMBAR applies a z-score transformation on each feature as in Equation 4.5.

$$Z_{il} = \frac{A_{il} - \mu_l}{\sigma_l} \quad (4.5)$$

Where μ_l and σ_l are the mean and standard deviation of the scores in gene set g_l across all patients using Equation 4.6.

$$\mu_l = \frac{1}{n} \sum_{i=1}^n A_{il}, \quad \sigma_l = \sqrt{\frac{1}{n} \sum_{i=1}^n (A_{il} - \mu_l)^2} \quad (4.6)$$

This final matrix $\mathbf{Z} \in \mathbb{R}^{n \times k}$ contains standardised functional scores and is used for downstream analyses, such as clustering, subtyping, and survival modelling.

SAMBAR thus enables the reduction of high-dimensional, sparse mutation data into a lower-dimensional, functionally enriched feature space, improving statistical power and biological interpretability.

4.2.3 Deep Representation Learning and Clustering

Following desparsification via SAMBAR, we obtained a dense, normalised feature matrix $\mathbf{Z} \in \mathbb{R}^{n \times k}$, where n is the number of patients and k is the number of gene sets (functional units). To further reduce dimensionality and capture non-linear representations of mutational signatures, we trained a deep autoencoder neural network.

An autoencoder consists of an encoder $\mathcal{E} : \mathbb{R}^k \rightarrow \mathbb{R}^d$ and a decoder $\mathcal{D} : \mathbb{R}^d \rightarrow \mathbb{R}^k$ such that the reconstruction loss is minimized as in Equation 4.7 [GBC16].

$$\mathcal{L}_{\text{AE}} = \frac{1}{n} \sum_{i=1}^n \|\mathbf{Z}_i - \mathcal{D}(\mathcal{E}(\mathbf{Z}_i))\|_2^2 \quad (4.7)$$

where $\mathcal{E}(\cdot)$ maps the input to a latent space of dimension $d \ll k$. The autoencoder architecture consisted of fully connected layers with Rectified Linear Unit (ReLU) activations in both the encoding and decoding branches. It was trained using the Adam optimiser with a Mean Squared Error (MSE) loss.

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be the matrix of encoded representations produced by the encoder. We applied Spectral Clustering on $\mathbf{X}$ to partition patients into molecular subtypes. Spectral Clustering operates by constructing an affinity graph $\mathcal{A}$ and computing the Laplacian matrix as in Equation 4.8.

$$\mathcal{L} = \mathcal{D} - \mathcal{A} \quad (4.8)$$

where $\mathcal{D}$ is the degree matrix and $\mathcal{A}$ is the affinity matrix based on k -nearest neighbors. The top k eigenvectors of $\mathcal{L}$ form a low-dimensional embedding that is clustered using k -means.

For a range of cluster numbers $k \in \{2, 3, 4\}$, we evaluated the clustering quality using Equation 4.9 [Rou87] and 4.10 [DB09].

- **Silhouette Score** (Sil), which measures intra-cluster cohesion and inter-cluster separation:

$$\text{Sil} = \frac{1}{n} \sum_{i=1}^n \frac{b_i - a_i}{\max(a_i, b_i)} \quad (4.9)$$

where a_i is the average intra-cluster distance for sample i , and b_i is the lowest average inter-cluster distance to another cluster.

- **Davies-Bouldin Index** (DB), which assesses the average "similarity" be-

tween each cluster and its most similar one:

$$DB = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d_{ij}} \right) \quad (4.10)$$

where σ_i is the average distance of points in cluster i to its centroid, and d_{ij} is the distance between centroids of clusters i and j .

The optimal number of clusters k^* was selected by maximising the difference as in Equation 4.11.

$$\text{Score}(k) = \text{Sil}(k) - \text{DB}(k) \quad (4.11)$$

The best clustering solution was used to annotate samples with subtype labels, and the latent space was visualised using UMAP two-dimensional representation.

The final subtype-labelled dataset $\mathbf{X} \in \mathbb{R}^{n \times d}$, along with clustering assignments, was saved for downstream analysis, including survival modelling and subtype characterisation.

4.2.4 Cluster Validation via Survival Analysis

To validate the discovered clusters biologically, we performed survival analysis using clinical metadata. The clinical data was merged with the mutation data using patient sample barcodes. Only common samples present in both datasets were retained for analysis.

For each patient, two clinical attributes were extracted:

- **Days-to-last-follow-up** (T_i): Time in days from diagnosis to last recorded follow-up or death.
- **Vital status** (δ_i): Binary indicator of event occurrence 1 if the patient is deceased, zero if censored (i.e., alive at last follow-up).

These were used to construct the survival object $\mathcal{S}_i = (T_i, \delta_i)$ for each patient i .

4.2.4.1 Kaplan–Meier Estimator

The non-parametric Kaplan–Meier estimator was used to estimate the survival probability $\hat{S}(t)$ over time. Given a sequence of observed event times $t_1 < t_2 < \dots$

$\dots < t_m$, the estimator is defined as in Equation 4.12 [KM58].

$$\hat{S}(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{n_j}\right) \quad (4.12)$$

Where:

- d_j : Number of events (deaths) at time t_j ,
- n_j : Number of individuals at risk just before t_j .

We plotted $\hat{S}(t)$ for each discovered cluster label to assess differences in survival distributions visually.

4.2.4.2 Log-Rank Test

To evaluate the difference in survival distributions among clusters statistically, we applied the log-rank test, which compares the observed and expected number of events in each group as in Equation 4.13 [M⁺66].

$$\chi^2 = \frac{(O_1 - E_1)^2}{V_1} \quad (4.13)$$

Where:

- O_1 : Total number of events observed in group 1,
- E_1 : Expected number of events under the null hypothesis,
- V_1 : Variance of $O_1 - E_1$.

A low p -value indicates that survival differs significantly across clusters.

4.2.4.3 Cox Proportional Hazards Model

To adjust for covariates and quantify the hazard ratio (HR) associated with cluster membership, we fitted a Cox proportional hazards model as in Equation 4.14 [D⁺72].

$$h(t|X_i) = h_0(t) \exp(\beta_1 \cdot \text{Class}_i + \beta_2 \cdot \text{Age}_i) \quad (4.14)$$

Where:

- $h(t|X_i)$: Hazard function for patient i at time t ,

- $h_0(t)$: Baseline hazard function (unspecified),
- Class_i : Cluster assignment of patient i ,
- Age_i : Age at initial pathological diagnosis,
- β_1, β_2 : Model coefficients to be estimated.

Model fitting was performed using the partial likelihood as in Equation 4.15.:

$$\mathcal{L}(\beta) = \prod_{i=1}^n \left(\frac{\exp(\beta^\top X_i)}{\sum_{j \in \mathcal{R}(T_i)} \exp(\beta^\top X_j)} \right)^{\delta_i} \quad (4.15)$$

where $\mathcal{R}(T_i)$ is the risk set at time T_i .

Significant coefficients ($p < 0.05$) for the `Class` variable confirm that the molecular subtypes discovered by clustering are predictive of survival outcomes, even after adjusting for confounders such as age.

4.2.5 Classification of Molecular Subtypes

To predict the molecular subtypes derived from spectral clustering, we trained a supervised deep neural network classifier on the SAMBAR-desparsified mutation data. Each sample was labelled with its respective cluster class, obtained from previous steps. The model formulation, including the network architecture and associated loss functions presented in Equations 4.16–4.18, is based on the foundational deep learning frameworks described by Goodfellow et al. [GBC16] and Aggarwal [A⁺18].

4.2.5.1 Model Architecture and Training

We designed a fully connected feedforward neural network consisting of multiple hidden layers with ReLU activations, batch normalisation, and dropout for regularisation. The architecture can be described as a function in Equation 4.16.

$$f_\theta(\mathbf{x}) = \text{Softmax}(W_4 \phi_3(W_3 \phi_2(W_2 \phi_1(W_1 \mathbf{x} + b_1) + b_2) + b_3) + b_4) \quad (4.16)$$

where W_i and b_i represent the weights and biases of the i -th layer, and ϕ_i represents the nonlinear transformation (ReLU). For binary classification, the final activation becomes a sigmoid function, while for multi-class classification, the final activation function becomes softmax. The loss (error) between predicted

and actual class labels during training is computed for the binary classification using Equation 4.17, and multi-class classification using 4.18. The visual architecture of our model is shown in Figure 4.2

$$\mathcal{L}_{\text{binary}} = -\frac{1}{n} \sum_{i=1}^n [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (4.17)$$

$$\mathcal{L}_{\text{categorical}} = -\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^C y_{ij} \log(\hat{y}_{ij}) \quad (4.18)$$

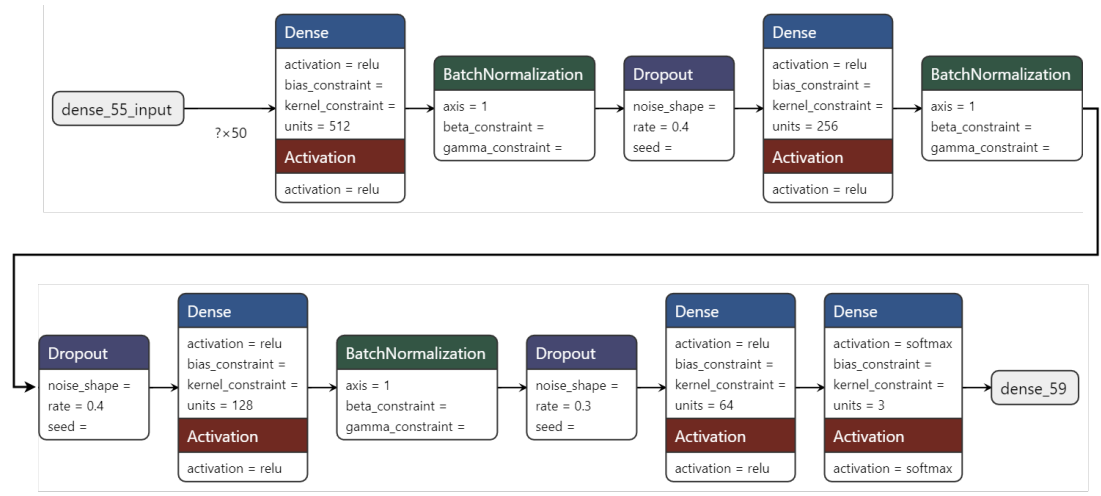


Figure 4.2: Model Architecture for Cancer Multi-Class Classification

Callbacks and training The early stopping criteria are designed to terminate training when the validation loss fails to improve for 20 consecutive epochs. The model's weights are saved when the validation loss improves in the model checkpoints. The learning rate is reduced by a factor of 0.2 if the validation loss does not improve for five consecutive epochs, with a minimum learning rate of 10^{-3} . The model is trained for 100 epochs with a batch size of 32, and the Adam optimiser with an adaptive learning rate scheduler.

To address class imbalance, we employed the SMOTE on the training set. Additionally, class weights w_c were calculated and applied during training as in Equation 4.19 [CBHK02].

$$w_c = \frac{n}{C \cdot n_c} \quad (4.19)$$

where n is the total number of samples, C is the number of classes, and n_c is the number of samples in class c .

4.2.5.2 Explainability via LIME

To interpret the decisions of the classifier and identify key features influencing subtype prediction, we utilised the LIME algorithm. It constructs a locally weighted linear approximation $\hat{f}$ of the original model f around a point $\mathbf{x}_0$ as in Equation 4.20 [RSG16].

$$\hat{f}(\mathbf{x}) = \arg \min_{g \in G} \mathcal{L}(f, g, \pi_{\mathbf{x}_0}) + \Omega(g) \quad (4.20)$$

Where:

- G is the family of interpretable models (e.g., linear models),
- $\pi_{\mathbf{x}_0}$ is a locality-aware kernel around $\mathbf{x}_0$,
- $\mathcal{L}$ is a loss function measuring fidelity to f ,
- $\Omega(g)$ is a regularization term to control complexity.

The top k features contributing to each prediction were visualised, and the most recurrent features across samples were selected for biological interpretation.

4.2.5.3 Visualisation and Feature Importance Analysis

To assess the structure and distribution of important features, we generated:

- **Clustermaps** to visualise hierarchical relationships between discriminative features. Technical Details of Clustermap are given in 5.2.4.3.
- **Violin plots** showing the distribution of key features across clusters. Technical Details of violin plots are given in 5.2.4.5.

These visualisations enhanced our understanding of mutation-driven subtypes and provided biological context to model predictions.

4.3 Results and Discussion

In this section, we present the setup of our experiments and analyse the results.

4.3.1 Experimental Setup

These experiments were performed using the 13th generation Intel Core i9-13900k (32 CPUs), 64 GB RAM, and 2 x NVIDIA 4080 Graphical Processing Units. The model architecture is shown in Figure 4.2. We used Python 3.9.18 and TensorFlow 2.10.1 to perform these experiments.

4.3.2 Dataset Description

We downloaded the somatic mutation data from TCGA, which contains 1,389 profiles and 18,230 genes. Mutation data was also curated with the gold-standard list of genes from COSMIC and the GISTIC algorithm. For mutation data, sample scores were calculated and normalised by mutation rate. Finally, we obtained three matrices after preprocessing, comprising 948 profiles. We merged the gene mutation data with pathways data from KEGG and MSigDB using the desparsification method 4.2.2. This method aggregated the scores of the genes into pathways with related genes. We removed all patient profiles with a null pathway, as well as the pathways that do not exist in any of the patient profiles. Finally, we got an ensemble matrix of 948 BRCA profiles with 50 hallmarks.

4.3.3 Evaluation Measures

We employed the following measures shown from 4.21 to 4.27 to assess the performance of our methodology for cancer data classification. The formulations of these evaluation metrics are inspired by the standard definitions presented by Bishop et al. [BN06].

Accuracy: The proportion of correctly classified cancer stages out of the total number of samples.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (4.21)$$

Precision: The proportion of correctly identified cancer samples among all samples classified as cancer.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (4.22)$$

Recall: The proportion of actual cancer stage samples that the model correctly

identified.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4.23)$$

F1 Score: The harmonic mean of precision and recall.

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4.24)$$

Geometric Mean (G-Mean): Evaluates the balance between sensitivity and specificity.

$$\text{G-Mean} = \sqrt{\text{Sensitivity} \cdot \text{Specificity}} \quad (4.25)$$

where

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4.26)$$

and

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (4.27)$$

where:

- TP = Number of samples correctly predicted as a given cancer stage,
- TN = Number of samples correctly predicted as not belonging to that stage,
- FP = Number of samples from other stages incorrectly predicted as this stage,
- FN = Number of samples from this stage incorrectly predicted as other stages.

4.3.4 Clustering of desparsed data

The ensemble matrix of BRCA, prepared by the desparsed algorithm, was clustered to identify cancer subtypes, i.e., initial, middle, or late-stage. We utilised the k-means Clustering (k-means) algorithm, and our designed autoencoder along with the spectral algorithm. The biological data with a high number of genes is usually sparse or low variance, and k-means assumes equal-sized clusters, which does not fit biological data well. To show the nature of our data, we projected it using UMAP (shown in the Figure 4.3). Spectral clustering is suitable for identifying connected components, clusters of equal size, and non-convex clusters. We designed an autoencoder to find meaningful data repre-

sentations for clustering, and embedded autoencoders with spectral clustering. Details of our clusters are shown in Table 4.1.

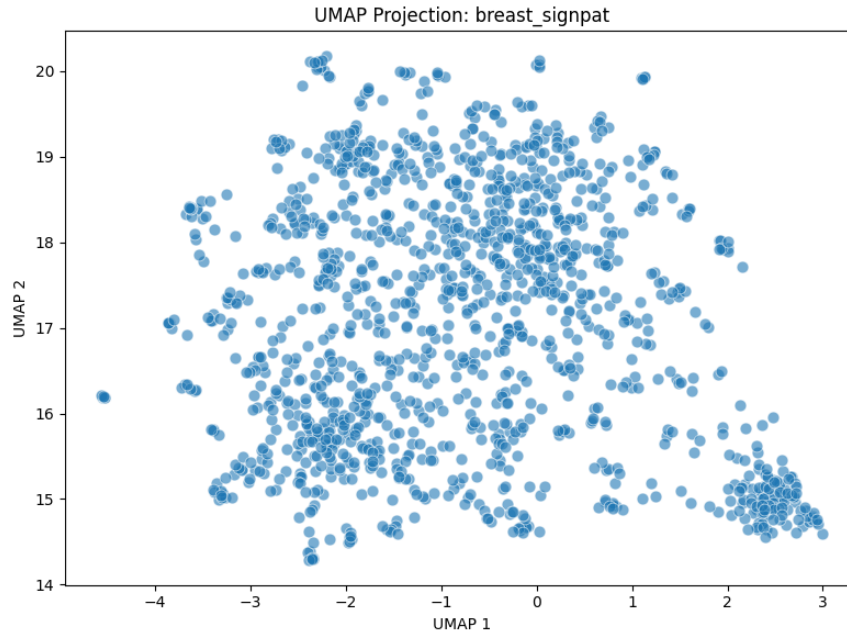


Figure 4.3: Representation of our desparsed data

Table 4.1: Comparison of Cluster Memberships for Different Values of k Using Autoencoder + Spectral Clustering vs. k-means

K	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Autoencoder + Spectral Clustering				
2	929	19	–	–
3	921	11	16	–
4	912	13	9	14
k-means Clustering				
2	933	15	–	–
3	891	15	42	–
4	891	13	42	2

Figure 4.3 shows how difficult it is to cluster the data, as there are no clear patterns in the graph. However, our proposed clustering algorithm can effectively separate this type of biological data. As shown in Table 4.1, our Autoencoder and Spectral Clustering approach consistently produced more compact and separable cluster assignments across all values of k . The separable clusters for $k=3$ are shown in Figure 4.4. The superiority of our proposed method was further supported by higher silhouette scores shown in Table 4.2, which were later validated through supervised classification.

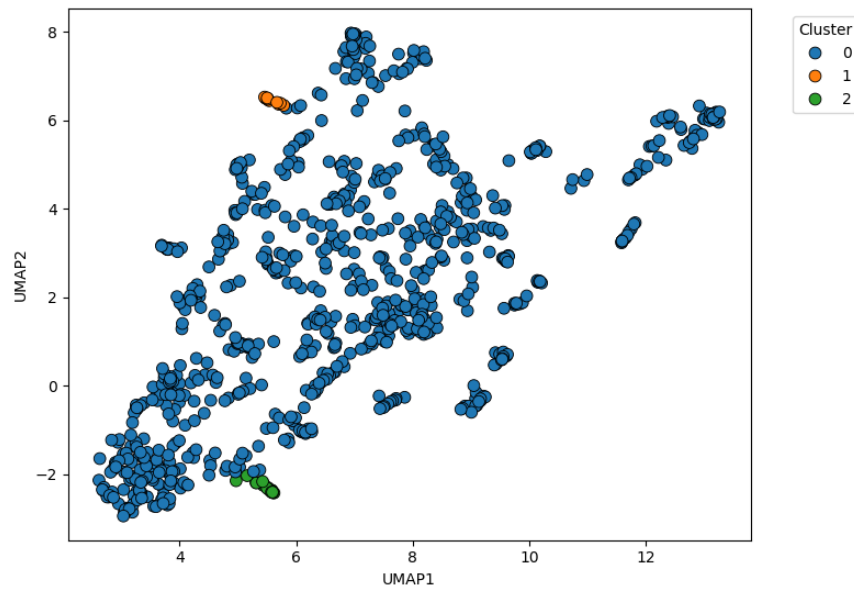


Figure 4.4: Representation of clustered data by our Autoencoder + Spectral clustering method

4.3.5 Survival and Statistical Analysis

To validate the clusters generated by CLARITY, we assumed that patients in groups 1, 2, and 3 were in the initial, last, and middle stages, respectively. For the validation of patients' cancer stage, we performed a survival analysis of patients by integrating their clinical information with our cluster information. We performed Kaplan-Meier Survival analysis on the clinical data of the same patients. Clustering the BRCA sample clinical data in three groups produced subtypes of patients, which are exactly equal to our clustered data. Our survival analysis is shown in Figure 4.5. Figure 4.5 visualises the survival probabilities over time (in days) for three different patient groups identified through clustering. The plot shows survival distributions for three clusters with sample sizes $n = 921$ (red), $n = 11$ (blue), and $n = 16$ (green). The x-axis represents time in days, while the y-axis shows the survival proportion; this proportion is the estimated probability of patients surviving beyond a certain time point. The separation between curves suggests potential differences in survival outcomes between these groups, which may reflect distinct biological subtypes or clinical stages. However, the small sample sizes for the blue and green clusters should be interpreted cautiously due to limited statistical power. We also performed statistical analysis using Cox Proportional Hazards to validate our clustering. The results of our methods are shown in Table 4.2.

Both methods yielded statistically significant results with $p < 10^{-4}$, indicating

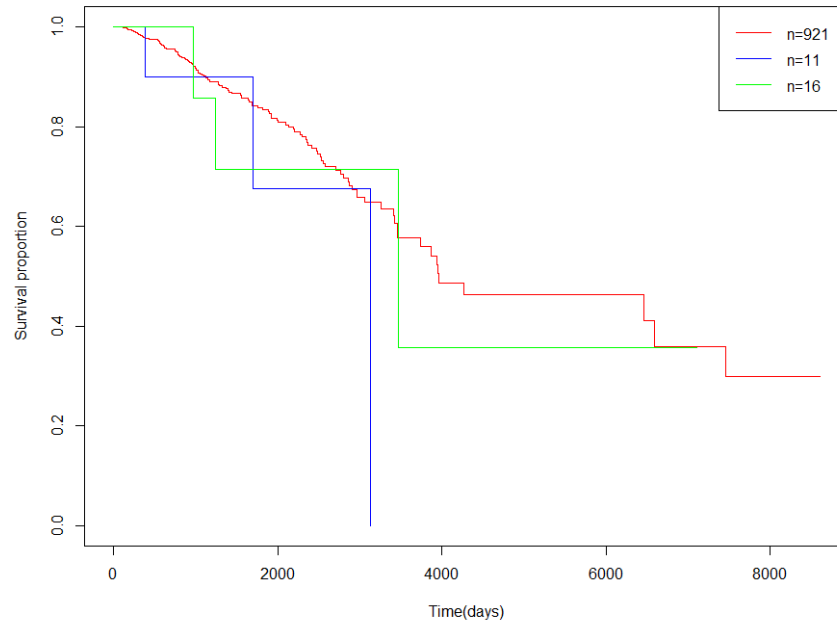


Figure 4.5: Kaplan–Meier survival curves for patients grouped by molecular subtypes discovered through Autoencoder + Spectral clustering (for $k = 3$).

Table 4.2: Comparison of Cox Proportional Hazards Model Results for Cluster-Based Survival Analysis

Clustering Method	Likelihood Ratio Test	Degrees of Freedom (df)	p-value	Silhouette
Autoencoder + Spectral Clustering	19.11	2	7.081×10^{-5}	0.17
k-means Clustering	19.73	2	5.195×10^{-5}	0.11

that the clusters are meaningfully associated with patient survival outcomes. Although the k-means model yielded a slightly higher likelihood ratio statistic (19.73 vs. 19.11), this does not necessarily imply better model fit or utility. The classification framework that followed confirmed the predictive stability and robustness of our clusters, something not assessed in the k-means pipeline.

4.3.6 Robustness of our methodology

After validation of our cluster, we performed classification to automate the pipeline. The architecture that we employed for classifying cancer subtype data is based on a dense layer and was carefully designed to address the high dimensionality and sparsity typically encountered in somatic mutation datasets. We also used machine learning classifiers to show the robustness of our hallmark selection process. Even with the use of simple classifiers, we achieved very satisfactory results. These results are shown in Table 4.3

Table 4.3: Performance Comparison of Classifiers with k-means and Autoencoder + Spectral Clustering (AES). Bold values indicate the better result for each classifier.

Classifier	Method	Accuracy (%)	Precision	Recall	F1 Score	G-Mean
KNeighbors	k-means	93.99	0.8943	0.9399	0.9158	0.9168
	AES	99.05	0.9913	0.9905	0.9898	0.9909
SVC	k-means	93.99	0.8834	0.9399	0.9107	0.9112
	AES	97.15	0.9439	0.9715	0.9575	0.9576
Decision Tree	k-means	93.25	0.9262	0.9325	0.9284	0.9293
	AES	98.52	0.9872	0.9852	0.9848	0.9862
Random Forest	k-means	95.57	0.9468	0.9557	0.9410	0.9512
	AES	99.16	0.9916	0.9916	0.9903	0.9916
AdaBoost	k-means	92.09	0.9225	0.9209	0.9192	0.9216
	AES	98.42	0.9780	0.9842	0.9807	0.9811
Gradient Boosting	k-means	95.99	0.9521	0.9599	0.9506	0.9560
	AES	98.84	0.9903	0.9884	0.9889	0.9893
GaussianNB	k-means	80.07	0.9430	0.8007	0.8544	0.8688
	AES	58.97	0.9499	0.5897	0.7199	0.7483
LDA	k-means	94.31	0.9394	0.9431	0.9394	0.9412
	AES	98.73	0.9909	0.9873	0.9882	0.9891
DNN	k-means	94.21	0.9455	0.9421	0.9437	0.9438
	AES	99.47	0.9965	0.9947	0.9953	0.9956

In the table, we demonstrate that groups created using our methodology outperform the traditional clustering method. Even though the cluster generated by the k-means algorithm has a good Cox proportional hazard score, the biological data context is more important. Our proposed model is not just a classifier; it is tightly integrated with autoencoder-based representation learning. It captures nonlinear patterns in high-dimensional, sparse, and noisy somatic mutation data, something classical machine learning algorithms struggle with unless carefully feature-engineered. Furthermore, deep learning can soundly adapt to additional data modalities and scale with increasing dataset complexity, while classical machine learning models often degrade or require re-tuning under these circumstances. The accuracy and loss results for our designed network are shown in Figure 4.6

In our model, dropout layers, which randomly deactivate a percentage of the neurons during training and force the model to generalise rather than rely on specific neurons, are essential to the robustness of the model. Considering that most cancer research has limited sample sizes, this regularisation technique is crucial. Furthermore, the Adam optimiser was selected because, in contrast to conventional gradient descent, it enables the model to converge more rapidly and consistently. This rapid convergence is because of the algorithm's adjustable learning rates. With slight changes, the model's architecture

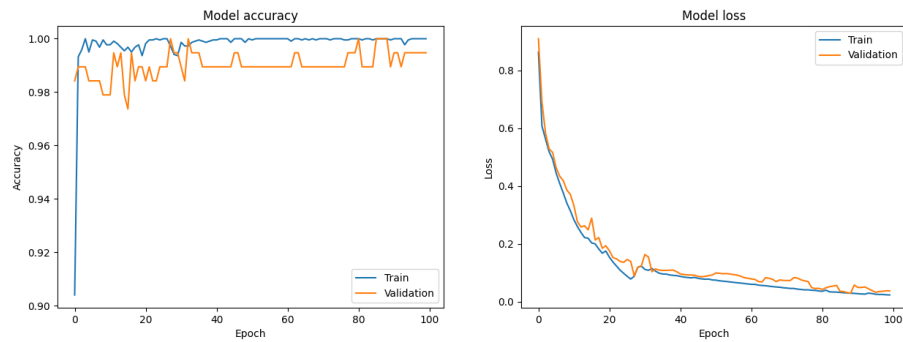


Figure 4.6: Accuracy and Loss Curves of our Model

can handle problems involving both binary and multiclass categorisation. The output layer uses a softmax activation function for multiclass classification and a sigmoid activation function for binary classification.

In summary, our model design successfully strikes a balance between efficiency and complexity. Dense layers, batch normalisation, dropout, and fully connected layers work together to minimise overfitting and guarantee that the model extracts both high- and low-level characteristics from the somatic mutation data. When dealing with genetic data, which frequently has a small sample size but a high number of features (genes), this method works particularly well. Furthermore, the architecture is a solid option for applications such as cancer classification, where it is crucial to identify minute differences in gene expression. This is because it can extract biologically significant patterns from the data, employing regularisation techniques and a robust optimiser.

The XAI produce transparency and interpretability to the deep learning model's predictions. Using LIME, we managed to examine the contribution of individual genes to the categorisation decision. This allowed us to understand the role of specific genes in driving the enrichment of cancer-related pathways and how they contribute to the model's decision-making process. We utilised LIME to interpret our model's decision and identified the top 10 hallmarks that play a significant role in breast cancer and can serve as biomarkers for the stage classification process. The abundance values of the top-selected hallmarks for BRCA are shown in Figure 4.7. These figures demonstrate the effectiveness of our methodology in handling high variance, outliers, and uneven data distributions, all of which are common in the identification and quantification of hallmarks.

The cluster map of top hallmarks for BRCA is shown in Figure 4.8. These im-

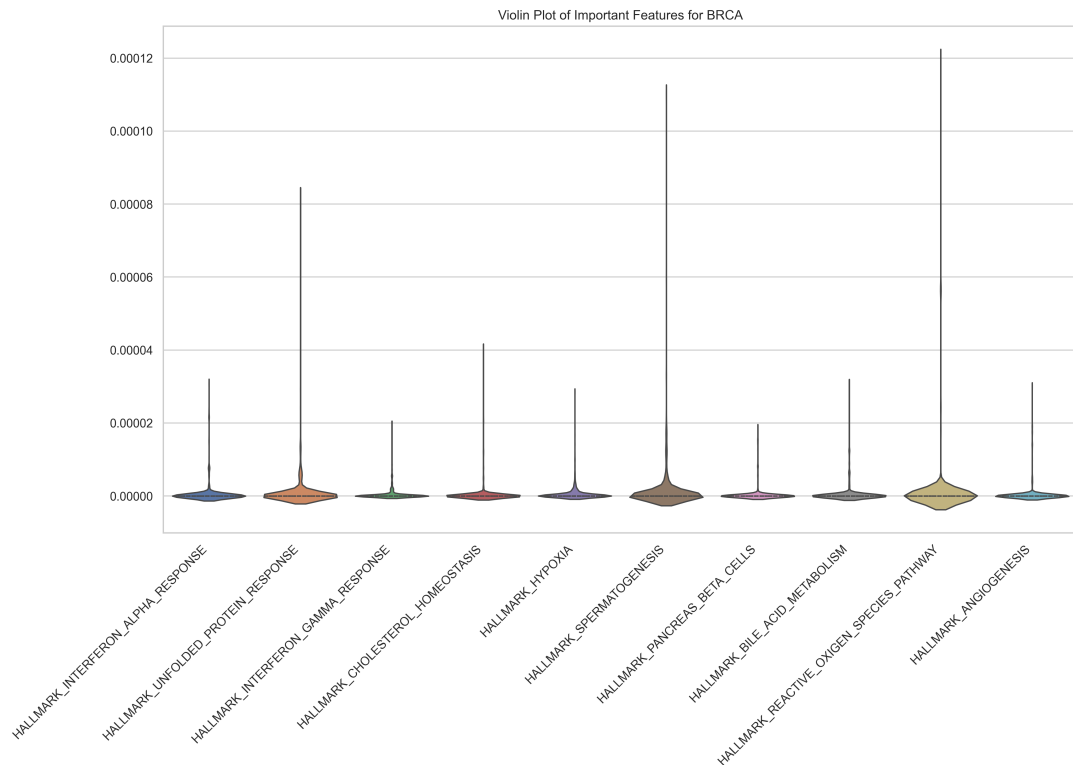


Figure 4.7: Abundance Value Distribution of Top Hallmarks in BRCA

ages demonstrate the robustness of our model in classifying mixed types of correlation. Our methodology highlights genes with high correlation and those with low correlation; our model classifies based on a combination of features, making it more reliable.

The violin plot for abundance levels shown in 4.7 provides a good representation of the distribution of expression values across a range of hallmarks. The ample range in the shapes of the violins implies that the model is excellent at capturing both highly variable and smoother expression patterns. The various violin shapes across different hallmarks, as well as the spread of expression values, support the model's capacity to discriminate across expression levels, which is critical for discovering cancer-specific biomarkers. The robust performance of the model in this image indicates its capacity to generalise across varied hallmark expression patterns while maintaining sensitivity to essential features, supporting its success in cancer classification tasks.

Figure 4.8 gives a strong representation of the hierarchical clustering of hallmarks based on their expression patterns, demonstrating links between gene clusters. The prominent clusters of related hallmarks illustrate the model's robustness in finding patterns of gene co-expression. Overall, these visualisations

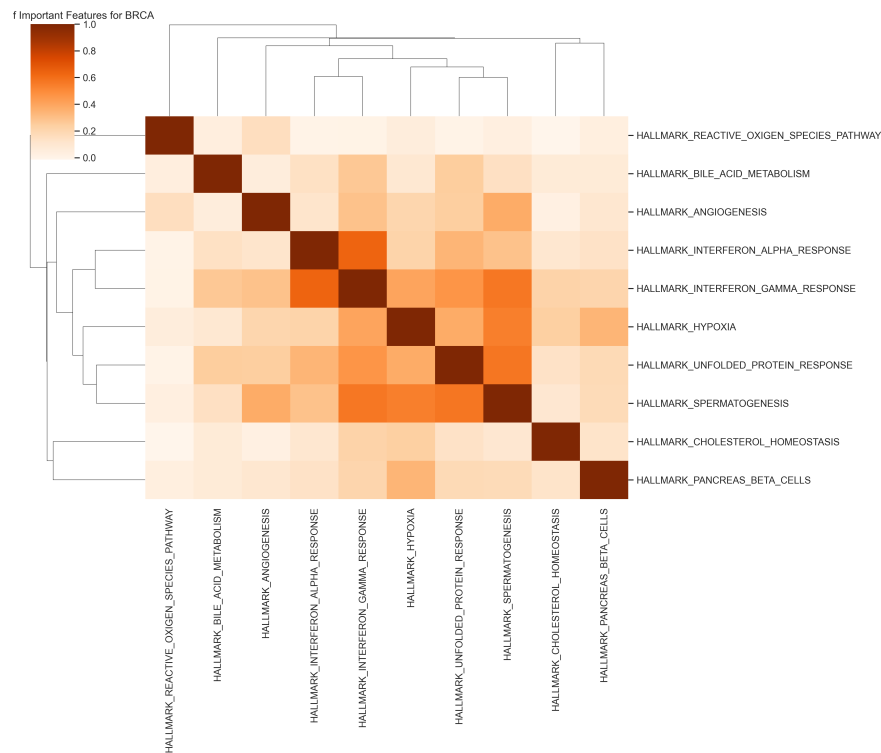


Figure 4.8: Cluster Map of Top Hallmarks

collectively highlight the model’s robustness in processing high-dimensional data and provide a detailed analysis of the categorisation and identification processes.

In this chapter, we proposed CLARITY, a pipeline combining deep learning, XAI, statistical analysis and data visualisation to both classify cancer stages based on somatic mutation data and to trace cancer hallmarks. In the process, we integrated autoencoders, spectral clustering, DNN, and LIME into a single interpretable workflow. Our experiments demonstrated superior subtype separation compared to k-means, both in clustering compactness and survival relevance. The clustering mechanism not only works with complex DNNs but also works on lightweight traditional machine learning algorithms. Although mutation data provide crucial insights into the genetic basis of cancer, they represent only one layer of the disease’s molecular complexity. To capture downstream functional consequences, Chapter 5 shifts focus to transcriptomic data. Here, deep learning models and XAI methods are applied to uncover patterns of GE and potential biomarkers.

Chapter 5

Transcriptomic Profiling and Deep Learning for Cancer Type and Subtype Classification

5.1 Introduction

This chapter builds upon our published works [YM25], focusing on RNA-seq data and [YBBM24, YARM24] addressing microarray data. In this chapter, we are dealing with limitations (see chapter 3) of comparative classification studies for RNA-seq and microarray data by proposing a unified methodology for different cancer subtypes classification and another deep learning-based algorithm for different cancer types classification, additionally producing an identification of relevant genes. First, we designed a deep learning and XAI-based lightweight model to classify cancer subtypes for multiple tissues and cancer types. This single model can be used for single or multiple-class classification. Second, with the use of XAI, we can preserve the gene information that impacts the classification result. These genes are potential biomarker candidates for cancer. For the cancer type classification, we developed another deep learning-based algorithm to classify the 7 primary cancer types against 1 normal type; it classified 8 classes within the neighbourhood of 87% of validation accuracy, precision, recall, F1 Score, and G-Mean. The major motivation behind developing a cancer type classifier is to identify the primary as well as metastatic cancer. Meanwhile, the cancer subtype classifier can help analyse the identified cancer types with 100% accuracy, precision, recall, F1 Score, and G-Mean. We also analysed

the 99 identified genes using the KEGG, GO, and visual analysis techniques to validate and demonstrate the robustness of our methodology. The visual analysis of the selected features also shows how well our methodology can handle the genes with high variance abundance values. We used 2 databases, including CuMiDa, which contains 78 cancer microarray datasets and BARRA:CuRDa, which contains 17 RNA-seq datasets. The technical details of our methodology are given in section 5.2, while detailed results, our findings, results comparison, and visual analyses are given in the 5.3 Section.

5.2 Methodology

This section describes the technical details of our methodology, including methods, equations, and illustrations.

The summarised flow of our approach is shown in Figure 5.1. The diagram shows the elements of the overall approach. Cancer subtype classification represents the classification within tissue type, e.g. normal or tumour profile classification; Cancer Type classification represents the overall classification, which is the classification of tissue types such as breast cancer or lung cancer. In the proposed pipeline, we start with data processing of both RNA-seq and microarray data. The data is pre-processed using bioinformatics and machine learning techniques (details in section 5.2.1). After the pre-processing of data, for cancer subtype classification, synthetic data was generated using SMOTE in order to deal with the class imbalance problem. Subsequently, data was passed to our designed deep learning models for classification and XAI model for the interpretability. For cancer type classification, different datasets are merged, and dimensions are reduced using PCA and passed to the cancer subtype classification pipeline. For the validation of the genes shortlisted by our interpretable models, Pathway Enrichment Analysis was performed, and a visual analysis was done to demonstrate the robustness of our methodology. Each of the elements of the diagram is individually explained below.

5.2.1 Data Preprocessing

In our experiments, we employed the aforementioned RNA-seq database named BARRA:CuRDa [FPD21]. We describe here the processes performed by the authors to create the dataset for completeness. They curated the dataset using the GEO query tool [DM07] for the R platform. Data for several cancer subtypes

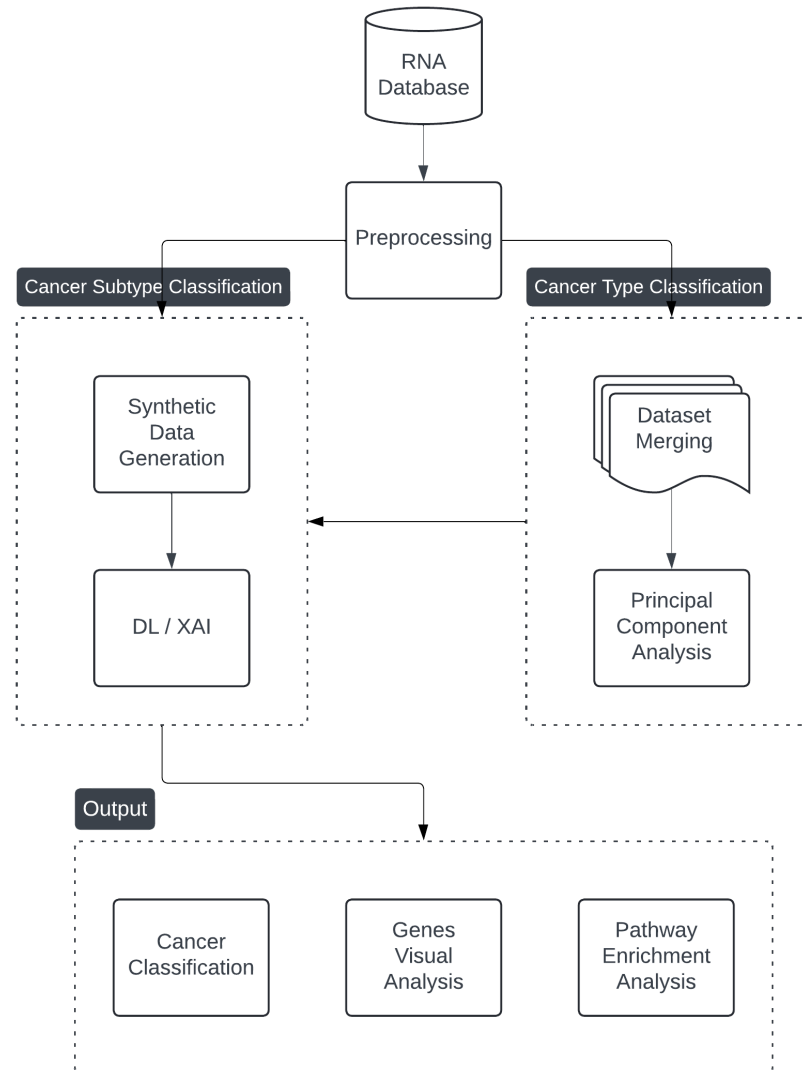


Figure 5.1: Overview of Our Transcriptomic Methodology

were obtained from the GEO database [CB16] to create numerous RNA-seq datasets named Geo Series (GSEs). Following data collection, a quality analysis was performed on the raw collected data using the FastQC software [WA18]. Utilising the Trimmomatic 0.35 tool [BLU14], trimming was performed to eliminate low-quality bases, poly-N sequences, residual ribosomal RNA, and the sequences of adapters. The arguments `slidingwindow` was done, which trims the text when the average quality throughout the window (in our example, four nucleotides) goes below 15 bp; then `minlen` was executed, which rejects the read if it falls below a predetermined length of 65 bp. All data sets were treated using this procedure, save GSE6511, whose original read length was 40 bp. The data that was generated was aligned against the Homo sapiens reference

sequence (Ensembl version GRCh38.94). Using the default parameters of the program, transcript-level abundance quantification was performed via aligning spliced transcripts to a reference (STAR) v2.6.0a [DDS⁺13] and RNA-seq using expectation-maximisation (RSEM) v1.3.1 [LD11]. The count data was converted utilising variance stabilising procedures from DESeq2 [LHA14], and the outcomes were imported and summarised as matrices utilising the tximport package [SLR15].

The dataset has unbalanced and low numbers of profiles, discussed in section 5.3.2. For the evaluation, data needed to be split into train and test sets, which further lowers the number of profiles for each set. To deal with this problem, we used SMOTE over-sampling technique (details in 5.2.2.1) on the training dataset to make sure the model can learn adequately and does not suffer from underfitting. While dealing with the cancer type classification, data was merged and its dimension reduced using PCA to make the model lightweight.

5.2.2 Cancer Classification

5.2.2.1 SMOTE for Synthetic Data Generation

SMOTE [CBHK02] tackles class imbalance by generating synthetic samples for the minority class. Synthetic data points are generated by interpolating between a minority class instance and one of its KNNs.

Minority Class Sampling Let x_i be a sample from the minority class. The set of its k -nearest neighbors, denoted as $\mathcal{N}_k(x_i)$, is obtained by selecting the k points closest to x_i in terms of Euclidean distance. This formula is represented mathematically as in Equation 5.1.

$$\mathcal{N}_k(x_i) = \{x_j \in \mathcal{X} \setminus \{x_i\} \mid x_j \text{ is among the } k \text{ closest samples to } x_i \text{ under } \|x_j - x_i\|\}$$
(5.1)

where $\|x_j - x_i\|$ denotes the Euclidean distance between samples x_i and x_j .

Once the nearest neighbours are determined, a synthetic sample x_{new} is generated by linearly interpolating between x_i and one of its neighbours x_j . The generation formula is given as in Equation 5.2.

$$x_{\text{new}} = x_i + \alpha (x_j - x_i), \quad \alpha \in (0, 1) \quad (5.2)$$

where α is a random number drawn from a uniform distribution.

To ensure reproducibility in the generation of synthetic samples, the value of α is drawn from a uniform distribution controlled by a random seed. This is defined as in Equation 5.3.

$$\alpha \sim \mathcal{U}(0, 1) \quad (5.3)$$

where $\mathcal{U}(0, 1)$ denotes the uniform distribution over the interval $[0, 1]$.

5.2.2.2 Model for Binary and Multiclass Cancer Subtype Classification

The model architecture is designed for both binary and multiclass classification using a one-dimensional CNN (Conv1D) [IS24]. The architecture can be divided into two parts: binary classification and multiclass classification. The key difference lies in the output layer, where sigmoid activation is used for binary classification, and softmax activation is used for multiclass classification. The model formulation for cancer type and subtype classification, including the network architecture and associated loss functions presented in Equations, is based on the foundational deep learning frameworks described by Goodfellow et al. [GBC16] and Aggarwal [A⁺18].

Convolution Operation The model begins with a Conv1D layer, which applies a 1D convolution operation to the input data. Given an input sequence $\mathbf{X} \in \mathbb{R}^{T \times d}$, where T is the sequence length and d is the feature dimension, the convolution operation is defined as in Equation 5.4.

$$\mathbf{h}_t = f \left(\sum_{i=1}^d \mathbf{W}_i * \mathbf{X}_{t+i-1} + b_i \right) \quad (5.4)$$

where:

- $\mathbf{h}_t$ is the output of the convolution at time step t ,
- $\mathbf{W}_i$ is the kernel (filter) of the convolution for the i -th feature,
- $\mathbf{X}_{t+i-1}$ is the input at time step $t + i - 1$,
- b_i is the bias term,

- $f(\cdot)$ is the activation function, in this case, ReLU.

Max-Pooling Operation After the convolution, a max-pooling operation is performed to reduce the dimensionality and retain the most important features. The max-pooling operation can be written as in Equation 5.5.

$$\mathbf{p}_t = \max(\mathbf{h}_{t:t+P-1}) \quad (5.5)$$

where P is the size of the pooling window and $\mathbf{h}$ represent the feature map from the previous layer.

Dropout Dropout is performed to prevent overfitting by randomly changing a fraction p of input units to 0 at every update throughout training. Mathematically, dropout can be described as in Equation 5.6.

$$\mathbf{h}^{\text{drop}} = \mathbf{h} \odot \mathbf{r} \quad (5.6)$$

where $\odot$ denotes element-wise multiplication, and $\mathbf{r} \sim \text{Bernoulli}(p_{\text{drop}})$ is a mask generated from a Bernoulli distribution with probability p_{drop} .

Fully Connected Layers After flattening the output of the max-pooling layer, the flattened vector is passed through one or more fully connected (dense) layers. The output of a dense layer is given by Equation 5.7.

$$\mathbf{z} = f(\mathbf{W}\mathbf{h} + \mathbf{b}) \quad (5.7)$$

where:

- $\mathbf{W}$ is the weight matrix of the dense layer,
- $\mathbf{h}$ is the input to the dense layer,
- $\mathbf{b}$ is the bias vector,
- $f(\cdot)$ is the activation function (ReLU for hidden layers).

Output Layer for Binary Classification For binary classification, the output layer contains a single unit with a sigmoid activation function, representing the probability of the positive class. The output is given by Equation 5.8.

$$\hat{y} = \sigma(\mathbf{W}\mathbf{h} + b) = \frac{1}{1 + e^{-(\mathbf{W}\mathbf{h} + b)}} \quad (5.8)$$

where $\sigma(\cdot)$ is the sigmoid activation function and b is the scalar bias associated with the output unit.

The binary cross-entropy loss function is used to optimise the model as in Equation 5.9.

$$\mathcal{L}_{\text{binary}} = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) \quad (5.9)$$

where y is the true label and $\hat{y}$ is the predicted probability.

Output Layer for Multiclass Classification For multiclass classification, the output layer consists of C units (one for each class) with a softmax activation function as in Equation 5.10.

$$\hat{y}_i = \frac{e^{z_i}}{\sum_{j=1}^C e^{z_j}} \quad (5.10)$$

where $\hat{y}_i$ is the predicted probability for class i , and z_i is the input to the softmax layer for class i .

The categorical cross-entropy loss function is used to optimise the model as in Equation 5.11.

$$\mathcal{L}_{\text{categorical}} = - \sum_{i=1}^C y_i \log(\hat{y}_i) \quad (5.11)$$

where y_i is the true label (one-hot encoded) and $\hat{y}_i$ is the predicted probability for class i . The final model architecture for cancer-subtype classification is shown in Figure 5.2.

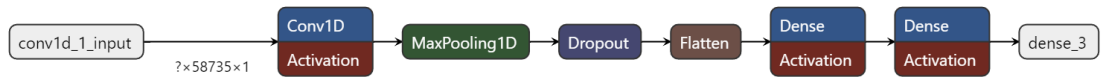


Figure 5.2: Model Architecture for Cancer Subtype Classification (Figure Generated using Netron [Roe25])

5.2.2.3 Early Stopping Criteria

An early stopping criterion is implemented to halt training once the validation accuracy reaches a target value. The stopping criterion can be described as

in 5.12.

$$\text{Stop training if } \text{val_accuracy} \geq \text{target_accuracy} \quad (5.12)$$

A patience parameter P is used to determine how many epochs to wait after reaching the target accuracy before stopping. If the validation accuracy does not improve after P epochs, training is stopped.

5.2.3 Cancer Type Classification using RNA Data

As all our data sets are combined into one, the resulting dimensions are very large. We reduce the dimensions by employing PCA.

5.2.3.1 PCA on RNA Dataset

PCA is a dimensionality reduction technique that projects data from a high-dimensional space into a lower-dimensional subspace while preserving as much variance as possible. In the context of RNA-seq data, which typically has tens of thousands of features (genes), applying PCA helps to reduce the dimensionality and capture the key patterns in the data.

For an RNA-seq dataset with p features (genes), let the input data matrix be $X \in \mathbb{R}^{n \times p}$, where n is the number of samples and p is the number of genes. In our case, $p = 58,735$, and n corresponds to the number of RNA samples. With $n_components = 640$, the dataset is projected into a 640-dimensional subspace. The mathematical details are outlined below, following the formulation described by Mackiewicz and Ratajczak [MR93].

Centring the Data First, each feature (GE value) is centred by subtracting the mean of the feature across all samples as in Equation 5.13.

$$X_{\text{centered}} = X - \mu \quad (5.13)$$

where $\mu \in \mathbb{R}^p$ is the vector of the mean values of each feature.

Covariance Matrix The covariance matrix $C \in \mathbb{R}^{p \times p}$ is computed from the centered data matrix as in Equation 5.14.

$$C = \frac{1}{n-1} X_{\text{centered}}^{\top} X_{\text{centered}} \quad (5.14)$$

The covariance matrix C captures the pairwise covariance between different genes.

Eigenvalue Decomposition PCA involves computing the eigenvectors and eigenvalues of the covariance matrix C . The eigenvalue decomposition is as in Equation 5.15.

$$C = V \Lambda V^{\top} \quad (5.15)$$

where:

- $V \in \mathbb{R}^{p \times p}$ is the matrix of eigenvectors (principal components),
- $\Lambda \in \mathbb{R}^{p \times p}$ is the diagonal matrix of eigenvalues, where each eigenvalue represents the variance captured by its corresponding principal component.

Dimensionality Reduction To reduce the dimensionality of the RNA data to 640 components, the top 640 eigenvectors corresponding to the largest eigenvalues are selected. This forms the projection matrix $V_{640} \in \mathbb{R}^{p \times 640}$, which is used to project the original data into the lower-dimensional space as in Equation 5.16.

$$X_{\text{PCA}} = X_{\text{centered}} V_{640} \quad (5.16)$$

where $X_{\text{PCA}} \in \mathbb{R}^{n \times 640}$ is the transformed data in the 640-dimensional subspace.

Variance Retained The amount of variance retained by the top 640 principal components is proportional to the sum of the top 640 eigenvalues relative to the total sum of all eigenvalues, as in Equation 5.17.

$$\text{Variance Retained} = \frac{\sum_{i=1}^{640} \lambda_i}{\sum_{i=1}^p \lambda_i} \quad (5.17)$$

where λ_i represents the eigenvalues in descending order.

Application in RNA Data By applying PCA to the RNA-seq dataset with $n_components = 640$, we reduce the original dataset from 58,735 dimensions (genes) to 640 dimensions while retaining most of the dataset's variance. This procedure reduces the complexity of the model and helps prevent overfitting when used for classification tasks.

In our deep learning models, the PCA-transformed data can be used as input to improve the model's performance by focusing on the most significant components of the gene expression data. Additionally, PCA facilitates visualisations and downstream analyses by simplifying the data to a more manageable number of dimensions.

5.2.3.2 Multiclass Cancer Type Classification Model with CNN

This subsection describes the architecture and mathematical formulation for the CNN used for multiclass classification of RNA-seq data. The architecture consists of three convolutional layers, followed by batch normalisation, max pooling, dropout layers, and dense layers. The AdamW optimiser is used to minimise the categorical cross-entropy loss. The model leverages regularisation techniques such as L2 weight regularisation, dropout, and learning rate adjustments to improve generalisation. The visual representation of the model architecture for the cancer-type classification is shown in Figure 5.3

The model is trained on input data $X \in \mathbb{R}^{n \times p \times 1}$, where n is the number of samples, p is the number of features after dimensionality reduction, and the input has a single channel since the data is univariate.

First Convolutional Layer - The first layer applies a 1D convolution to the input as in Equation 5.18.

$$Z_1 = \text{ReLU}(X * W_1 + b_1) \quad (5.18)$$

where $*$ represents the convolution operation, $W_1 \in \mathbb{R}^{5 \times 1 \times 64}$ are the filters (64 filters of size 5), and $b_1 \in \mathbb{R}^{64}$ is the bias term. A regularisation term is added using L2 regularisation as in Equation 5.19.

$$\mathcal{L}_{\text{reg}}(W_1) = \frac{\lambda}{2} \sum W_1^2 \quad (5.19)$$

where $\lambda = 0.01$ is the L2 regularization parameter.

Batch Normalization - Batch normalisation is applied to normalise the activations as in Equation 5.20.

$$Z_1^{\text{norm}} = \frac{Z_1 - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \gamma + \beta \quad (5.20)$$

where μ_B and σ_B^2 are the batch mean and variance, ϵ is a small constant to prevent division by zero, and γ and β are learnable parameters.

Max Pooling - Max pooling is used to reduce the spatial dimensions of the output as in Equation 5.21.

$$Z_1^{\text{pool}} = \max(Z_1^{\text{norm}}, \text{pool_size} = 2) \quad (5.21)$$

Dropout - Dropout is applied to prevent overfitting by randomly setting 50% of the activations to zero as in Equation 5.22.

$$Z_1^{\text{drop}} = Z_1^{\text{pool}} \odot \mathbf{r}_1 \quad (5.22)$$

where $\odot$ represents element-wise multiplication and $\mathbf{r}_1 \sim \text{Bernoulli}(0.5)$.

Second and Third Convolutional Layers - The second and third convolutional layers are applied similarly to the first layer, with the following transformations as in Equation 5.23.

$$Z_2 = \text{ReLU}(Z_1^{\text{drop}} * W_2 + b_2), \quad Z_3 = \text{ReLU}(Z_2^{\text{drop}} * W_3 + b_3) \quad (5.23)$$

where $W_2 \in \mathbb{R}^{3 \times 64 \times 128}$ and $W_3 \in \mathbb{R}^{3 \times 128 \times 256}$, and batch normalization, max pooling, and dropout are applied after each layer.

Flattening Layer - After the last convolutional layer, the output is flattened into a 1D vector as in Equation 5.24.

$$Z_{\text{flat}} = \text{Flatten}(Z_3^{\text{drop}}) \quad (5.24)$$

This transformation prepares the data for fully connected layers.

Fully Connected Layers - Two fully connected layers are applied, each with ReLU activation and L2 regularisation as in Equation 5.25.

$$Z_{\text{dense1}} = \text{ReLU}(Z_{\text{flat}}W_4 + b_4), \quad Z_{\text{dense2}} = \text{ReLU}(Z_{\text{dense1}}W_5 + b_5) \quad (5.25)$$

where $W_4 \in \mathbb{R}^{d_{\text{flat}} \times 512}$ and $W_5 \in \mathbb{R}^{512 \times 256}$, and dropout is applied after each layer as in Equation 5.26.

$$Z_{\text{dense1}}^{\text{drop}} = Z_{\text{dense1}} \odot \mathbf{r}_4, \quad Z_{\text{dense2}}^{\text{drop}} = Z_{\text{dense2}} \odot \mathbf{r}_5 \quad (5.26)$$

Output Layer - The final dense layer produces the class scores with softmax activation as in Equation 5.27.

$$\hat{y} = \text{softmax}(Z_{\text{dense2}}^{\text{drop}}W_6 + b_6) \quad (5.27)$$

where $W_6 \in \mathbb{R}^{256 \times c}$ and c is the number of classes. The softmax function ensures that the output is a valid probability distribution as in Equation 5.28. The final model architecture for cancer type classification is shown in Figure 5.3.

$$\hat{y}_i = \frac{\exp(Z_i)}{\sum_{j=1}^c \exp(Z_j)}, \quad i \in \{1, \dots, c\} \quad (5.28)$$

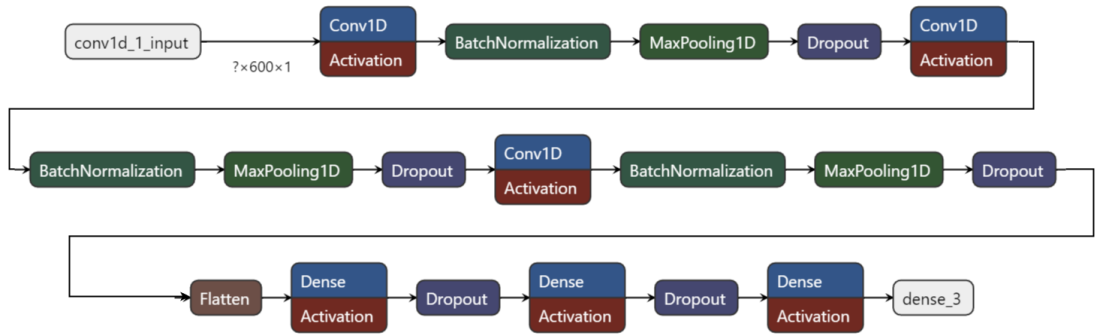


Figure 5.3: Model Architecture for Cancer Type Classification (Figure Generated using Netron [Roe25])

5.2.3.3 Optimisation and Class Balancing

To address class imbalance during training, class weights are computed and used to adjust the loss function. These weights ensure the model gives appropriate importance to each class, especially the minority classes.

Computation of Class Weights Let $\mathcal{C} = \{c_1, c_2, \dots, c_k\}$ be the set of k classes, where n_i denotes the number of samples in class c_i , and N is the total number of samples in the training set as in Equation 5.29.

$$N = \sum_{i=1}^k n_i \quad (5.29)$$

The frequency f_i of each class c_i is calculated as in Equation 5.30.

$$f_i = \frac{n_i}{N} \quad (5.30)$$

The class weight w_i for each class c_i is computed as in Equation 5.31.

$$w_i = \frac{\frac{1}{f_i}}{\sum_{j=1}^k \frac{1}{f_j}} \cdot k \quad (5.31)$$

The computed class weights are used to adjust the loss function during training.

Loss Function with Class Weights The categorical cross-entropy loss function with class weights is defined as in Equation 5.32.

$$\mathcal{L}_{\text{weighted}} = - \sum_{i=1}^k w_i \cdot y_i \cdot \log(\hat{y}_i) \quad (5.32)$$

where:

- w_i is the weight for class c_i ,
- y_i is the true label for class c_i (one-hot encoded),
- $\hat{y}_i$ is the predicted probability for class c_i .

Optimization The model is optimised using the AdamW optimiser, which is an Adam optimiser with decoupled weight decay as in Equation 5.33.

$$W_{t+1} = W_t - \eta \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} + \lambda W_t \right) \quad (5.33)$$

where $\hat{m}_t$ and $\hat{v}_t$ are the biased-corrected first and second moment estimates, η is the learning rate, and λ is the weight decay parameter.

Regularization L2 regularisation is applied to the weights in the convolutional and dense layers to prevent overfitting. The regularisation term is added to the loss function as in Equation 5.34.

$$\mathcal{L}_{\text{reg}} = \frac{\lambda}{2} \sum W^2 \quad (5.34)$$

Callbacks The early stopping criteria are designed to stop the training when the validation loss does not improve for 20 epochs. The model's weights are saved when the validation loss improves in the model checkpoints. The learning rate is reduced by a factor of 0.5 if the validation loss does not improve for 5 consecutive epochs, with a minimum learning rate of 10^{-6} .

Training The model is trained for 500 epochs with a batch size of 32, using class weights to handle class imbalance in the RNA data. The history of the training and validation metrics is stored for analysis.

5.2.4 Explainability and Visual Analysis

5.2.4.1 Explainable AI using LIME

LIME [RSG16] is a method that explains the predictions of a complex, black-box model by approximating it with an interpretable model in the local neighbourhood of the instance being explained. In our case, LIME is used to interpret the predictions of the RNA-seq deep learning model.

The core idea of LIME is to fit a local surrogate model g around the prediction of the complex model f for a specific instance x . The surrogate model is trained on perturbations of the original instance, weighted by their proximity to x .

Let $f : \mathbb{R}^d \rightarrow \mathbb{R}^K$ represent the complex model, where d is the number of features (genes), and K is the number of classes. The model outputs the predicted probabilities $f(x)$ for a given input x .

For a given instance x , LIME generates perturbed samples $\{x'_1, x'_2, \dots, x'_N\}$ from the original instance by altering some of its feature values. The corresponding predictions $f(x'_i)$ are recorded, and the proximity $\pi_x(x'_i)$ of each perturbed sample to the original instance x is computed using a kernel function, such as the exponential kernel as in Equation 5.35.

$$\pi_x(x'_i) = \exp\left(-\frac{d(x, x'_i)^2}{\sigma^2}\right) \quad (5.35)$$

where:

- $d(x, x'_i)$ is the distance between the original instance x and the perturbed instance x'_i (e.g., using Euclidean distance),
- σ controls the width of the kernel and the locality of the explanation.

Next, a simply interpretable model $g(x)$, such as a linear model, is trained on the perturbed data weighted by their proximity to x . The objective is to minimise the following weighted loss function as in Equation 5.36.

$$\mathcal{L}(f, g, \pi_x) = \sum_{i=1}^N \pi_x(x'_i) (f(x'_i) - g(x'_i))^2 + \lambda \cdot \Omega(g) \quad (5.36)$$

where:

- $f(x'_i)$ is the prediction of the original complex model on the perturbed instance x'_i ,
- $g(x'_i)$ is the prediction of the surrogate model on x'_i ,
- $\pi_x(x'_i)$ is the proximity weighting,
- $\Omega(g)$ is a regularisation term that controls the complexity of the surrogate model g ,
- λ is a hyperparameter that controls the regularisation strength.

Feature Importance The LIME explanation provides a list of feature importance values w_i , where each feature i contributes to the local decision. These weights represent the contribution of each gene i to the prediction for the instance x as in Equation 5.37.

$$g(x) = w_0 + \sum_{i=1}^d w_i \cdot x_i \quad (5.37)$$

where w_i is the weight assigned to gene i and x_i is the feature value of gene i for the instance x .

Application to RNA Data By applying LIME to our deep learning model trained on RNA-seq data, we can extract the most important genes contributing

to the prediction for a given sample. This extraction enables us to identify which genes drive the classification decisions, adding interpretability to the model.

The extracted feature importance values were subsequently used to generate plots, indicating the robustness of the model. Heatmaps, cluster maps, and violin plots were generated based on the genes identified as important by LIME, allowing for further validation of the model's behaviour. Visualisations can then be used to gain insight into the main features (genes) that led to the results.

5.2.4.2 Heatmap

A heatmap is a data visualisation technique that maps the values of a matrix to colours. In the context of RNA data, the rows can represent genes, and the columns can represent different samples or conditions. Each cell H_{ij} in the heatmap corresponds to the expression value of gene i in sample j , where colours represent the magnitude of the expression values.

Mathematically, the heatmap is based on a matrix $\mathbf{H} \in \mathbb{R}^{m \times n}$, where m is the number of genes, and n is the number of samples. Each element H_{ij} represents the value of GE, and it can be normalised using methods such as z-score normalisation as in Equation 5.38 [ESBB98].

$$Z_{ij} = \frac{H_{ij} - \mu_i}{\sigma_i} \quad (5.38)$$

where:

- H_{ij} is the expression value of gene i in sample j ,
- μ_i is the mean expression value of gene i across all samples,
- σ_i is the standard deviation of gene i across all samples.

The heatmap provides an immediate visual insight into the patterns of gene expression across samples, which can be used to observe clusters or correlations within the RNA data.

5.2.4.3 Cluster Map

A cluster map extends the heatmap by incorporating hierarchical clustering. It clusters both rows (genes) and columns (samples) based on similarity, allowing for better pattern detection in RNA data. The clustering is often performed using the Euclidean distance $d(\mathbf{x}, \mathbf{y})$ between two vectors $\mathbf{x}$ and $\mathbf{y}$, where each

vector represents either a gene's expression profile or a sample's expression across genes as in Equation 5.39 [ESBB98].

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (5.39)$$

To create the hierarchical tree, a linkage method such as “average” linkage or “complete” linkage is applied. The linkage function L can be written as in Equation 5.40.

$$L(A, B) = \frac{1}{|A||B|} \sum_{x \in A} \sum_{y \in B} d(x, y) \quad (5.40)$$

where A and B are clusters, and $d(x, y)$ is the distance between elements x and y .

5.2.4.4 UMAP (Uniform Manifold Approximation and Projection)

UMAP is a dimensionality reduction technique used to project high-dimensional RNA data into a lower-dimensional space while preserving its global and local structure. Let $\mathbf{X} \in \mathbb{R}^{m \times d}$ represent the RNA data, where m is the number of samples, and d is the number of features (genes). UMAP projects $\mathbf{X}$ into a lower-dimensional space $\mathbf{Y} \in \mathbb{R}^{m \times 2}$ or $\mathbf{Y} \in \mathbb{R}^{m \times 3}$.

UMAP relies on a weighted KNN graph. The weight w_{ij} between two points x_i and x_j is computed using the probability distribution based on local distances as in Equation 5.41 [RC22].

$$w_{ij} = \exp \left(-\frac{d(x_i, x_j) - \rho_i}{\sigma_i} \right) \quad (5.41)$$

where:

- $d(x_i, x_j)$ is the distance between points x_i and x_j ,
- ρ_i is the distance to the nearest neighbor of x_i ,
- σ_i is a normalization factor.

The low-dimensional projection is then optimised to preserve both local and global structures by minimising the cross-entropy between the original high-dimensional graph and the low-dimensional embedding.

UMAP is particularly useful in showing how well a dataset clusters by class, providing insights into the model's performance in segregating biological samples.

5.2.4.5 Violin Plot

A violin plot is a method of plotting numeric data and comparing distributions. It is similar to a box plot, but it includes a Kernel Density Estimate (KDE) of the data's probability density. For RNA data, a violin plot is useful to visualise the distribution of gene expression levels across different conditions or classes.

Given a set of expression values $X = \{x_1, x_2, \dots, x_n\}$, the KDE is defined as in Equation 5.42 [HN98].

$$f(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) \quad (5.42)$$

where:

- $K(\cdot)$ is the kernel function (often Gaussian),
- h is the bandwidth parameter,
- n is the number of data points.

The violin plot shows the KDE on both sides of the central line, making it easier to see the density and distribution of RNA expression levels. It helps demonstrate the robustness of the model by visually comparing the spread of gene expression data across different classification results or labelled conditions.

5.2.5 KEGG Pathway Enrichment Analysis

To complete our analysis and find physiologically relevant pathways, we performed a KEGG pathway enrichment analysis using the *clusterProfiler* [YWHH12] package in R. Finding over-represented pathways based on differentially expressed genes in the dataset was the aim of the investigation. The following measures were implemented:

Gene Selection: Differentially expressed genes were extracted using the RNA-seq dataset. The *AnnotationDbi* and *org.Hs.eg.db* packages were utilised to map Entrez IDs [Ost12] to gene identifiers. In order to preserve alignment with the KEGG database, only genes with a valid Entrez ID were used for further analysis.

Pathway Enrichment: The KEGG pathway enrichment was performed using the *enrichKEGG* function of the *clusterProfiler* package. The gene list, which contained the Entrez IDs of the differentially expressed genes, served as the input to the process. The parameters selected for the enrichment analysis were as follows:

- **Organism:** *Homo sapiens*, symbolized by 'hsa' in KEGG.
- **p-value Adjustment Method:** Multiple comparisons were accounted for by applying the Benjamini-Hochberg (BH) method. This method is recommended for high-throughput analyses, such as RNA-seq, because it controls the false discovery rate (FDR).
- **q-value Cut-off:** A strict q-value threshold of 0.05 was applied to ensure that only highly enriched pathways were retained. Pathways with a q-value greater than 0.05 were considered non-significant and excluded from further analysis.

5.2.6 GO Enrichment Analysis

The *enrichGO* function from the *clusterProfiler* package was used for GO enrichment analysis. The gene list was used as input, and the following settings were applied:

- Organism: *Homo sapiens* (symbolized as 'hsa').
- P-value adjustment method: BH to control the false discovery rate (FDR).
- Threshold: Significant GO keywords were found with a p-adjusted value (p_{adj}) below 0.05.

5.2.6.1 Equations for P-value Adjustment

The p-values were modified using the BH approach to control the FDR. The adjusted p-values (p_{adj}) were computed using Equation 5.43 [BH95].

$$p_{adj} = \frac{p \cdot N}{rank(p)} \quad (5.43)$$

where:

- p is the original p-value.

- N is the total number of tests.
- $rank(p)$ is the rank of the p-value among all tests.

5.3 Results and Discussion

In this section, we present the setup of our experiments and analyses of the results.

5.3.1 Experimental Setup and Evaluation Measures

These experiments were performed using the 13th-generation Intel Core i9-13900k (32 CPUs), 64 GB RAM, and 2 X NVIDIA 4080 Graphical Processing Units. Due to the high dimension of the data, the number of parameters in the model was high and required high computational power to train the model. The model architectures are shown in figures 5.2 and 5.3. We used Python 3.9.18 and TensorFlow 2.10.1 to perform these experiments. We used the accuracy, precision, recall, F1 Score and G-Mean measures to evaluate the performance of our methodology for the cancer classification using RNA data. The details were previously presented in Section 4.3.3.

5.3.2 Dataset Description

The dataset [FPD21] contains 7 types of cancers that include Breast, Colon, Head/Neck, Kidney, Liver, Prostate, and Lung. There are 17 tissues included in the dataset, and for each tissue, there are 58735 genes except Colon_SRR2089755 and Colon_GSE50760, as these have 58148 genes. Colon_GSE50760 has 3 types of profiles, including Primary, Normal, and Metastasis, and Colon_SR2089755 has 4 types of profiles, including NormalColon, NormalLiver, Primary, and Metastasis. All other 15 cases have 2 types of profiles, including Normal and Tumour. Detailed information about the dataset is given in Table 5.1.

We did not need to merge the dataset to train the model for the cancer classification. For the cancer type classification, however, we merged the dataset. Since the two datasets have different numbers of features, we had to use padding to merge them. As a result of merging, our final dataset shape has 640 instances with 58735 features.

Table 5.1: Overview of RNA-seq datasets used for Cancer Classification.

Dataset	Samples [Class Distribution]	Genes	Classes
Kidney_GSE89122	13 [7 tumor, 6 normal]	58735	2
Liver_GSE55758	16 [8 tumor, 8 normal]	58735	2
Prostate_GSE22260	28 [19 tumor, 9 normal]	58735	2
Breast_GSE52194	20 [17 tumor, 3 normal]	58735	2
Breast_GSE69240	35 [25 tumor, 10 normal]	58735	2
Breast_GSE71651	33 [15 tumor, 18 normal]	58735	2
Colon_GSE50760	54 [18 primary, 18 metastasis, 18 normal]	58148	3
Colon_GSE72820	14 [7 tumor, 7 normal]	58735	2
Colon_SRR2089755	20 [5 primary, 5 metastasis, 5 normal, 5 normal]	58148	4
HeadNeck_GSE48850	11 [6 tumor, 5 normal]	58735	2
HeadNeck_GSE63511	16 [8 tumor, 8 normal]	58735	2
HeadNeck_GSE64912	22 [18 tumor, 4 normal]	58735	2
HeadNeck_GSE68799	45 [41 tumor, 4 normal]	58735	2
Lung_GSE37764	12 [6 tumor, 6 normal]	58735	2
Lung_GSE40419	164 [87 tumor, 77 normal]	58735	2
Lung_GSE60052	86 [79 tumor, 7 normal]	58735	2
Lung_GSE87340	51 [25 tumor, 26 normal]	58735	2

5.3.3 Robustness of Our Methodology

The architecture that we developed for classifying RNA-seq data is based on a 1D CNN (Conv1D) specifically built to handle the sequential nature of gene expression data. Conv1D layers are ideal for RNA-seq data, where the gene expression levels matter because they are especially good at extracting localised patterns from data. By applying convolutional filters throughout the input, the model uncovers subtle patterns in gene expression that carry substantial biological significance. These dependencies may be correlated with biological aspects like co-expressed genes or pathways implicated in disease processes. By adding a MaxPooling1D layer, the dimensionality of the data is reduced while the most important features are retained. This controls overfitting, which is a common problem when working with high-dimensional RNA-seq datasets and increases the computational efficiency of the model.

Dropout layers, which randomly deactivate a percentage of the neurons during training and force the model to generalise rather than rely on specific neurons, are essential to the robustness of the model. Considering that most RNA-seq research has limited sample sizes, this regularisation technique is crucial. Furthermore, the AdamW optimiser was selected because, in contrast to conventional gradient descent, it enables the model to converge more rapidly and consistently. This is because of its adjustable learning rates. With few changes, the model's architecture can handle problems involving both binary and multiclass categorisation. The output layer uses a softmax activation function for multiclass classification and a sigmoid activation function for binary classi-

fication. Because of its versatility, we show that the same base model may be applied to a larger variety of RNA-seq datasets and effectively handle different categorisation tasks.

In summary, our CNN-based model design successfully balances efficiency and complexity. Convolutional layers, max pooling, dropout, and fully connected layers work together to minimise overfitting and guarantee that the model extracts both high- and low-level characteristics from the RNA-seq data. When dealing with RNA-seq data, which frequently has a small sample size but a high number of features (genes), this method works especially well. Furthermore, the architecture is a very solid option for applications like cancer classification, where it is crucial to identify minute differences in gene expression. This is because it can extract biologically significant patterns from the data, employing regularisation techniques and a robust optimiser.

However, for the cancer-type classification, by adding more convolutional layers, more regularisation methods, and deeper learning layers to the basic Conv1D architecture, the model improves further, making it suitable to handle complex multiclass classification tasks with RNA-seq data. The network can gradually capture more complex and high-level characteristics thanks to the model's incorporation of three Conv1D layers with progressively larger filter sizes (64, 128, and 256). BatchNormalization is used in conjunction with each convolutional layer to standardise activations and hasten convergence while stabilising training. Large weight values are penalised by L2 regularisation, which is performed on both convolutional and dense layers to prevent overfitting. In order to further reduce dimensionality, MaxPooling1D layers trail each convolutional block, and Dropout layers, which randomly remove neurons during training, guarantee additional regularisation.

The model uses two dense layers (512 and 256 units) with ReLU activations to analyse high-level features learned by the convolutional layers. The last dense layer has a softmax activation function, making it perfect for multiclass classification, where it outputs the probability of each cancer kind. Weight decay is incorporated into the AdamW optimiser, which speeds up learning and enhances the model's generalisation. In order to prevent overfitting, early stopping and a learning rate scheduler termed *ReduceLROnPlateau* were used to ensure training ends when validation performance stagnates.

Class weights were also applied during training to account for any possible class imbalance in the dataset and make sure the model did not miss any less com-

mon kinds of cancer. To better balance the distribution of classes, resampling of the training data was added. With these improvements, the model becomes more resilient and can handle the intricacies of classifying cancer types by utilising the capabilities of deep CNN, sophisticated regularisation, and reliable optimisation methods. The reason behind using two different architectures is to keep the methodology lightweight. The first model for cancer classification is quite light and can handle data with binary and multiple classes. The second model has more parameters and is used to classify the multiple types of cancers.

The results of our model for the binary class for tissue GSE87340 are shown in Figure 5.4. The curves shown represent the accuracy and loss over epochs. The number of epochs was fixed at 200, but due to early stopping criteria, training stops automatically at different numbers of epochs. The best weights are saved to ensure the top accuracy achieved by our model, which is 100% for all datasets in the cancer classification task.

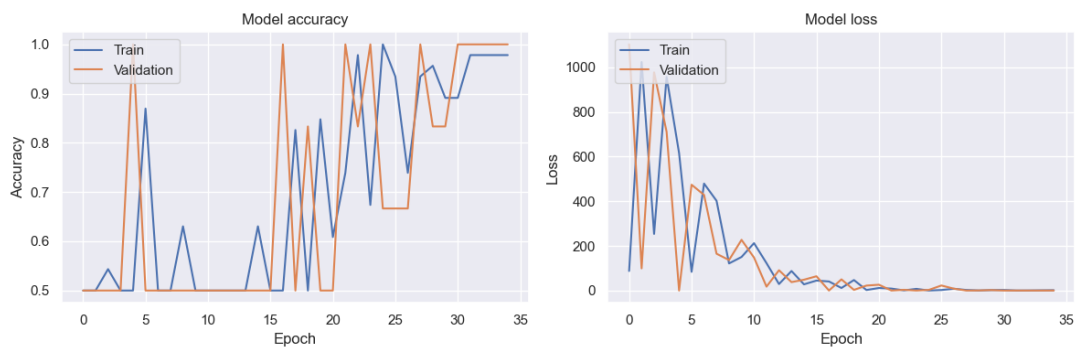


Figure 5.4: Results of Lung_GSE87340 Cancer Classification Model for Binary Class

The separation of data points done by our model for the tissue GSE87340 is shown in Figure 5.5. In these images, the normal profiles are represented in yellow, while other colours represent tumour profiles. These graphs show the capability of our methodology to separate the profiles regardless of the number and random distribution. These images show that our methodology is robust enough to handle linear and non-linear distributions of data. The abundance values of top genes for the tissue GSE87340 are shown in Figure 5.6. These figures show the power of our methodology to handle high variance, outliers and uneven distribution of data, all common in the identification and quantification of genes.

The cluster map of top genes for the tissue GSE87340 is shown in Figure 5.7.

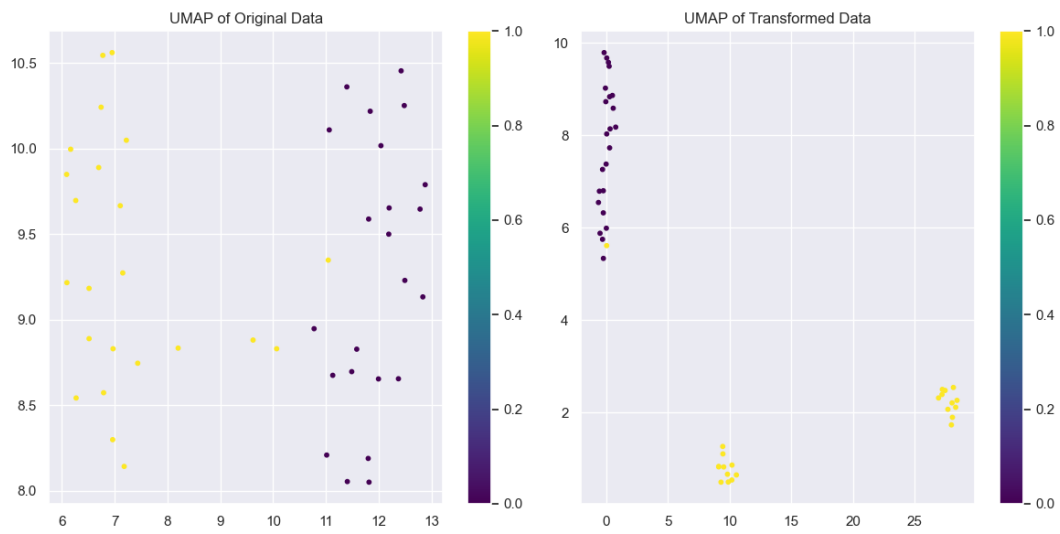


Figure 5.5: Data Point Distribution of Tumour and Normal profiles

These images show the classification robustness of our model with mixed kinds of correlation. Our methodology highlights the genes with high correlation and those with low correlation; our model classifies based on mixed features, which makes it more reliable.

The results of our model for the 4 classes of tissue SRR2089755 are shown in Figure 5.8. The separation of data points done by our model for the tissue SRR2089755 is shown in Figure 5.9. The abundance values of top genes for the tissue SRR2089755 are shown in Figure 5.10. The cluster map of top genes for the tissue SRR2089755 is shown in Figure 5.11.

The violin plot for gene expression levels provides a good representation of the distribution of expression values across a range of genes. The ample range in the shapes of the violins implies that the model is excellent at capturing both highly variable and smoother gene expression patterns. The various violin shapes across different genes, as well as the spread of expression values, support the model's capacity to discriminate across gene expression levels, which is critical for discovering cancer-specific biomarkers. The robust performance of the model in this image indicates its capacity to generalise across varied gene expression patterns while maintaining sensitivity to essential features, supporting its success in cancer classification tasks.

The shown cluster map gives a strong representation of the hierarchical clustering of genes based on their expression patterns, demonstrating key links between gene clusters. The obvious clusters of related genes illustrate the model's

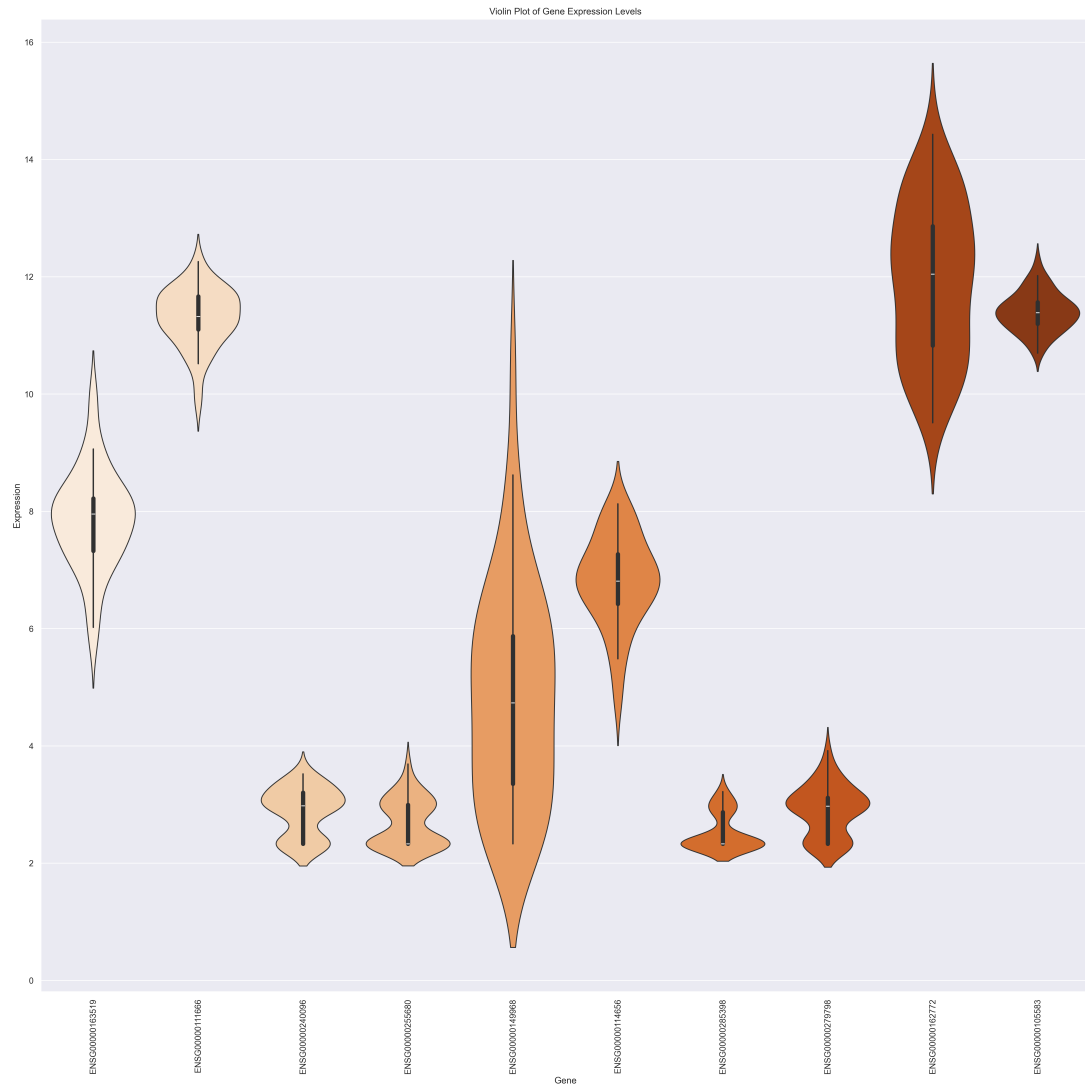


Figure 5.6: Abundance Value Distribution of Top Genes for Tissue GSE87340

robustness in finding patterns of gene co-expression. This robustness is especially relevant for cancer classification, as co-expressed genes generally lead to shared biological processes or pathways. The hierarchical structure in the dendrogram further highlights the model's ability to recognise complex interactions between genes. This level of deep clustering verifies the model's capacity to not only classify cancer kinds but also provide biological insights, particularly in the discovery of genes or gene sets that may operate as cancer biomarkers.

The results of our model for cancer type classification (8 classes) are shown in Figure 5.12. The separation of data points done by our model for 8 classes is shown in Figure 5.13. The abundance values of top genes for the 8 classes are shown in Figure 5.14. The violin plot for cancer-type classification shows the distribution of extracted characteristics, illustrating how effectively the model

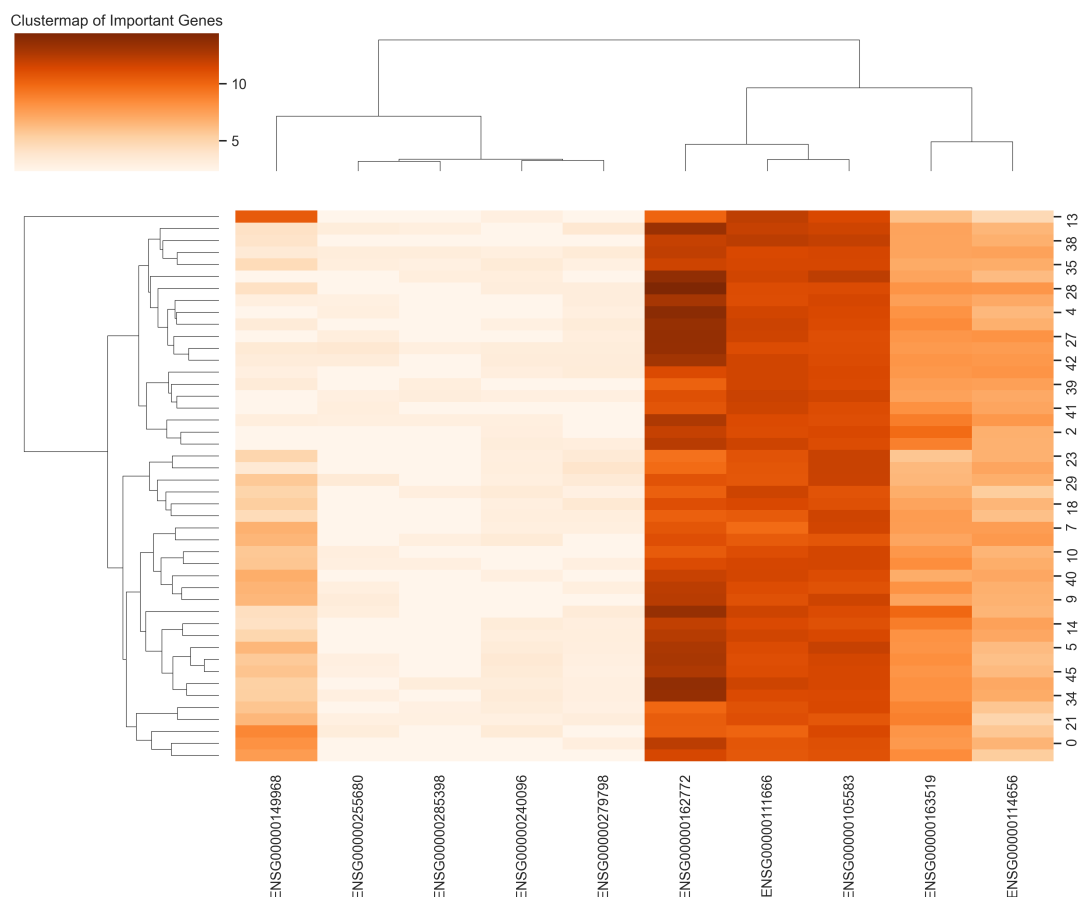


Figure 5.7: Cluster Map of Top Genes for Tissue GSE87340. Columns correspond to genes, and rows correspond to samples. The colour bar indicates normalised expression values.

captures essential properties from the RNA-seq data. The variety in the shape and size of the violins for different extracted data illustrates the model's capacity to catch distinct patterns across various cancer kinds. Some violins are significantly wider, showing a vast range of feature values that may correspond to diverse biological properties within cancer types. This diversity further confirms the model's robustness in retrieving and learning from the data, ensuring that it does not miss details that could be important for successful categorisation. The evident separations between the different extracted characteristics further underscore the model's capacity to generalise across numerous classes while maintaining performance across varied inputs.

The heatmap visualisation highlights the retrieved features and demonstrates a clear clustering of features that are strongly connected. The colour variations that extend from purple to yellow show the significance of these extracted features, with yellow signifying greater relevance. The distinct blocks of colour,

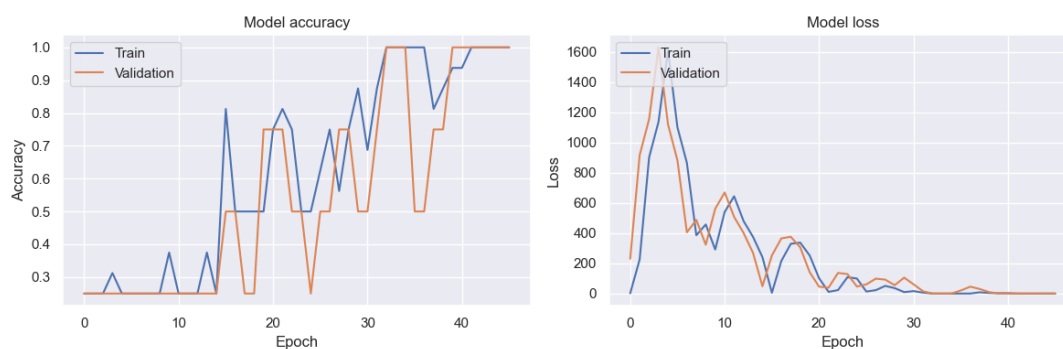


Figure 5.8: Results of Cancer Classification Model for Multi-Class Tasks

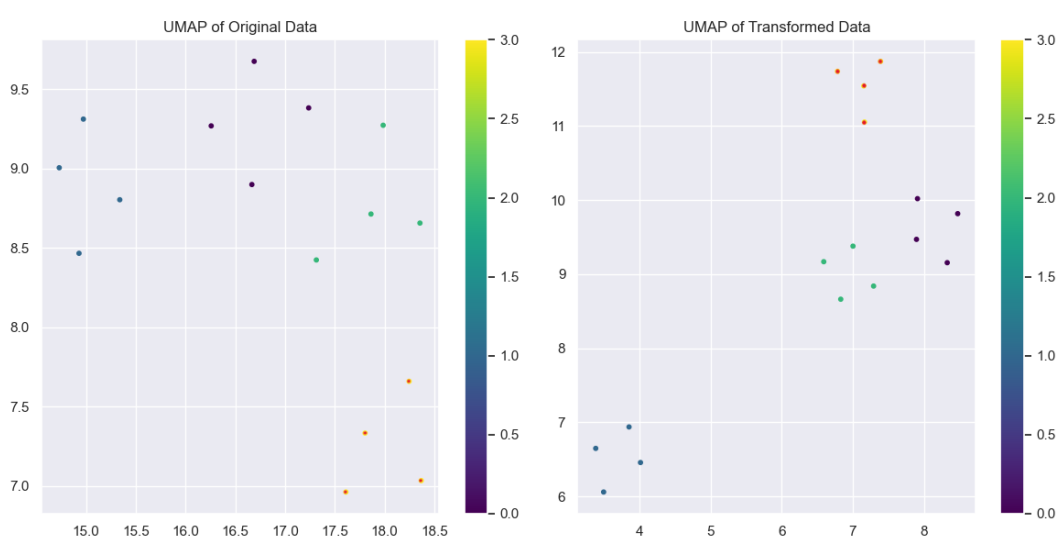


Figure 5.9: Data point distribution of Primary Tumour(Orange), Normal(Green), NormalLiver(Purple), Metastasis(Blue)

notably the clusters of brilliant yellow, suggest that the model effectively isolates crucial elements, further demonstrating its robustness. The results of the hierarchical cluster at both the row and column levels imply that the model identifies links between characteristics and data points. The clear clustering of these extracted characteristics illustrates the model's ability to group comparable patterns, which is crucial in difficult cancer-type classification tasks.

Overall, these visualisations collectively highlight the model's robustness in processing complicated, high-dimensional RNA-seq data and provide a detailed analysis of the categorisation and identification processes. The unique gene expression patterns and feature extraction results underline its ability to learn both high-level and nuanced patterns, ultimately increasing its classification performance.

The cluster map of top genes for the 8 Classes is shown in Figure 5.15. For the

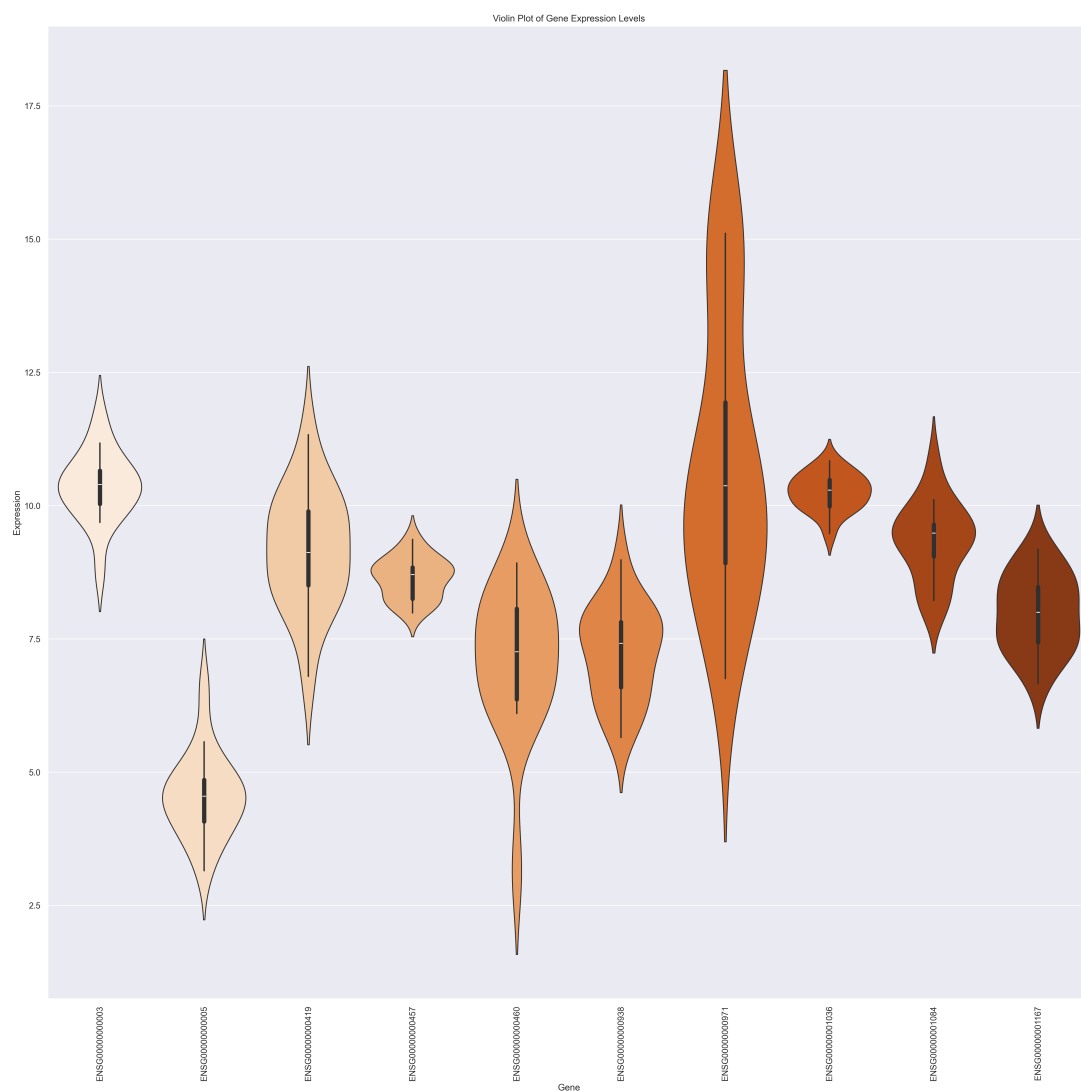


Figure 5.10: Abundance Value Distribution of Top Genes for Tissue SRR2089755

cancer type classification, the weighted average accuracy, precision, recall, F1-Score, and G-Mean on validation data are respectively 91%, 87%, 86%, 86%, and 84%, achieved on 8 classes. The distribution is given in Table 5.2. The confusion matrix for the 8 Classes is shown in Figure 5.16.

We also compared the results of our methodology with the baseline study [FPD21] shown in Figure 5.17. The results clearly show the improvements in the classification results and how our methodology outperformed the comparative study.

Our KEGG pathway enrichment analysis efficiently revealed a spectrum of cancer-associated genes, confirming our technology's efficiency in discover-

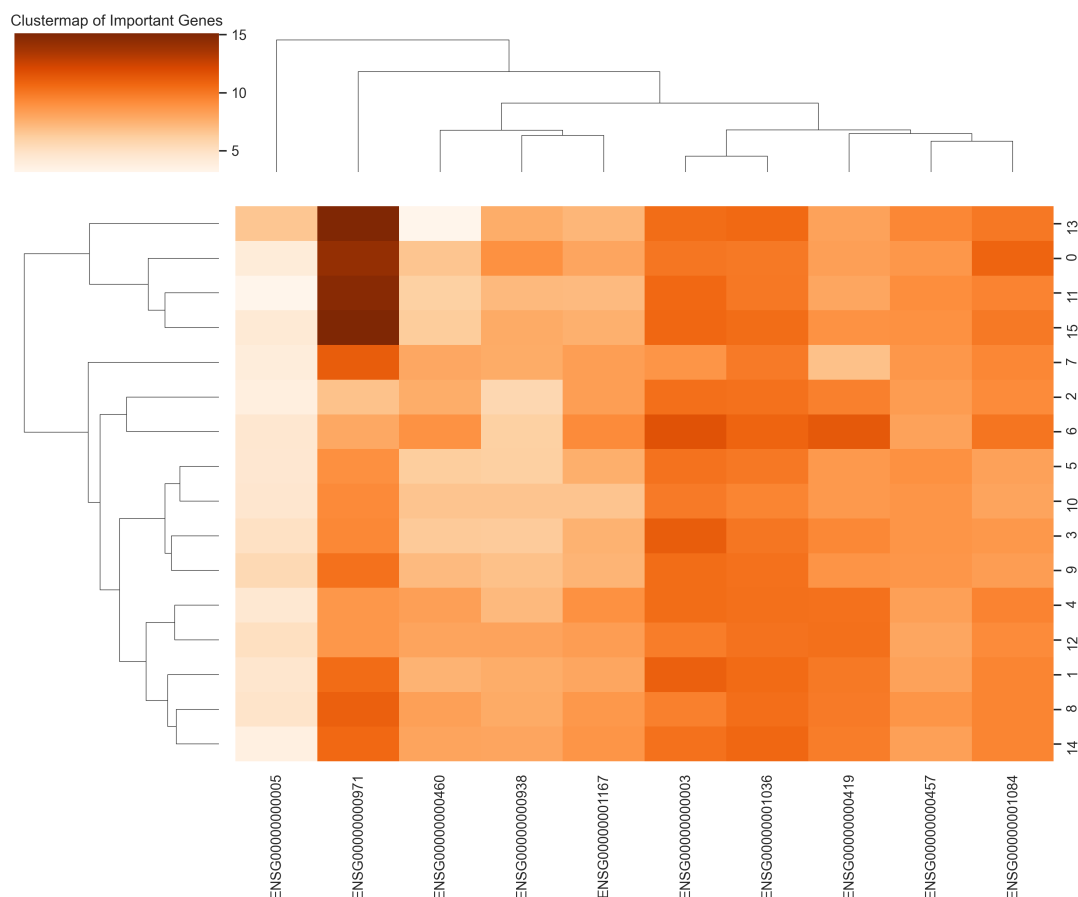


Figure 5.11: Cluster Map of Top Genes for Tissue SRR2089755. Columns correspond to genes, and rows correspond to samples. The colour bar indicates normalised expression values.

ing crucial cancer biomarkers [CZL⁺17]. Remarkably, genes associated in the “Transcriptional Misregulation in Cancer” [YCZ⁺21, CCR23] and “Acute Myeloid Leukemia” [LDB99, DEFW23] pathways were enriched, showing these genes’ critical role in cancer development [KL23]. These pathways comprise well-established tumour suppressors like TP53 [SHRyCP⁺23, MG23] and RB1[YLS⁺24], frequently altered in different malignancies. Their identification confirms our technique, indicating that it can dependably reveal vital genetic causes of cancer [JBT⁺24]. Furthermore, genes related to “Choline Metabolism in Cancer” [LHS⁺24], a route gaining notoriety for its participation in cancer cell proliferation and metabolism, were also discovered [LHS⁺24]. These findings emphasise our model’s capabilities, not simply to uncover well-known pathways but also to highlight the expanding areas of cancer research, such as cancer metabolism.

The bar plot in Figure 5.18 shows the significant biological processes enriched

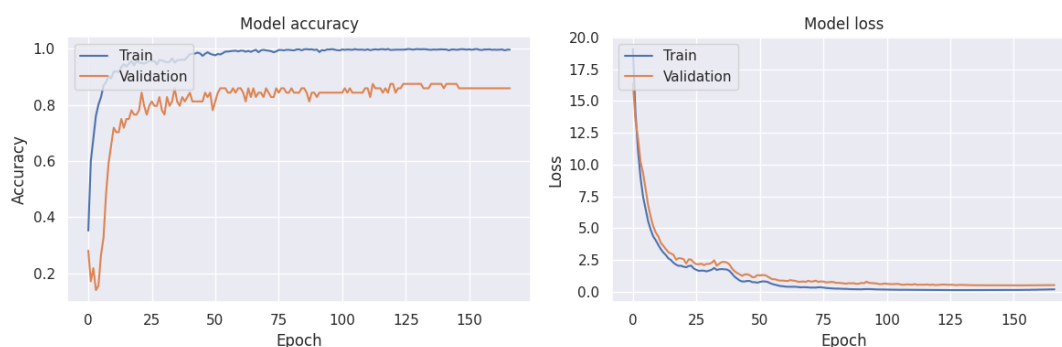


Figure 5.12: Results of Cancer Type Classification Model (8 Classes)

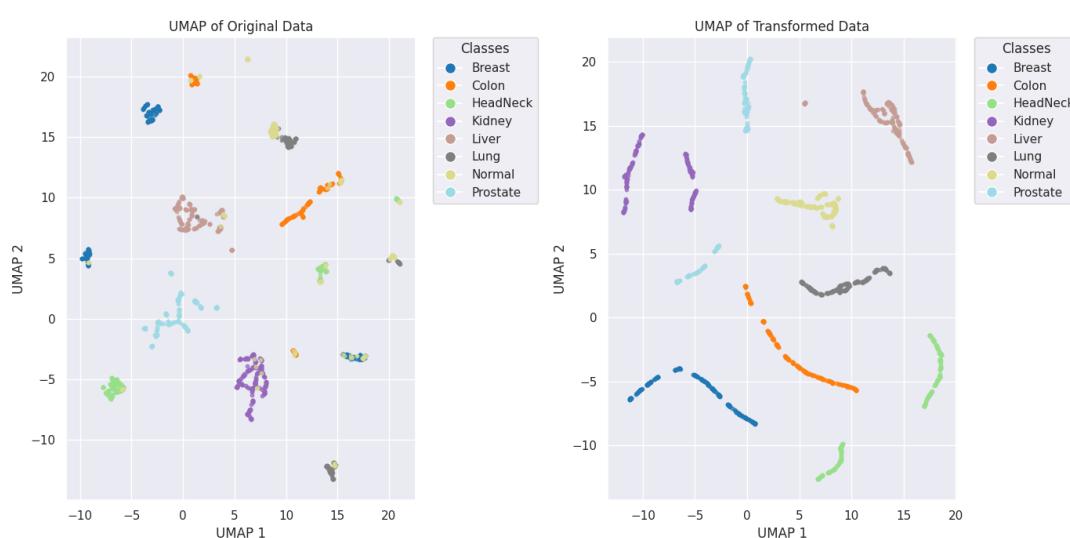


Figure 5.13: Data Point Distribution of Cancer Types and Normal

in our dataset, which were identified by GO enrichment analysis. The processes shown, such as the ether lipid biosynthetic process [PBCV23], TRAIL-activated apoptotic signalling pathway [GHG⁺23], and glycerol ether metabolic process [SB24], are crucial in comprehending the underlying molecular mechanisms in cancer. Lipid biosynthesis and metabolism, which are strongly featured in this analysis, are known to play critical roles in cancer cell survival and proliferation, as cancer cells often reorganise their metabolic pathways to favour rapid growth [HS24, FMZ⁺23]. Similarly, apoptotic mechanisms, such as the TRAIL-activated apoptotic signalling pathway, are crucial in regulating cell death, and their dysregulation is a hallmark of cancer [GM24]. The continuous relevance of these processes, as evidenced by the low adjusted p-values (p.adjust), ranges between 0.044 and 0.0045, demonstrating the robustness of our approach to identifying physiologically meaningful pathways. This underlines the efficacy of our technology in discovering possible biomarkers and

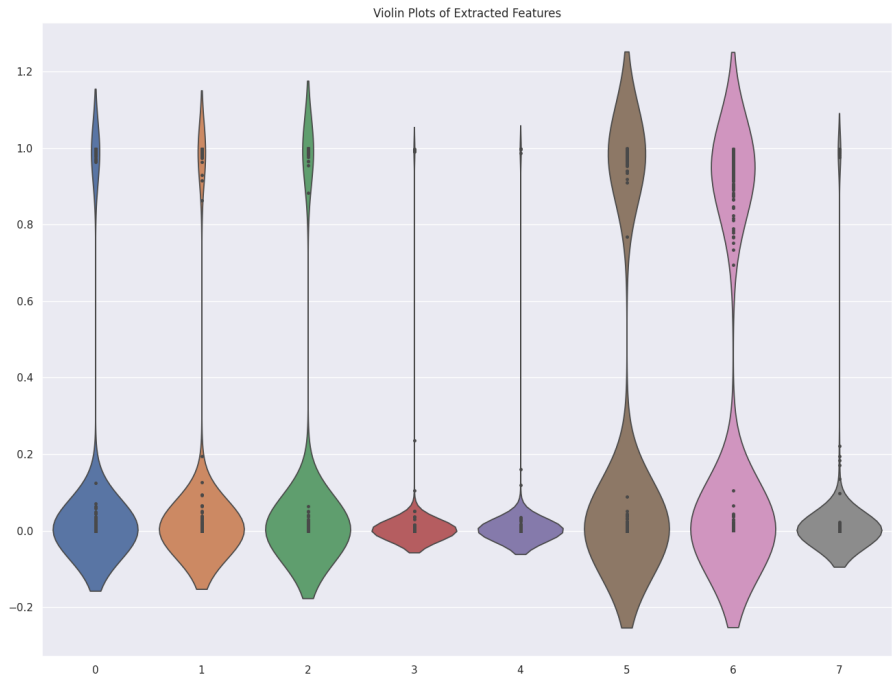


Figure 5.14: Abundance Value Distribution of Top Genes for 8 Classes

Table 5.2: Classification performance metrics (Precision, Recall, F1-Score, and Accuracy) for each cancer type on the validation dataset.

Class	Precision	Recall	F1-Score	Accuracy
Breast	0.86	1.00	0.92	1.00
Colon	0.60	0.60	0.60	0.60
HeadNeck	1.00	1.00	1.00	1.00
Kidney	1.00	1.00	1.00	1.00
Liver	1.00	1.00	1.00	1.00
Lung	0.95	0.95	0.95	0.95
Normal	0.84	0.73	0.78	0.73
Prostate	0.50	1.00	0.67	1.00

uncovering essential cancer-related cellular processes. By employing XAI, our technique further provides insights into the crucial molecular mechanisms driving these classifications, thereby revealing potential novel therapeutic targets.

The XAI adds transparency and interpretability to the deep learning model’s predictions. Using LIME, we managed to examine the contribution of indi-

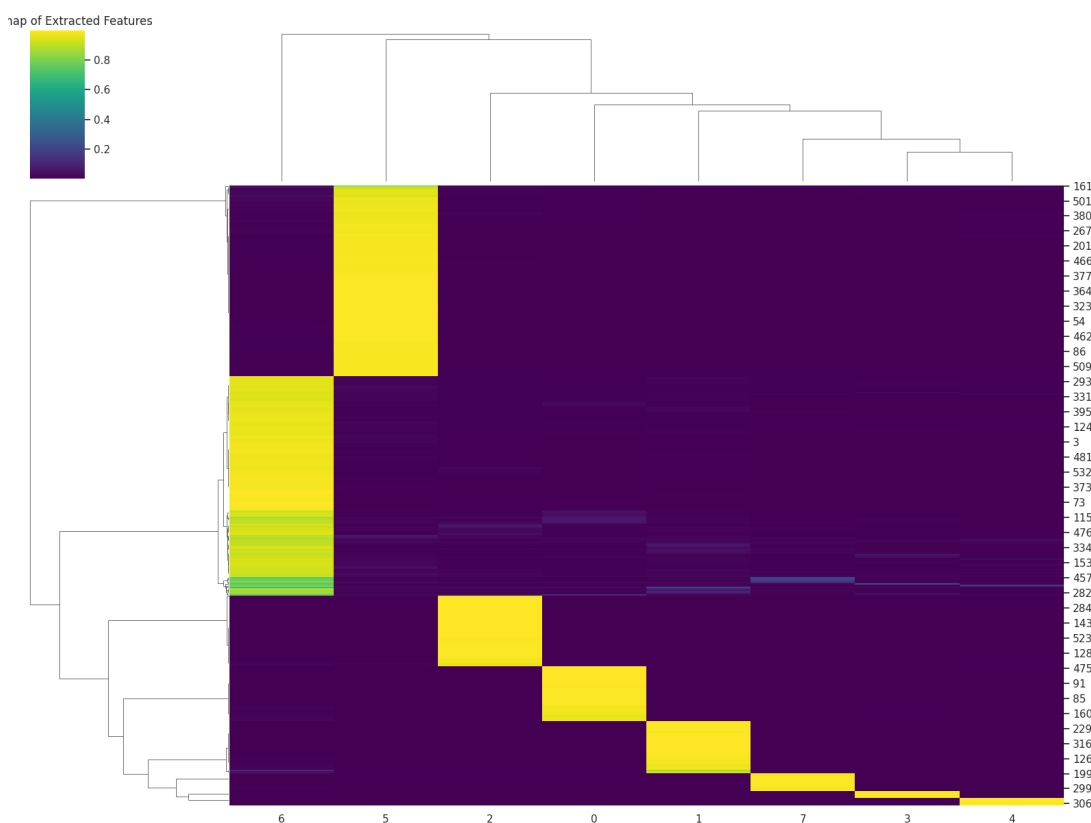


Figure 5.15: Cluster Map of Top Genes for 8 classes. Columns correspond to genes, and rows correspond to samples. The colour bar indicates normalised expression values.

vidual genes to the categorisation decision. This allowed us to understand the role of specific genes in driving the enrichment of cancer-related pathways and how they contribute to the model’s decision-making process. For instance, genes found in “MicroRNAs in Cancer” and “Nicotinate and Nicotinamide Metabolism”, although not typically associated with well-established cancer pathways [BC24, MZB24], were highlighted by XAI as significant contributors to classification. This indicates that these genes may have under-explored roles in cancer, making them possible new biomarkers for further research. The XAI framework thus not only increases our model’s prediction capacity but also helps validate the biological relevance of the genes identified, guaranteeing that the model’s output is biologically interpretable. By offering insights into how these genes influence classification outcomes, XAI offers a deeper understanding of our model’s predictions, helping to prioritise genes for experimental validation.

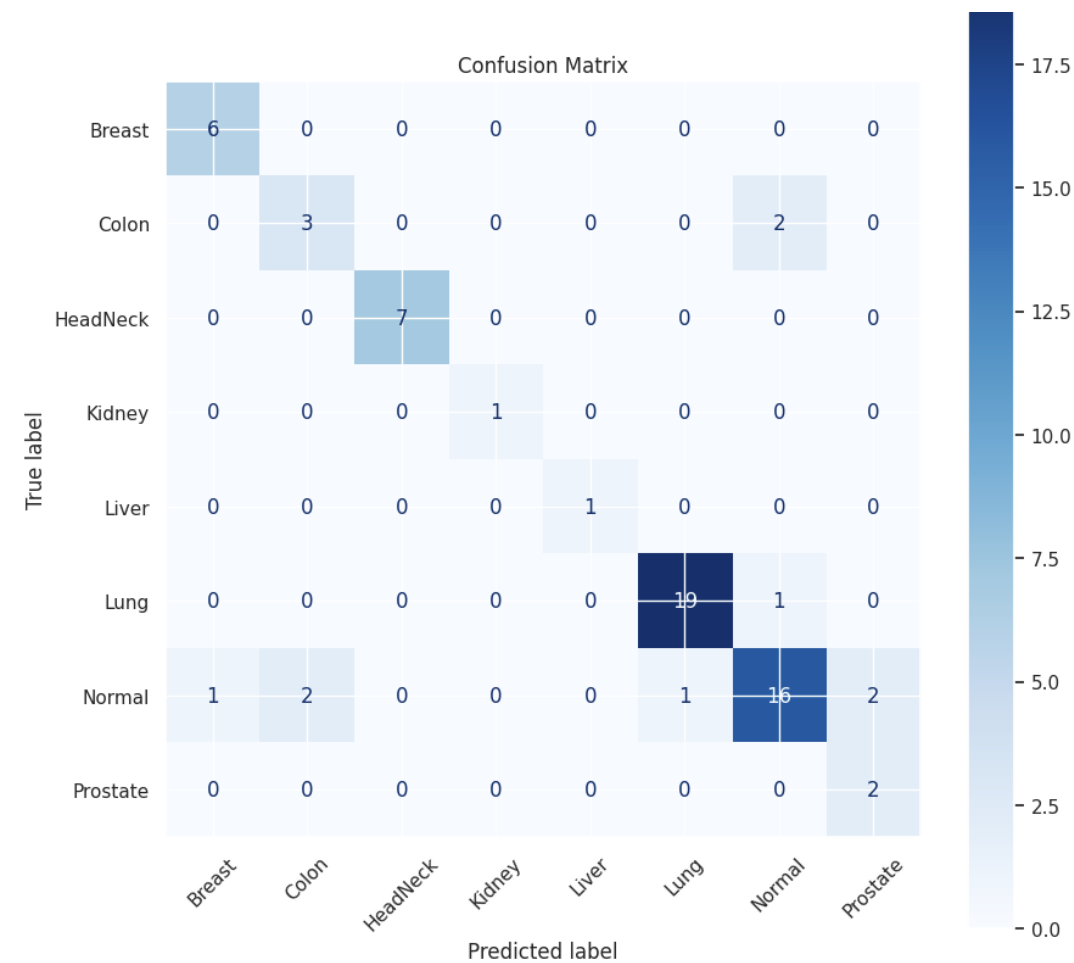


Figure 5.16: Confusion Matrix for Cancer Type Classification

5.3.3.1 Results for Microarray Datasets

For the CuMiDa database, we compared our results with [GFD19], and our methodology outperformed their methodology, as shown in Figure 5.19. Out of the 73, our cancer subtype classification model outperformed in 47, equalled in 19 and underperformed in 8 Datasets.

Our cancer classification model performed very well for the 14 classes and handled the CuMiDa data with 80% average accuracy. The results are shown in Figure 5.20, and the data point distribution done by our model is shown in Figure 5.21.

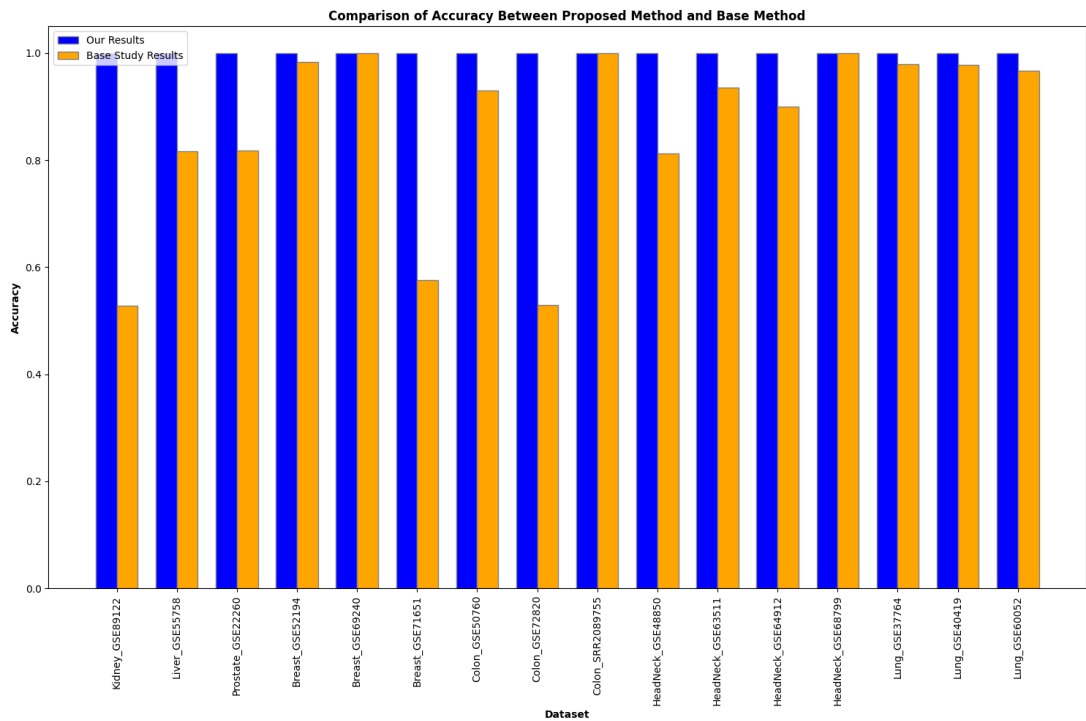


Figure 5.17: Comparison Between Proposed Method and Base Method

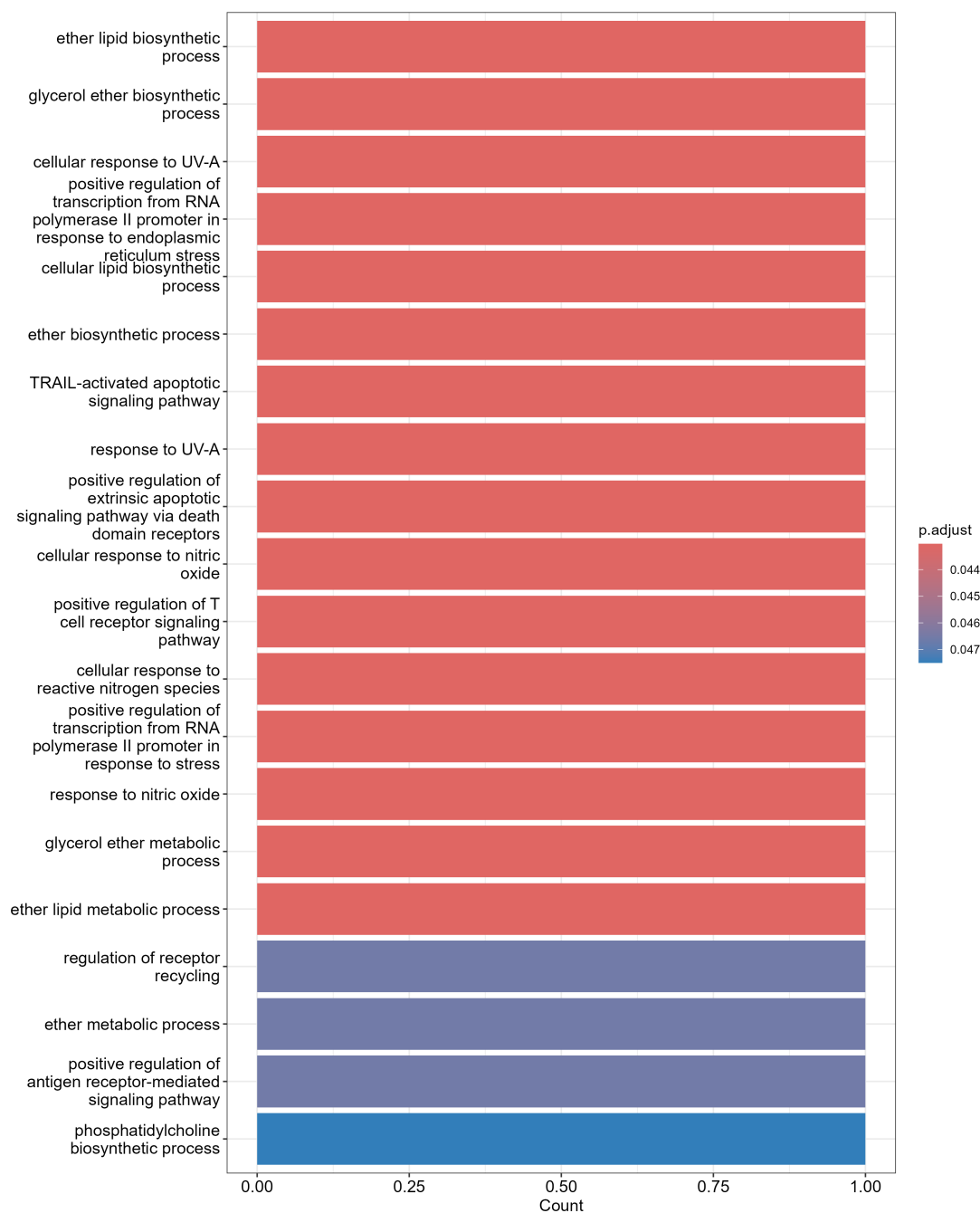


Figure 5.18: Gene Ontology (GO) Enrichment Analysis

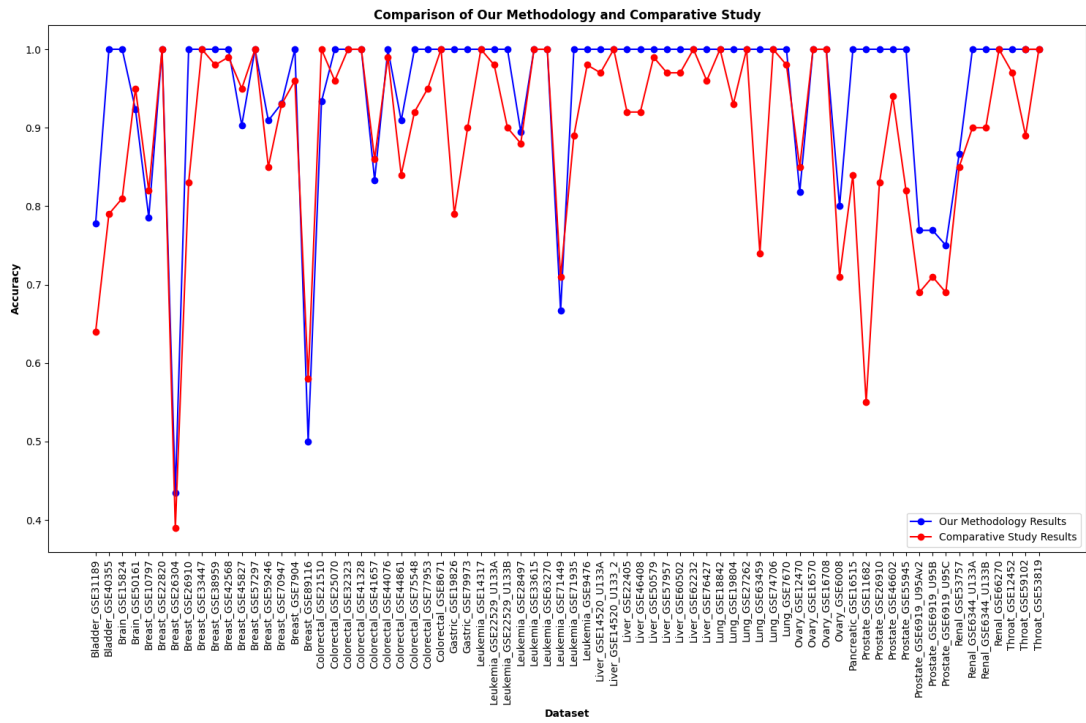


Figure 5.19: Comparison of our Methodology with Base Study [GFD19] on CuMiDa.

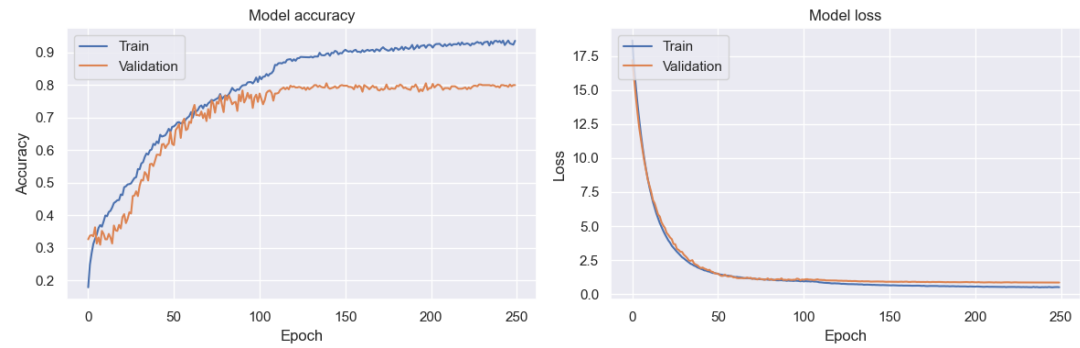


Figure 5.20: Results of Our Methodology on CuMiDa

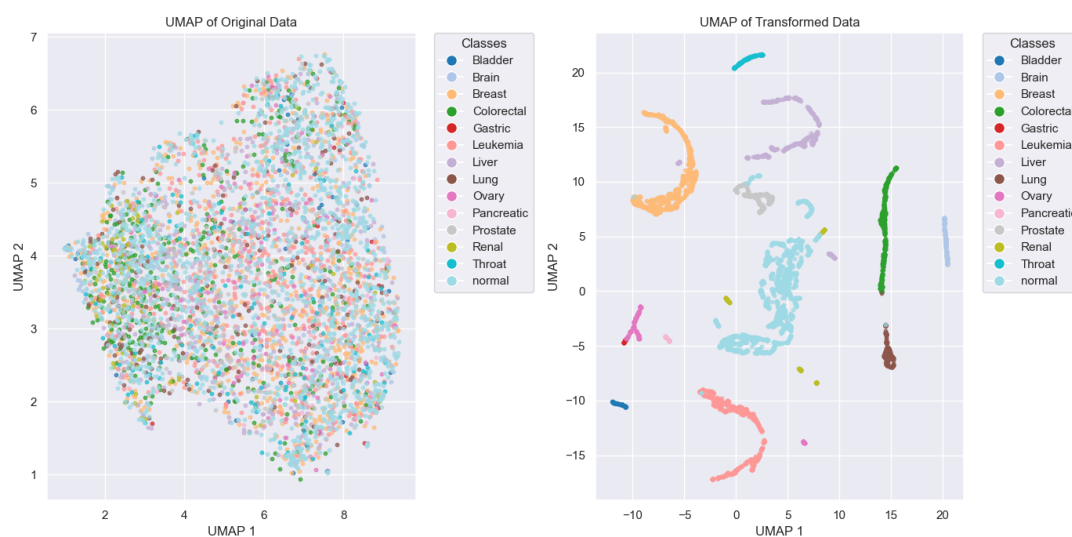


Figure 5.21: Data Point Distribution of Original Dataset and Model Separation on CuMiDa

In this chapter, we applied deep learning, XAI, KEGG Pathway enrichment analysis and data visualisation to offer a robust, efficient pipeline to both classify different cancers with the same model and to trace relevant genes to the high performance of the models. In the process, we provided deep learning-based solutions to deal with data with highly imbalanced classes, large dimensions and low instance numbers, all persistent issues when dealing with biological data. While transcriptomic profiles shed light on gene activity, they do not fully represent the functional state of a cell. Proteins serve as the direct effectors of cellular processes, and their disruption often drives disease phenotypes. Chapter 6 extends the analysis to proteomic data, leveraging MS to identify protein-complex level alterations relevant to cancer biology.

Chapter 6

Proteomic Complex Disruption Analysis for Biomarker Discovery in Oral Cancer

6.1 Introduction

This chapter presents an understanding of the proteome changes related to the progression of oral cancer by combining proteomic data, enhanced feature selection, machine learning classification, and PPIN analysis. We employ an integrative bioinformatics approach to dissect the proteomic landscape of oral cancer, investigating HMK, SCC-9, and HSC-3 cell lines. The approach strikes a balance between computational tractability and data depth by choosing 50, 100, and 200 protein attributes, ensuring that the most informative proteins are pursued for additional examination. These feature sets are then subjected to classification methods implemented in machine learning frameworks. They are into distinct groups using a variety of algorithms, including ensemble approaches, LR, and SVM. This classification provides insight into the proteome profiles unique to each form of cancer and acts as a standard for evaluating how well the selected features work. Analysing the protein interaction networks is a critical next step. The study explores the complex web of protein interactions using the Predicting Interactomes from Co-Elution (PRINCE) R package [SCSF21], a programme created especially for building and analysing PPINs. Using machine learning to forecast PPIs from co-elution data, PrInCE has a methodology that fits well with the high-throughput requirements of contemporary proteomics.

Cytoscape [SMO⁺03], a platform that makes these networks visually vibrant and offers insights into their topological and functional characteristics, is then used to visualise and analyse the resultant networks. Important proteins and their interactions are discovered, potentially providing biomarkers or therapeutic targets, especially those that are markedly highlighted in the aggressive cancer cell lines. The integration of proteomics and bioinformatics increases our knowledge of PPIs associated with oral cancer and highlights the players that can be further investigated as therapeutic targets or biomarkers for the disease.

6.2 Methodology

This section covers the materials, data sources, and detailed procedures applied in our work to classify oral cancer cell lines based on co-fractionation followed by Liquid Chromatography–Tandem Mass Spectrometry (LC-MS/MS) data. The study focuses on data obtained from three cell lines: HMK are normal cell lines, while SCC-9, and HSC-3 are oral cancer cell lines. We provide a detailed review of data preparation stages, data visualisation and feature selection methods, application of machine learning algorithms for classifying these cell lines based on protein features, and PPI to evaluate protein complexes. The following sections offer details of each stage of the methodology, ensuring clarity in our approach. An abstract diagram of the methodology is shown in Figure 6.1. In the proposed pipeline, data processing is done, which includes standardisation and cleaning of data. After pre-processing of data, feature selection methods were used to shortlist the relevant proteins that can play an important role in the classification of different oral cancer cell lines, as well as are enough to analyse the PPINs. Subsequently, machine learning algorithms were used to classify the shortlisted proteins and determine the importance of the selected proteins. Then, selected proteins were analysed by visual methods and used for the identification of co-eluting protein complexes. On the other hand, PPI was done and PPIN were generated. For the validation of generated PPIN, literature as well as CORUM dataset was used. Each of the elements of the diagram is individually explained below.

6.2.1 Data Collection and Preprocessing

The study utilised mass spectrometry data from oral cancer cell lines that were collected by Co-Fractionation (Co-Fractionation) combined with LC-MS/MS

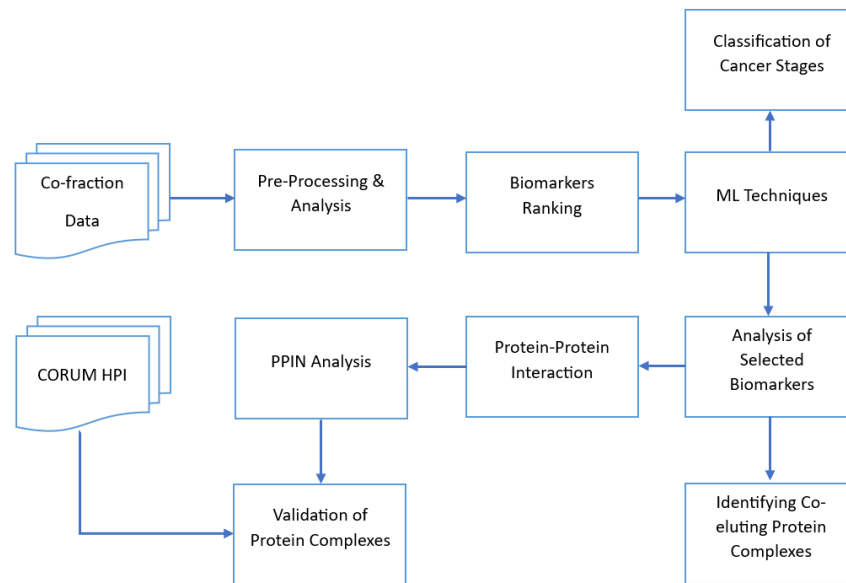


Figure 6.1: Flow Diagram of the proposed proteomic analysis workflow

analysis. The data was collected in the Mass Spectrometry Laboratory teams from the Brazilian Biosciences National Laboratory at the Centre for Research in Energy and Materials (LNBio/CNPEN). Details and availability of the data are provided in our partner's paper [TAGA⁺25]. Four datasets were considered:

- Peptides_without_match_between_runs_Unique
- Peptides_without_match_between_runs_Unique_Razors
- Protein_groups_without_match_between_runs_Unique
- Protein_groups_without_match_between_runs_Unique_Razors

The data preprocessing involved handling missing values: Any missing values were either imputed or excluded based on the dataset's characteristics. After that, the data were normalised to ensure consistent scales or standardised. We separated the data by type and generated a separate file for each category with protein names. These separated files are input to PPI and co-eluted complex identification. Using these files, we analysed our data with heatmaps to find the association between the data.

6.2.2 Data Visualisation

To show the structure of the data and the relationships between its many aspects, heat maps were created. These heatmaps made it easier to find patterns

and connections between the cell lines. Correlation heatmaps for classes, each comprised of instances, are utilised to visually represent the associations between features (proteins) within and among these classes. These heatmaps furnish a graphical depiction of the extent to which pairs of features are inter-related. When considering three distinct classes, designated as HMK, SCC-9, and HSC-3, each containing m instances, and a selection of proteins denoted as $\{P_1, P_2, P_3, \dots, P_n\}$, where P_i represents the i^{th} protein in a total of n selected proteins, the objective is to compute and visualize correlations both within each class and between the classes.

A data matrix (X) is constructed, where each row corresponds to an instance, and each column represents a feature (protein). The matrix dimensions are $(m \times n)$, with n signifying the number of selected proteins as in Equation 6.1.

$$X = [X_{\text{HMK}}; X_{\text{HSC-3}}; X_{\text{SCC-9}}] \quad (6.1)$$

- X_{HMK} represents the data for class HMK (m instances $\times$ n features).
- $X_{\text{HSC-3}}$ represents the data for class HSC-3 (m instances $\times$ n features).
- $X_{\text{SCC-9}}$ represents the data for class SCC-9 (m instances $\times$ n features).

The correlation matrix (R) is computed individually for each class. It illustrates the pairwise correlations between features within each class.

$$R_{\text{HMK}} = \text{corrcoef}(X_{\text{HMK}})$$

$$R_{\text{HSC-3}} = \text{corrcoef}(X_{\text{HSC-3}})$$

$$R_{\text{SCC-9}} = \text{corrcoef}(X_{\text{SCC-9}})$$

Where R_{HMK} represents the correlation matrix for class HMK, $R_{\text{HSC-3}}$ represents the correlation matrix for class HSC-3, and $R_{\text{SCC-9}}$ represents the correlation matrix for class SCC-9.

6.2.3 Biomarker Ranking

To rank the important biomarkers that play a significant role in classifying the cancer cell lines, four feature selection methods were employed. These four methods are outlined below, and their mathematical formulations and theoretical inspiration are derived from the works of Breiman et al. [BFOS17], Kursu

et al. [KJR10], and Dash and Liu [DL97].

6.2.3.1 ExtraTree Feature Selection

Extremely Randomized Trees (ExtraTrees), an ensemble technique, was used to evaluate feature importance based on the Gini impurity (see Equation 6.14). Features with higher importance were selected. In the realm of ensemble learning methods, ExtraTrees, also recognised as Extremely Randomised Trees, is a prominent approach employed for predictions. It operates by harnessing multiple DTs to make informed predictions. While akin to the RF technique, ExtraTrees distinguishes itself through a unique approach to constructing its constituent DTs.

In the context of ExtraTrees, each DT is forged from the original training data. Nevertheless, a noteworthy departure from conventional methods occurs at each decision node. Here, a random subset of features is selected, and the optimal split is chosen based on an established criterion, such as Gini impurity or information gain (see Equation 6.15). This methodical introduction of randomness plays a pivotal role in decorrelating the DT and mitigating the risk of overfitting. ExtraTrees offers an advantageous avenue for feature selection, a process where the relevance of individual features is assessed. Feature importance is established by evaluating the average reduction in impurity or information gain across all branching points within the DTs where a given feature is employed. Features that command higher importance scores are given a greater degree of significance in the context of the prediction task.

The calculation of the importance score for a specific feature denoted as f is structured as in Equation 6.2:

$$\text{Imp}(f) = \text{avg}(\Delta I_f) \quad (6.2)$$

Here, ΔI_f signifies the extent to which impurity is diminished or information gain is augmented by the utilisation of feature f .

For the Gini impurity criterion, the reduction is quantified as in Equation 6.3:

$$\Delta I_f = G(P) - \bar{G}(L, R) \quad (6.3)$$

In this equation, $G(\cdot)$ characterises the Gini impurity associated with a node. The weighted average $\bar{G}(L, R)$ is computed based on the sample counts in the

left (L) and right (R) child nodes. Alternatively, for the information gain criterion, the reduction is ascertained through the formula as in Equation 6.4:

$$\Delta I_f = H(P) - \bar{H}(L, R) \quad (6.4)$$

Within this equation, $H(\cdot)$ encapsulates the entropy measure of a node. The weights are proportionally determined in alignment with the sample distribution within each respective child node.

The applicability of ExtraTrees extends to the realm of protein feature selection. This approach is of immense utility in tasks linked to protein classification or the prediction of protein attributes. The features may encompass diverse aspects, such as amino acid sequences, physicochemical properties, or structural information. The importance scores derived through ExtraTrees are invaluable in discerning the most pertinent features tailored to the specific task. By effectively reducing the dimensionality of the data, this feature selection process considerably enhances the performance of machine learning models, particularly in the context of protein-related applications.

6.2.3.2 K-Best Feature Selection

The k-best feature selection method employed statistical tests to select the top k features based on their p-values. The k-best feature selection method is utilised to select the top “ k ” most pertinent features (proteins, in this context) from a dataset for further analysis. This technique relies on statistical tests that evaluate the significance of each feature concerning the target variable. A common statistical test employed for this purpose is the Analysis of Variance (ANOVA). Here, we will explain the k-best feature selection method and the mathematical equations involved, focusing on ANOVA. The F-statistic for a specific protein feature i is determined using the F-statistic formula. This statistic gauges the variance ratio between group means to the variance within each group. The formula for computing the F-statistic is as in Equation 6.5.

$$F_i = \frac{SSB_i/(k-1)}{SSW_i/(n-k)} \quad (6.5)$$

Where:

- F_i denotes the F-statistic for feature i ,
- SSB_i represents the sum of squares between groups for feature i ,

- SSW_i corresponds to the sum of squares within groups for feature i ,
- k is the number of groups (classes), in this case, the number of distinct types of oral cancer cell lines,
- n is the total count of instances (samples).
- Sum of Squares Between Groups (SSB_i) is calculated as in Equation 6.6.

$$SSB_i = \sum_{c=1}^k n_c (\mu_{ic} - \mu_i)^2 \quad (6.6)$$

where:

- n_c indicates the number of instances in class c ,
- μ_{ic} is the mean of feature i within class c ,
- μ_i denotes the overall mean of feature i across all classes.
- Sum of Squares Within Groups (SSW_i) is calculated as in Equation 6.7.

$$SSW_i = \sum_{c=1}^k \sum_{j=1}^{n_c} (x_{icj} - \mu_{ic})^2 \quad (6.7)$$

where:

- x_{icj} represents the value of feature i for the j^{th} instance in class c ,
- μ_{ic} is the mean of feature i for class c ,
- n_c corresponds to the number of samples in class c .

Once the ANOVA scores are computed for all protein features, they can be ranked in descending order based on their F-statistic values. Subsequently, the top k features with the highest F-statistic values are chosen to constitute the final set of pertinent protein features for the analysis. This feature selection approach permits a focus on a subset of proteins that exhibit the most substantial discriminatory power in classifying distinct types of oral cancer cell lines, potentially enhancing the efficiency and efficacy of the classification model.

6.2.3.3 Recursive Feature Elimination (RFE)

RFE used a machine learning model (e.g. SVM) to iteratively remove the least important features until the desired number of features was achieved. RFE is a feature selection method that systematically assesses the importance of each feature, with proteins as features in this context. This assessment is carried out through the iterative construction of models and the elimination of the least significant features. RFE is frequently employed in combination with machine learning models to identify the most informative features relevant to a particular task, such as the classification of oral cancer cell lines. The mathematical equations RFE primarily pertain to the ranking of features, the elimination of the least important feature, and the iterative procedures are shown in Equations 6.8 and 6.9, respectively.

– **Feature Ranking:**

$$GiniDecrease(i) = Gini(D) - \sum_{v=1}^V w_v Gini(D_v) \quad (6.8)$$

Where:

- * $GiniDecrease(i)$ denotes the Gini impurity decrease for feature i ,
- * $Gini(D)$ represents the Gini impurity of the parent dataset D ,
- * D_v refers to the subset of data resulting from the v^{th} split on feature i ,
- * $w_v = \frac{|D_v|}{|D|}$ is the proportion (weight) of samples in subset D_v relative to the parent dataset,
- * $Gini(D_v)$ indicates the Gini impurity of subset D_v ,
- * V is the total number of splits (branches) created by the feature i .

• **Feature Elimination:**

$$FeatureSet = FeatureSet - \{Feature_i\} \quad (6.9)$$

Where:

- *FeatureSet* refers to the current set of features,
- *Feature_i* designates the feature with the lowest importance score.
- Model Rebuilding: After eliminating the least important feature, a fresh model is trained utilising the reduced feature set. This process is reiterated in a continuous manner.
- Stopping Criterion: The RFE process concludes when a specified number of features k is attained, or when a stopping criterion is met, such as the absence of significant enhancement in model performance.

RFE allows for the systematic identification and selection of the most informative proteins for a classification task, contributing to the refinement of the feature set and potentially enhancing the efficiency and performance of the machine learning model.

6.2.3.4 Boruta Feature Selection

Boruta is a wrapper method that extends RF to identify relevant features by comparing them with a shadow set of random features. It is used to identify the most relevant proteins for a given classification or prediction task.

The mathematical equations central to Boruta pertain to comparing real feature importance to shadow features and determining feature relevance. Feature importance scores depend on model-specific criteria, such as Gini impurity decrease in the case of a RF classifier, as in Equation 6.10.

$$GiniDecrease(i) = Gini(parent) - \text{Weighted_Average}(Gini(left_child), Gini(right_child)) \quad (6.10)$$

Where:

- *GiniDecrease(i)* signifies the Gini impurity decrease for feature i ,
- *Gini(parent)* represents the Gini impurity of the parent node,
- *Gini(left_child)* and *Gini(right_child)* correspond to the Gini impurities of the left and right child nodes.

A user-defined threshold (*Threshold*) is established for feature importance. Features with scores exceeding this threshold are relevant, while those below it are unimportant and removed from the dataset. Boruta's approach, comparing real features to shadow features, helps select informative proteins, reducing

dataset dimensionality. Threshold adjustment balances feature selection and model performance. Different sets of proteins are selected using these feature selection methods, and machine learning classification algorithms are used to assess the importance of these feature sets.

6.2.4 Machine Learning Methods

Several machine learning algorithms were applied to classify cancer cell lines based on the fractions and calibrate the importance of these features in finding the protein-protein interaction. The algorithms used are the following:

6.2.4.1 k-Nearest Neighbours (k-NN)

The KNN Classifier [LL14], a supervised machine learning algorithm, is employed for classification tasks, utilising proteins as features and classifying data into HMK, SCC-9, and HSC-3 categories. Classification of an instance is based on the majority class among its “k” nearest neighbours in the feature space. The KNN algorithm primarily relies on calculating Euclidean distances between instances and determining the majority class among the “k” nearest neighbours. The Euclidean distance between two instances, represented by vectors A and B in a feature space with n features, can be calculated as in Equation 6.11.

$$\text{Distance}(\mathbf{A}, \mathbf{B}) = \sqrt{\sum_{i=1}^n (A_i - B_i)^2} \quad (6.11)$$

Where:

- $\text{Distance}(\mathbf{A}, \mathbf{B})$ is the Euclidean distance between instances $\mathbf{A}$ and $\mathbf{B}$,
- A_i and B_i represent the values of the i -th feature for instances $\mathbf{A}$ and $\mathbf{B}$, respectively.

To classify an instance, the KNN algorithm identifies the k instances from the training dataset that are closest to the instance being classified, based on a chosen distance metric. These instances are referred to as the k nearest neighbours, as expressed in Equation 6.12.

$$\mathcal{N}_k(\mathbf{x}) = \arg \min_{S \subset \mathcal{D}, |S|=k} \sum_{\mathbf{x}_i \in S} \text{dist}(\mathbf{x}, \mathbf{x}_i) \quad (6.12)$$

where:

- $\mathcal{N}_k(\mathbf{x})$ denotes the set of k nearest neighbours of the query instance $\mathbf{x}$,
- $\mathcal{D}$ represents the complete training dataset,
- $\text{dist}(\mathbf{x}, \mathbf{x}_i)$ is a distance metric (commonly Euclidean, Manhattan, or cosine distance),
- $\mathcal{S}$ is a subset of $\mathcal{D}$ containing exactly k instances that minimise the total distance to $\mathbf{x}$.

Once the nearest neighbours are identified, the algorithm determines the majority class among these neighbours and assigns that class to the instance being classified. This step can be expressed as in Equation 6.13.

$$\text{Class}(\mathbf{x}) = \text{mode}(\{y_i \mid \mathbf{x}_i \in \mathcal{N}_k(\mathbf{x})\}) \quad (6.13)$$

Where:

- $\text{Class}(\mathbf{x})$ represents the predicted class for the instance $\mathbf{x}$,
- $\text{mode}(\{y_i \mid \mathbf{x}_i \in \mathcal{N}_k(\mathbf{x})\})$ is the class that occurs most frequently among the “ k ” nearest neighbours.

Effective operation of the KNN algorithm is achieved by selecting an appropriate value for “ k ” and choosing a suitable distance metric. Considering nearest neighbours’ classes provides a straightforward and intuitive approach to classification tasks, including classifying data into the HMK, SCC-9, and HSC-3 categories based on protein features.

6.2.4.2 Decision Tree Classifier

The DT Classifier, a supervised machine learning algorithm, is also employed for classification tasks, utilising proteins as features and classifying data into HMK, SCC-9, and HSC-3 categories. DTs create a hierarchical tree-like structure to make decisions based on feature values. The mathematical equations for the DT Classifier primarily revolve around the criteria used to select the best feature for splitting and the determination of the majority class in a leaf node. The theoretical inspiration and mathematical formulations for tree-based classifiers used in this study are derived from the works of Breiman et al. [BFOS17]

- **Gini Impurity:** The Gini impurity measures the probability of misclassifying an instance. For a given node t with C classes and $p(i)$ denoting

the probability of class i , the Gini impurity $G(t)$ is calculated as in Equation 6.14.

$$G(t) = 1 - \sum_{i=1}^C (p(i))^2 \quad (6.14)$$

- **Information Gain:** Information gain quantifies the reduction in entropy achieved after a split. For a given node t , the information gain $IG(t)$ is computed as in Equation 6.15.

$$IG(t) = H(t) - \sum_{v=1}^V \frac{|D_v|}{|D|} H(v) \quad (6.15)$$

where:

- $H(t)$ is the entropy of node t ,
- D_v represents the subset of instances in the v^{th} child node,
- $|D_v|/|D|$ is the proportion of samples in that child node,
- $H(v)$ is the entropy of child node v ,
- V is the total number of child nodes (e.g., $V = 2$ for binary splits).

To classify a new instance, the decision tree is traversed from the root node to a leaf node based on the feature values of the instance. The majority class in the leaf node is assigned as the predicted class for the instance.

These mathematical equations describe the key components of the DT Classifier, which constructs a tree-like structure to make decisions based on feature values and is used for classifying data into HMK, SCC-9, and HSC-3 categories based on protein features.

6.2.4.3 Random Forest Classifier

The RF Classifier, an ensemble machine learning algorithm, is employed for classification tasks, utilising proteins as features and classifying data into HMK, SCC-9, and HSC-3 categories. It comprises an ensemble of DTs, where multiple DTs are trained, and their predictions are combined to enhance overall accuracy and reduce overfitting. The RF algorithm primarily involves the use of multiple DTs and aggregation techniques. The mathematical representation of the core components is described in Equation 6.16.

- **Bootstrapping:** To create a random subsample, a process called bootstrapping is employed. This involves randomly selecting instances from the dataset with replacement. The probability of an instance being selected is given in Equation 6.16.

$$\text{Probability(Selected)} = 1 - \left(\frac{1}{n}\right)^n \quad (6.16)$$

where “n” is the number of instances in the dataset.

- **Feature Subsampling:** A random subset of features is also selected for each tree to promote diversity among the trees. This selection is often done by randomly selecting a subset of features.

Each decision tree is constructed similarly to the DT Classifier, employing criteria like Gini impurity or information gain to select the best feature for splitting. The tree-building process is based on recursive splitting.

In classification tasks, the predictions of all DTs are combined by taking a majority vote. The class with the most votes among the trees is assigned as the predicted class for the instance.

The RF Classifier leverages the power of ensemble learning to improve accuracy and reduce overfitting. It is a versatile algorithm that can classify data into the HMK, SCC-9, and HSC-3 categories based on protein features.

6.2.4.4 Adaboost Classifier

The Adaptive Boosting (AdaBoost) Classifier is utilised for classification tasks, with data represented by proteins as features and classification into the HMK, SCC-9, and HSC-3 categories. In our work, predictions are created by combining the outputs of multiple weak classifiers in an ensemble fashion. The AdaBoost algorithm utilises weighted instances, weighted error calculations, and the weighted aggregation of weak classifier outputs. The subsequent mathematical equations from 6.17 to 6.19 encapsulate the fundamental components based on the studies [FS97, JL10].

- **Weighted Error (ε):**

$$\varepsilon(t) = \frac{\sum(w(i) \cdot \text{IsMisclassified}(i, t))}{\sum w(i)} \quad (6.17)$$

Here, “w(i)” represents the weight of instance “i”, and “IsMisclassified(i,

t)” equals 1 if weak classifier “ t ” misclassifies instance “ i ” and 0 otherwise.

- **Classifier Weight (α):**

$$\alpha(t) = 0.5 \cdot \ln \left(\frac{1 - \varepsilon(t)}{\varepsilon(t)} \right) \quad (6.18)$$

- **Instance Weight Update ($w(i)$):**

$$w(i) = w(i) \cdot \exp(\alpha(t) \cdot \text{IsMisclassified}(i, t)) \quad (6.19)$$

- **Classifier Importance:** - The importance of the weak classifier “ t ” is directly related to its weight “ $\alpha(t)$ ”. A higher weight signifies greater importance.
- **Weighted Voting:** - In the final classification, the outputs of weak classifiers are weighted according to their respective “ $\alpha(t)$ ” values, and a weighted majority vote is employed to make the ultimate prediction.

AdaBoost emerges as a potent algorithm for classification tasks, leveraging the strengths of numerous weak classifiers to achieve robust results. It can effectively classify data into the HMK, SCC-9, and HSC-3 categories based on protein features.

6.2.4.5 Gradient Boosting Classifier

The Gradient Boosting Classifier (GBC) is employed for classification tasks, with data represented by proteins as features and classification into the HMK, SCC-9, and HSC-3 categories. In this method, predictions are refined through the sequential building of DTs, with each tree addressing the errors made by the previous ones. The GBC encompasses the utilisation of loss functions and gradients to refine predictions. The ensuing mathematical equations from 6.20 to 6.22 delineate the fundamental elements.

- **Loss Function (L):**

$$L(y, f(x)) = -[y \log(f(x)) + (1 - y) \log(1 - f(x))] \quad (6.20)$$

Here, “ y ” signifies the true label (0 or 1), and “ $f(x)$ ” represents the forecasted probability of the instance belonging to Class 1.

- **Residuals (r):**

$$r = -\frac{\partial L(y, f(x))}{\partial f(x)} \quad (6.21)$$

These residuals represent the errors that require correction by the subsequent DT.

- **Learning Rate (η):** - The learning rate serves as a hyperparameter that regulates the step size for adjusting predictions in each iteration. It scales the contribution of each DT to the ensemble.
- **Update Rule:** - In each iteration, a new DT is trained to predict the residuals. The forecast from this tree is subsequently incorporated into the existing prediction, with a scaled step size determined by the learning rate:

$$f(x)_{\text{new}} = f(x)_{\text{old}} + \eta \cdot \text{tree_prediction} \quad (6.22)$$

The process is reiterated for a predefined number of iterations, with the ultimate prediction being the cumulative outcome of all iterations.

GBC combines multiple DT to create a robust classifier, systematically minimising the loss function and enhancing predictions. It can be deployed to classify data into the HMK, SCC-9, and HSC-3 categories based on protein features, effectively addressing intricate classification endeavours.

6.2.4.6 Gaussian Naïve Bayes Classifier

The Gaussian Naive Bayes Classifier is utilised for classification tasks, with data represented by proteins as features and classification into the HMK, SCC-9, and HSC-3 categories. This method is grounded in Bayes' theorem, and the feature independence assumption is calculated using Equations from 6.23 to 6.25 are based on Machine learning: a probabilistic perspective [Mur12].

- **Prior Probability ($P(C)$):**

$$P(C) = \frac{\text{Number of instances of class C}}{\text{Total number of instances}} \quad (6.23)$$

The prior probability of a class “C” corresponds to the probability of encountering that class without considering any features. It is estimated based on the relative frequency of the class in the dataset.

- **Class-Conditional Probability ($P(X|C)$):**

$$P(X|C) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(X-\mu)^2}{2\sigma^2}} \quad (6.24)$$

- “X” corresponds to the feature value,
- “C” represents the class,
- “ μ ” designates the mean of the feature for Class “C”,
- “ σ ” signifies the standard deviation of the feature for Class “C”.

The class-conditional probability signifies the probability of observing a particular set of feature values “X” given a specific Class “C”. In Gaussian Naive Bayes, this probability is assumed to follow a Gaussian (normal) distribution.

- **Posterior Probability ($P(C|X)$):**

$$P(C|X) = \frac{P(C) \cdot P(X|C)}{P(X)} \quad (6.25)$$

- “ $P(C|X)$ ” stands for the posterior probability of Class “C”,
- “ $P(C)$ ” denotes the prior probability of Class “C”,
- “ $P(X|C)$ ” represents the class-conditional probability of feature values “X” given Class “C”,
- “ $P(X)$ ” signifies the total probability of observing feature values “X” across all classes.

The posterior probability refers to the probability of a Class “C” given the observed feature values “X”. It is calculated through Bayes’ theorem.

- **Prediction:** The class with the highest posterior probability is assigned as the predicted class for an instance.

Gaussian Naive Bayes emerges as a straightforward yet effective classifier; it is particularly suitable when the feature independence assumption is valid. It can be employed for classifying data into the HMK, SCC-9, and HSC-3 categories based on protein features, making it a pertinent choice for diverse classification tasks.

6.2.4.7 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is employed for classification tasks, with data represented by proteins as features and classification into the HMK, SCC-9, and HSC-3 categories. LDA aims to discover a linear combination of features that optimally separates the classes. This algorithm involves linear transformations and eigenvalue calculations to determine the optimal linear discriminants. The core components are described by the following mathematical Equations 6.26 and 6.27. The mathematical formulations of LDA and QDA follow the standard discriminant analysis framework [HTF09].

- **Within-Class Scatter Matrix (S_w):** The within-class scatter matrix measures the dispersion of samples around their respective class means, as defined in Equation 6.26.

$$S_w = \sum_{i=1}^C \sum_{j=1}^{N_i} (\mathbf{x}_{ij} - \mu_i)(\mathbf{x}_{ij} - \mu_i)^T \quad (6.26)$$

where:

- C is the total number of classes,
 - N_i is the number of instances in class i ,
 - $\mathbf{x}_{ij}$ denotes the feature vector of the j^{th} instance in class i ,
 - μ_i is the mean vector of class i .
- **Between-Class Scatter Matrix (S_b):** The between-class scatter matrix quantifies the dispersion of class means around the overall mean, as shown in Equation 6.27.

$$S_b = \sum_{i=1}^C N_i (\mu_i - \mu)(\mu_i - \mu)^T \quad (6.27)$$

where:

- N_i is the number of instances in class i ,
- μ_i is the mean vector of class i ,
- μ represents the overall mean vector across all classes.

- **Eigenvalue and Eigenvector Calculation:** LDA identifies the linear combinations of features that maximise the ratio of between-class scatter to within-class scatter. This identification is accomplished by computing the eigenvalues and eigenvectors of the matrix $(S_w)^{-1} \cdot S_b$. The eigenvectors represent the directions in the feature space that maximise class separation.
- **Projection to Reduced Space:** The eigenvectors corresponding to the largest eigenvalues are chosen as the directions in the reduced space. Data is projected onto this lower-dimensional space to create the linear discriminants.
- **Prediction:** To classify a new instance, the data is projected onto the reduced space and a decision is made based on its proximity to the centroid of a class in the reduced space.

LDA proves to be a robust method for dimensionality reduction and classification tasks, particularly when classes are well separated in the feature space. It can effectively be used to classify data into the HMK, SCC-9, and HSC-3 categories based on protein features, enabling the discovery of linear combinations of features that maximise class separation.

6.2.4.8 Quadratic Discriminant Analysis Classifier

Quadratic Discriminant Analysis (QDA) is employed for classification tasks, with data represented by proteins as features and classification into the HMK, SCC-9, and HSC-3 categories. Unlike LDA, Quadratic Discriminant Analysis (QDA) does not assume that all classes have the same covariance matrix, allowing for more flexibility in modelling. The Discriminant Functions and Covariance Matrix are computed using Equation 6.28 and Equation 6.29, respectively.

- **Discriminant Function ($D_k(\mathbf{x})$):**

The discriminant function for class k in Quadratic Discriminant Analysis (QDA) is defined as:

$$D_k(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k) - \frac{1}{2} \ln |\boldsymbol{\Sigma}_k| + \ln P(C_k) \quad (6.28)$$

where:

- $D_k(\mathbf{x})$ denotes the discriminant score for class k ,

- $\mathbf{x}$ is the feature vector of the sample being classified,
- $\boldsymbol{\mu}_k$ is the mean vector of class k ,
- $\boldsymbol{\Sigma}_k$ is the covariance matrix of class k ,
- $P(C_k)$ is the prior probability of class k .

- **Class-Specific Covariance Matrix ($\boldsymbol{\Sigma}_k$):**

The covariance matrix for class k is computed as:

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} (\mathbf{x}_{ki} - \boldsymbol{\mu}_k)(\mathbf{x}_{ki} - \boldsymbol{\mu}_k)^T \quad (6.29)$$

where:

- $\mathbf{x}_{ki}$ represents the i^{th} feature vector in class k ,
- N_k is the total number of samples in class k ,
- $\boldsymbol{\mu}_k$ is the mean vector of class k .

- **Prediction Rule:**

For a new observation $\mathbf{x}$, QDA computes $D_k(\mathbf{x})$ for all classes k and assigns the instance to the class with the maximum discriminant score:

$$\hat{C} = \arg \max_k D_k(\mathbf{x})$$

6.2.5 Protein-Protein Interaction

We comprehensively analysed co-elution data to predict PPIs within the context of oral cancer cell lines. Our analysis commenced with data preprocessing. To enhance the quality of the protein profiles, we addressed missing values by imputing them with the mean of adjacent data points. Profiles with at most five consecutive data points were excluded from further consideration. Subsequently, we applied a moving average filter to smooth the profiles.

To model the protein profiles effectively, we employed Gaussian mixtures, ranging from one to five components, to fit each profile. Model selection was carried out to retain the best-fitting Gaussian model. Proteins that could not adequately fit with a Gaussian model were eliminated, as indicated by a minimum

R-squared value.

Feature Selection and Protein-Protein Interaction Analysis: We focused on various feature sets by selecting 50, 100, and 200 proteins using feature selection algorithms. Ensuring the selected features were equally important, we applied feature selection techniques. Subsequently, we harnessed the PRINCE R package for PPI analysis. The PRINCE package encompassed several key functions:

- **build_gaussians:** This function encompasses a series of essential steps, including profile preprocessing through data filtering and missing value imputation, Gaussian model fitting to protein profiles, and the selection of the most appropriate model based on criteria such as Corrected Akaike Information Criterion (AICc) and Bayesian Information Criterion (BIC). Parameters, including the minimum number of consecutive points, imputation settings, and curve-fitting criteria, were considered.
- **calculate_features:** Following the fitting of Gaussian models to protein profiles, we computed six distinct features for each pair of co-elution profiles. These features encompassed Pearson correlations between raw and cleaned profiles, p-values, Euclidean distances, co-peak scores, and co-apex scores. These features served as crucial inputs for the classifier.
- **predict_interactions:** To rank protein pairs, we utilised a machine-learning classifier. Cross-validation was implemented, and the reference data were divided into ten folds to score interactions. The PRINCE package offered a range of classifiers, including naive Bayes, SVM, RF, and LR, along with the option to aggregate results from multiple classifiers. Parameters for classifier type, model ensemble, cross-validation folds, and the number of trees were considered.
- **threshold_precision:** High-confidence interactions were identified based on a 0.7 precision threshold, enabling the creation of a PPIN.

In addition to the classifier-based interaction prediction, we employed a permutation-based method inspired by approaches used in other interactome profiling techniques to assess the statistical significance of co-elution patterns among known protein complexes. This comprehensive methodology allowed us to achieve our primary goal of identifying PPI and classifying oral cancer cell lines based on co-elution data, specifically focusing on discovering novel interactions and an in-depth examination of complex co-elution patterns within

these cell lines.

6.3 Results and Discussion

This section describes the results and analysis of our methodology.

6.3.1 Dataset analysis

In these experiments, four datasets were used, whose names and dimensions of those datasets are as follows:

- **Peptides_without_match_between_runs_Unique**
 - Dimension: 20,597, 38
- **Peptides_without_match_between_runs_Unique_Razors**
 - Dimension: 24,469, 38
- **Protein_groups_without_match_between_runs_Unique**
 - Dimension: 3,046, 37
- **Protein_groups_without_match_between_runs_Unique_Razors**
 - Dimension: 3,344, 37

In the dimensions descriptor above, the first element represents the number of proteins, or gene names, and the second represents the profiles. The peptide files have 38 columns, of which 12 columns are for each profile type, 1 column for protein names, and one is for the peptide sequences. Protein files do not have the peptide column, so there are 37 columns in that file.

First of all, we separated the data by type and generated a separate file for each category with protein names. These separated files are input to PPI and co-eluted complex identification. Using these files, we analysed our data with heatmaps to find the association between the attributes. The association finding was important for the classification; the detailed results are shown in Figures 6.2 to 6.5.

These heatmaps seem to depict the Label-Free Quantitation (LFQ) intensity of proteins across different cell lines, with HMK likely representing control cell lines, HSC-3 aggressive cancer cell lines, and SCC-9 less aggressive cancer cell lines. The pattern of colour intensity across the grid suggests the expression

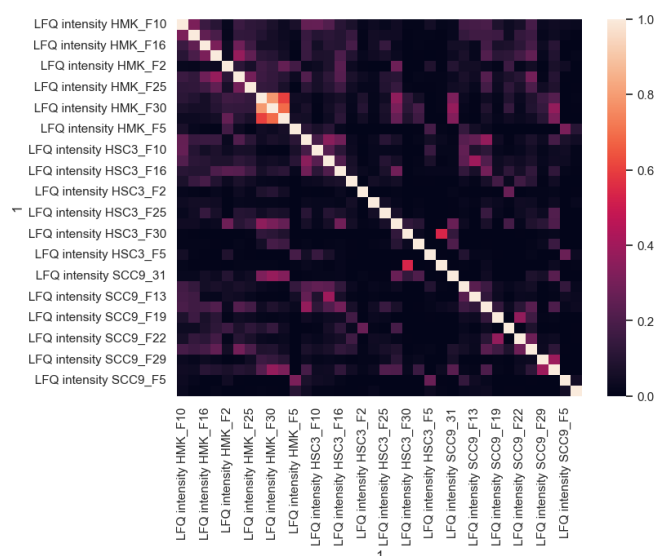


Figure 6.2: Correlation between Peptides

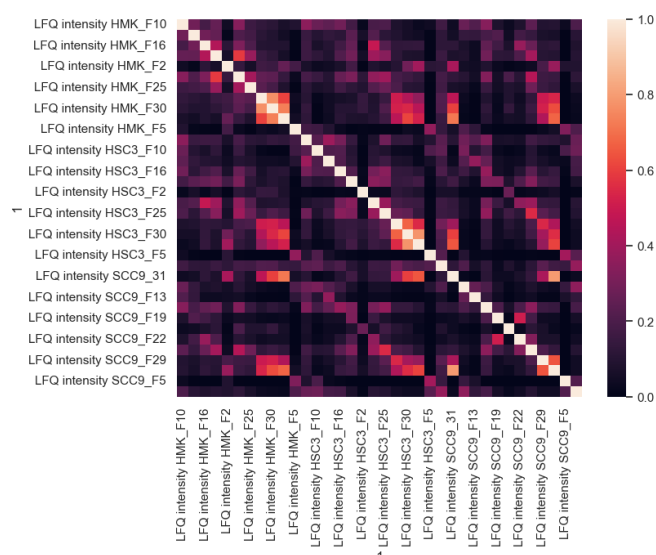


Figure 6.3: Correlation between Peptides Razors

levels of different proteins (rows) in different fractions (columns) during the experiment. The darker regions indicate higher expression or abundance, while lighter regions indicate lower levels. Such a heatmap is critical for identifying differentially expressed proteins across conditions, which could serve as potential biomarkers or therapeutic targets. It also provides insights into the similarities and differences in protein profiles between cancerous and non-cancerous cell lines. After analysing the dataset, we found that there needs to be more correlation between the data, which is good for classification.

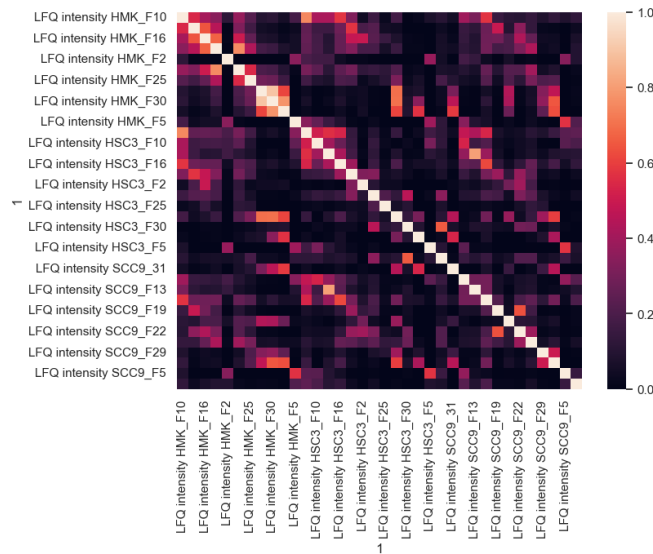


Figure 6.4: Correlation between Proteins

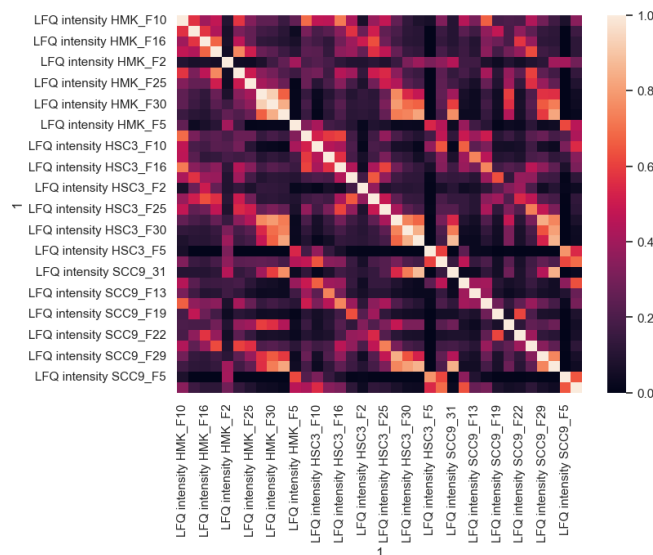


Figure 6.5: Correlation between Proteins

6.3.2 Machine Learning for Biomarker ranking and importance

As the dimensions of the dataset were large, it was computationally expensive to find the protein interactions, as there were a lot of irrelevant protein complexes that did not play a vital role in the cancer progression. The cancer biomarkers were ranked and shortlisted without losing their significance in finding the relevant protein that plays a significant role in different cancer stages. For the ranking, the different relevant algorithms mentioned in the methodology section were used. To know the importance of the selected biomarkers, the

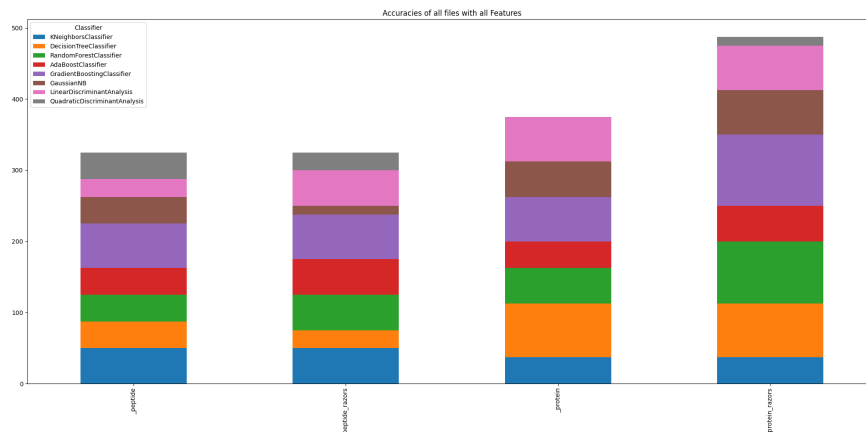


Figure 6.6: Classification Accuracies with all Biomarkers

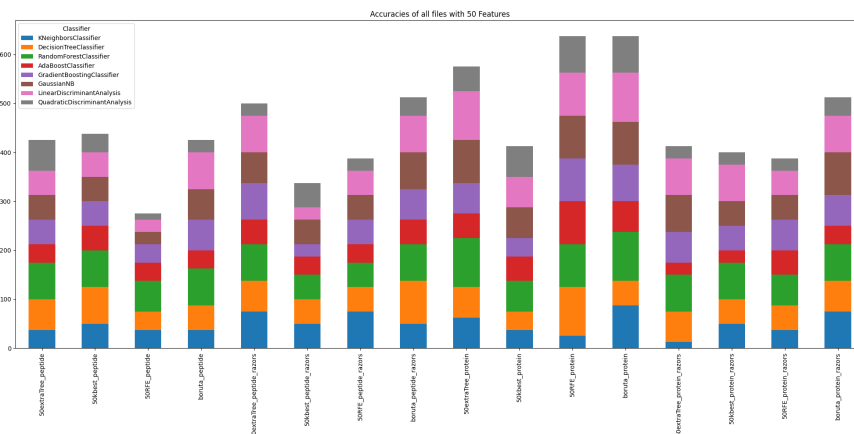


Figure 6.7: Classification Accuracies with top 50 Biomarkers

ranking algorithms were classified using machine learning models. The same machine learning models were applied to the whole dataset and suffered from the “curse of dimensionality” problem with insufficient results, as shown in Figure 6.6. To overcome this problem, the machine learning models were applied to the selected feature sets (50, 100, and 200) proteins that provided good results and proved that these proteins are significant in classifying the cancer types. The results are shown in Figures 6.7 to 6.9. The described protein sets are selected to ensure that only important biomarkers are chosen. These sets were considered to find the PPI, and the set with 50 biomarkers did not have enough proteins to make complexes. So, we had to move to the 100 biomarkers, which were also not enough, so we shifted to 200 biomarkers. We could not go further than that because it led to low results.

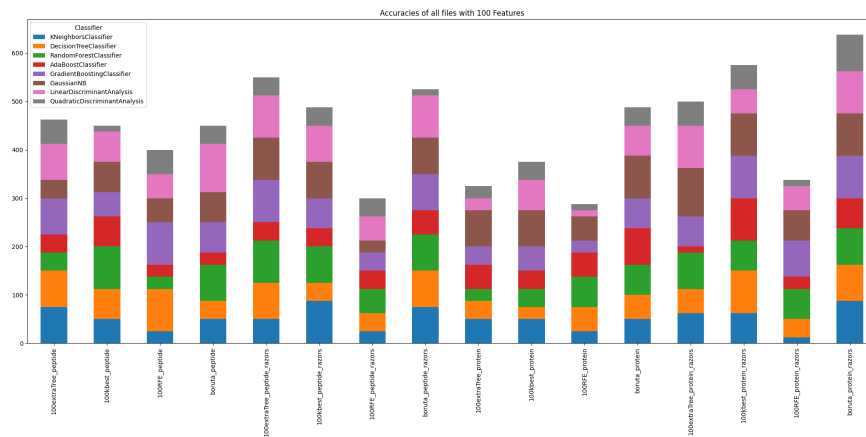


Figure 6.8: Classification Accuracies with top 100 Biomarkers

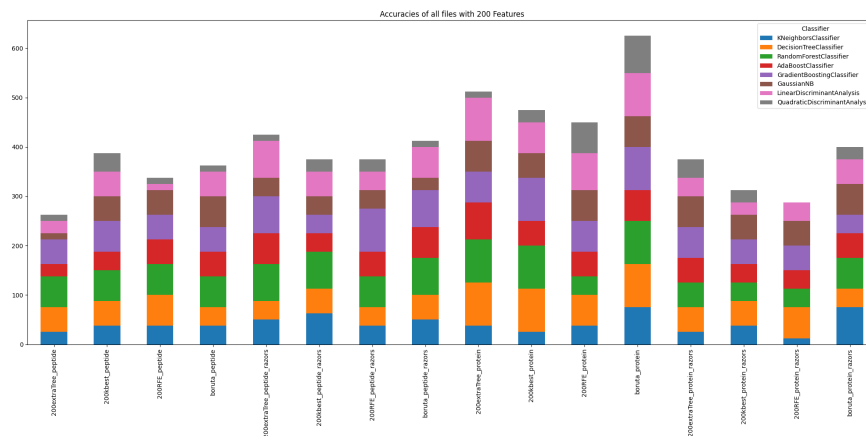


Figure 6.9: Classification Accuracies with top 200 Biomarkers

6.3.3 Analysis of Selected Biomarkers

To analyse the selected biomarkers, we iterated them using the algorithms and created the three sets according to the different set sizes we gathered in Section 6.3.2. Then, we displayed these common biomarkers using a word cloud and analysed them using a Venn diagram outlining the intersection of these biomarkers. The word cloud of common biomarkers selected from the three sets and the Venn diagram are shown in Figures 6.10 and 6.11, respectively.

The total number of selected proteins by all the algorithms was 3761, out of which 1211 were unique in all the sets. The distribution of unique proteins in all the sets, along with the intersection, is shown in Figure 6.11. After the analysis, 251 proteins were identified, of which protein IDs “Q09666”, “P07355”, and “Q9Y490” were at the top. The protein abundance distribution for each of the normal, less aggressive, and aggressive cell lines were analysed to monitor the value level of the particular category, along with the candidate proteins. The

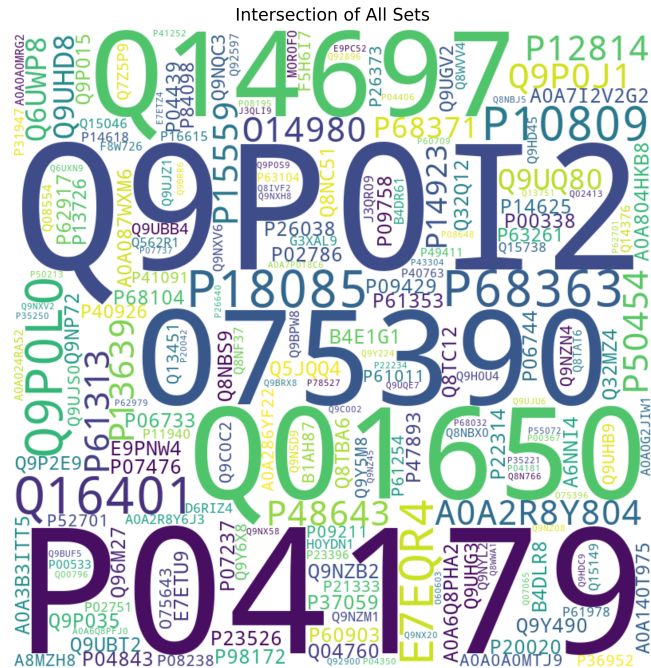


Figure 6.10: Common Proteins Within the Selected Sets

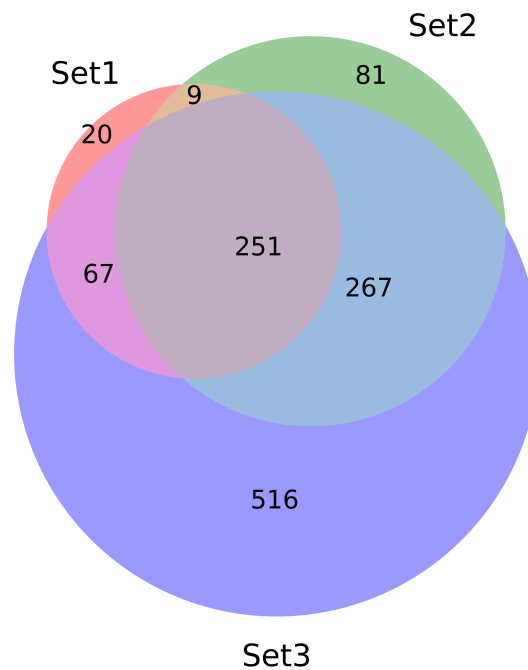


Figure 6.11: Distribution of Proteins Between Selected Sets

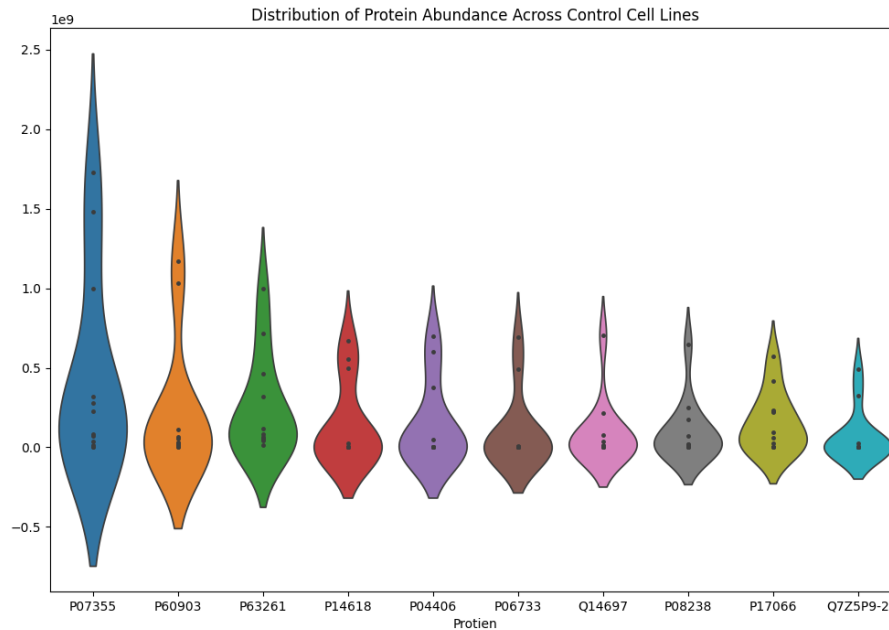


Figure 6.12: Distribution of Protein Abundance Across Control Cell Lines

top 10 proteins with high variance were selected and shown in Figure 6.12 to Figure 6.14. The proteins shown in these Figures are reported as the cancer hallmarks, including the strong candidate for oral cancer [SD21]. The protein numbers “P63261”, “P04406”, “P06733”, “P7355”, “P68032”, and “P14618” are common in the aggressive and less aggressive cell lines but have different distributions. Meanwhile, “P07355”, “P04406”, “P06733”, and “P14618” are common in the normal and aggressive cell lines; they are also all common in normal and less aggressive, with the exception of “P07355”. However, the abundance distribution level is different and can be observed in the abundance distribution Figures.

6.3.4 Protein-Protein Interaction and Network Analysis

After analysing data for all the datasets, we consider the dataset protein groups without a match between runs, “Unique Razors” to perform the PPI and network analysis. To find the interaction level between the proteins, PPI (see section 6.2.5) was used for all selected sets with a 70% precision threshold. The unweighted PPINs with a gold-standard dataset are shown in Figure 6.15 to Figure 6.17.

The precision curves show the number of complexes made by the PPI algo-

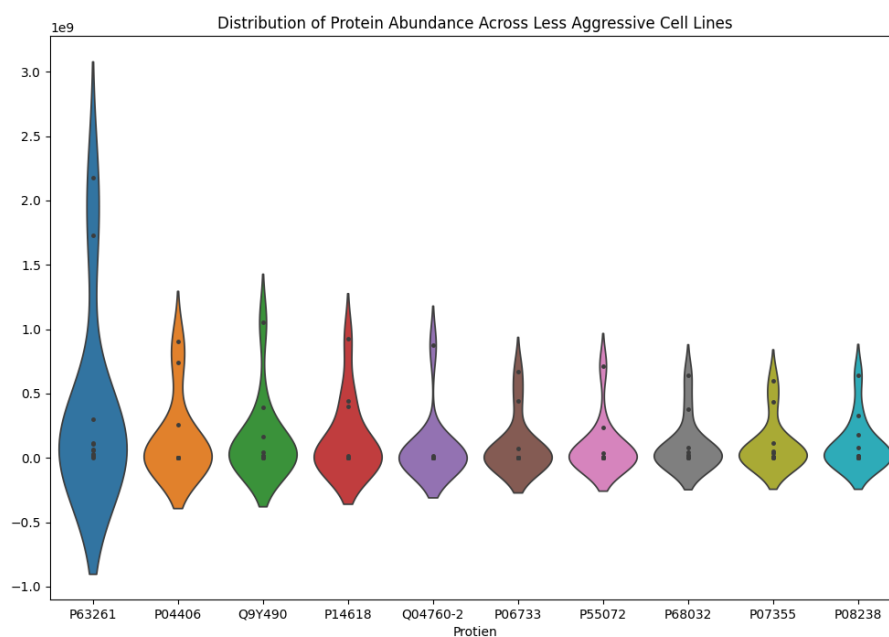


Figure 6.13: Distribution of Protein Abundance Across Less Aggressive Cell Lines

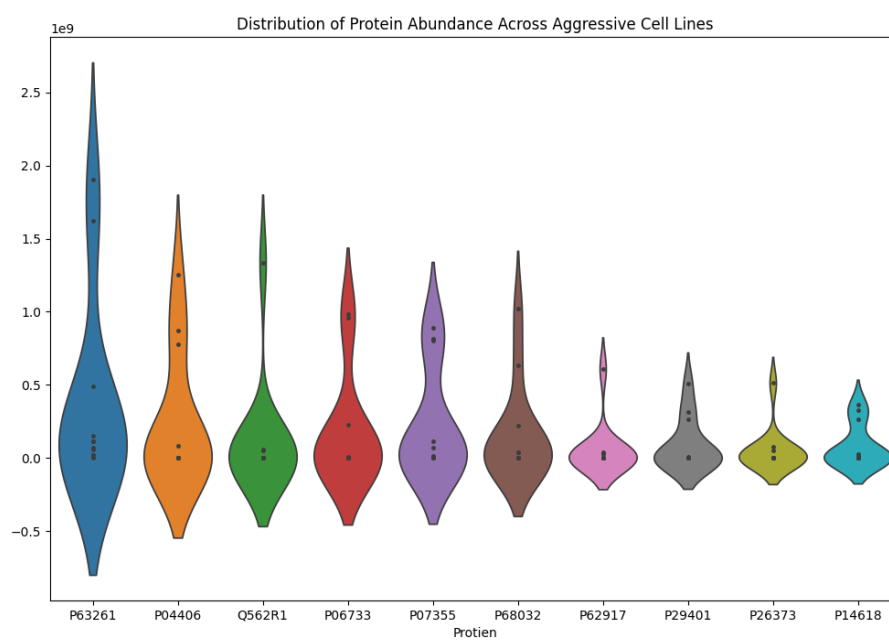


Figure 6.14: Distribution of Protein Abundance Across Aggressive Cell Lines

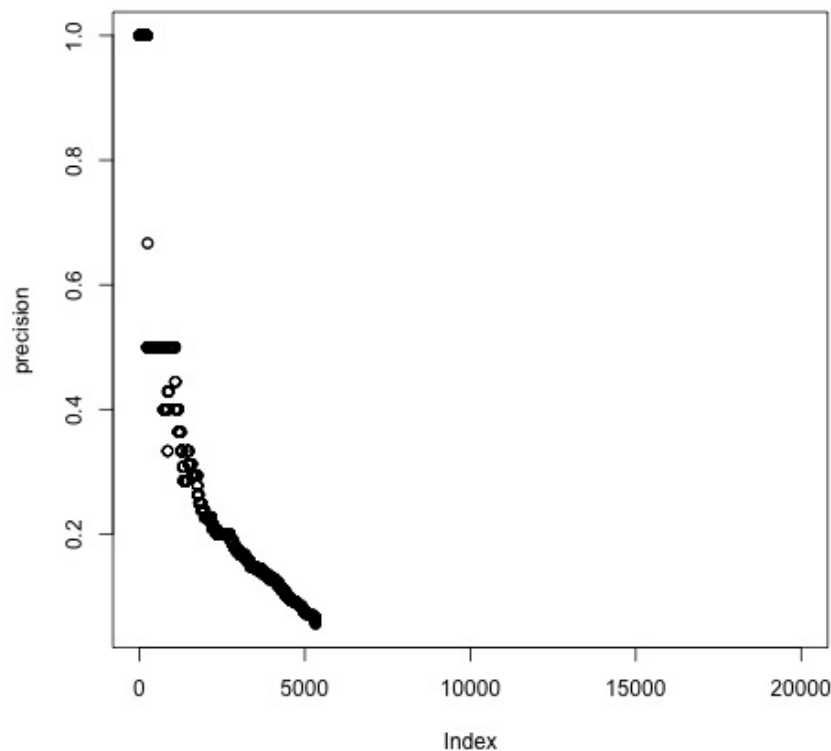


Figure 6.15: Precision Curve for Control Complexes

rithm, and we filtered out all the complexes that have less than 70% precision to ensure stability. The PPI was captured by the descriptors using the Kbest algorithm. With the 70% precision and 200 descriptors, we could identify 104 proteins and 236 interactions for control, 90 proteins and 219 interactions for aggressive, and 79 proteins and 317 interactions for less aggressive.

6.3.5 Validation of Protein Complexes

The protein complexes were compared with the Comprehensive Resource of Mammalian protein complexes (CORUM) Human complexes [STB⁺25], which are standard protein complexes for humans. The number of protein complexes extracted by our methodology for normal was 36, for less aggressive was 46, while for the aggressive was 35. We filtered the complexes for all three categories that have at least 2 proteins common in CORUM. For the control, we found CCT complex (chaperonin containing TCP1 complex) and the key protein “P48643” have over two proteins (“P78371”, “P40227”) in common, CCT complex (chaperonin containing TCP1 complex) and the key protein “P78371”

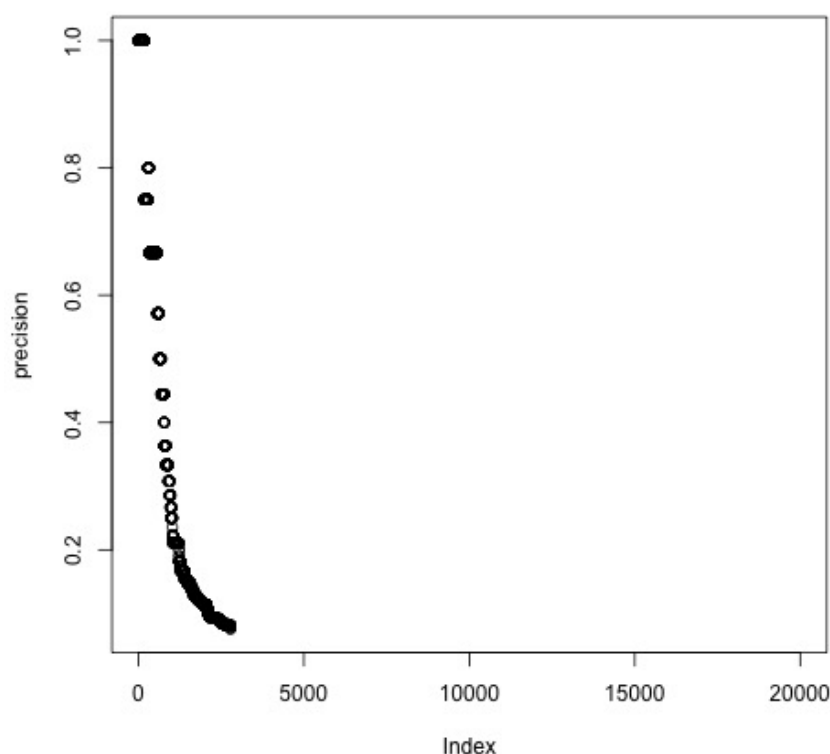


Figure 6.16: Precision Curve for Less Aggressive Complexes

have one protein (“P40227”) in common. For the aggressive, we found only one CCT complex (chaperonin containing TCP1 complex), and the key protein “P48643” has over two proteins (“P78371”, “P40227”) in common. However, for the less aggressive, we found 5 common complexes; 40S ribosomal subunit, cytoplasmic and the key protein “P23396” have over two proteins (“P61247”, “P60866”) in common, CCT complex (chaperonin containing TCP1 complex) and the key protein “P40227” have over two proteins (“P48643”, “P49368”, “P50990”) in common, CCT complex (chaperonin containing TCP1 complex) and the key protein “P50990” have over two proteins (“P48643”, “P49368”) in common, 40S ribosomal subunit, cytoplasmic and the key protein “P60866” have one protein (“P62249”) in common, 40S ribosomal subunit, cytoplasmic and the key protein “P62249” have one protein (“P61247”) in common. For the shortlisted protein IDs, we extracted the gene ID to find out the relevant studies that have reported these genes contributing to cancers. The genes are shown in the Table 6.1.

Based on the gene’s name against the proteins, we found the interaction that

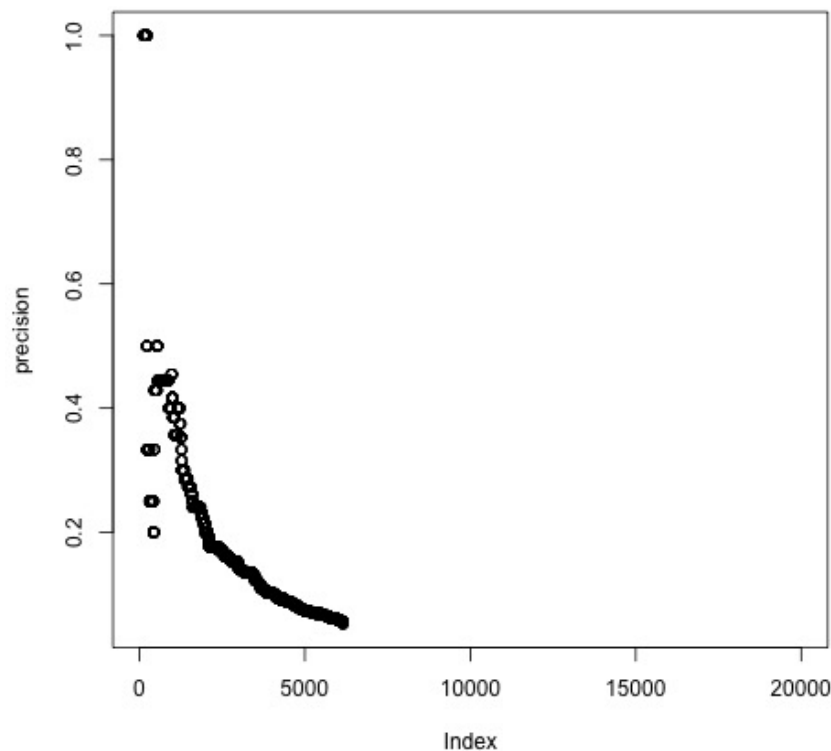


Figure 6.17: Precision Curve for Aggressive Complexes

plays an important role in cancer growth, as reported by different studies. For example, the protein EGFR, for which the protein code is “P00533”, is a component of cell signalling pathways that regulate cell survival and division. Sometimes, certain types of cancer cells produce more EGFR proteins than usual due to mutations. Furthermore, Junction Plakoglobin (JUP), for which the protein code is “P14923”, plays an important role in different kinds of cancers [YHB⁺23]. The interaction of both Genes is reported as a cancer cell map initiative [ZKR⁺21], and the same interaction was found in the shortlisted proteins by our methodology for the less aggressive cell lines shown in Figure 6.18. Similarly, CCT8, for which the protein code is “P50990”, is responsible for many cancer developments and progression [LRY⁺21]. The CCTA6 also play the same role [ZZQ⁺24], and the interaction of CCTA6 with CCT8 is reported to cause cancer [ZKR⁺21]. Our methodology is also capable of finding the interaction between these genes, as shown in Figure 6.19.

For the aggressive cancer cell lines, our methodology shortlisted 35 proteins, out of which 12 were reported to play an important role in thyroid cancer if they

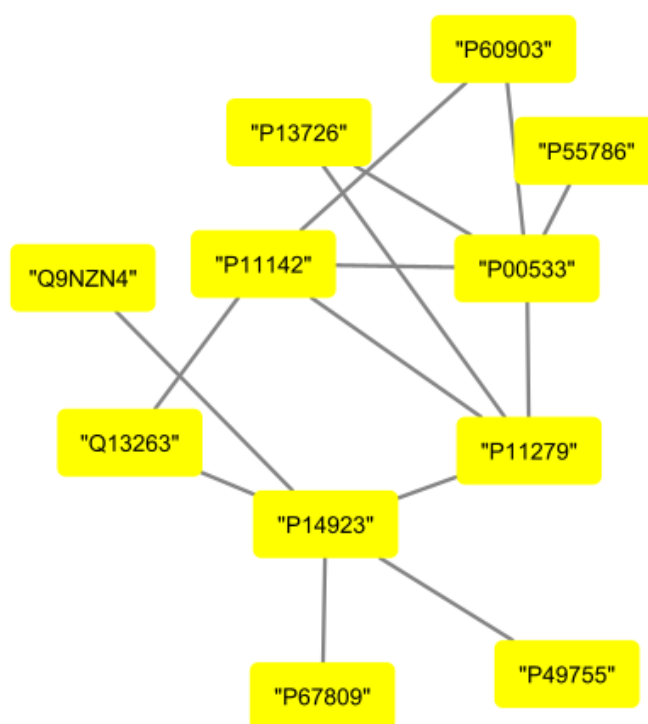


Figure 6.18: Interaction network of EGFR and JUP in Less Aggressive Categories

interact with each other [CZZ19]. The interactions of proteins for the aggressive cell lines are shown in Figure 6.20. These interactions are also reported in oral cancer [KMM⁺15]. Furthermore, Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) plays a regularity mechanism in cancer cells [ZZH⁺15], PKM is also overexpressed in cancers [DMX⁺16] and their interaction is reported in cancer metabolic pathways [CRP18, SSD⁺17, MPEMR12, VMSC14, AR17], prostate cancer [TAP17], and kidney cancer [WHM⁺15]. These identified complexes confirm that our methodology has the capability to identify new complexes and biomarkers to classify the cancer type using the co-fraction data.

In this chapter, we have an integrative bioinformatics approach to uncover significant proteomic variations associated with OSCC progression. By leveraging mass spectrometry, biomarker ranking strategies, protein-protein interaction network analyses, and robust machine learning classifiers, we have identified specific protein complexes and biomarkers that distinguish normal, less aggressive, and highly aggressive oral cancer cell lines with remarkable accu-

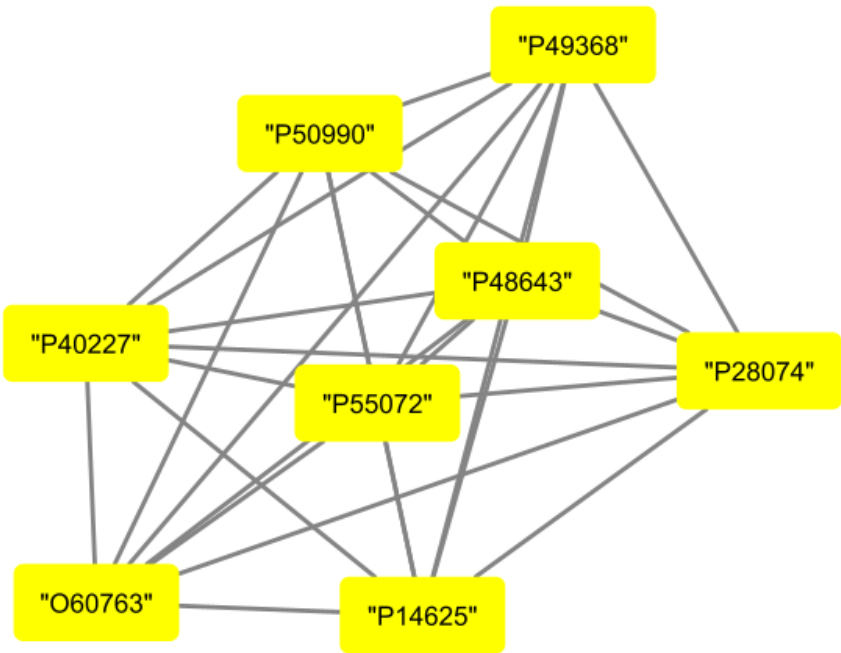


Figure 6.19: Interaction network of CCT8 and CCT6A in Less Aggressive Categories

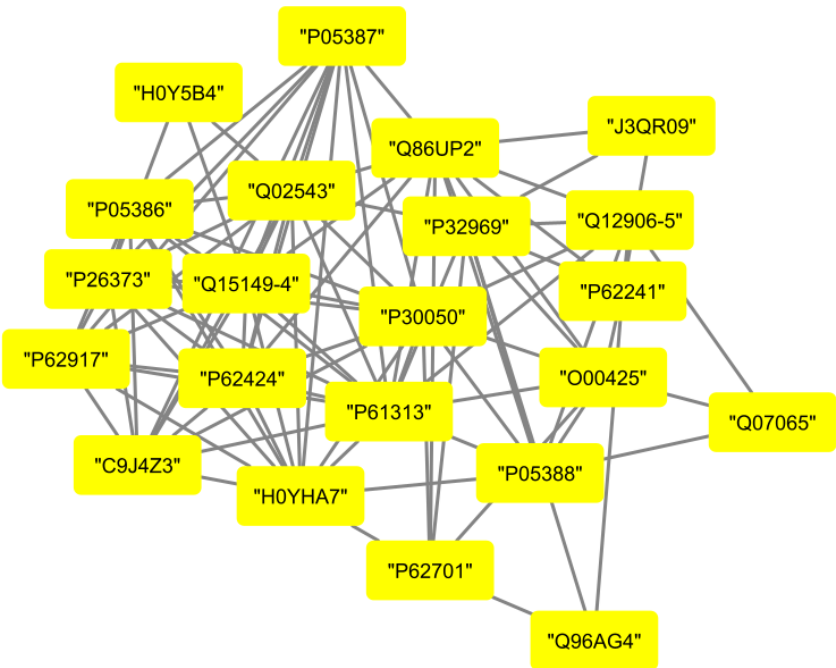


Figure 6.20: Interaction network of RPLs and RPSs in Aggressive Categories

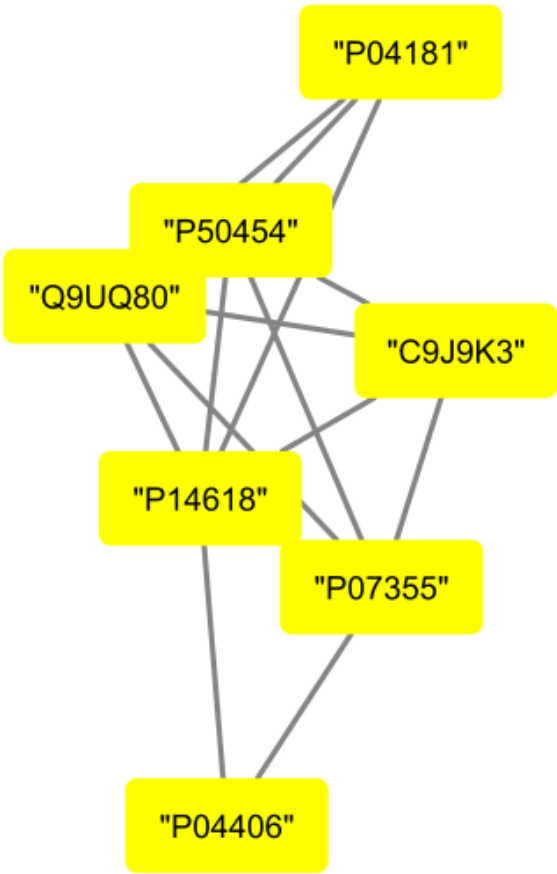


Figure 6.21: Interaction network of GAPDH and PKM in Aggressive Categories

racy. Together with preceding chapters, it demonstrates how somatic, transcriptomic, and proteomic perspectives contribute complementary insights into cancer biology. Having presented these approaches, Chapter 7 synthesises the findings, reflects on their implications, and outlines future directions for integrating multi-omics data in cancer research.

Table 6.1: Gene Classification into Control, Less Aggressive, and Aggressive Categories

Control Genes (HMK)	“Less Aggressive Genes (SCC-9)”	“Aggressive Genes (HSC-3)”
HSPA1B	RPL17	RPL19
HSPA9	NCL	VCP
P4HB	HSP90B1	RPL37A
HSP90B1	VCP	RPL18
VCP	RPL23A	GAPDH
RPL23A	PLOD2	RPN1
PDLIM1	RPS5	RPLP2
EGFR	PLOD3	RPLP0
HSPB1	EGFR	ANXA2
RPN2	HSPB1	PKM
ANXA2	ANXA2	UBA1
TUBB	HSP90AB1	RPS3
FH	HSPD1	RPL13
SLC3A2	HSPA8	RPS19
HSPD1	F3	CCT5
HSPA5	DARS	SERPINH1
ACTN1	JUP	RPS20
JUP	HSPA6	S100A10
UBA1	RPS3	RPL15
AHCY	ATP5A1	RPS8
MSN	STOM	RPL7A
STIP1	PSMB5	RPS4X
CCT5	PDIA3	RPL8
TUFM	CCT6A	CAP1
SERPINH1	TUFM	CAP1
S100A10	SERPINH1	RPL18A
YWHAZ	CCT8	CKAP4
ACTG1	BCAP31	AHNAK
ACTC1	RPS20	ILF3
CCT2	S100A10	GANAB
AHNAK	HNRNPK	GANAB
PRDX4	RPS8	PLEC
GANAB	RPS16	KTN1
PLEC	RPS4X	PA2G4
FARSB	RPL23	TLN1
	RPL8	
	EEF1A1	
	CKAP4	
	AHNAK	
	LAMB3	
	LAMC2	
	KTN1	
	LRRC59, RUVBL2	
	STOML2, TLN1	

Chapter 7

Conclusion and Future Directions

This chapter summarises the work completed in this doctoral thesis, which focuses on the use of independent omics data, genomic, transcriptomic, and proteomic, for cancer classification and biomarker discovery. It also addresses the research question of this thesis, outlines the ethical considerations of data, and provides future directions that can build upon the contributions of this research.

7.1 Conclusion

The primary objective of this doctoral research was to explore the potential of three independent omics modalities: somatic mutation data, RNA microarray GE data, and protein CF-MS data, to improve cancer type and subtype classification, along with identifying potential biomarkers. The research was divided into three main investigations, one for each omics modality, employing deep learning and explainable machine learning models. This thesis has made significant progress in understanding, analysing, and applying machine learning methods for cancer biomarker identification using independent omics data. Through a series of detailed studies and analyses, it offers valuable insights and solutions to cancer classification, reliability of methods, and performance in early prognosis and diagnosis.

7.1.1 Somatic Mutation Analysis

Somatic mutation data are sparse and heterogeneous, which presents a major challenge for classification tasks. This study proposed “CLARITY”, a desparsification strategy using biological pathway scores and clustering to label cancer

stages and to classify cancer. This pipeline combines deep learning, XAI, statistical analysis and data visualisation to both classify cancer stages based on somatic mutation data and to trace cancer hallmarks. In the process, we integrated autoencoders, spectral clustering, DNN, and LIME into a single interpretable workflow. Our experiments demonstrated superior subtype separation compared to traditional clustering algorithms, both in clustering compactness and survival relevance. The clustering mechanism not only works with complex DNNs but also works on lightweight traditional machine learning algorithms. This work demonstrates that properly processed genomic data can support reliable cancer classification and subtype detection. The approaches were tested in breast cancer data, but we have shown evidence that the pipeline is sound for other similar datasets.

7.1.2 Transcriptomic-Based Classification

Transcriptomic data, particularly from RNA-seq and microarrays, provide a rich view of gene activity in cancer. We applied deep learning, XAI, KEGG pathway enrichment analysis and data visualisation to offer a robust, efficient and more reproducible pipeline to both classify different cancers with the same model and to trace relevant genes to the high performance of the models. In the process, we provided deep learning-based solutions to deal with data with highly imbalanced classes, large dimensions and low instance numbers, all persistent issues when dealing with biological data. Our approach uses a CNN architecture tuned for both binary and multiclass classification. The architecture, incorporating numerous Conv1D layers, was designed to capture complicated patterns in high-dimensional gene expression data, successfully discriminating between various cancer kinds with a single model. Regularisation techniques such as dropout and L2 regularisation, paired with powerful optimisers like AdamW, guaranteed that the model remained robust and generalised well, even with a relatively limited dataset. Through KEGG pathway enrichment analysis, we effectively identified genes and pathways that are significantly related to cancer, including well-established genes involved in transcriptional misregulation and cancer metabolism. Additionally, our research identified lesser-known genes as relevant, suggesting their potential as novel cancer biomarkers. Results showed that deep learning can outperform traditional models and generalise well across datasets. Key biomarkers were identified and validated through pathway and enrichment analysis, contributing to potential clinical applications. We have tested our approach for RNA-seq and microarray datasets, but have shown evi-

dence that the models and procedures proposed have the potential to apply to other datasets of the same nature.

7.1.3 Proteomic Complex-Based Biomarker Discovery

The proteomic study used CF-MS data to analyse protein abundance and complex formation across oral cancer cell lines. By leveraging MS, biomarker ranking strategies, PPIN analyses, and robust machine learning classifiers, we have identified specific protein complexes and biomarkers that distinguish HMK, SCC-9, and HSC-3 cell lines with remarkable accuracy. This work showed that proteomic signatures, particularly from protein complexes, can distinguish cancer phenotypes and suggest therapeutic targets.

7.1.4 Addressing the Research Questions

This section revisits the main research question and its subcomponents, summarising how they have been addressed through the findings presented in this thesis.

RQ1: *How can computational methods be improved to effectively classify cancers with state-of-the-art performance and discover biomarkers from independent omics datasets?*

This thesis demonstrates that independent, modality-specific pipelines integrating machine learning, data visualisation, and statistical methods can achieve high-performance and interpretable cancer classification. Each omics modality, genomics, transcriptomics, and proteomics, was analysed using a tailored computational approach that addressed its specific data characteristics, such as sparsity, dimensionality, and heterogeneity. The integration of visual and statistical validation throughout the analysis ensured that the identified models and biomarkers were both computationally robust and biologically meaningful. Collectively, these pipelines represent a structured framework for improving classification accuracy and enhancing interpretability across diverse omics datasets, which, taken together, offer the backbone of future omics integration.

sub-RQ1: *How can sparse, high-dimensional somatic mutation data be transformed into biologically meaningful features for cancer subtype discovery and validated using statistical methods?*

This question was addressed through the development of the **CLARITY** pipeline,

which desparsified somatic mutation data by converting raw mutation profiles into biologically interpretable pathway-level scores. Dimensionality reduction using autoencoders and clustering via spectral methods enabled the identification of clinically relevant cancer subtypes. Statistical validation using survival analysis (Kaplan–Meier and log-rank tests) confirmed that the identified clusters were significantly associated with patient outcomes, demonstrating that the transformed features retained biological relevance and prognostic value.

sub-RQ2: *How can transcriptomic data from RNA-seq and microarray platforms be modelled using deep learning architectures to achieve robust and interpretable subtype classification, and validated using bioinformatics methods?*

For this question, deep learning architectures including convolutional neural networks (CNNs) and transformer-based models were implemented to classify cancer subtypes across RNA-seq and microarray datasets. Class imbalance was managed using the SMOTE, while model interpretability was achieved using LIME. The resulting gene importance scores were subjected to functional enrichment and pathway analysis to ensure biological consistency. These models not only achieved high classification accuracy but also provided insights into gene-level biomarkers and functional pathways involved in cancer progression.

sub-RQ3: *How can proteomic co-fractionation data be analysed better to identify protein-level biomarkers and disrupted complexes linked to cancer aggressiveness?*

This objective was fulfilled by constructing a feature selection and classification pipeline for MS-based co-fractionation proteomic data. Machine learning algorithms were employed to rank proteins based on their discriminative power. Data visualisation through violin plots, hierarchical clustering, and network mapping enabled clear interpretation of protein abundance patterns across normal, less aggressive, and highly aggressive oral cancer cell lines. PPIN analysis and literature validation revealed disrupted complexes associated with aggressiveness, highlighting potential biomarker candidates for oral cancer progression.

Summary. Together, these investigations show that integrating data-driven modelling with visual and statistical validation can improve both the accuracy and interpretability of cancer classification across omics modalities. The pipelines developed in this thesis provide flexible and biologically validated frameworks that can be extended to future integrative multi-omics studies aimed at advancing precision oncology.

7.1.5 Cross-Platform Impact

Each omics modality contributed uniquely to understanding cancer biology. Genomics helped classify stages and identify driver mutations, transcriptomics enabled accurate subtype prediction, and proteomics provided insight into functional protein interactions. These independent omics studies form a foundation for future integrated models and offer practical insights for clinical applications.

7.2 Ethical Considerations and Data Access

All datasets used in this research were handled in accordance with ethical standards and institutional data governance requirements. The genomic and transcriptomic datasets used in this work were obtained from publicly available repositories (details are provided in Chapters 4 and 5). These datasets contain no personally identifiable information, ensuring full compliance with data privacy and ethical research practices.

The proteomic dataset used in this thesis was provided by our collaborators at LNbio (Brazil) and was generated from controlled laboratory experiments on oral keratinocyte cell lines. The collection, processing, and sharing of these data adhered to the institutional and national ethical guidelines of the partner organisation. Further details on the data collection protocol can be found in our partner's publication [TAGA⁺25].

In general, the ethical use of cancer-related datasets in AI research extends beyond data handling to issues of algorithmic security, transparency, and clinical responsibility. The integration of AI models into oncology requires careful consideration of model bias, interpretability, and reproducibility to avoid potential misclassification and ensure patient safety. In this context, XAI methods adopted in this thesis not only enhance scientific understanding but also contribute to the responsible deployment of computational models in clinical environments. As AI-driven frameworks continue to evolve, maintaining data security, fairness, and interpretability will remain crucial for bridging the gap between computational research and real-world medical practice. All algorithms, code and data generated in this thesis is made publicly available and referenced as such in our open access publications

7.3 Recommendations for Future Work

Future studies could explore more robust graph-based methods or attention models to capture higher-order mutation patterns. Integration with clinical variables may also enhance patient stratification and survival predictions. Furthermore, we aim to improve our work by integrating other omics (e.g. gene expression, methylation) and cross-cancer subtype generalisation. Expanding to pan-cancer cohorts could improve generalizability.

Though this study focused on microarray data, future work can apply similar architectures to RNA-seq data or single-cell expression profiles. Including domain adaptation techniques, which may help with cross-platform normalisation and enhance model robustness.

Proteomic studies can be extended to include post-translational modification data. Future work may also involve integrating predicted complex disruption with drug response profiles to identify actionable interventions.

While this thesis focused on independent omics analysis, future research should aim to integrate these layers using multi-modal deep learning. Models like cross-omics transformers or multi-branch networks can capture cross-domain patterns for improved classification and biomarker interpretation. We will use Graph Neural Networks to perform prediction and interpret the multi-omics data. Furthermore, our focus will also be on developing foundation models that can help predict, generate reasoning behind their predictions, identify biomarkers, validate them using literature and bioinformatics methods, and provide recommendations for precision oncology.

7.4 Final Words

This doctoral work provides a comprehensive investigation into the potential of independent omics data for cancer classification and biomarker discovery. Through rigorous methodological design, robust machine learning techniques, and biological validation, this research contributes significantly to the fields of bioinformatics and computational oncology. It lays a strong foundation for future integrative and translational cancer research.

References

- [A⁺18] Charu C Aggarwal et al. *Neural networks and deep learning*, volume 10. Springer, 2018.
- [AAA⁺22a] Hanan Aljuaid, Nazik Alturki, Najah Alsubaie, Lucia Cavallaro, and Antonio Liotta. Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning. *Computer Methods and Programs in Biomedicine*, 223:106951, 2022.
- [AAA⁺22b] Abdulrhman M Alshareef, Raed Alsini, Mohammed Alsieni, Fadwa Alrowais, Radwa Marzouk, Ibrahim Abunadi, and Nadehem Nemri. Optimal deep learning enabled prostate cancer detection using microarray gene expression. *Journal of Healthcare Engineering*, 2022(1):7364704, 2022.
- [ABAFS18] Mohamed Abdel-Basset, Laila Abdel-Fatah, and Arun Kumar Sangaiah. Metaheuristic algorithms: A comprehensive review. *Computational intelligence for multimedia big data on the cloud with engineering applications*, pages 185–231, 2018.
- [ABB⁺00] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.
- [ACQ⁺20] Elaheh Alizadeh, Jordan Castle, Analia Quirk, Cameron DL Taylor, Wenlong Xu, and Ashok Prasad. Cellular morphological features are predictive markers of cancer cell state. *Computers in biology and medicine*, 126:104044, 2020.

- [All12] Kimberly H Allison. Molecular pathology of breast cancer: what a pathologist needs to know. *American journal of clinical pathology*, 138(6):770–780, 2012.
- [Alt05] Naomi Altman. Replication, variation and normalisation in microarray experiments. *Applied bioinformatics*, 4:33–44, 2005.
- [AM03] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, 2003.
- [APPS22] Yash Amethiya, Prince Pipariya, Shlok Patel, and Manan Shah. Comparative analysis of breast cancer detection using machine learning and biosensors. *Intelligent Medicine*, 2(2):69–81, 2022.
- [AR17] Gabriela Andrejeva and Jeffrey C Rathmell. Similarities and distinctions of cancer and immune metabolism in inflammation and tumors. *Cell metabolism*, 26(1):49–70, 2017.
- [ASF⁺22] Ganjar Alfian, Muhammad Syafrudin, Imam Fahrurrozi, Norma Latif Fitriyani, Fransiskus Tatas Dwi Atmaji, Tri Widodo, Nurul Bahiyah, Filip Benes, and Jongtae Rhee. Predicting breast cancer from risk factors using svm and extra-trees-based feature selection method. *Computers*, 11(9):136, 2022.
- [ASMB24] Jordan Anaya, John-William Sidhom, Faisal Mahmood, and Alexander S Baras. Multiple-instance learning of somatic mutations for the classification of tumour type and the prediction of microsatellite status. *Nature biomedical engineering*, 8(1):57–67, 2024.
- [AVZ⁺25] Fadi Alharbi, Aleksandar Vakanski, Boyu Zhang, Murtada K Elbashir, and Mohanad Mohammed. Comparative analysis of multi-omics integration using graph neural networks for cancer classification. *IEEE Access*, 2025.
- [Azi22] Rabia Musheer Aziz. Nature-inspired metaheuristics model for gene selection and classification of biomedical microarray data. *Medical & Biological Engineering & Computing*, 60(6):1627–1646, 2022.
- [BAB05] Ralf Bender, Thomas Augustin, and Maria Blettner. Generating survival times to simulate cox proportional hazards models. *Statistics in medicine*, 24(11):1713–1723, 2005.

- [BBBH⁺23] David F Burke, Patrick Bryant, Inigo Barrio-Hernandez, Danish Memon, Gabriele Pozzati, Aditi Shenoy, Wensi Zhu, Alistair S Dunham, Pascal Albanese, Andrew Keller, et al. Towards a structurally resolved human protein interaction network. *Nature Structural & Molecular Biology*, 30(2):216–225, 2023.
- [BC24] Candace Benjamin and Rebecca Crews. Nicotinamide mononucleotide supplementation: Understanding metabolic variability and clinical implications. *Metabolites*, 14(6):341, 2024.
- [BFOS17] Leo Breiman, Jerome Friedman, Richard A Olshen, and Charles J Stone. *Classification and regression trees*. Chapman and Hall/CRC, 2017.
- [BGE01] Dietmar H Blohm and Anthony Guiseppi-Elie. New developments in microarray technology. *Current Opinion in Biotechnology*, 12(1):41–47, 2001.
- [BH95] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.
- [BL13] Rok Blagus and Lara Lusa. Smote for high-dimensional class-imbalanced data. *BMC bioinformatics*, 14(1):106, 2013.
- [Bla03] Eric M Blalock. *A beginner's guide to microarrays*. Springer Science & Business Media, 2003.
- [BLU14] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.
- [BMD21] Christoph J Burckhardt, John D Minna, and Gaudenz Danuser. Co-immunoprecipitation and semi-quantitative immunoblotting for the analysis of protein-protein interactions. *STAR protocols*, 2(3):100644, 2021.
- [BN06] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.
- [BNR⁺25] Semeh Bejaoui, Sofie Holtsmark Nielsen, Astrid Rasmussen, John Eugenio Coia, Dorte Terp Andersen, Tobias Bruun Ped-

- ersen, Martin Vad Møller, Marc Trunjer Kusk Nielsen, Dorte Frees, and Søren Persson. Comparison of illumina and oxford nanopore sequencing data quality for *clostridioides difficile* genome analysis and their application for epidemiological surveillance. *BMC genomics*, 26(1):92, 2025.
- [Bov08] Theodor Boveri. Concerning the origin of malignant tumours by theodor boveri. translated and annotated by henry harris. *Journal of Cell Science*, 121(Supplement_1):1–84, 01 2008.
- [Bro02] Terence A Brown. Mutation, repair and recombination. In *Genomes. 2nd edition*. Wiley-Liss, 2002.
- [Bro23] Terry A Brown. *Genomes 5*. CRC Press, 2023.
- [CB16] Emily Clough and Tanya Barrett. The gene expression omnibus database. *Statistical Genomics: Methods and Protocols*, pages 93–110, 2016.
- [CBCB19] Eric A Collisson, Peter Bailey, David K Chang, and Andrew V Biankin. Molecular subtypes of pancreatic cancer. *Nature reviews Gastroenterology & hepatology*, 16(4):207–220, 2019.
- [CBHK02] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.
- [CBJP23] Laura Calvet, Sergio Benito, Angel A Juan, and Ferran Prados. On the role of metaheuristic optimization in bioinformatics. *International Transactions in Operational Research*, 30(6):2909–2944, 2023.
- [CCR23] Amelia Casamassimi, Alfredo Ciccodicola, and Monica Rienzo. Transcriptional regulation and its misregulation in human diseases, 2023.
- [CCZ22] Miao Cui, Chao Cheng, and Lanjing Zhang. High-throughput proteomics: a methodological mini-review. *Laboratory investigation*, 102(11):1170–1181, 2022.
- [CKP⁺22] Chiranji Lal Chowdhary, Neelu Khare, Harshita Patel, Srinivas Koppu, Rajesh Kaluri, and Dharmendra Singh Rajput. Past,

- present and future of gene feature selection for breast cancer classification—a survey. *International Journal of Engineering Systems Modelling and Simulation*, 13(2):140–153, 2022.
- [CMC⁺17] Shouvik Chakraborty, Kalyani Mali, Sankhadeep Chatterjee, Soumen Banerjee, Kaustav Guha Mazumdar, Mainak Debnath, Pikorab Basu, Soumyadip Bose, and Kyamelia Roy. Detection of skin disease using metaheuristic supported artificial neural networks. In *2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON)*, pages 224–229, 2017.
- [CMF⁺21] Priscila DR Cirillo, Katia Margiotti, Marco Fabiani, Mateus C Barros-Filho, David Sparacino, Antonella Cima, Salvatore A Longo, Marina Cupellaro, Alvaro Mesoraca, and Claudio Giorlandino. Multi-analytical test based on serum mirnas and proteins quantification for ovarian cancer early detection. *PLoS One*, 16(8):e0255804, 2021.
- [CPLZ22] Zhaoxiang Cai, Rebecca C Poulos, Jia Liu, and Qing Zhong. Machine learning for multi-omics data integration in cancer. *iScience*, 25(2), 2022.
- [CRP18] Ferdinando Chiaradonna, Francesca Ricciardiello, and Roberta Palorini. The nutrient-sensing hexosamine biosynthetic pathway as the hub of cancer metabolic rewiring. *Cells*, 7(6):53, 2018.
- [CS14] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & electrical engineering*, 40(1):16–28, 2014.
- [CSCW21] Heping Cao, Kandan Sethumadhavan, Fangping Cao, and Thomas TY Wang. Gossypol decreased cell viability and down-regulated the expression of a number of genes in human colon cancer cells. *Scientific Reports*, 11(1):5922, 2021.
- [CSO⁺16] Antonio Colaprico, Tiago C Silva, Catharina Olsen, Luciano Garofano, Claudia Cava, Davide Garolini, Thais S Sabedot, Tathiane M Malta, Stefano M Pagnotta, Isabella Castiglioni, et al. Tcgabiolinks: an r/bioconductor package for integrative

- analysis of tcga data. *Nucleic acids research*, 44(8):e71–e71, 2016.
- [CSSS22] Sanjoy Chakraborty, Sushmita Sharma, Apu Kumar Saha, and Ashim Saha. A novel improved whale optimization algorithm to solve numerical optimization and real-world applications. *Artificial Intelligence Review*, pages 1–112, 2022.
- [CZL⁺17] Lei Chen, Yu-Hang Zhang, Guohui Lu, Tao Huang, and Yu-Dong Cai. Analysis of cancer-related lncrnas using gene ontology and kegg pathways. *Artificial intelligence in medicine*, 76:27–36, 2017.
- [CZZ19] Li Chen, Jing Zhu, and Ling-Jie Zhang. Long non-coding rna small nucleolar rna host gene 7 is upregulated and promotes cell proliferation in thyroid cancer. *Oncology letters*, 18(5):4726–4734, 2019.
- [D⁺72] Cox R David et al. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, 34(2):187–220, 1972.
- [DAAC20] Tonmoy Das, Geoffroy Andrieux, Musaddeque Ahmed, and Sajib Chakraborty. Integration of online omics-data resources for cancer research. *Frontiers in Genetics*, 11:578345, 2020.
- [DB09] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 2009.
- [DBS06] Marco Dorigo, Mauro Birattari, and Thomas Stutzle. Ant colony optimization. *IEEE computational intelligence magazine*, 1(4):28–39, 2006.
- [DBSP16] Mario Deng, Johannes Brägelmann, Joachim L Schultze, and Sven Perner. Web-tcga: an online platform for integrated analysis of molecular cancer data sets. *BMC bioinformatics*, 17:1–7, 2016.
- [DDN⁺23] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9):1–33, 2023.

- [DDS⁺13] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.
- [DEFW23] Courtney D DiNardo, Harry P Erba, Sylvie D Freeman, and Andrew H Wei. Acute myeloid leukaemia. *The Lancet*, 401(10393):2073–2086, 2023.
- [DGH17] Padideh Danaee, Reza Ghaeini, and David A Hendrix. A deep learning approach for cancer detection and relevant gene identification. In *Pacific symposium on biocomputing 2017*, pages 219–229. World Scientific, 2017.
- [Dik22] Nikolaos Dikaïos. Sparse-input neural networks to differentiate 32 primary cancer types on the basis of somatic point mutations. *Onco*, 2(2):56–68, 2022.
- [DJR12] Vincent T DeVita Jr and Steven A Rosenberg. Two hundred years of cancer research. *New England Journal of Medicine*, 366(23):2207–2214, 2012.
- [DL97] Manoranjan Dash and Huan Liu. Feature selection for classification. *Intelligent data analysis*, 1(1-4):131–156, 1997.
- [DLDW22] Xiongshi Deng, Min Li, Shaobo Deng, and Lei Wang. Hybrid gene selection approach using xgboost and multi-objective genetic algorithm for cancer classification. *Medical & Biological Engineering & Computing*, 60(3):663–681, 2022.
- [DLH⁺17] Kevin Drew, Chanjae Lee, Ryan L Huizar, Fan Tu, Blake Borgeason, Claire D McWhite, Yun Ma, John B Wallingford, and Edward M Marcotte. Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Molecular systems biology*, 13(6):932, 2017.
- [DM07] Sean Davis and Paul S Meltzer. Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor. *Bioinformatics*, 23(14):1846–1847, 2007.
- [DMBM17] Kevin Drew, Christian L Müller, Richard Bonneau, and Edward M Marcotte. Identifying direct contacts between protein complex subunits from their conditional depen-

- dence in proteomics datasets. *PLoS computational biology*, 13(10):e1005625, 2017.
- [DMX⁺16] Gaochao Dong, Qixing Mao, Wenjie Xia, Youtao Xu, Jie Wang, Lin Xu, and Feng Jiang. Pkm2 and cancer: The function of pkm2 beyond glycolysis. *Oncology letters*, 11(3):1980–1986, 2016.
- [dNO20] Renan Gomes do Nascimento and Kaléu Mormino Otoni. Histological and molecular classification of breast cancer: what do we know? *Mastology*, 30:1–8, 2020.
- [Doo06] Erin E Dooley. Ehpnet: Wellcome trust sanger institute, 2006.
- [DP23] S Sathiya Devi and K Prithiviraj. Breast cancer classification with microarray gene expression data based on improved whale optimization algorithm. *International Journal of Swarm Intelligence Research (IJSIR)*, 14(1):1–21, 2023.
- [DRIE⁺08] Michał Damiński, Alvaro Rada-Iglesias, Stefan Enroth, Claes Wadelius, Jacek Koronacki, and Jan Komorowski. Monte carlo feature selection for supervised classification. *Bioinformatics*, 24(1):110–117, 2008.
- [DST⁺24] Rebecca J. Deyell, Yaoqing Shen, Emma Titmuss, Katherine Dixon, Laura M. Williamson, Erin Pleasance, Jessica M. T. Nelson, Sanna Abbasi, Martin Krzywinski, Linlea Armstrong, Melika Bonakdar, Carolyn Ch’ng, Eric Chuah, Chris Dunham, Alexandra Fok, Martin Jones, Anna F. Lee, Yussanne Ma, Richard A. Moore, Andrew J. Mungall, Karen L. Mungall, Paul C. Rogers, Kasmintan A. Schrader, Alice Virani, Kathleen Wee, Sean S. Young, Yongjun Zhao, Steven J. M. Jones, Janessa Laskin, Marco A. Marra, and Shahrard R. Rassekh. Whole genome and transcriptome integrated analyses guide clinical care of pediatric poor prognosis cancers. *Nature Communications*, 15(1):4165, May 2024.
- [DWC16] William N Dudley, Rita Wickham, and Nicholas Coombs. An introduction to survival statistics: Kaplan-meier analysis. *Journal of the advanced practitioner in oncology*, 7(1):91, 2016.

- [EB00] Charles B Epstein and Ronald A Butow. Microarray technology—enhanced versatility, persistent challenge. *Current opinion in biotechnology*, 11(1):36–41, 2000.
- [EEMS19] Murtada K Elbashir, Mohamed Ezz, Mohanad Mohammed, and Said S Saloum. Lightweight convolutional neural network for breast cancer classification using rna-seq gene expression data. *IEEE Access*, 7:185338–185348, 2019.
- [ENM15] Issam El Naqa and Martin J Murphy. What is machine learning? In *Machine learning in radiation oncology: theory and applications*, pages 3–11. Springer, 2015.
- [ENS24] Amir Esmaeili Nezhad and Mohammad Hamed Samimi. A review of the applications of machine learning in the condition monitoring of transformers. *Energy Systems*, 15(1):463–493, 2024.
- [EOT⁺15] Nathan J Edwards, Mauricio Oberti, Ratna R Thangudu, Shuang Cai, Peter B McGarvey, Shine Jacob, Subha Madhavan, and Karen A Ketchum. The cptac data portal: a resource for cancer proteomics research. *Journal of proteome research*, 14(6):2707–2713, 2015.
- [EP19] Absalom E Ezugwu and Doddy Prayogo. Symbiotic organisms search algorithm: theory, recent advances and applications. *Expert Systems with Applications*, 119:184–209, 2019.
- [ESBB98] Michael B Eisen, Paul T Spellman, Patrick O Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, 1998.
- [Fag15] Guy B. Faguet. A brief history of cancer: Age-old milestones underlying our current knowledge database. *International Journal of Cancer*, 136(9):2022–2036, 2015.
- [FCGD19] Bruno Cesar Feltes, Eduardo Bassani Chandelier, Bruno Iochins Grisci, and Márcio Dorn. Cumida: An extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research. *Journal of Computational Biology*, 26(4):376–386, 2019.

- [FCS⁺16] Fiona Fouhy, Adam G Clooney, Catherine Stanton, Marcus J Claesson, and Paul D Cotter. 16s rna gene sequencing of mock microbial populations-impact of dna extraction method, primer choice and sequencing platform. *BMC microbiology*, 16:1–13, 2016.
- [FdPP21] Mattia Furlan, Stefano de Pretis, and Mattia Pelizzola. Dynamics of transcriptional and post-transcriptional regulation. *Briefings in Bioinformatics*, 22(4):bbaa389, 2021.
- [FMZ⁺23] Tzu-Yu Feng, Stephanie J Melchor, Xiao-Yu Zhao, Haider Ghuman, Mark Kester, Todd E Fox, and Sarah E Ewald. Tricarboxylic acid (tca) cycle, sphingolipid, and phosphatidylcholine metabolism are dysregulated in t. gondii infection-induced cachexia. *Heliyon*, 9(7), 2023.
- [FPD21] Bruno Cesar Feltes, Joice De Faria Poloni, and Marcio Dorn. Benchmarking and testing machine learning approaches with barra: Curda, a curated rna-seq database for cancer research. *Journal of Computational Biology*, 28(9):931–944, 2021.
- [FS97] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.
- [GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep feedforward networks. *Deep learning*, 1:161–217, 2016.
- [GCM10] Tamar Geiger, Juergen Cox, and Matthias Mann. Proteomic changes resulting from gene copy number variations in cancer cells. *PLoS genetics*, 6(9):e1001090, 2010.
- [GFBG06] Pablo M Granitto, Cesare Furlanello, Franco Biasioli, and Flavia Gasperi. Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products. *Chemometrics and intelligent laboratory systems*, 83(2):83–90, 2006.
- [GFD14] Manizheh Ghaemi and Mohammad-Reza Feizi-Derakhshi. Forest optimization algorithm. *Expert systems with applications*, 41(15):6676–6687, 2014.
- [GFD19] Bruno Iochins Grisci, Bruno César Feltes, and Marcio Dorn. Neuroevolution as a tool for microarray gene expression pat-

- tern identification in cancer research. *Journal of biomedical informatics*, 89:122–133, 2019.
- [GHG⁺23] Claudia Geismann, Charlotte Hauser, Frauke Grohmann, Christian Schneeweis, Nico Bölter, Jan-Paul Gundlach, Günter Schneider, Christoph Röcken, Christian Meinhardt, Heiner Schäfer, et al. Nf- κ b/rela controlled a20 limits trail-induced apoptosis in pancreatic cancer. *Cell Death & Disease*, 14(1):3, 2023.
- [GHL⁺24] Mohamed Amine Guerroudji, Zineb Hadjadj, Mohamed Lichouri, Kahina Amara, and Nadia Zenati. Efficient machine learning-based approach for brain tumor detection using the cad system. *IETE Journal of Research*, 70(4):3664–3678, 2024.
- [GI23] Aishwarya Gondane and Harri M Itkonen. Revealing the history and mystery of rna-seq. *Current Issues in Molecular Biology*, 45(3):1860–1874, 2023.
- [GJKS10] Karly P Garnock-Jones, Gillian M Keating, and Lesley J Scott. Trastuzumab: a review of its use as adjuvant treatment in human epidermal growth factor receptor 2 (her2) positive early breast cancer. *Drugs*, 70(2):215–239, 2010.
- [GJW⁺21] Paul Gamble, Ronnachai Jaroensri, Hongwu Wang, Fraser Tan, Melissa Moran, Trissia Brown, Isabelle Flament-Auvigne, Emad A Rakha, Michael Toss, David J Dabbs, et al. Determining breast cancer biomarker status and associated morphological features using deep learning. *Communications medicine*, 1(1):14, 2021.
- [GM24] Abderrahmane Guerrache and Olivier Micheau. Tnf-related apoptosis-inducing ligand: Non-apoptotic signalling. *Cells*, 13(6):521, 2024.
- [GMB⁺02] Helen R Griffiths, Lennart Møller, Grzegorz Bartosz, Aalt Bast, Carlo Bertoni-Freddari, Andrew Collins, Marcus Cooke, Stefan Coolen, Guido Haenen, Anne-Mette Hoberg, et al. Biomarkers. *Molecular aspects of medicine*, 23(1-3):101–208, 2002.
- [GS11] Vishal V Gandhi and David C Samuels. A review comparing deoxyribonucleoside triphosphate (dntp) concentrations in the

- mitochondrial and cytoplasmic compartments of normal and transformed cells. *Nucleosides, Nucleotides and Nucleic Acids*, 30(5):317–339, 2011.
- [GSB20] Filippo Geraci, Indrajit Saha, and Monica Bianchini. Rna-seq analysis: methods, applications and challenges. *Frontiers in Genetics*, 11:220, 2020.
- [GSP⁺06] Adam M Gustafson, Evan S Snitkin, Stephen CJ Parker, Charles DeLisi, and Simon Kasif. Towards the identification of essential genes using targeted genome sequencing and comparative analysis. *Bmc Genomics*, 7:1–16, 2006.
- [GSS02] John M Grange, John L Stanford, and Cynthia A Stanford. Campbell de morgan’s ‘observations on cancer’, and their relevance today. *Journal of the Royal Society of Medicine*, 95(6):296–299, 2002. PMID: 12042378.
- [GT08] Virendra S Gomase and Somnath Tagore. Epigenomics. *Current drug metabolism*, 9(3):232–237, 2008.
- [GYA13] Amir Hossein Gandomi, Xin-She Yang, and Amir Hossein Alavi. Cuckoo search algorithm: a metaheuristic approach to solve structural optimization problems. *Engineering with computers*, 29:17–35, 2013.
- [HAH⁺21] Essam H Houssein, Daa Salama Abdelminaam, Hager N Hassan, Mustafa M Al-Sayed, and Emad Nabil. A hybrid barnacles mating optimizer algorithm with support vector machines for gene selection of microarray cancer classification. *IEEE Access*, 9:64895–64905, 2021.
- [HBP⁺17] Edward L Huttlin, Raphael J Bruckner, Joao A Paulo, Joe R Cannon, Lily Ting, Kurt Baltier, Greg Colby, Fana Gebreab, Melanie P Gygi, Hannah Parzen, et al. Architecture of the human interactome defines protein communities and disease networks. *Nature*, 545(7655):505–509, 2017.
- [HCMD21] Taishan Hu, Nilesh Chitnis, Dimitri Monos, and Anh Dinh. Next-generation sequencing technologies: An overview. *Human Immunology*, 82(11):801–811, 2021.

- [HD17] Seyed Hossein Hassanpour and Mohammadamin Dehghani. Review of cancer from perspective of molecular. *Journal of cancer research and practice*, 4(4):127–129, 2017.
- [HH12] N Lynn Henry and Daniel F Hayes. Cancer biomarkers. *Molecular oncology*, 6(2):140–146, 2012.
- [HHN⁺12] Pierre C Havugimana, G Traver Hart, Tamás Nepusz, Haixuan Yang, Andrei L Turinsky, Zhihua Li, Peggy I Wang, Daniel R Boutz, Vincent Fong, Sadhna Phanse, et al. A census of human soluble protein complexes. *Cell*, 150(5):1068–1081, 2012.
- [HHP⁺15] Marco Y Hein, Nina C Hubner, Ina Poser, Jürgen Cox, Nagarjuna Nagaraj, Yusuke Toyoda, Igor A Gak, Ina Weisswange, Jörg Mansfeld, Frank Buchholz, et al. A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell*, 163(3):712–723, 2015.
- [HKDL⁺12] Benjamin Haibe-Kains, Christine Desmedt, Sherene Loi, Aedin C Culhane, Gianluca Bontempi, John Quackenbush, and Christos Sotiriou. A three-gene model to robustly identify breast cancer molecular subtypes. *Journal of the National Cancer Institute*, 104(4):311–325, 2012.
- [HL22] August Yue Huang and Eunjung Alice Lee. Identification of somatic mutations from bulk and single-cell sequencing data. *Frontiers in Aging*, 2:800380, 2022.
- [HN98] Jerry L Hintze and Ray D Nelson. Violin plots: a box plot-density trace synergism. *The American Statistician*, 52(2):181–184, 1998.
- [HS07] Daniel A Haber and Jeff Settleman. Drivers and passengers. *Nature*, 446(7132):145–146, 2007.
- [HS24] Huijie Han and Hélder A Santos. Nano-and micro-platforms in therapeutic proteins delivery for cancer therapy: Materials and strategies. *Advanced Materials*, page 2409522, 2024.
- [HTB⁺15] Edward L Huttlin, Lily Ting, Raphael J Bruckner, Fana Gebreab, Melanie P Gygi, John Szpyt, Stanley Tam, Gabriela Zarraga, Greg Colby, Kurt Baltier, et al. The bioplex network: a system-

- atic exploration of the human interactome. *Cell*, 162(2):425–440, 2015.
- [HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition, 2009.
- [HTT17] Radmila Hrdlickova, Masoud Toloue, and Bin Tian. Rna-seq methods for transcriptome analysis. *Wiley Interdisciplinary Reviews: RNA*, 8(1):e1364, 2017.
- [HZ18] Carolyn Hutter and Jean Claude Zenklusen. The cancer genome atlas: creating lasting value beyond its data. *Cell*, 173(2):283–285, 2018.
- [HZYZ21] Zongzhen He, Junying Zhang, Xiguo Yuan, and Yuanyuan Zhang. Integrating somatic mutations for breast cancer survival prediction using machine learning methods. *Frontiers in genetics*, 11:632901, 2021.
- [HZZ24] Yiran Huang, Pingfan Zeng, and Cheng Zhong. Classifying breast cancer subtypes on multi-omics data via sparse canonical correlation analysis and deep learning. *BMC bioinformatics*, 25(1):132, 2024.
- [IS24] Ayokunle Olalekan Ige and Malusi Sibiyi. State-of-the-art in 1d convolutional neural networks: A survey. *IEEE Access*, 2024.
- [JBT⁺24] Dirk Jäger, Aileen Berger, Alexandra Tuch, Claudia Luckner-Minden, Rosa Eurich, Mario Hlevnjak, Andreas Schneeweiss, Peter Lichter, Sebastian Aulmann, Claus-Peter Heussel, et al. Novel chimeric antigen receptors for the effective and safe treatment of ny-br-1 positive breast cancer. *Clinical and Translational Medicine*, 14(7), 2024.
- [JCC⁺21] Dongfang Jia, Cheng Chen, Chen Chen, Fangfang Chen, Ningrui Zhang, Ziwei Yan, and Xiaoyi Lv. Breast cancer case identification based on deep learning and bioinformatics analysis. *Frontiers in genetics*, 12:628136, 2021.
- [JFGS17] Mark A Jensen, Vincent Ferretti, Robert L Grossman, and Louis M Staudt. The nci genomic data commons as an engine

- for precision medicine. *Blood, The Journal of the American Society of Hematology*, 130(4):453–459, 2017.
- [JL10] ZQ John Lu. The elements of statistical learning: data mining, inference, and prediction, 2010.
- [JPP⁺21] Sol A Jeon, Jong Lyul Park, Seung-Jin Park, Jeong Hwan Kim, Sung-Ho Goh, Ji-Youn Han, and Seon-Young Kim. Comparison between mgi and illumina sequencing platforms for whole genome sequencing. *Genes & Genomics*, 43:713–724, 2021.
- [JPZ14] Peilin Jia, William Pao, and Zhongming Zhao. Patterns and processes of somatic mutations in nine major cancers. *BMC medical genomics*, 7:1–11, 2014.
- [JS24] G JagadeeswaraRao and A Sivaprasad. An integrated ensemble learning technique for gene expression classification and biomarker identification from rna-seq data for pancreatic cancer prognosis. *International Journal of Information Technology*, 16(3):1505–1516, 2024.
- [JWC⁺14] Peilin Jia, Quan Wang, Qingxia Chen, Katherine E Hutchinson, William Pao, and Zhongming Zhao. Msea: detection and quantification of mutation hotspots through mutation set enrichment analysis. *Genome biology*, 15:1–16, 2014.
- [Kar10] Dervis Karaboga. Artificial bee colony algorithm. *scholarpedia*, 5(3):6915, 2010.
- [Kas14] Ali Husseinzadeh Kashan. League championship algorithm (lca): An algorithm for global optimization inspired by sport championships. *Applied Soft Computing*, 16:171–200, 2014.
- [KASD20] Satnam Kaur, Lalit K Awasthi, Amrit Lal Sangal, and Gaurav Dhiman. Tunicate swarm algorithm: A new bio-inspired based metaheuristic paradigm for global optimization. *Engineering Applications of Artificial Intelligence*, 90:103541, 2020.
- [KAU⁺24] Hanen Karamti, Raed Alharthi, Muhammad Umer, Hadil Shaiba, Abid Ishaq, Nihal Abuzinadah, Shtwai Alsubai, and Imran Ashraf. Breast cancer detection employing stacked ensemble model with convolutional features. *Cancer Biomarkers*, 40(2):155–170, 2024.

- [KG00] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.
- [KJR10] Miron B Kursa, Aleksander Jankowski, and Witold R Rudnicki. Boruta—a system for feature selection. *Fundamenta informaticae*, 101(4):271–285, 2010.
- [KK16] Oliver Kramer and Oliver Kramer. Scikit-learn. *Machine learning for evolution strategies*, pages 45–53, 2016.
- [KK17] Ali Kaveh and Ali Kaveh. Imperialist competitive algorithm. *Advances in Metaheuristic Algorithms for Optimal Design of Structures*, pages 353–373, 2017.
- [KL23] Sabine Kayser and Mark J Levis. The clinical impact of the molecular landscape of acute myeloid leukemia. *Haematologica*, 108(2):308, 2023.
- [KM58] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.
- [KM23] Mehrdad Kaveh and Mohammad Saadi Mesgari. Application of meta-heuristic algorithms for training neural networks and deep learning architectures: A comprehensive review. *Neural Processing Letters*, 55(4):4519–4622, 2023.
- [KMAS⁺22] Mahmood Khalsan, Lee R Machado, Eman Salih Al-Shamery, Suraj Ajit, Karen Anthony, Mu Mu, and Michael Opoku Agyeman. A survey of machine learning approaches applied to gene expression analysis for cancer prediction. *IEEE Access*, 10:27522–27534, 2022.
- [KMM⁺15] Manish Kumar, Ajay Matta, Olena Masui, Gunjan Srivastava, Jatinder Kaur, Alok Thakar, Nootan Kumar Shukla, Ajoy Roy-Choudhury, Meherchand Sharma, Paul G Walfish, et al. Nuclear heterogeneous nuclear ribonucleoprotein d is associated with poor prognosis and interactome analysis reveals its novel binding partners in oral cancer. *Journal of translational medicine*, 13:1–15, 2015.

- [KP04] A. Karpozilos and N. Pavlidis. The treatment of cancer in greek antiquity. *European Journal of Cancer*, 40(14):2033–2040, 2004.
- [KPS⁺18] Marieke Lydia Kuijjer, Joseph Nathaniel Paulson, Peter Salzman, Wei Ding, and John Quackenbush. Cancer subtype identification using somatic mutation data. *British journal of cancer*, 118(11):1492–1501, 2018.
- [LAS07] Gert Lubec and Leila Afjehi-Sadat. Limitations and pitfalls in protein identification by mass spectrometry. *Chemical reviews*, 107(8):3568–3584, 2007.
- [LaV08] Michael P LaValley. Logistic regression. *Circulation*, 117(18):2395–2399, 2008.
- [LBC⁺11] Brian D Lehmann, Joshua A Bauer, Xi Chen, Melinda E Sanders, A Bapsi Chakravarthy, Yu Shyr, Jennifer A Pietenpol, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *The Journal of clinical investigation*, 121(7):2750–2767, 2011.
- [LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
- [LCC⁺12] Hugo YK Lam, Michael J Clark, Rui Chen, Rong Chen, Georges Natsoulis, Maeve O’huallachain, Frederick E Dewey, Lukas Habegger, Euan A Ashley, Mark B Gerstein, et al. Performance comparison of whole-genome sequencing platforms. *Nature biotechnology*, 30(1):78–82, 2012.
- [LCW⁺17] Jundong Li, Kewei Cheng, Suhan Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 50(6):1–45, 2017.
- [LD11] Bo Li and Colin N Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12:1–16, 2011.
- [LDB99] Bob Lowenberg, James R Downing, and Alan Burnett. Acute myeloid leukemia. *New England Journal of Medicine*, 341(14):1051–1062, 1999.

- [LHA14] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with *deseq2*. *Genome biology*, 15:1–21, 2014.
- [LHS⁺24] Xinrong Li, Zhixiang Hu, Qili Shi, Wenying Qiu, Yizhe Liu, Yanfang Liu, Shenglin Huang, Linhui Liang, Zhiao Chen, and Xianghuo He. Elevated choline drives klf5-dominated transcriptional reprogramming to facilitate liver cancer progression. *Oncogene*, pages 1–16, 2024.
- [Liu20] Zhenqiu Liu. Visualizing single-cell rna-seq data with semisupervised principal component analysis. *International journal of molecular sciences*, 21(16):5797, 2020.
- [LL14] Daniel T. Larose and Chantal D. Larose. *k-nearest neighbor algorithm*, pages 149–164. Wiley Data and Cybersecurity, 2014.
- [LL18] Wei Vivian Li and Jingyi Jessica Li. Modeling and analysis of rna-seq data: a review from a statistical perspective. *Quantitative Biology*, 6(3):195–209, 2018.
- [LLY⁺21] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 33(12):6999–7019, 2021.
- [LMZV17] Marine Le Morvan, Andrei Zinovyev, and Jean-Philippe Vert. Netnorm: Capturing cancer-relevant information in somatic exome mutation data with gene networks for cancer stratification and prognosis. *PLoS computational biology*, 13(6):e1005573, 2017.
- [LÖJ17] Wenwen Li, Ender Özcan, and Robert John. A learning automata-based multiobjective hyper-heuristic. *IEEE Transactions on Evolutionary Computation*, 23(1):59–73, 2017.
- [LPL23] Pengzhi Li, Yan Pei, and Jianqiang Li. A comprehensive survey on design and application of autoencoder in deep learning. *Applied Soft Computing*, 138:110176, 2023.
- [LRC⁺22] Quan Li, Zilin Ren, Kajia Cao, Marilyn M Li, Kai Wang, and Yunyun Zhou. Cancervar: An artificial intelligence–empowered

- platform for clinical interpretation of somatic mutations in cancer. *Science advances*, 8(18):eabj1624, 2022.
- [LRY⁺21] Qing Liao, Yun Ren, Yuyi Yang, Xiaohui Zhu, Yunfei Zhi, Yujie Zhang, Yi Chen, Yanqing Ding, and Liang Zhao. Cct8 recovers wtp53-suppressed cell cycle evolution and emt to promote colorectal cancer progression. *Oncogenesis*, 10(12):84, 2021.
- [LSM⁺14] Michael S Lawrence, Petar Stojanov, Craig H Mermel, James T Robinson, Levi A Garraway, Todd R Golub, Matthew Meyerson, Stacey B Gabriel, Eric S Lander, and Gad Getz. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484):495–501, 2014.
- [LSP⁺11] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.
- [LST⁺18] Xiaonan Liu, Kari Salokas, Fitsum Tamene, Yaming Jiu, Rigbe G Weldatsadik, Tiina Öhman, and Markku Varjosalo. An ap-ms- and bioid-compatible mac-tag enables comprehensive mapping of protein interactions and subcellular localizations. *Nature communications*, 9(1):1188, 2018.
- [LWZ12] Yanli Liu, Yourong Wang, and Jian Zhang. New machine learning algorithm: Random forest. In *Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings 3*, pages 246–252. Springer, 2012.
- [M⁺66] Nathan Mantel et al. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, 50(3):163–170, 1966.
- [MC15] Iñigo Martincorena and Peter J Campbell. Somatic mutation in cancer and normal cells. *Science*, 349(6255):1483–1489, 2015.
- [MCA⁺24] Raafat M Munshi, Lucia Cascone, Nazik Alturki, Oumaima Saidani, Amal Alshardan, and Muhammad Umer. A novel approach for breast cancer detection using optimized ensemble

- learning framework and xai. *Image and Vision Computing*, 142:104910, 2024.
- [MD20] Akshaya Kumar Mandal and Satchidananda Dehuri. A survey on ant colony optimization for solving some of the selected np-hard problem. In *Biologically Inspired Techniques in Many-Criteria Decision Making: International Conference on Biologically Inspired Techniques in Many-Criteria Decision Making (BITMDM-2019)*, pages 85–100. Springer, 2020.
- [MEFD⁺23] Tehnan IA Mohamed, Absalom E Ezugwu, Jean Vincent Fonou-Dombeu, Abiodun M Ikotun, and Mohanad Mohammed. A bio-inspired convolution neural network architecture for automatic breast cancer detection and classification using rna-seq gene expression data. *Scientific Reports*, 13(1):14644, 2023.
- [MG21] Otília Menyhárt and Balázs Gyórfy. Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. *Computational and structural biotechnology journal*, 19:949–960, 2021.
- [MG23] Marcela Braga Mansur and Mel Greaves. Convergent tp53 loss and evolvability in cancer. *BMC Ecology and Evolution*, 23(1):54, 2023.
- [MJ22] Shino Murakami and Samie R Jaffrey. Hidden codes in mrna: control of gene expression by m6a. *Molecular cell*, 82(12):2236–2251, 2022.
- [MKM21] Nida Meddouri, Hela Khoufi, and Mondher Maddouri. Dfc: A performant dagging approach of classification based on formal concept. *International Journal of Artificial Intelligence and Machine Learning (IJAIML)*, 11(2):38–62, 2021.
- [MKZ⁺22] DR Mani, Karsten Krug, Bing Zhang, Shankha Satpathy, Karl R Clauser, Li Ding, Matthew Ellis, Michael A Gillette, and Steven A Carr. Cancer proteogenomics: current impact and future prospects. *Nature Reviews Cancer*, 22(5):298–313, 2022.
- [MM04] George L Gabor Miklos and Ryszard Maleszka. Microarray reality checks in the context of a complex disease. *Nature biotechnology*, 22(5):615–621, 2004.

- [MMM⁺21] Mohanad Mohammed, Henry Mwambi, Innocent B Mboya, Murtada K Elbashir, and Bernard Omolo. A stacking ensemble deep learning approach to cancer type classification based on tcga data. *Scientific reports*, 11(1):15626, 2021.
- [MMSAMN20] Habib MotieGhader, Yosef Masoudi-Sobhanzadeh, Saman Hosseini Ashtiani, and Ali Masoudi-Nejad. mrna and microrna selection for breast cancer molecular subtype stratification using meta-heuristic based algorithms. *Genomics*, 112(5):3207–3217, 2020.
- [MNA⁺22] Amichay Meirovitz, Benjamin Nisman, Tanir M Allweis, Einat Carmon, Luna Kadouri, Bella Maly, Ofra Maimon, and Tamar Peretz. Thyroid hormones and morphological features of primary breast cancer. *Anticancer Research*, 42(1):253–261, 2022.
- [MON⁺00] Elizabeth Mallon, Pinchas Osin, Nasar Nasiri, Iain Blain, Beatrice Howard, and Barry Gusterson. The basic pathology of human breast cancer. *Journal of mammary gland biology and neoplasia*, 5:139–163, 2000.
- [MPEMR12] Cristina Muñoz-Pinedo, N El Mjiyad, and Jean-Ehrland Ricci. Cancer metabolism: current perspectives and future directions. *Cell death & disease*, 3(1):e248–e248, 2012.
- [MR93] Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342, 1993.
- [MRB05] Adele Murrell, Vardhman K Rakyan, and Stephan Beck. From genome to epigenome. *Human molecular genetics*, 14(suppl_1):R3–R10, 2005.
- [MSH⁺11] Craig H Mermel, Steven E Schumacher, Barbara Hill, Matthew L Meyerson, Rameen Beroukhim, and Gad Getz. Gistic2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome biology*, 12:1–14, 2011.
- [MSM16] Yosef Masoudi-Sobhanzadeh and Habib Motieghader. World competitive contests (wcc) algorithm: A novel intelligent opti-

- mization algorithm for biological and non-biological problems. *Informatics in Medicine Unlocked*, 3:15–28, 2016.
- [MT15] Joseph A Marsh and Sarah A Teichmann. Structure, dynamics, assembly, and evolution of protein complexes. *Annual review of biochemistry*, 84:551–575, 2015.
- [Mur12] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- [MX22] Qin Ma and Dong Xu. Deep learning shapes single-cell data analysis. *Nature Reviews Molecular Cell Biology*, 23(5):303–304, 2022.
- [MZB24] Marie E Migaud, Mathias Ziegler, and Joseph A Baur. Regulation of and challenges in targeting nad⁺ metabolism. *Nature Reviews Molecular Cell Biology*, pages 1–19, 2024.
- [NBS⁺22] Mario C. Nurchis, Valerio Bordino, Martina Sapienza, Gerardo Altamura, Luca A. Rea, Americo Cicchetti, and Davide Carrozzini. Cancers: What are the costs in relation to disability-adjusted life years? a systematic review and meta-analysis. *International Journal of Environmental Research and Public Health*, 19(8):4862, 2022.
- [NK10] Pauline C Ng and Ewen F Kirkness. Whole genome sequencing. *Genetic variation: Methods and protocols*, pages 215–226, 2010.
- [NLZJ23] Ruth Nussinov, Yonglan Liu, Wengang Zhang, and Hyunbum Jang. Cell phenotypes can be predicted from propensities of protein conformations. *Current opinion in structural biology*, 83:102722, 2023.
- [OBA23] Onyinyechi Obidiro, Gantumur Battogtokh, and Emmanuel O Akala. Triple negative breast cancer treatment options and limitations: future outlook. *Pharmaceutics*, 15(7):1796, 2023.
- [OE22] Olaide N Oyelade and Absalom E Ezugwu. A novel wavelet decomposition and transformation convolutional neural network with data augmentation for breast cancer detection using digital mammogram. *Scientific Reports*, 12(1):5913, 2022.

- [OM01] Godfrey C Onwubolu and Michael Mutingi. A genetic algorithm approach to cellular manufacturing systems. *Computers & industrial engineering*, 39(1-2):125–144, 2001.
- [OMD⁺22] Roseline Oluwaseun Ogundokun, Sanjay Misra, Mychal Douglas, Robertas Damaševičius, and Rytis Maskeliūnas. Medical internet-of-things based breast cancer diagnosis using hyperparameter-optimized neural networks. *Future Internet*, 14(5):153, 2022.
- [ORC⁺21] Rose Oughtred, Jennifer Rust, Christie Chang, Bobby-Joe Breitkreutz, Chris Stark, Andrew Willems, Lorrie Boucher, Genie Leung, Nadine Kolas, Frederick Zhang, et al. The biogrid database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science*, 30(1):187–200, 2021.
- [Ost12] James M Ostell. Entrez: The ncbi search and discovery engine. In *Data Integration in the Life Sciences: 8th International Conference, DILS 2012, College Park, MD, USA, June 28-29, 2012. Proceedings* 8, pages 1–4. Springer, 2012.
- [PABH⁺24] Antonio Passaro, Maise Al Bakir, Emily G Hamilton, Maximilian Diehn, Fabrice André, Sinchita Roy-Chowdhuri, Giannis Mountzios, Ignacio I Wistuba, Charles Swanton, and Solange Peters. Cancer biomarkers: emerging trends and clinical implications for personalized treatment. *Cell*, 187(7):1617–1635, 2024.
- [PBCV23] Marion Papin, Ana Maria Bouchet, Aurélie Chantôme, and Christophe Vandier. Ether-lipids and cellular signaling: A differential role of alkyl-and alkenyl-ether-lipids? *Biochimie*, 2023.
- [PM15] Julia R Pon and Marco A Marra. Driver and passenger mutations in cancer. *Annual Review of Pathology: Mechanisms of Disease*, 10(1):25–50, 2015.
- [PMDLB19] Aurora Paiano, Azzurra Margiotta, Maria De Luca, and Cecilia Bucci. Yeast two-hybrid assay to identify interacting proteins. *Current protocols in protein science*, 95(1):e70, 2019.

- [PRT⁺20] Stefano Piana, Paul Robustelli, Dazhi Tan, Songela Chen, and David E Shaw. Development of a force field for the simulation of single-chain proteins and protein–protein complexes. *Journal of chemical theory and computation*, 16(4):2494–2507, 2020.
- [PRZD⁺19] Yasset Perez-Riverol, Andrey Zorin, Gaurhari Dass, Manh-Tu Vu, Pan Xu, Mihai Glont, Juan Antonio Vizcaíno, Andrew F Jar-nuczak, Robert Petryszak, Peipei Ping, et al. Quantifying the impact of public omics data. *Nature communications*, 10(1):3512, 2019.
- [PS15] Vivek K Patel and Vimal J Savsani. Heat transfer search (hts): a novel optimization algorithm. *Information sciences*, 324:217–246, 2015.
- [PS20] Derek A Pisner and David M Schnyer. Support vector machine. In *Machine learning*, pages 101–121. Elsevier, 2020.
- [Qui96] J. Ross Quinlan. Learning decision tree classifiers. *ACM Computing Surveys (CSUR)*, 28(1):71–72, 1996.
- [QZY⁺21] Hui Qu, Mu Zhou, Zhennan Yan, He Wang, Vinod K Rustgi, Shaoting Zhang, Olivier Gevaert, and Dimitris N Metaxas. Genetic mutation and biological pathway prediction based on whole slide images in breast carcinoma using deep learning. *NPJ precision oncology*, 5(1):87, 2021.
- [Ray22] Atish Ray. Dna mutation, repair, and recombination. In *Genetics Fundamentals Notes*, pages 433–490. Springer, 2022.
- [RC22] Arif Ahmad Rather and Manzoor Ahmad Chachoo. Umap guided topological analysis of transcriptomic data for cancer subtyping. *International Journal of Information Technology*, 14(6):2855–2865, 2022.
- [RF09] Jorge S Reis-Filho. Next-generation sequencing. *Breast cancer research*, 11(Suppl 3):S12, 2009.
- [RHP⁺20] Matthew A Reyna, David Haan, Marta Paczkowska, Lieven PC Verbeke, Miguel Vazquez, Abdullah Kahraman, Sergio Pulido-Tamayo, Jonathan Barenboim, Lina Wadi, Priyanka Dhingra, et al. Pathway and network analysis of more than 2500 whole cancer genomes. *Nature communications*, 11(1):729, 2020.

- [Roe25] Lutz Roeder. Netron: Visualizer for neural network, deep learning and machine learning models. <https://github.com/lutzroeder/netron>, 2025. Accessed on: 22 July 2025.
- [Rou87] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.
- [RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.
- [RSO⁺21] Md Ashikur Rahman, Rajalingam Sokkalingam, Mahmod Othman, Kallol Biswas, Lazim Abdullah, and Evizal Abdul Kadir. Nature-inspired metaheuristic techniques for combinatorial optimization problems: Overview and recent advances. *Mathematics*, 9(20):2633, 2021.
- [RTM14] Bahareh Rabbani, Mustafa Tekin, and Nejat Mahdieh. The promise of whole-exome sequencing in medical genetics. *Journal of human genetics*, 59(1):5–15, 2014.
- [RTSB18] Muhammad Farooq Rai, Eric D Tycksen, Linda J Sandell, and Robert H Brophy. Advantages of rna-seq compared to rna microarrays for transcriptome profiling of anterior cruciate ligament tears. *Journal of Orthopaedic Research®*, 36(1):484–497, 2018.
- [RVH⁺05] Jean-François Rual, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amélie Dricot, Ning Li, Gabriel F Berriz, Francis D Gibbons, Matija Dreze, Nono Ayivi-Guedehoussou, et al. Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, 437(7062):1173–1178, 2005.
- [SB24] Raghu Solanki and Dhiraj Bhatia. Stimulus-responsive hydrogels for targeted cancer therapy. *Gels*, 10(7):440, 2024.
- [SCSF21] Michael A Skinnider, Charley Cai, R Greg Stacey, and Leonard J Foster. Prince: an r/bioconductor package for protein–protein

- interaction network inference from co-fractionation mass spectrometry data. *Bioinformatics*, 37(17):2775–2777, 2021.
- [SD21] Rahul Suresh and Roberto J Diaz. The remodelling of actin composition as a hallmark of cancer. *Translational Oncology*, 14(6):101051, 2021.
- [SDBLM18] Zhen Rong Siow, Richard H De Boer, Geoffrey J Lindeman, and G Bruce Mann. Spotlight on the utility of the oncotype dx® breast cancer assay. *International journal of women’s health*, pages 89–100, 2018.
- [SFS⁺21] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021.
- [SGN⁺21] Damian Szklarczyk, Annika L Gable, Katerina C Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T Doncheva, Marc Legeay, Tao Fang, Peer Bork, et al. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, 49(D1):D605–D612, 2021.
- [SH23] Mariwan Hama Saeed and Jihad Ibrahim Hama. Cardiac disease prediction using ai algorithms with selectkbest. *Medical & Biological Engineering & Computing*, 61(12):3397–3408, 2023.
- [SHHG18] Devin K Schweppe, Edward L Huttlin, J Wade Harper, and Steven P Gygi. Bioplex display: an interactive suite for large-scale ap–ms protein–protein interaction data. *Journal of proteome research*, 17(1):722–726, 2018.
- [SHRyCP⁺23] Ana Beatriz Sánchez-Heras, Teresa Ramon y Cajal, Marta Pineda, Elena Aguirre, Begoña Graña, Isabel Chirivella, Judit Balmaña, Joan Brunet, SEOM Hereditary Cancer Working Group, and AEGH Hereditary Cancer Committee. Seom clinical guideline on heritable tp53-related cancer syndrome (2022). *Clinical and Translational Oncology*, 25(9):2627–2633, 2023.

- [SHZ⁺19] Zhengshuai Song, Yu Huang, Ye Zhao, Hailong Ruan, Hongmei Yang, Qi Cao, Di Liu, Xiaoping Zhang, and Ke Chen. The identification of potential biomarkers and biological pathways in prostate cancer. *Journal of Cancer*, 10(6):1398, 2019.
- [SK24] Shinnichi Sakamoto and Kentaro Kikuchi. Expanding the cytological and architectural spectrum of mucoepidermoid carcinoma: The key to solving diagnostic problems in morphological variants. In *Seminars in Diagnostic Pathology*. Elsevier, 2024.
- [SLAN20] Chloé B Steen, Chih Long Liu, Ash A Alizadeh, and Aaron M Newman. Profiling cell type abundance and expression in bulk tissues with cibersortx. *Stem Cell Transcriptional Networks: Methods and Protocols*, pages 135–157, 2020.
- [SLR15] Charlotte Soneson, Michael I Love, and Mark D Robinson. Differential analyses for rna-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 4, 2015.
- [SMAN10] Boe S Sorensen, Lise S Mortensen, Jørn Andersen, and Ebba Nexø. Circulating her2 dna after trastuzumab treatment predicts survival and response in breast cancer. *Anticancer research*, 30(6):2463–2468, 2010.
- [SMK⁺23] Prima Sanjaya, Katri Maljanen, Riku Katainen, Sebastian M Waszak, Lauri A Aaltonen, Oliver Stegle, Jan O Korbel, and Esa Pitkänen. Mutation-attention (muat): deep representation learning of somatic mutations for tumour typing and subtyping. *Genome medicine*, 15(1):47, 2023.
- [SMO⁺03] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.
- [SMS⁺14] Sriganesh Srihari, Piyush B Madhamshettiwar, Sarah Song, Chao Liu, Peter T Simpson, Kum Kum Khanna, and Mark A Ragan. Complex-based analysis of dysregulated cellular processes in cancer. *BMC Systems Biology*, 8:1–15, 2014.

- [SMS⁺18] Mohd Herwan Sulaiman, Zuriani Mustaffa, Mohd Mawardi Saari, Hamdan Daniyal, Ismail Musirin, and Mohd Razali Daud. Barnacles mating optimizer: an evolutionary algorithm for solving optimization. In *2018 IEEE international conference on automatic control and intelligent systems (I2CACIS)*, pages 99–104. IEEE, 2018.
- [SNC⁺19] Nicolas Sompairac, Petr V Nazarov, Urszula Czerwinska, Laura Cantini, Anne Biton, Askhat Molkenov, Zhaxybay Zhumadilov, Emmanuel Barillot, Francois Radvanyi, Alexander Gorban, et al. Independent component analysis for unraveling the complexity of cancer omics datasets. *International Journal of molecular sciences*, 20(18):4414, 2019.
- [SREG⁺24] Ahmed M Salih, Zahra Raisi-Estabragh, Ilaria Boscolo Galazzo, Petia Radeva, Steffen E Petersen, Karim Lekadir, and Gloria Menegaz. A perspective on explainable artificial intelligence methods: Shap and lime. *Advanced Intelligent Systems*, page 2400304, 2024.
- [SS19] Debasree Sarkar and Sudipto Saha. Machine-learning techniques for the prediction of protein–protein interactions. *Journal of biosciences*, 44(4):104, 2019.
- [SSD⁺17] Aurélie Strickaert, Manuel Saiselet, Geneviève Dom, Xavier De Deken, Jacques Emile Dumont, Olivier Feron, Pierre Sonveaux, and Carine Maenhaut. Cancer heterogeneity is not compatible with one unique cancer cell metabolic map. *Oncogene*, 36(19):2637–2642, 2017.
- [Sta04] John A Stamatoyannopoulos. The genomics of gene expression. *Genomics*, 84(3):449–457, 2004.
- [STB⁺25] Ralph Steinkamp, George Tsitsiridis, Barbara Brauner, Corinna Montrone, Gisela Fobo, Goar Frishman, Sorin Avram, Tudor I Oprea, and Andreas Ruepp. Corum in 2024: protein complexes as drug targets. *Nucleic Acids Research*, 53(D1):D651–D657, 2025.
- [STRC20] Alok Kumar Shukla, Diwakar Tripathi, B Ramachandra Reddy, and D Chandramohan. A study on metaheuristics approaches

- for gene selection in microarray data: algorithms, applications and open challenges. *Evolutionary intelligence*, 13(3):309–329, 2020.
- [SWGZ21] Zhiao Shi, Bo Wen, Qiang Gao, and Bing Zhang. Feature selection methods for protein biomarker discovery from proteomics or multiomics data. *Molecular & Cellular Proteomics*, 20, 2021.
- [TAGA⁺25] Luciana D Trino Albano, Daniela C Granato, Luiz GS Albano, Fábio MS Patroni, Aline G Santana, Guilherme A Câmara, Davi HS de Camargo, Ana L Mores, Thaís B Brandão, Ana C Prado-Ribeiro, et al. Noninvasive and sensitive biosensor for the detection of oral cancer prognostic biomarkers. *Small*, page e04278, 2025.
- [TAP17] Claire Tonry, John Armstrong, and Stephen R Pennington. Probing the prostate tumour microenvironment i: Impact of glucose deprivation on a cell model of prostate cancer progression. *Oncotarget*, 8(9):14374, 2017.
- [TBJ⁺19] John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, et al. Cosmic: the catalogue of somatic mutations in cancer. *Nucleic acids research*, 47(D1):D941–D947, 2019.
- [TBLA20] Said Tkatek, O Bahti, Y Lmzouari, and J Abouchabaka. Artificial intelligence for improving the optimization of np-hard problems: a review. *International Journal of Advanced Trends Computer Science and Applications*, 9(5), 2020.
- [TBN⁺15] Martine Tetreault, Eric Bareke, Javad Nadaf, Najmeh Alirezaie, and Jacek Majewski. Whole-exome sequencing as a diagnostic tool: current challenges and future opportunities. *Expert review of molecular diagnostics*, 15(6):749–760, 2015.
- [TDZ⁺22] Alexandra Triantafyllou, Nikolaos Dovrolis, Eleni Zografos, Charalampos Theodoropoulos, George C Zografos, Nikolaos V Michalopoulos, and Maria Gazouli. Circulating mirna expression profiling in breast cancer molecular subtypes: Applying

- machine learning analysis in bioinformatics. *Cancer Diagnosis & Prognosis*, 2(6):739, 2022.
- [TM18] Hind Taud and Jean-Francois Mas. Multilayer perceptron (mlp). *Geomatic approaches for modeling land change scenarios*, pages 451–455, 2018.
- [VdBHS⁺19] Koen Van den Berge, Katharina M Hembach, Charlotte Sone-son, Simone Tiberi, Lieven Clement, Michael I Love, Rob Patro, and Mark D Robinson. Rna sequencing data: hitchhiker’s guide to expression analysis. *Annual Review of Biomedical Data Science*, 2(1):139–173, 2019.
- [VF05] Michel Verleysen and Damien François. The curse of dimensionality in data mining and time series prediction. In *International work-conference on artificial neural networks*, pages 758–770. Springer, 2005.
- [VLL⁺22] Douglas Voet, Pei Lin, LRZ Lichtenstein, Saksena G II, M Noble, K Cibulskis, et al. Firehose: An analysis infrastructure. *Firehose Broad GDAC, Broad Inst*, 1, 2022.
- [VLP⁺18] Garima Verma, Maria Laura Luciani, Alessandro Palombo, Linda Metaxa, Giovanna Panzironi, Federica Pediconi, Alessandro Giuliani, Mariano Bizzarri, and Virginia Todde. Microcalcification morphological descriptors and parenchyma fractal dimension hierarchically interact in breast cancer: A diagnostic perspective. *Computers in biology and medicine*, 93:1–6, 2018.
- [VM12] Christine Vogel and Edward M Marcotte. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature reviews genetics*, 13(4):227–232, 2012.
- [VMSC14] Dries Verdegem, Stijn Moens, Peter Stapor, and Peter Carmeliet. Endothelial cell metabolism: parallels and divergences with cancer cell metabolism. *Cancer & metabolism*, 2:1–19, 2014.
- [WA18] Steven W Wingett and Simon Andrews. Fastq screen: A tool for multi-genome mapping and quality control. *F1000Research*, 7, 2018.
- [WBP⁺15] Cuihong Wan, Blake Borgeson, Sadhna Phanse, Fan Tu, Kevin Drew, Greg Clark, Xuejian Xiong, Olga Kagan, Julian Kwan,

- Alexandr Bezginov, et al. Panorama of ancient metazoan macromolecular complexes. *Nature*, 525(7569):339–344, 2015.
- [WBRF10] Britta Weigelt, Frederick L Baehner, and Jorge S Reis-Filho. The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, 220(2):263–280, 2010.
- [WD18] Neal K Williams and Bernhard Dichtl. Co-translational control of protein complex formation: a fundamental pathway of cellular organization? *Biochemical Society Transactions*, 46(1):197–206, 2018.
- [WHH⁺24] Jun Wang, Jinyong Huang, Yunlong Hu, Qianwen Guo, Shasha Zhang, Jinglin Tian, Yanqin Niu, Ling Ji, Yuzhong Xu, Peijun Tang, et al. Terminal modifications independent cell-free rna sequencing enables sensitive early cancer detection and classification. *Nature Communications*, 15(1):156, 2024.
- [WHM⁺15] Hiromi I Wettersten, A Ari Hakimi, Dexter Morin, Cristina Bianchi, Megan E Johnstone, Dallas R Donohoe, Josephine F Trott, Omran Abu Aboud, Steven Stirdivant, Bruce Neri, et al. Grade-dependent metabolic reprogramming in kidney cancer revealed by combined proteomics and metabolomics analysis. *Cancer research*, 75(12):2541–2552, 2015.
- [WK23] Jael Sanyanda Wekesa and Michael Kimwele. A review of multi-omics data integration through deep learning approaches for disease diagnosis, prognosis, and treatment. *Frontiers in genetics*, 14:1199087, 2023.
- [WKB24] Niyaz Ahmad Wani, Ravinder Kumar, and Jatin Bedi. Deep-xplainer: An interpretable deep learning based approach for lung cancer detection using explainable artificial intelligence. *Computer Methods and Programs in Biomedicine*, 243:107879, 2024.
- [WLH⁺22] Kaimin Wei, Tianqi Li, Feiran Huang, Jinpeng Chen, and Zefan He. Cancer classification with data augmentation based on

- generative adversarial networks. *Frontiers of Computer Science*, 16:1–11, 2022.
- [WLL20] Zhouren Wu, Qing Liao, and Bin Liu. A comprehensive review and evaluation of computational methods for identifying protein complexes from protein–protein interaction networks. *Briefings in bioinformatics*, 21(5):1531–1548, 2020.
- [WLZ⁺18] Deling Wang, Jia-Rui Li, Yu-Hang Zhang, Lei Chen, Tao Huang, and Yu-Dong Cai. Identification of differentially expressed genes between original breast cancer and xenograft using machine learning algorithms. *Genes*, 9(3):155, 2018.
- [Wor25a] World Health Organization. Cancer Fact Sheet – Key Facts. <https://www.who.int/news-room/fact-sheets/detail/cancer>, 2025. Accessed 13 May 2025.
- [Wor25b] World Health Organization. Early cancer diagnosis saves lives, cuts treatment costs. <https://www.who.int/news/item/03-02-2017-early-cancer-diagnosis-saves-lives-cuts-treatment-costs>, 2025. Accessed 19 May 2025.
- [WTZQ25] Ying Wang, Yun Tie, Dalong Zhang, and Chenxu Quan. A novel vision neural network for pan-cancer classification by constructing somatic mutation map with feature selection optimization. *IEEE Transactions on Computational Biology and Bioinformatics*, pages 1–9, 2025.
- [XWL⁺21] Xiaoliang Xie, Xulin Wang, Yuebin Liang, Jingya Yang, Yan Wu, Li Li, Xin Sun, Pingping Bing, Binsheng He, Geng Tian, et al. Evaluating cancer-related biomarkers based on pathological images: a systematic review. *Frontiers in Oncology*, 11:763527, 2021.
- [XZD⁺25] Xin Xiong, Le-Wei Zheng, Yu Ding, Yu-Fei Chen, Yu-Wen Cai, Lei-Ping Wang, Liang Huang, Cui-Cui Liu, Zhi-Ming Shao, and Ke-Da Yu. Breast cancer: pathogenesis and treatments. *Signal transduction and targeted therapy*, 10(1):49, 2025.
- [YARM24] Haseeb Younis, Muhammad Azeem, Isabel Ronan, and Rosane Minghim. Manet: A deep learning framework for multi-cancer microarray analysis, classification, and visualization. In *2024*

- International Conference on Frontiers of Information Technology (FIT)*, pages 1–6. IEEE, 2024.
- [YBA19] Haseeb Younis, Muhammad Hamza Bhatti, and Muhammad Azeem. Classification of skin cancer dermoscopy images using transfer learning. In *2019 15th International Conference on Emerging Technologies (ICET)*, pages 1–4, 2019.
- [YBBM24] Haseeb Younis, Horiya Imane Brahmi, Jonathan Byrne, and Rosane Minghim. Unccer: Unified network for cancer classification and efficient representation using microarray data. In *2024 18th International Conference on Open Source Systems and Technologies (ICOSST)*, pages 1–6. IEEE, 2024.
- [YCZ⁺21] Yin Yuan, Wen Cao, Hongbing Zhou, Haixin Qian, and Honggang Wang. H2a. z acetylation by lincznf337-as1 via kat5 implicated in the transcriptional misregulation in cancer signaling pathway in hepatocellular carcinoma. *Cell Death & Disease*, 12(6):609, 2021.
- [YHB⁺23] Long Yang, Wen Huang, Xiaoyu Bai, Haoyu Wang, Xiaolei Wang, Huiyuan Xiao, and Yanlei Li. Androgen dihydrotestosterone promotes bladder cancer cell proliferation and invasion via eppk1-mediated mapk/jup signalling. *Cell Death & Disease*, 14(6):363, 2023.
- [YJ22] Samir S Yadav and Shivajirao M Jadhav. Thermal infrared imaging based breast cancer diagnosis using machine learning techniques. *Multimedia Tools and Applications*, pages 1–19, 2022.
- [YK15] In Seok Yang and Sangwoo Kim. Analysis of whole transcriptome sequencing data: workflow and software. *Genomics & informatics*, 13(4):119, 2015.
- [YLS⁺24] Arseniy E Yuzhalin, Frank J Lowery, Yohei Saito, Xiangliang Yuan, Jun Yao, Yimin Duan, Jingzhen Ding, Sunil Acharya, Chenyu Zhang, Abigail Fajardo, et al. Astrocyte-induced cdk5 expedites breast cancer brain metastasis by suppressing mhc-i expression to evade immune recognition. *Nature Cell Biology*, pages 1–17, 2024.

- [YLZ20] Fei Yuan, Lin Lu, and Quan Zou. Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1866(8):165822, 2020.
- [YM25] Haseeb Younis and Rosane Minghim. Enhancing cancer classification from rna sequencing data using deep learning and explainable ai. *Machine Learning and Knowledge Extraction*, 7(4):114, 2025.
- [YTM22] Haseeb Younis, Paul Trust, and Rosane Minghim. Understanding high dimensional spaces through visual means employing multidimensional projections. *International Journal on Engineering, Science and Technology*, 4(2):170–179, 2022.
- [YWHH12] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, 16(5):284–287, 2012.
- [Zar15] Gökmen Zararsız. Development and application of novel machine learning approaches for rna-seq data classification. *Int. J. Comput. Trends Tech.(IJCT)*, 2017:62–64, 2015.
- [ZBA⁺19] Junjun Zhang, Rosita Bajari, Dusan Andric, Francois Gerthofert, Alexandru Lepsa, Hardeep Nahal-Bose, Lincoln D Stein, and Vincent Ferretti. The international cancer genome consortium data portal. *Nature biotechnology*, 37(4):367–369, 2019.
- [ZCC⁺22] Yiqun Zhang, Fengju Chen, Darshan S Chandrashekar, Sooryanarayana Varambally, and Chad J Creighton. Proteogenomic characterization of 2002 human cancers reveals pan-cancer molecular subtypes and associated pathways. *Nature Communications*, 13(1):2669, 2022.
- [ZFLB⁺14] Shanrong Zhao, Wai-Ping Fung-Leung, Anton Bittner, Karen Ngo, and Xuejun Liu. Comparison of rna-seq and microarray in transcriptome profiling of activated t cells. *PloS one*, 9(1):e78644, 2014.
- [ZGT⁺22] Wengang Zhang, Xin Gu, Libin Tang, Yueping Yin, Dongsheng Liu, and Yanmei Zhang. Application of machine learning, deep

- learning and optimization algorithms in geoengineering and geoscience: Comprehensive review and future challenge. *Gondwana Research*, 109:1–17, 2022.
- [Zho21] Zhi-Hua Zhou. *Machine learning*. Springer nature, 2021.
- [ZKR⁺21] Fan Zheng, Marcus R Kelly, Dana J Ramms, Marissa L Heintschel, Kai Tao, Beril Tutuncuoglu, John J Lee, Keiichiro Ono, Helene Foussard, Michael Chen, et al. Interpretation of cancer mutations using a multiscale map of protein systems. *Science*, 374(6563):eabf3067, 2021.
- [ZLX⁺16] Qiaosheng Zhang, Jie Li, Haozhe Xie, Hanqing Xue, and Yadong Wang. A network-based pathway-expanding approach for pathway analysis. *BMC bioinformatics*, 17(Suppl 17):536, 2016.
- [ZRM⁺23] Mara Zilocchi, Matineh Rahmatbakhsh, Mohamed Taha Moutaoufik, Kirsten Broderick, Alla Gagarinova, Matthew Jesulat, Sadhna Phanse, Hiroyuki Aoki, Khaled A Aly, and Mohan Babu. Co-fractionation–mass spectrometry to characterize native mitochondrial protein assemblies in mammalian neurons and brain. *Nature Protocols*, 18(12):3918–3973, 2023.
- [ZWWC11] Hanhong Zhu, Yi Wang, Kesheng Wang, and Yun Chen. Particle swarm optimization (pso) for the constrained portfolio optimization problem. *Expert Systems with Applications*, 38(8):10161–10169, 2011.
- [ZZH⁺15] Jin-Ying Zhang, Fan Zhang, Chao-Qun Hong, Armando E Giuliano, Xiao-Jiang Cui, Guang-Ji Zhou, Guo-Jun Zhang, and Yu-Kun Cui. Critical protein gapdh and its regulatory mechanisms in cancer cells. *Cancer biology & medicine*, 12(1):10, 2015.
- [ZZQ⁺24] Han Zhang, Taotao Zheng, Chuan Qin, Xinyue Zhang, Han Lin, Xiaoping Huang, Qiang Liu, Shichuan Chang, Li Zhang, Jing Guo, et al. Cct6a promotes cell proliferation in colon cancer by targeting birc5 associated with p53 status. *Cancer Gene Therapy*, pages 1–13, 2024.

Appendix A

Glossary of Terms

Adaptive Boosting A machine learning ensemble technique that combines multiple weak learners (typically decision trees) to create a strong classifier by focusing on misclassified instances in each iteration. xiii, 107

Affinity Purification–Mass Spectrometry A technique used to study protein–protein interactions by isolating protein complexes using affinity tags and identifying the components via mass spectrometry. xiii, 31

AI Artificial Intelligence – machines or programs that simulate human thinking to solve problems or make decisions. xiii

Analysis of Variance A statistical method used to compare the means of three or more groups to determine if at least one group mean is significantly different from the others. xiii, 100

ANN Artificial Neural Network – a computer model inspired by the brain, used to find patterns in data. xiii, 17

Area Under the Curve A performance metric for classification models, representing the area under the receiver operating characteristic (ROC) curve; higher AUC indicates better model discrimination. xiii, 22

Bayesian Information Criterion A model selection criterion based on the likelihood function and model complexity, penalising models with more parameters to avoid overfitting. xiii, 114

Benjamini-Hochberg A statistical procedure used to control the false discovery rate (FDR) when performing multiple hypothesis tests, reducing the chance of type I errors. xiii, 75

Breast Cancer Gene Refers to the BRCA1 and BRCA2 genes, which produce proteins that help repair damaged DNA; mutations in these genes significantly increase the risk of breast, ovarian, and other cancers. xiii

Catalogue Of Somatic Mutations In Cancer A comprehensive database that curates information on somatic mutations implicated in human cancer, widely used for cancer genomics and mutation analysis. xiii, 33

CNN Convolutional Neural Network – a deep learning model especially good at analysing images and structured data. ix, xiii

Co-Fractionation An experimental technique used to study protein-protein interactions by separating cellular components into fractions and analysing co-elution profiles, often combined with mass spectrometry. xiii, 96

Co-Fractionation–Mass Spectrometry A method for studying protein complexes by separating them through biochemical fractionation, followed by mass spectrometry analysis to infer protein associations based on co-elution profiles. xiii, 3

complementary DNA A DNA molecule synthesised from a messenger RNA (mRNA) template using the enzyme reverse transcriptase; commonly used in gene expression studies and cloning. xiii, 25

Comprehensive Resource of Mammalian protein complexes A curated database of experimentally verified protein complexes in mammals, particularly human, used to support studies of protein interactions and complex organization. xiii, 123

Corrected Akaike Information Criterion An adjusted version of the Akaike Information Criterion (AIC) which accounts for small sample sizes when comparing the goodness of fit of statistical models. xiii, 114

CuMiDa Curated Microarray Database – a comprehensive and standardised collection of preprocessed microarray gene expression datasets for benchmarking and analysis. xiii, 16

Decision Tree A flowchart-like tree structure used in machine learning where

internal nodes represent tests on features, branches represent outcomes, and leaf nodes represent class labels or regression values. xiii, 29

Deep Neural Network A type of neural network architecture with multiple hidden layers that allows modelling of complex non-linear relationships in data. xiii

Deoxynucleotide Triphosphates The building blocks of DNA, consisting of a nitrogenous base, deoxyribose sugar, and three phosphate groups; essential for DNA replication and PCR. xiii, 25

DNA Deoxyribonucleic Acid – the molecule that contains all genetic instructions in living organisms. xiii, 11

DNA Methylation An epigenetic mechanism involving the addition of a methyl group to DNA, typically acting to repress gene transcription and playing a key role in development, ageing, and disease. xiii, 13

Driver Gene A gene that contains mutations contributing to the development or progression of cancer by conferring growth advantage to cells. xiii

Epidermal Growth Factor Receptor A transmembrane protein that, when mutated or overexpressed, promotes cell proliferation and survival; frequently implicated in various cancers and targeted by specific therapies. xiii

ER Estrogen Receptor – a protein found inside cells that binds to estrogen; commonly tested in breast cancer to guide treatment decisions. xiii, 11

Extremely Randomised Trees An ensemble machine learning method similar to Random Forest, but introduces more randomness by choosing cut-points for splits at random, which can lead to faster training and potentially better generalisation. xiii, 99

GAN Generative Adversarial Network – a deep learning model that generates new data similar to what it learned. xiii, 27

Gene Expression The process by which information from a gene is used to synthesise functional gene products like proteins or RNA, often measured to study cellular function and disease states. xiv, 13

Gene Expression Omnibus A public functional genomics data repository supporting MIAME-compliant data submissions.. xiv

- Gene Expression Profiling** A technique used to measure the activity (expression levels) of thousands of genes at once to understand cellular functions, identify disease subtypes, or predict treatment responses. xiv, 24
- Gene Ontology** A framework for the standardised representation of gene and gene product attributes across species and databases, encompassing biological processes, cellular components, and molecular functions. xiv, 41
- Genomic Data Commons** A data platform developed by the National Institute of Health (NIH) to provide access to cancer genomic datasets, including TCGA. xiv, 38
- Genomic Identification of Significant Targets in Cancer** A computational method used to identify genes that are significantly amplified or deleted across a set of cancer genomes, helping to pinpoint potential driver genes. xiv, 38
- GLOBOCAN** Global Cancer Observatory – an initiative by the International Agency for Research on Cancer (IARC) providing global cancer statistics and estimates of incidence and mortality. xiv, 9
- Gradient Boosting Classifier** An ensemble machine learning technique that builds a strong predictive model by sequentially adding weak learners, typically decision trees, while minimising the error through gradient descent. xiii, 108
- HER2** Human Epidermal Growth Factor Receptor 2 – a protein often tested in breast cancer as a growth factor. xiv, 11
- Human Genome Project** An international scientific initiative that successfully mapped and sequenced the entire human genome, providing a foundational reference for biomedical research. xiv, 9
- Human Oral Keratinocyte** A type of epithelial cell found in the oral mucosa, often used as a normal control in studies of oral diseases and cancers. xiv, 32
- Human Squamous Carcinoma-3** A highly invasive oral squamous cell carcinoma cell line used in cancer biology research, especially for studying mechanisms of invasion and metastasis. xiv, 32
- IHC** Immunohistochemistry – a lab method that uses antibodies to detect proteins in tissue samples. xiv

- k-means Clustering** An unsupervised learning algorithm used to partition data into distinct groups based on feature similarity. xiv, 49
- k-Nearest Neighbour** A simple, instance-based machine learning algorithm that classifies new data points based on the majority class among their k closest neighbours in the feature space. xiv, 29
- Kernel Density Estimate** A non-parametric way to estimate the probability density function of a random variable by averaging over kernel functions centred at observed data points. xiv, 74
- Kyoto Encyclopedia of Genes and Genomes** A comprehensive database resource that integrates genomic, chemical, and systemic functional information, widely used for understanding high-level functions and utilities of biological systems through pathway mapping. xiv, 41
- Linear Discriminant Analysis** A statistical and machine learning technique used for dimensionality reduction and classification by finding the linear combinations of features that best separate two or more classes. xiv, 111
- Linear Regression** A statistical method for modelling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. xiv, 28
- Liquid Chromatography–Tandem Mass Spectrometry** An analytical technique that combines the physical separation capabilities of liquid chromatography with the mass analysis capabilities of tandem mass spectrometry, widely used for identifying and quantifying complex mixtures of biomolecules. xiv, 96
- Local Interpretable Model-agnostic Explanations** An explainable AI technique that provides simple and interpretable explanations for predictions of complex models. ix, xiv
- Long Non-Coding RNA** A type of RNA transcript longer than 200 nucleotides that does not encode proteins but plays regulatory roles in gene expression, chromatin modification, and various cellular processes. xiv, 25
- Luminal A** A molecular subtype of breast cancer characterised by high expression of hormone receptors (ER+ and/or PR+), low proliferation (low Ki-67), and typically HER2-negative; associated with a favourable prognosis. xiv

Luminal B A more aggressive molecular subtype of breast cancer that is also hormone receptor-positive but shows higher proliferation (high Ki-67) and may be HER2-positive; generally has a less favourable prognosis than Luminal A. xiv

Mean Squared Error A common loss function used in regression tasks that measures the average of the squares of the differences between predicted and actual values. xv, 42

Molecular Signatures Database A comprehensive resource of annotated gene sets used for gene set enrichment analysis (GSEA), helping researchers interpret gene expression data by identifying biological pathways and processes. xv, 38

mRNA Messenger Ribonucleic Acid – a molecule that carries genetic instructions from DNA to make proteins. xiv, 20

MS Mass Spectrometry – a technique used to identify and quantify molecules by measuring their mass. ix, xv

Multilayer Perceptron A class of feedforward artificial neural networks consisting of at least three layers (input, hidden, and output) that can model complex relationships through non-linear transformations. xiv, 29

Multiple Instance Learning A type of weakly supervised learning where labels are provided for groups of instances rather than individual ones.. xiv, 22

Multiple Kernel Learning A machine learning technique that combines multiple kernels to improve model flexibility and performance, often used in classification tasks involving heterogeneous data. xiv, 22

NGS Next Generation Sequencing – advanced methods for quickly reading large amounts of DNA or RNA. xv

Pancreatic Adenocarcinoma A type of cancer that forms in the pancreas' exocrine cells, representing the most common and aggressive form of pancreatic cancer; abbreviated as PAAD in genomic studies. xv, 29

PCA Principal Component Analysis – a mathematical technique to simplify and visualise large datasets. xv, 18

PR Progesterone Receptor – a protein in cells that binds to the hormone progesterone; often assessed in breast cancer alongside ER status. xv, 11

Predicting Interactomes from Co-Elution A computational approach for inferring protein–protein interaction networks by analysing co-elution profiles from biochemical fractionation experiments. xv, 95

Protein–Protein Interaction The physical contact between two or more proteins, often essential for cellular processes such as signalling, structural support, and enzymatic activity; commonly studied to understand molecular mechanisms and disease pathways. xv, 31

Protein–Protein Interaction Network A graphical representation of the physical or functional interactions between proteins within a cell, used to understand cellular processes, signalling pathways, and disease mechanisms. x, xv

PSA Prostate Specific Antigen – a protein measured in blood to help detect prostate cancer. xv, 10

Quadratic Discriminant Analysis A classification algorithm that models each class with its own covariance matrix, allowing for quadratic decision boundaries between classes; useful when class distributions differ in shape. xv, 112

Random Forest An ensemble machine learning method that constructs multiple decision trees and combines their predictions to enhance accuracy and robustness. xv, 18

Rectified Linear Unit A widely used activation function in neural networks, defined as the positive part of its input, which introduces non-linearity and helps mitigate the vanishing gradient problem. xv, 42

Recursive Feature Elimination A feature selection technique that iteratively removes the least important features to improve model accuracy. xv

RNA Ribonucleic Acid – a molecule involved in coding, decoding, and regulating genes. xv, 12

RNA-seq RNA Sequencing – a technique to analyse the quantity and sequences of RNA in a sample. ix, xv

SHAP SHapley Additive exPlanations – a method to explain feature importance in machine learning models. xv, 18

- SNV** Single Nucleotide Variant – a mutation involving a change in a single DNA base. xv, 33
- Squamous Cell Carcinoma-9** A human tongue squamous cell carcinoma cell line commonly used in oral cancer research to study tumour behaviour and response to treatments. xv, 32
- Subtyping Agglomerated Mutations By Annotation Relations** An algorithm that desparsifies somatic mutation matrices by aggregating gene-level mutations into functional gene sets, improving downstream clustering. xv, 24
- Support Vector Machine** A classical machine learning algorithm used to classify data by finding the optimal boundary separating different classes. xv, 18
- Synthetic Minority Over-sampling Technique** A method to balance datasets by creating artificial samples of underrepresented classes to enhance predictive accuracy. ix, xv
- TCGA** The Cancer Genome Atlas – a large-scale project mapping genetic changes in different cancers. xv, 9
- Tumor Protein 53** A tumour suppressor gene that encodes the p53 protein, which regulates the cell cycle and prevents genome mutation; commonly mutated in many cancers. xv, 9
- UMAP** Uniform Manifold Approximation and Projection – a tool for visualising complex datasets in 2D or 3D. xv, 18
- WES** Whole Exome Sequencing – a method to sequence only the protein-coding parts of the genome. xvi
- World Health Organization** A specialised agency of the United Nations responsible for international public health, setting health standards, and coordinating global health responses. xvi, 9
- XAI** Explainable Artificial Intelligence – AI systems designed to be understandable and transparent in their decisions. x, xvi