Gene Set Anomaly Score: A Genomic Data and Knowledge Driven Approach for Analysing Anomalous Gene Expression in Cancer Patients

by Md Sarwar Kamal

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

UNDER THE SUPERVISION OF

OF

Principal Supervisor
Professor Daniel Catchpoole

Co-Supervisor
Professor Barry Drake

University of Technology Sydney

Faculty of Engineering and Information Technology

June, 2023

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Md Sarwar Kamal, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note: Signature removed prior to publication.

Date: 14-June-2023

Dedication

To my family and supervisors.

Abstract

Genomics research often uses Gene Set Enrichment Analysis (GSEA) to rank genes that correlate with the presence of phenotypical traits and to interpret how variations in gene expression influence those traits. GSEA provides an explanation of found genes through their associations with gene sets. As gene sets represent different biological mechanisms, they can indicate overall shifts in expression values in relation to their biology.

This thesis investigated the relationships between patients and diseases by using gene sets, integrating gene expression data and gene set ontologies to develop a new analytics method called gene expression anomaly scores. These scores measure the deviation of expression values from expected values.

This thesis investigated the representation of patient biology as two-dimensional graphs derived from anomaly scores. There are thousands of patient gene sets relating to a given disease, such as cancer. To identify strongly associated gene sets, this thesis apply principal component analysis (PCA) and maximum relevance and minimum redundancy (MRMR), selecting the two most prominent dimensions. Thus, PCA and MRMR were each used to embed patients into a 2-dimensional anomaly score space. Embedding patients using anomaly scores revealed relationships between patients and patient biology through clustering and feature selection in this space. Moreover, this thesis applied explainable AI (XAI) to understand patients' biology (gene sets) responsible for prediction by predictive models or AI algorithms. This thesis applied Local Interpretation-Driven Abstract Bayesian Network (LINDA-BN) which extracts patients biology and shows the relationships between biologies responsible for a prediction.

The proposed method was used to analyse gene expression data of cancer patients from four different data sets. More specifically, anomaly scores followed by PCA or MRMR showed groups of cancer patients in scatter plots. These groups appeared to be related to treatment outcomes. In addition, MRMR was able to identify potential gene sets with meaningful biological implications. Comparatively, when raw and state-of-the-art gene expression scores were analyzed, only genes patterns were apparent. The outcomes of the distributions showed that the distribution of anomaly scores varied significantly between patients who relapsed and those who did not. In addition, the k-means algorithm revealed that the anomaly score performs better clustering than state-of-the-art methodologies.

Furthermore, anomaly scores uncovered novel cancer biology in contrast to gene set enrichment analysis (GSEA) and state-of-the-art approaches. Finally, the outcomes of instance-based LINDA-BN showed an interpretable and explicable method for predicting medical condition a cancer patient.

Contribution

This thesis has developed a combined knowledge-and data-driven approach to understand gene expression, which determines an anomaly score that captures the combined shift in gene expression for all genes in a gene set. This thesis has demonstrated that constructing a data model based on gene set “anomaly scores” can effectively capture a combined shift in gene expression for all genes in a gene set. Anomaly scores serve as a counterpart to gene expression values, enabling new classifications of disease states based on gene expression data. This thesis has shown that this principle of combining knowledge-driven and data-driven analysis can lead to insights into genetic responses to disease. In summary, the proposed approach introduces a new analytical tool and new directions for understanding the genetic causes of diseases.

Acknowledgments

Everything is owed to God, who is all-powerful and all-merciful. I express my gratitude to Him for bestowing upon me the ability to complete this work.

There are no words to express my gratitude to my principal supervisor **Professor Barry Drake** and co-supervisor **Professor Daniel Catchpoole** who have tirelessly supported, guided and encouraged me throughout my thesis. I was better equipped to finish this thesis because to their professional guidance, psychological support and insightful study ideas.

I would also like to thank **A/Professor Wei Liu**, the chair of my candidature review committee, and **A/Professor Dominik Beck**, the expert on the committee, for their advice and comments during the evaluation process.

Throughout my PhD journey, I would want to express my gratitude to my partner for her constant attention and tremendous efforts to support me.

Table of Contents

Nomenclature

Abbreviations

ALL	Acute Lymphoblastic Leukemia
GSEA	Gene Set Enrichment Analysis
IBL	Instance Based Learning
KNN	K-Nearest Neighbor
LINDA-BN	Local Interpretation-Driven Abstract Bayesian Network
LIME	Local Interpretable Model-agnostic Explanations
MSigDB	Molecular Signatures Database
MRMR	Minimum Redundancy Maximum Relevance
PCA	Principle Component Analysis
t-SNE	t-Stochastic Neighbor Embedding
RF	Random Forest
ROC	Receiver Operating Characteristic
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Oversampling Technique
SVC	Support Vector Classifier
XAI	Explainable AI

Symbols		**Chapter**
p	Patient	Ch3
$G(p)$	Set of genes	Ch3

$x(p, g)$	Gene expression values	Ch3
$z(p, g)$	z-score values	Ch3
$\mu(g)$	Mean of gene expression value	Ch3
$\sigma(g)$	Variance of gene expression value	Ch3
s	Gene sets	Ch3
$a(p, s)$	Anomaly scores	Ch3
$b(p, s)$	Average anomaly scores	Ch3
A	Optimal anomaly scores	Ch3
$I(s, c)$	Mutual information	Ch3
c	Target class	Ch3
D	Maximum relevance	Ch3
R	Minimum redundancy	Ch3
$f(d, \alpha, \beta)$	Histogram distribution function	Ch3
α	Shape parameter	Ch3
β	Inverse scale parameter	Ch3
S^-	Minority class	Ch4
S^+	Majority class	Ch4
h^{th}	Maximum degree of imbalance	Ch4
$exp(f)$	Explainable features by LIME	Ch4
$\varphi_f(d)$	Shapely Value	Ch4
$p(G\|d)$	Structure learning	Ch4
$P(\phi\|G, d)$	Parameter learning	Ch4
$SCore(G, d)$	Bayesian information criterion	Ch4

List of Figures

List of Tables

Chapter 1

Introduction

"Genes that underlie the capacity to receive, use and transmit information are the evolving properties". -Peter R. Grant

This chapter provides an overview of the problem to solve, the background for conducting research, its objectives, research questions, key contributions and organisation of this thesis.

1.1 Background on gene set enrichment analysis

A genome is the collection of all genes of an organism, and genomics is the study of genomes. The more understandable the genome of an organism and how it affects the organism, the easier it will be to learn about health and make informed decisions about health and disease. Some diseases and syndromes are associated with mutations in a gene or group of genes during a human life cycle. Studies of genomic data enable genetic information associated with human health and disease to be understood [1].

Gene expression is part of a process by which a gene is used to synthesize a protein. Subsequently, the protein controls cellular functions in humans. The synthesis of a gene begins with transfer RNA (tRNA), which is an intermediate step in protein synthesis, then tRNA used to create messenger RNA (mRNA), and finally, a protein [2]. The amount of gene expression (mRNA) in a given cell controls what that cell can do [3]. Understanding gene expression allows us to know the behaviour of genomes and provide insights into disease-patient relationships, how organisms behave, adapt, evolve, and reproduce a new organism. For example, gene expression has been used to distinguish acute myoblastic leukaemia (AML) and acute lymphoblastic leukaemia (ALL), which are alternative analysis to fully characterise them using conventional molecular and cellular analyses [4].

Gene sets are groups of genes that represent the biological functionalities of patients and their diseases [5]. Biological functionalities may include cell cycle, cell locations, DNA replication (a process that duplicates a molecule of DNA), cell proliferation (a process that divides cells), and pathways (a chain of genes, e.g. KEGG (Kyoto Encyclopedia of Genes and Genomes) [6, 7]. Gene expression and gene set analyses allow the investigation of genetic variation and cancer progression in human tumour genomes. One popular approach is gene set enrichment analysis (GSEA). A gene is considered enriched when the level of expression of that gene correlates with a chosen phenotype trait.

GSEA is a method to understand the functional relationships between genes and a chosen phenotypical features in genomic data [8]. Subramanian *et al.* [9] proposed a gene set enrichment score, which is a measure of overexpressed genes that are either at the top or bottom of a ranked list of genes in a gene set. An enriched gene set indicates that the number of its genes at the other end of the ranking list are similar to the genes from the gene expression values. GSEA measures enrichment scores (ES) using the Kolmogorov-Smirnov (K-S) statistics.

GSEA works on a selection of genes that are ranked by a classifier or model. The functionality represented within the gene selection is predicted by measuring the ES which quantifies cumulated fold changes between matched genes, where a fold change is the ratio between an initial and a final gene expression value [10]. Genes correlating with the chosen phenotypical features tend to the top or the bottom of the ranking order within a gene set. Several studies and reviews explore GSEA over different foci, for example, phenotype similarities [11–13], biological functions [14, 15], clinical outcomes [16–20], knowledge-based gene expression analysis [21], and gene set discovery [22]. Tian *et al.* [23] proposed a gene set analysis using a t-test or Wilcoxon rank-sum test statistics to create an enrichment score that measures overexpressed genes. Kim and Volsky [24] presented PAGE (Parametric Analysis of Gene Set Enrichment), which generates an enrichment score using the z-score. The z-score takes into account fold changes, which is the ratio between a patient's initial and final gene expression values. PAGE achieves high sensitivity but has low specificity [25]. Irizarry *et al.* [26] proposed a t-test statistic for the gene set score to determine the degree of association between each gene and phenotype information. However, they ignored gene-gene correlation [27]. Tamayo *et al.* [27] proposed a method that combines both parametric and GSEA in terms of gene-gene correlation. Moreover, there is extensive literature on methods of gene set enrichment [28–43], each emphasizing different mathematical formulations and computational approaches.

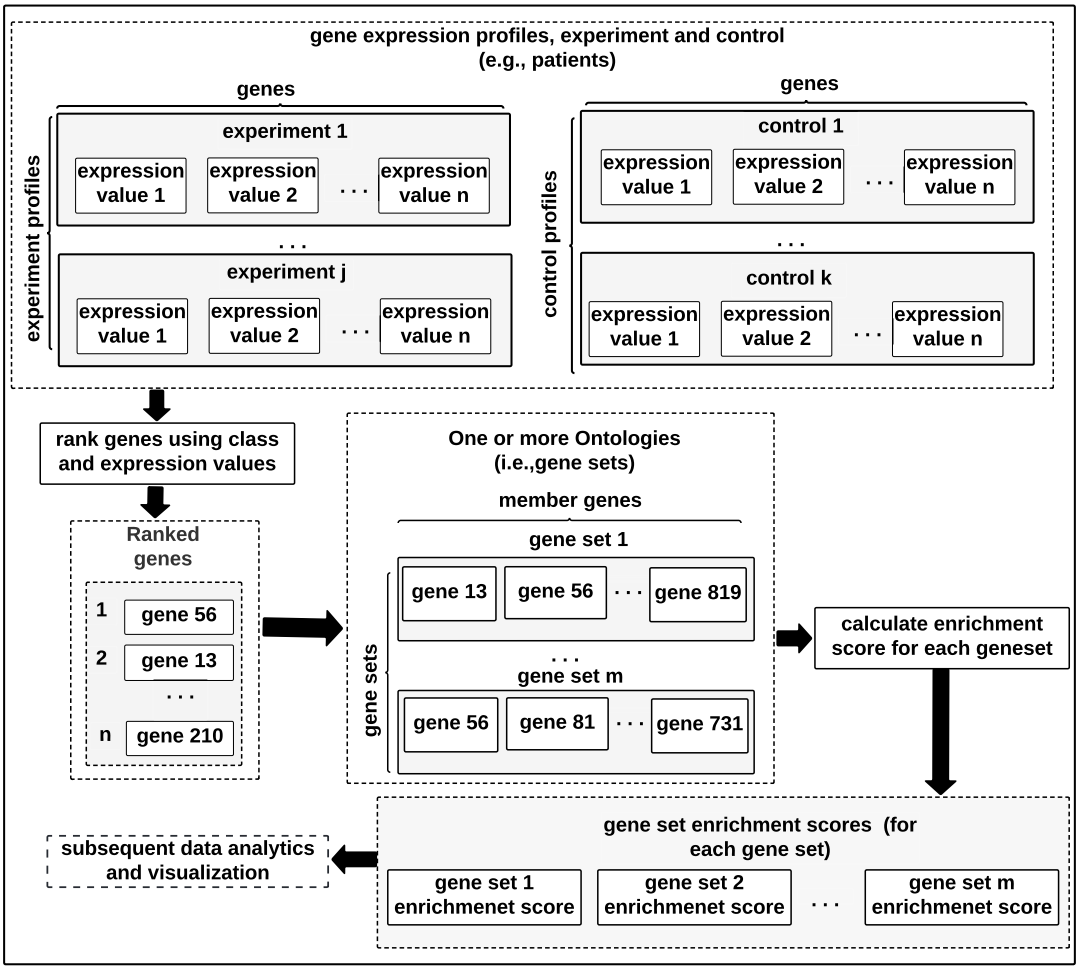


Figure 1.1: Schematic diagram showing the application of GSEA.

Figure 1.1 shows the processing involved to perform GSEA. The process starts with gene expression profiles giving an expression value for each gene, for each profile. An expression profile is typically associated with a patient. GSEA can work with other quantified data, as can our proposed method, but here we concentrate on gene expression data. In GSEA, the profiles are grouped into two classes (experiment and control) which is important when using the method. In addition to gene expression profiles, GSEA uses gene sets. A small number of gene sets are selected by the researcher based on hypotheses that the selected gene sets are relevant to explaining the expression values of the two classes.

The first processing step of GSEA uses the gene expression profiles to rank the genes. There are multiple methods for doing this and the selection of a ranking method is largely from common practice rather than theoretical considerations. The idea is to use a method so that high-ranked genes have higher expression values in the experiment class and low-ranked genes have higher expression values in the control class.

The next processing step produces an enrichment score for each gene set. An enrichment score measures how well the ranks of the genes of the gene set cluster together (particularly at the start and end of the ranking). An enrichment score is calculated from a running sum by processing each gene in its rank order. If a gene is in the gene set, the running sum is increased, otherwise it is decreased. The enrichment score is the maximum/minimum of the running sum. Each enrichment score can also have an associated probability of being produced by chance. GSEA permits further analysis of the selected gene sets by plotting the running sum, in so-called enrichment plots, and relating this to each of the two classes.

Gene sets characterize the biological functionalities of human diseases. Thus, gene sets are hypothesized to be valuable knowledge to understand the behaviour of diseases such as leukaemia and blood cancer, breast cancer, and colon cancer. If a relationship can be established between patients, their disease and gene sets, it would be possible to understand which biological functionalities are responsible for a particular patient in relation to a particular disease. In particular, if it is possible to investigate the relationship with disease abnormalities in patients, this would provide insight into understanding the biology behind the diseases. The relationship would also be valuable for investigating and designing treatment planning and patient prediction for a particular disease.

In this research, the relationships between genomics and diseases specifically to patient biology as defined by gene sets are investigated by integrating gene expression data and gene set knowledge. Specifically, the relationships between patients and diseases are examined in relation to gene set anomalies. Also, this thesis investigates how anomaly distributions varies between cancer patient groups related to their gene sets. In addition, this thesis investigates how the anomalies may improve predictions about a patient. Finally, this thesis shows how prediction results may be analysed that makes sense to humans, not just computers, specifically, techniques from explainable artificial intelligence are explored.

1.2 Problem statement

Genomic data analysis is a topic of importance in understanding cancer and other diseases. One challenge is to identify anomalies in gene expression data which relate to tumour behaviour. Detecting anomalous gene expression leads to insights into the cellular mechanisms underlying disease by clarifying the functional impact of anomalous genes. Gene Set Enrichment Analysis (GSEA) uses gene sets that group genes according to their functional interrelationships. The functionality represented with gene selection is predicted by measuring an enrichment score that quantifies the cumulative fold changes between matched genes, where a fold change is a ratio between an initial and a final gene expression value. GSEA ranks genes in gene sets according to their correlation with the presence of a phenotypic trait and thus can be used to interpret the influence of variation in gene expression on traits. Genes that correlate with the selected phenotypic traits are usually at the top or bottom of the rank order within a gene set. GSEA considers only a small number of genes that are top or bottom of the rank and does not consider variations in gene expression values. However, this thesis sees that gene sets also embody knowledge about how a group of genes is related to each other, so the group may be used to understand more deeply the effects of anomalous gene expression.

1.3 Research questions

This thesis poses the following hypothesis

That aggregation of gene expression values into their respective gene sets will provide opportunities for building a knowledge-based classification.

The hypothesis is divided into the following four research questions.

- **Research question 1**

  What technique(s) can assess the variation of gene expression in each gene set for each patient and build a model using this assessment?

  To answer this question, it is necessary to find an appropriate approach to gene expression data and gene set analyses. Gene set anomalies could be an important consideration to find the characteristics of cancers. Statistical and machine learning approaches could facilitate the study of gene set anomaly analyses.

- **Research question 2**

  What approach can visually map a patient, with respect to the biology of the patient, to determine the response to cancer treatment?

  To answer this question, it is necessary to investigate the relationships between patients and the patients' biology as patient embedding or mapping with respect to anomaly scores.

- **Research question 3**

  How do the distributions of gene set anomaly scores vary across different groups of patients?

  To answer this question, it is necessary to understand how anomaly scores may change, conditioned on whether a cancer patient relapsed or not.

- **Research question 4**

  How can an explainable and interpretable method predict a patient status (healthy or cancerous or relapse or non-relapse) built on an individual instance in relation to anomaly scores?

To answer this question, it is necessary to find an approach that considers a different set of knowledge from the training data to reach a decision for each different test instance. Moreover, it is necessary to find a method that explains the process of predictive methods and makes it interpretable in a trustworthy way. To evaluate the prediction results, it is important to explain which genes or gene sets are important for cancer prediction to understand the whole process of prediction. Appropriate explainable methods increase confidence in the prediction results.

Figure 1.2 shows relationships between hypothesis and research questions.

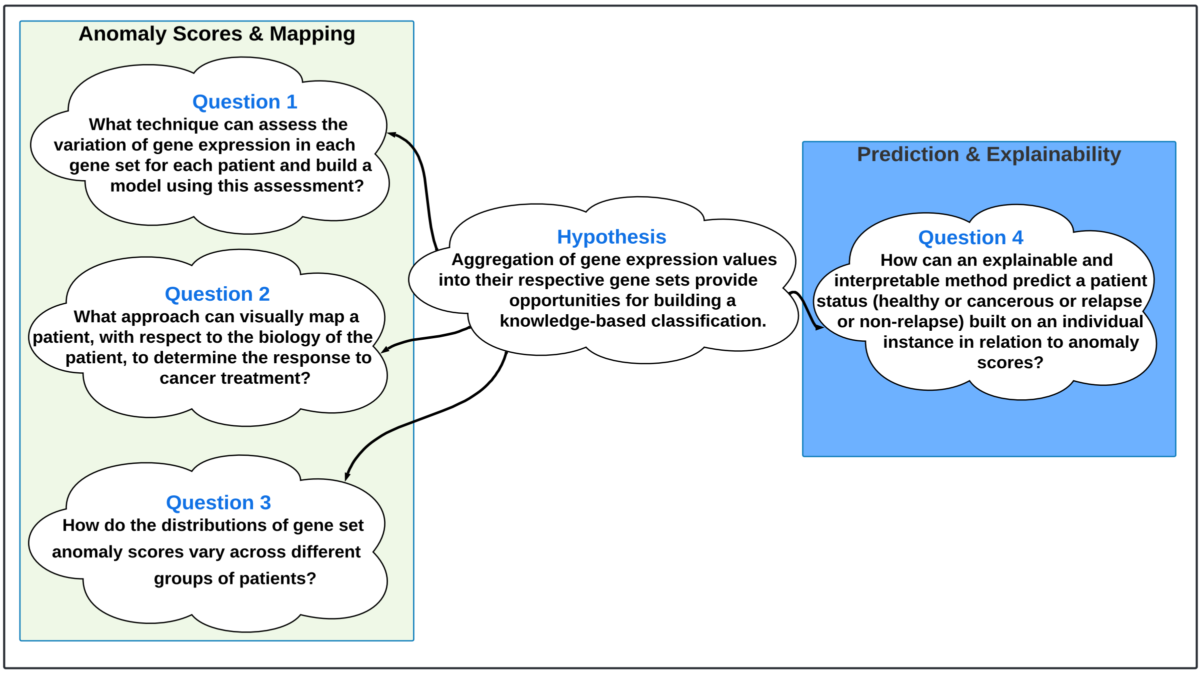

Anomaly Scores & Mapping
Question 1
What technique can assess the variation of gene expression in each gene set for each patient and build a model using this assessment?
Question 2
What approach can visually map a patient, with respect to the biology of the patient, to determine the response to cancer treatment?
Question 3
How do the distributions of gene set anomaly scores vary across different groups of patients?
Hypothesis
Aggregation of gene expression values into their respective gene sets provide opportunities for building a knowledge-based classification.
Prediction & Explainability
Question 4
How can an explainable and interpretable method predict a patient status (healthy or cancerous or relapse or non-relapse) built on an individual instance in relation to anomaly scores?

Figure 1.2: Research questions.

The four research questions presented above are designed to address various aspects of the main hypothesis. These questions break down the overarching hypothesis into smaller, more manageable components, allowing for a targeted investigation of each aspect. By addressing these specific questions, this thesis aims to gather the necessary information to determine whether aggregating gene expression values into gene sets will be valuable for understanding the underlying biology responsible for cancer treatment and classification in patients.

For example, research question 1 will investigate development of a method for combining gene expression values into a gene set. Research question 2 will investigate the mapping of cancer patients with respect to the variation of gene expression values. In this case, patients will be grouped together to identify their respective cohorts. Research question 3 will examine how anomaly score distributions may vary across multiple patient groups, such as high-risk, medium-risk, and low-risk patient groups. The final question will ensure the generation of interpretable and explainable prediction analyses concerning anomaly scores. In the figure 1.2, the arrows demonstrate the subdivision of the main hypothesis into four distinct parts. Each arrow points to a specific research question that addresses a particular aspect of the hypothesis.

1.4 Thesis organization

The aim of this thesis is to advance our understanding of the relationship between patient biology (gene sets) and disease with respect to anomaly scores. The thesis consists of five chapters. The structure of the thesis and the flow of the chapters are shown in Figure 1.3.

Chapter 1 provides the background to the research, defines the research questions and problem statement.

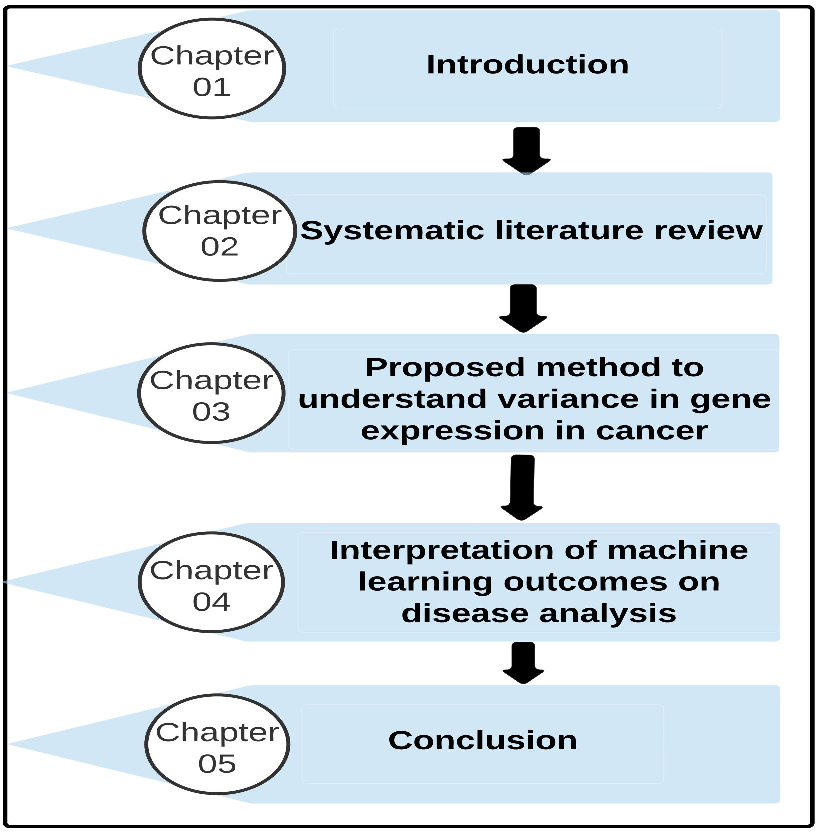

Chapter 01
Introduction
Chapter 02
Systematic literature review
Chapter 03
Proposed method to understand variance in gene expression in cancer
Chapter 04
Interpretation of machine learning outcomes on disease analysis
Chapter 05
Conclusion

Figure 1.3: Organization of thesis chapters.

Chapter 2 presents a literature review. This chapter provides a brief overview of the main aspects of the thesis. It presents general information about gene set enrichment analysis. It then discusses the modelling approaches for mapping, dimensionality reduction, prediction method and explainable AI.

Chapter 3 answers research questions one, two and three and reports on the development of an anomaly score for gene set analysis. In this chapter, the processes to generate the anomaly score are described and dimensionality reduction is explained. Then the gene set is mapped with the patients and the validation process of the patient mapping is described.

Chapter 4 presents methods for generating cancer biology and instance-based interpretable machine learning for predicting medical condition of a cancer patient. This chapter also includes outcomes of new cancer biology. Finally, this chapter shows predictive outcomes of instance-based explainable AI.

Finally, Chapter 5 presents a summary of the thesis and its findings with suggestions for how the techniques presented in this thesis can be developed and applied further.

1.5 Key Contributions

The contributions of this thesis are as follows:

Contribution 1

statistical and machine learning approach for generating anomaly scores by integrating gene sets and gene expression data. A patient embedding approach to identify patient cohorts in terms of highly associated patient biology or gene sets has been developed.

This thesis proposes that careful aggregating gene expression values into gene sets can lead to valuable insights from data analysis of gene expression profiles. Specifically, it posits that understanding variations in gene expression values within gene sets can improve the visibility and detectability of patterns across different expression profiles, rather than using expression values directly.

To achieve this, the thesis develops an approach by considering both a data and knowledge-driven approach that starts with the integration of gene expression profiles and gene sets. Key consideration is that the proposed is designed to work with large numbers of gene sets, such as those found in one or more ontologies. The method generates an “anomaly score” for each gene set within each gene expression profile, measuring variations in gene expression values.

Gene set anomaly scores provide a new way to represent each profile, making it possible to use different data processing methods from statistical analytics and machine learning. This approach combines the benefits of the knowledge found within gene sets while still allowing for profile-based data analysis. Additionally, it supports methods that don’t require splitting profiles into only two classes (i.e., no classes or more than two classes), providing more flexibility in classification for users who may be new to the subject.

In conclusion, this thesis introduces the concept of calculating anomaly scores for each gene set is presented, opening up the possibility to use various data processing methods from statistics and machine learning. This novel approach takes advantage of the information found in gene sets and keeps the benefits of analysing data based on profiles, making the investigation of gene expression values more complete and flexible.

Contribution 2

generating understanding of new cancer biology.

Chapter 2

Systematic literature reviews

"What do researchers know? What do they not know? What has been researched and what has not been researched? Is the research reliable and trustworthy? Where are the gaps in the knowledge? When you compile all that together, you have yourself a literature review".-Jim Ollhoff

2.1 Introduction

This thesis has a purpose of undertaking a literature review on gene set enrichment analysis (GSEA), gene expression data analysis, disease prediction algorithms, and gene ontology. A question is how to start a literature review on these broad areas? To search a paper, read it, search another, read it and continue to search and read. This is an unstructured and tedious process. In addition, there is a possibility of missing out some section of the relevant literature that could be useful in discovering a problem or a suitable idea. Therefore, in order to carry out a comprehensive literature review, it would be useful to undertake a literature review that is both structured and systematic. A structured and methodological literature review would allow an individual to undertake a complete literature review [44].

A systemic literature review (SLR) is a research strategy that involves conducting a thorough and methodical analysis of published research in a particular area of study. An SLR process includes contents found in literature, search strategies used by an individual, and how and where an individual searched. In addition, an SLR includes evaluation criteria that an individual uses to choose which pieces of literature should be included or excluded from the review. An SLR, like any other literature review, provides a broad overview of the study area. However, an SLR differs from other literature reviews in the way it is compiled [45]. Fink [46] defined an SLR as a *"process that is methodical, precise, and replicable for finding, analyzing, and summarizing the preexisting amount of finished and published work generated by researchers, scholars, and practitioners"*.

2.2 Overview of the chapter

The aim of this chapter is to provide a systematic literature review of genomic data mining techniques, including algorithms, software, models, and tools for uncovering relationships between genomic data and phenotypic traits, related to genetic diseases [47]. Specially this review will focus on genomic data refers to gene expression, gene sets, and genomics [48]. Gene expression is the conversion of DNA sequence to mRNA and represent the process of “transcription”, “mRNA” represents the gene and the amount of mRNA is it’s expression [49]. Gene sets are groups of genes that represent biological functionalities. Modern technology generates huge amounts of genomic data. Therefore, data scientists and researchers are interested in analysing the large amounts of genomic data processed by data analysis techniques.

Data mining and predictive modelling play an important role in extracting insights from data. This thesis considers two approaches for data analysis, namely data-driven and knowledge-driven approaches. A data-driven approach is the process of building a model using sample data. A data-driven approach is effective when there are large amounts of sample data from the domain and one’s understanding of the domain is weak [50]. A knowledge-driven approach is the process of building one’s understanding of the domain. Knowledge-driven modelling is useful when it is difficult to obtain sample data and one’s understanding of the domain is strong [51].

Genomic data analysis typically forms a data processing pipeline, such as data acquisition, preprocessing, mapping, presentation, visualisation, and publication. The data processing pipeline can be performed in a linear fashion, but this is not always the case.

The pipeline requires combinations of multiple steps to solve a complex genomic task, such as gene expression analysis [52]. Appropriate combinations of pipeline steps lead to meaningful results from a large amount of genomic data. However, this review shows that current genomic data processing analysis faces the following challenges.

- Large amounts of genomic data are difficult to process due to the limitations of conventional computer hardware and algorithms. (i.e., a gene set contains many genes or traits.) [53, 54].
- Existing techniques for analyzing genomic data predict only genes or gene sets for a disease without considering any meaningful relationship between genes, i.e., gene sets, diseases, and patients [55]. An approach that examines the relationships between genes, gene sets, and gene expression is called gene set analysis (GSA) or gene set enrichment analysis (GSEA) [56].
- Sometimes, high-dimensional gene expression data are difficult to process when it's imbalance of class labels of sample instances. Moreover, cross-validation does not explain why some predictions are wrong while others are correct.

In GSA, knowledge-driven approaches integrate gene expression and gene sets that characterise biological functions, such as phenotype correlations, molecular interactions, or regulations of gene extraction [57]. However, existing research mostly focuses on the identification of genes or gene groups from the integration of genes and gene sets. Thus, there is an opportunity to study the relationships between patient characteristics, patient gene expression, and gene sets to understand disease. Similarly, many of the data-driven approaches developed for GSA can help in the study of diseases [58–62]. In addition to predictive modelling, GSA is used to identify genes or gene sets from a large amount of genomic data on breast cancer, obesity, bipolar disorder, schizophrenia, and arthritis [61–69].

In addition to GSEA and GSA, this review explored the use of AI approaches to better comprehend gene expression data analysis in complicated diseases. These diseases include leukemia, breast cancer, colorectal cancer, and adrenal cancer, among others. Figure 2.1) shows that GSEA has association with multiple datasets (gene expression data, microarray data, and genomics data) that span multiple GSEA domains. Specifically, the goal of the review is to examine current predictive modelling, data-driven, and knowledge-driven approaches to identify their strengths, limitations, and the opportunities for GSA and GSEA.

This review is arranged as follows: Section 2.3 describes the methods of SLR with six subsections, namely scope (2.3.1), research objective (2.3.2), inclusion criteria (2.3.4), evaluation

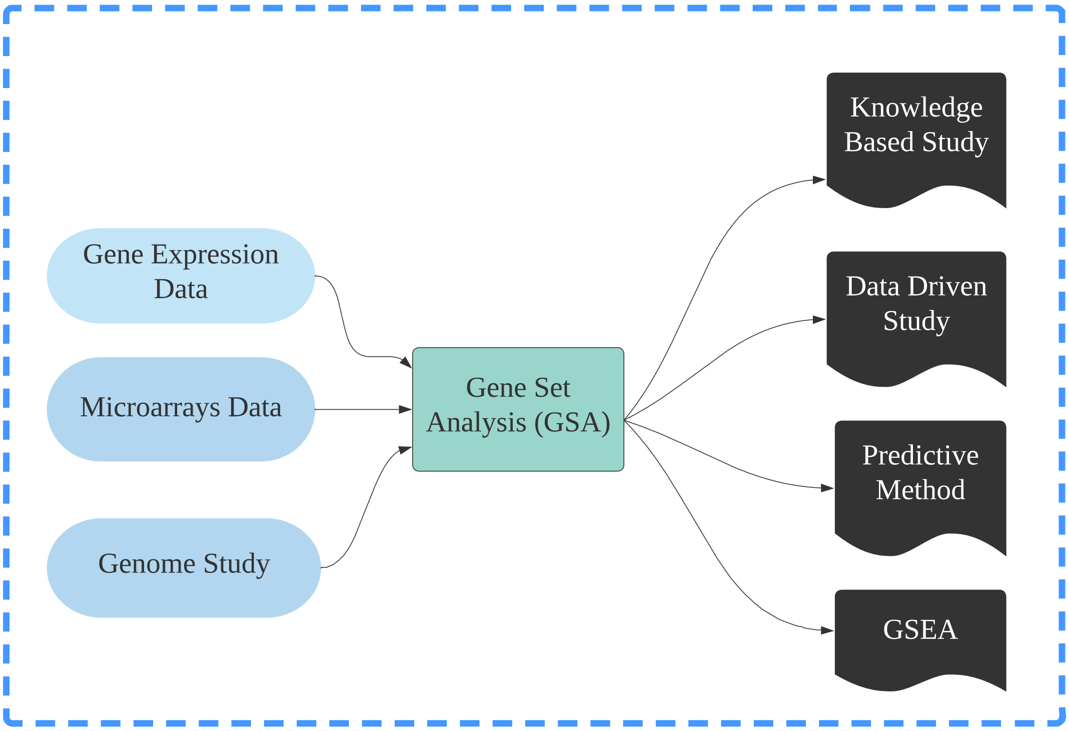


Figure 2.1: Literature review of three themes (left) for gene set analysis with respect to data-driven, knowledge-driven, predictive models (right), and gene set enrichment analysis (GSEA).

criteria (2.3.5), literature search (2.3.3), and results and taxonomy (2.3.6). The last subsection describes the taxonomy of the listed literature. The taxonomy considers associations between biological functionalities and gene expression data in complex disease analyses. Sections 2.4, 2.5, 2.6, 2.7 and 2.8 provide an overview of the classified works. Section 2.9 recognises the research gaps and proposes solutions to improve GSEA analysis. Section 2.12 concludes the review.

2.3 Paper selection method

2.3.1 Scope

The aim of this review is to explain GSA methods and suggest areas where existing methods can be improved to provide a deeper understanding of the relationships between genomics and diseases.

2.3.2 Research objective

The research objective (RO) of this review is to gain deeper understanding of gene expression data, diseases that are related with gene expression data, the impact of patient biological functions on diseases, and impact of variations in gene expression values on diseases. In addition to that, this thesis will explore anomalies in gene expression data, the processes of disease prediction, and approaches for dealing with imbalanced data. Finally, this thesis will explore how disease analysis could benefit from explainable artificial intelligence (explainable AI), which is a paradigm for interpretable artificial intelligence.

To explore the contents of the literature, the research objective (RO) has been broken down into three research questions (RQs) (Table 2.1). These questions will be reviewed while reading various literature to see whether any of the literature satisfy these requirements. A key motivation for these research questions is to gather relevant information and ideas while reading a research paper or any related article. Therefore, this thesis defined these research questions as review research questions (RRQs). These RRQs differ from the research question posed in chapter 1 as these research questions only focused on gathering desired information from an article. Where research questions in chapter 1 are the quarries from a global context of all literature.

2.3.3 Literature search in Google Scholar

The literature was reviewed using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) model [70]. This thesis searched journals and publishers such as Nature, Bioinformatics, PubMed and preprint archives for research articles published as recently as October 2020. The search strings were: “GSEA”, “Data-Driven approaches for GSEA”,

Table 2.1: Research objective and questions.

Reference	Research questions	Domain
RO	This thesis focused to understand impact of integrating gene sets with gene expression data under different circumstances (anomaly testing, disease status prediction) and analyse the results with explainable AI.	Machine learning with genomic data
RRQ1	How can biological functionalities integrate into gene expression data to improve the gene set enrichment analysis leading to anomalies in gene expression data?	Genomic data
RRQ2	How does predictive modelling classify disease status and propose treatment planning by considering imbalanced training samples?	Machine learning
RRQ3	How can results of cross-validation evaluate that are incorrect and correct prediction?	Machine learning

"Knowledge-Driven approaches for GSEA", "Predicting model for cancer status", "Sampling Models in Gene Set Enrichment Analysis", "Enricher tools", "Cluster and visualisation GSA", and "Microarray data analysis for GSEA".

2.3.4 Inclusion criteria

This thesis explored literature that focused on at least one of the following parameters:

- tools of gene or gene set prediction.
- ranking of genes or gene sets exploring biological perturbations
- visualization or clustering tools for visualizing gene clusters and gene ranking.
- prediction methods or tools on microarray and RNA sequences.
- enricher tools for the analysis of coding and non-coding regions for microarray data
- statistical prediction methods for different cancers
- knowledge-driven GSA

This thesis reviewed the abstracts and full texts of the selected studies with the inclusion criteria as shown in Table 2.2.

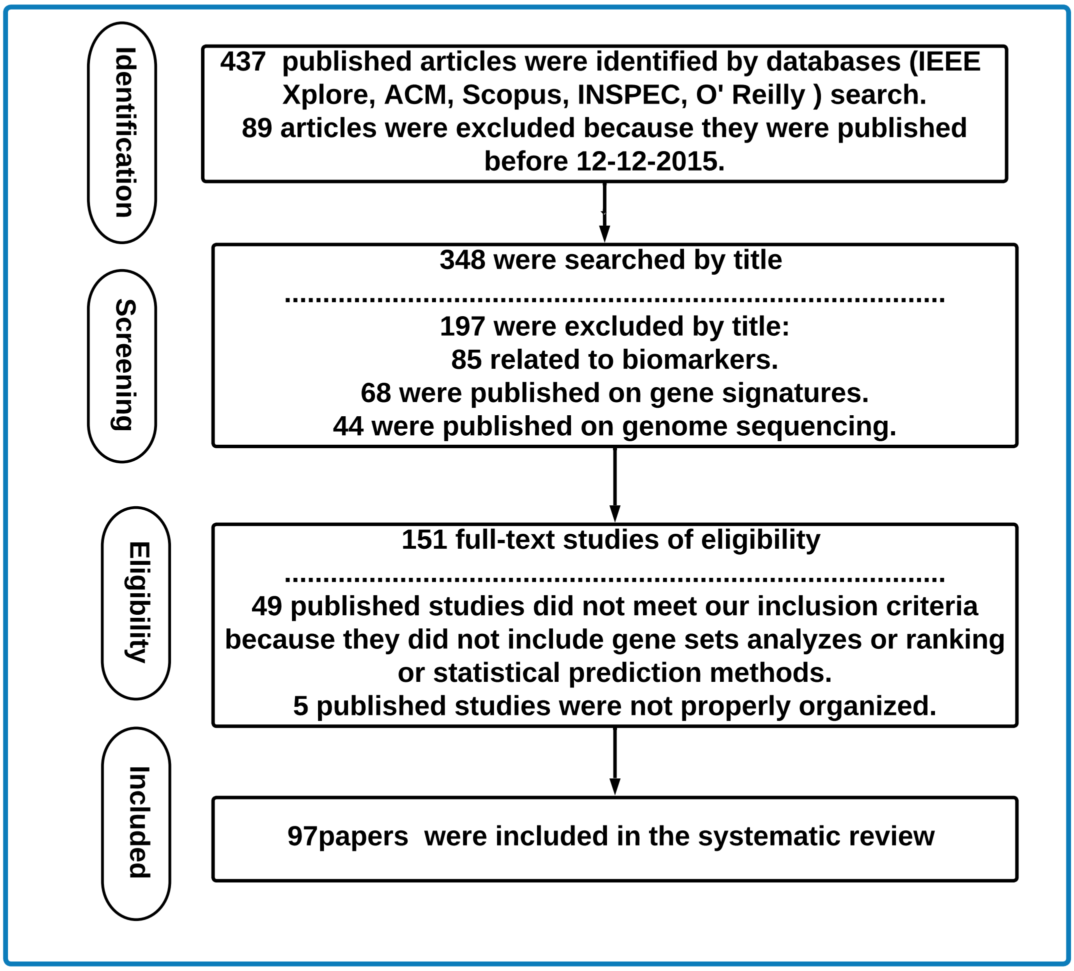

Identification
437 published articles were identified by databases (IEEE Xplore, ACM, Scopus, INSPEC, O' Reilly) search.
89 articles were excluded because they were published before 12-12-2015.
Screening
348 were searched by title
197 were excluded by title:
85 related to biomarkers.
68 were published on gene signatures.
44 were published on genome sequencing.
Eligibility
151 full-text studies of eligibility
49 published studies did not meet our inclusion criteria because they did not include gene sets analyzes or ranking or statistical prediction methods.
5 published studies were not properly organized.
Included
97papers were included in the systematic review

Figure 2.2: Steps in the PRISMA model to search and identify articles or research papers related to gene set enrichment analysis (GSEA).

Table 2.2: Standards for selecting literature.

Ref.	Description
Grouping rules	
GR1	Exploration strings appear in the title
GR2	Written in English
GR3	Published in a journal or conference
GR4	Books
Elimination rules	
ER1	Out of the scope of genes or GSA
ER2	Citations
ER3	Irrelevant to research questions

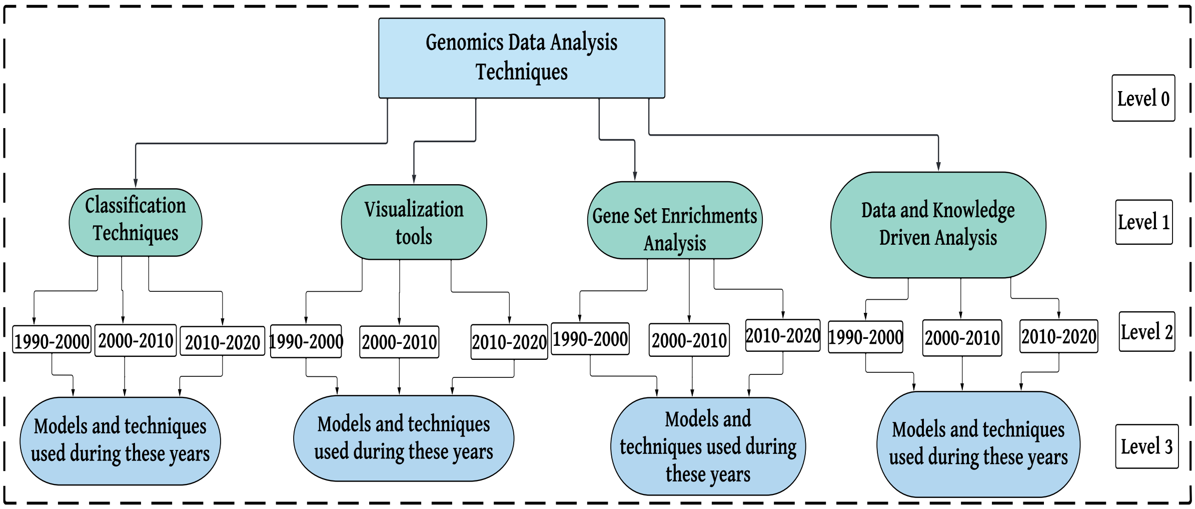

Genomics Data Analysis Techniques
Level 0
Classification Techniques
Visualization tools
Gene Set Enrichments Analysis
Data and Knowledge Driven Analysis
Level 1
1990-2000
2000-2010
2010-2020
Level 2
Models and techniques used during these years
Level 3

Figure 2.3: Taxonomy of genomics data processing approaches.

2.3.5 Evaluation criteria

To achieve objective, this thesis will explore the following requirements to understand possible contributions to GSA and GSEA:

Requirement 1 (R1): This review will result in a better understanding of the integration of biological functionalities into gene expression data to improve the perception of complex disease analyses. In addition, genomic data analysis techniques should explore the association of biological functions with gene expression data and between instances (patients) and biological functions with a concentration on gene expression data.

Requirement 2 (R2): This review will result in a better understanding of the impact of machine learning on the association of genes, gene sets and patients to improve existing gene set enrichment analyses as machine learning becomes more widely used. In addition, the interpretation of genomic data using only known statistical analyses should apply predictive modelling, data-driven and knowledge-driven approaches to explain the relationships between related biological traits and patients in complex disease analysis.

Requirement 3 (R3): This review will result in a better understanding of how to classify (predict) a disease state by looking at genomic data samples integrated with their functionalities (gene sets) and demonstrate the results with explainable AI [71]. Furthermore, predictive modelling should explore ”trustworthy explainability” that focuses on predicting patients’ class labels. Current cross-validation techniques that predict target attributes do not track why some predictions are wrong while others are correct.

Following the evaluation criteria of R1-R3, this review highlights ways to improve current genomic data analysis by integrating genes and gene sets (biological functionalities) into an AI-based framework. Many research articles and applications focus on gene expression data analysis (Table 2.3), while some studies [72–75] explore ways to achieve R1-R3. Even with approaches that focus directly on gene set enrichment, most do not meet R1 – R3 which relate to the integration of genes and gene sets (biological functionalities) in genomic data analysis for disease instance mapping.

The relationships between the four components (i.e., genes, gene sets [biological functions], instances [patients], and explainable AI) have not yet been investigated. Therefore, it is unclear which relationships are significant. *Therefore, this review explores the implications of integrating gene sets with gene expression data, leading to complex disease analyses with predictive modelling, data-driven and knowledge-driven approaches, particularly in relation to leukaemia and cancer.*

2.3.6 Results and taxonomy

Figure 2.2 shows literature searching mechanism using a PRISMA model [76]. Titles and abstracts of the individual articles were reviewed and evaluated using taxonomy of PRISMA model shown in Figure 2.3. Of these articles, 344 could not be classified into the taxonomy and were discarded from further analysis. The remaining 97 papers were thoroughly reviewed.

Table 2.3 presents a group of research publications on cancer disease analysis using gene expression data, RNA-seq data, and protein data. The second column of the table labelled ”Have any gene integration,” and the third column, labelled ”Have any anomaly analysis,” are criteria that indicate whether or not these papers focused on these criteria. If an article not meet them, it will be a cross sign, otherwise, it will be a check mark. Last column of the table represents whether these articles meet the requirements mentioned above.

Table 2.3: An evaluation of the selected literature review of the research publications on cancer genomics datasets from the following categories: gene expression data, RNA-seq data, and protein data.

Paper	Year	Have any gene set integration?	Have any anomaly analysis?	Is their focus on these requirements?		
				R1	R2	R3
Chang *et al.* [77]	2016	✓	×	✓	×	×
Geistlinger *et al.* [78]	2020	✓	×	✓	×	×
Maleki *et al.* [79]	2020	✓	×	✓	×	×
Hu *et al.* [80]	2014	✓	×	✓	×	×
Joly *et al.* [81]	2020	✓	×	✓	×	×
Rahmatallah *et al.* [82]	2014	×	×		×	×
Tiong *et al.* [83]	2019	✓	×	✓	×	×
Eupa *et al.* [84]	2019	×	×	×	×	×
Meng *et al.* [85]	2019	✓	×	✓	×	×
Neupane *et al.* [86]	2018	✓	×	✓	×	×
Kong *et al.* [87]	2018	✓	×	✓	×	×
Arloth *et al.* [88]	2020	×	×	×	×	×
Roy *et al.* [67]	2020	×	×	×	×	×
Allahyar *et al.* [89]	2019	×	×	×	×	×
Zhou *et al.* [90]	2019	×	×	×	×	×
Tong *et al.* [91]	2020	×	×	×	×	×
Soltis *et al.* [92]	2013	×	×	×	×	×
Xin *et al.* [93]	2020	×	×	×	×	×
Walter *et al.* [94]	2015	×	×	×	×	×
Ge *et al.* [95]	2020	✓	×	✓	×	×
Yusuf *et al.* [96]	2005	✓	×	✓	×	×
Rho *et al.* [97]	2011	✓	×	✓	×	×
Kang *et al.* [98]	2014		×	✓	×	×
Ewing *et al.* [99]	2020	✓	×	✓	×	×
Wang *et al.* [100]	2020	×	×	×	×	×
Ong *et al.* [101]	2019	×	×	×	×	×
Perampalam *et al.* [102]	2020	×	×	×	✓	×
Reyes *et al.* [103]	2019	×	×	×	✓	×
Yousif *et al.* [104]	2020	×	×	×	×	×
Zhu *et al.* [105]	2019	×	×	×	×	×
Netanely *et al.* [106]	2019	✓	×	✓	×	×

2.4 Gene set enrichment analysis

GSEA is an approach that analyses gene expression data to identify sets of genes that relate to specific biological functions (e.g., chromosomal location, molecular function, or gene regulation).

GSEA uncovers gene sets from an ontology that are strongly associated with a named phenotypic trait. Samples with gene expression data are classified into two groups: those with the trait and those without the trait (e.g., cancer vs. normal or male vs. female) [56]. This thesis explored a large body of literature on GSEA that considers relationships between genetics and disease. The papers listed in (Table 2.4) describe data mining and statistical methods on GSEA that explain the relationships between gene sets and gene expression data.

Before describing the computational steps of GSEA, it would be useful to be familiar with several terms related to GSEA, namely differential gene expressions, mRNA and false discovery rate.

The computational steps (Figure 2.4) of GSEA are as follows:

- The first step of GSEA is to rank the genes in the gene expression data. The ranking is created by measuring the fold change (FC) of a gene, which is a ratio between two states of that gene. For example, the ranking of a gene list is created by measuring the fold change of gene expression values between tumour and normal genes.

- The second step is to find a matching gene for each gene set which finds the position of each gene in the gene set within the ranked gene list.

- An enrichment score (ES) is calculated for each gene set using weighted Kolmogorov Smirnov statistics [107]. ES measures overexpressed genes in gene sets that are either at the top or bottom of the rankings. ES is used to calculate the variation between genes in each gene set by assigning the ranking and the number of matched genes in the gene set (Figure 2.5). Figure 2.5 shows FC, hit and miss, i.e. fold change, matched genes between gene expression and gene set, unmatched genes in gene sets and gene expression data. The running total in the figure 2.5 shows the summation of the total hits and misses. For example, ES for a gene set is 0.59 for the given data in Figure 2.5.

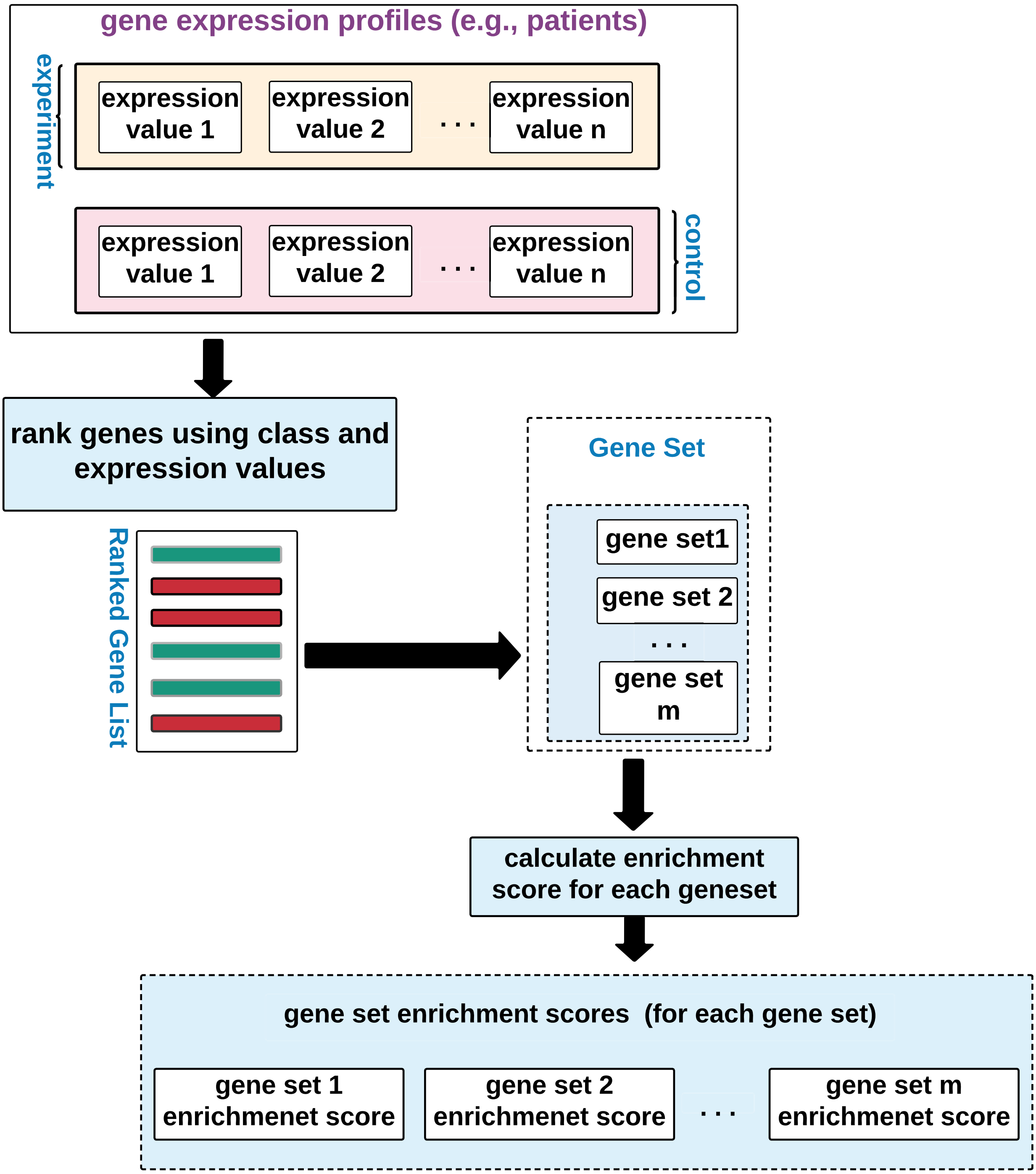
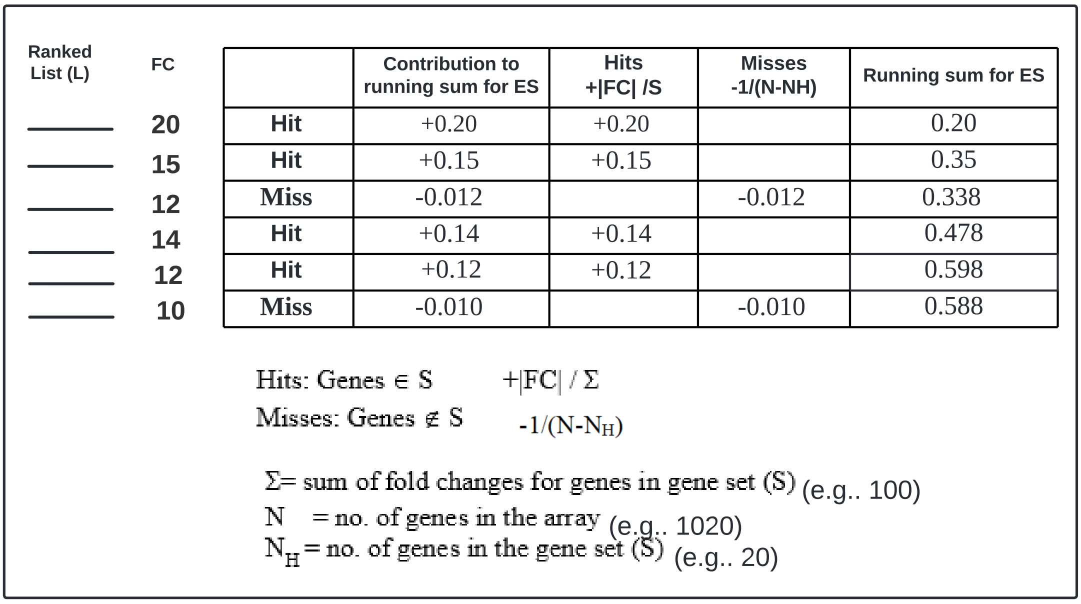

gene expression profiles (e.g., patients)
experiment
expression value 1
expression value 2
. . .
expression value n
expression value 1
expression value 2
. . .
expression value n
control
rank genes using class and expression values
Ranked Gene List
Gene Set
gene set1
gene set 2
. . .
gene set m
calculate enrichment score for each geneset
gene set enrichment scores (for each gene set)
gene set 1 enrichmenet score
gene set 2 enrichmenet score
. . .
gene set m enrichmenet score

Figure 2.4: Computational steps for GSEA.

- Repeat steps 1-3 for all gene sets, noting that each gene set from an ontology is processed in order of gene list ranking.

- The final step of GSEA is to measure the false positive rate or false discovery rate (equation 2.1)) for multiple gene sets using statistical tests (e.g., t-test) [108].

$$\text{FDR} = \frac{m_0}{m} \tag{2.1}$$



Ranked List (L)	FC		Contribution to running sum for ES	Hits +|FC| /S	Misses -1/(N-NH)	Running sum for ES
	20	Hit	+0.20	+0.20		0.20
	15	Hit	+0.15	+0.15		0.35
	12	Miss	-0.012		-0.012	0.338
	14	Hit	+0.14	+0.14		0.478
	12	Hit	+0.12	+0.12		0.598
	10	Miss	-0.010		-0.010	0.588

Hits: Genes $\in$ S $+|\text{FC}| / \Sigma$

Misses: Genes $\notin$ S $-1/(\text{N}-\text{N}_\text{H})$

Σ= sum of fold changes for genes in gene set (S) (e.g.. 100)

N = no. of genes in the array (e.g.. 1020)

N_H = no. of genes in the gene set (S) (e.g.. 20)

Figure 2.5: Calculation of enrichment scores for GSEA.

GSA examines the relationship of correlated genes between multiple sets of genes and can provide the context for genomic changes in these genes [40–43, 115]. In addition, GSEA has been used to understand the biological functions of genes using gene expression data to analyse complex diseases, such as schizophrenia [116], bipolar disorder [117], Crohn's disease [118], rheumatoid arthropathy [119], breast tumour [120], obesity [121] and gossypium [33]. There are three broad categories of GSEA [122]:

- over-representation analysis,
- functional class scoring,
- pathway-topological-based methods.

This thesis will explain and explore these categories.

Table 2.4: Literature review of the research publications on existing approaches to GSEA.

Technique	R1	R2	R3
Permutation-based gene set analysis [77]	✓	×	×
Biological reasoning on the relevance of enriched gene sets [78]	✓	×	×
Integrative differential expression [68]	✓	×	×
Gene set enrichment analysis for DNA methylation [109]	✓	×	×
Benchmarking of gene set [110]	✓	×	×
Method to classify gene set analysis[79]	✓	×	×
Comparative simulation analysis for breast cancer [80]	×	×	×
Differential gene set enrichment analysis to quantify the relative enrichment of two gene sets [81]	✓	×	×
Multivariate differential co-expression test to compare gene signatures [82]	×	×	×
Combinatorial relations of feature scores [83]	✓	×	×
Alternative assessments of enrichment [84]	×	×	×
A framework which combines both gene expression changes and gene set analysis approaches [85]	✓	×	×
A network analysis to identify gene sets [86]	✓	×	×
Method for analysis of repeatedly measured phenotype data [111]	×	×	×
A method for differentially expressed genes and mutated cancer genes [112]	×	×	×
Gene Ontology (GO) enrichment analysis tool for tissue-specific information [113]	×	×	×
An innovative statistical approach for interpreting gene expression data [62]	✓	×	×
Gene set analysis algorithm for biomarker identification [114]	✓	×	×

2.4.1 Over-representation analysis

Over-representation analysis (ORA) examines whether genes from gene sets are represented more than expected in gene expression data [123]. For example, the transcriptome of acute lymphoblastic leukaemia (ALL) has 12,000 genes. Suppose 300 genes are characterised as 'axon direction'and in an analysis an individual observe 1,000 genes that are differentially expressed. If 200 of these genes are in the 'axon direction'class, the over-representation analysis determines the significance of the 'axon direction'class at ALL. ORA uses conventional statistical tests like the t-test, hypergeometric test, binomial test and chi-square test.

Many publications on genomic disease analysis use ORA, which depicts the relationships between genomic data and diseases in their methodology. In Figure 2.7, this thesis divide ORA into two broad categories: (1) gene ontology and differential expression and (2) pathway tools. All existing approaches for these two categories are included.

2.4.2 Gene ontology and differential expression

Two Greek words, “onto” and “logia”, form the word onotology [124], where onto means real and logia means science [125]. Philosophically, ontology describes the existing world and non-philosophically, it says something about a particular domain. In bioinformatics, it is a combination of genes and gene products. Moreover, GO describes the relationships between genes and gene sets. In addition, GO maintains and develops its control vocabulary for genes and gene product attributes [105, 126].

This GO is divided into three categories: molecular function ontology, biological process ontology, and cellular component ontology [127]. Molecular function describes the function of a gene at the molecular level. Biological process describes the function of specific genes integrated into cells, tissues, and organs. The cellular component describes cellular functionalities with their environment. These ontologies are used for classification problems with binary labels or for multi-label classification, which tests whether classifiers can identify class labels for a gene or not, using gene annotation information defined in the following [128–132]. Gene annotation is a process of identifying the location of genes in a genome.

2.4.3 Pathway analysis

Pathway analysis (PA) is a set of interactions between gene sets and differentially expressed genes; a gene is differentially expressed if a variation in expression level is observed between two experimental counts. Figure 2.6 illustrates a pathway analysis where the blue circles represent differentially expressed genes and the green circles represent gene sets that interact for gene set enrichment. In addition, Figure 2.6b shows how enrichment and non-enrichment occur in the GSEA. Enrichment analysis within the PA determines the genes responsible for cancer [133] and helps compare tumour cells vs. normal cells [134].

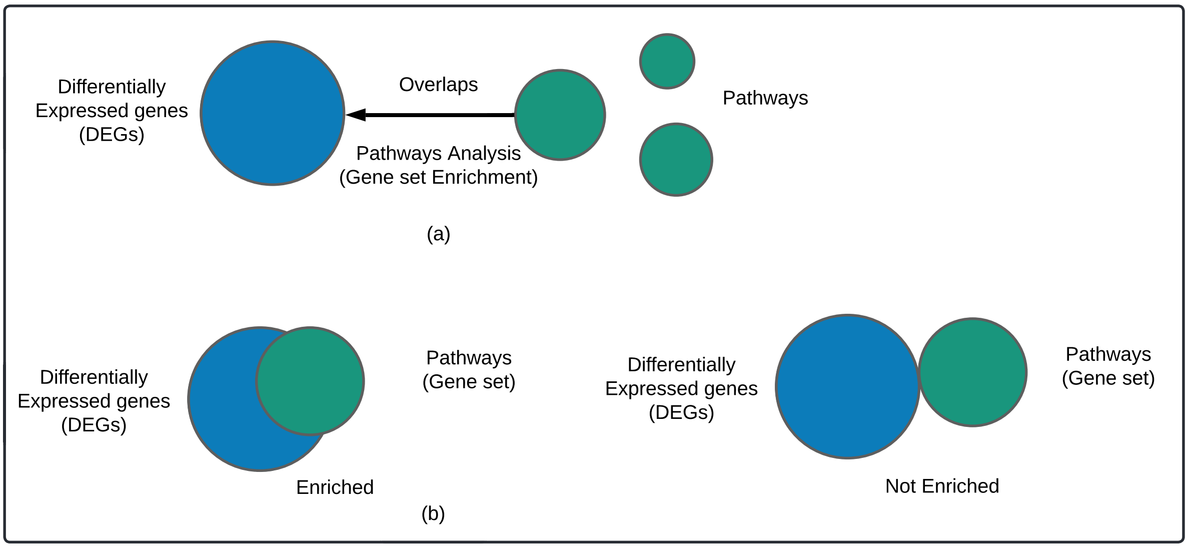

Differentially Expressed genes (DEGs)
Overlaps
Pathways Analysis (Gene set Enrichment)
Pathways
(a)
Differentially Expressed genes (DEGs)
Pathways (Gene set)
Enriched
Differentially Expressed genes (DEGs)
Pathways (Gene set)
Not Enriched
(b)

Figure 2.6: Interactions between pathways and differentially expressed genes. (a) shows the initial phase in which pathways and overexpressed genes interact with similar genes, and (b) shows the final phase of interactions in which matched genes are labeled as enriched and mismatched genes are labeled as not enriched.

2.4.4 Functional class scoring of GSEA

Functional class scoring (FCS) generates enrichment scores from a gene expression matrix containing all the information from genes to explain the relationships between genomic data and diseases [27]. FCS is divided into two classes: univariate and multivariate. Univariate applies a signal-to-noise ratio (SNR), which is a deviation between gene expression measured in training and test samples, to calculate an enrichment score. Multivariate FCS calculates enrichment scores directly without intermediate steps using differential expression (DE), which is an approach to determine whether genes are expressed at significantly different levels between two mRNA (messenger RNA) sample groups.

A comparative review of these three GSEA approaches is presented in Table 2.5.

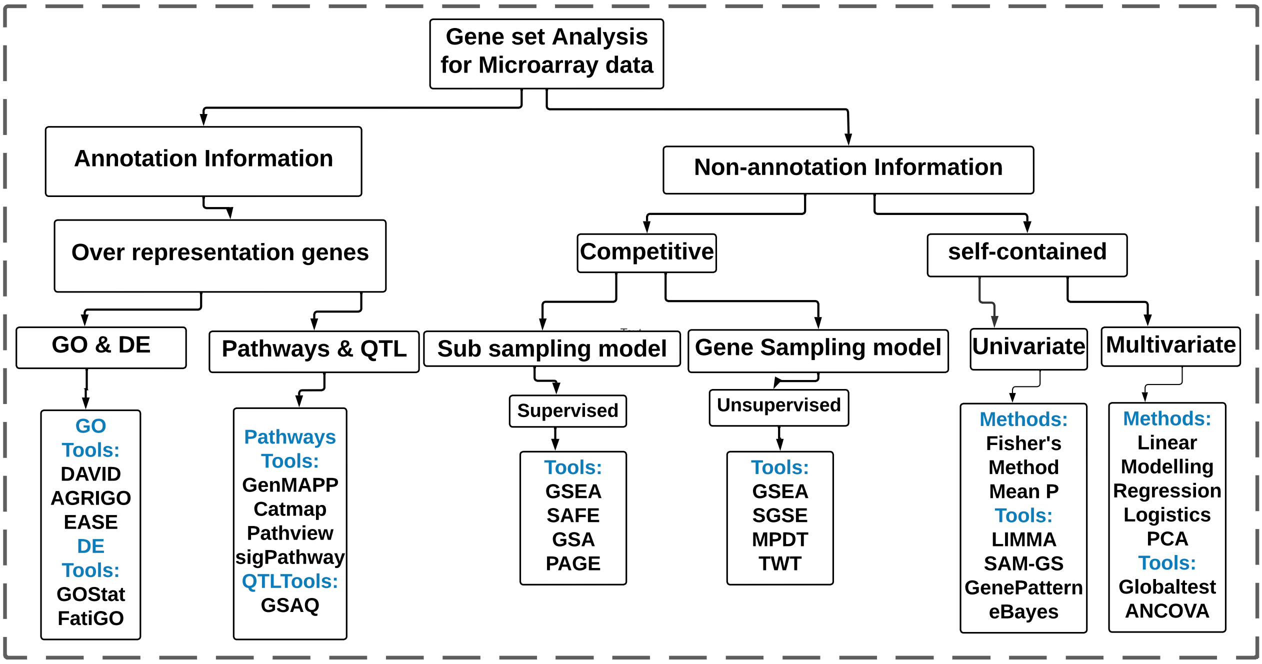

Gene set Analysis for Microarray data
Annotation Information
Non-annotation Information
Over representation genes
Competitive
self-contained
GO & DE
Pathways & QTL
Sub sampling model
Gene Sampling model
Univariate
Multivariate
GO Tools: DAVID AGRIGO EASE
DE Tools: GOStat FatiGO
Pathways Tools: GenMAPP Catmap Pathview sigPathway
QTLTools: GSAQ
Supervised
Unsupervised
Tools: GSEA SAFE GSA PAGE
Tools: GSEA SGSE MPDT TWT
Methods: Fisher's Method Mean P
Tools: LIMMA SAM-GS GenePattern eBayes
Methods: Linear Modelling Regression Logistics PCA
Tools: Globaltest ANCOVA

Figure 2.7: Taxonomy of GSEA approaches.

Table 2.5: Advantages and limitations of over-representation analysis, functional class score and topological analysis.

Over-Representation Analysis	Functional Class Score	Topological Analysis
Advantages: 1. Simple and powerful. 2. Requires less input data	Advantages: 1. More Accurate than ORA. 2. Uses entire list of genes measured.	Advantages: 1. Considers the role, position, magnitude and interactions of each gene. 2. Can make predictions.
Limitations: 1. Discards 90% of data. 2. Assumes all genes are independent. 3. Evaluates only the number of significant genes.	Limitations: 1. Ignores interaction among genes. 2. Analyzes each pathway independently. 3. Many false positives.	Limitation: 1. Requires more data 2. Takes slightly longer 3. Not currently applicable to metabolic pathways.

Table 2.6: Algorithms and tools for gene set analysis using microarray data (gene expression data, RNA-seq data, and protein data).

Approaches	Algorithm/Tools
Over-representation Analysis	DAVID [135], GenMAPP [136], GoMiner [137], FatiGO [138], GOToolBox [139], GeneMerge [140], GOEAST [141], ClueGO [142], FunSpec [143], GARBAN [144], WebGestalt [145],GOFFA [146], WEGO [147], GOTM [148], Pathview [149], Wholepathwayscope [150]
Functional Class Score	Random set [151], PCOT2 [152], SAM-GS [153], LIMMA [154], Catmap [155], T-profiler [156], FunCluster [157],GeneTrail [158], Gazer [159], CAMERA [160], GAGE [161], SGSE [162], GSNCA [82],GSA-SDR [163], GenePattern [164]
Topological Analysis	PathwayExpress [149], ScorePAGE [165], SPIA [166], NetGSA [167], TopoGSA [168], CliPPER [169]

2.4.5 RNA-Seq technologies and micro-array genomics data

Microarray technology is a widely used method in genomics for investigating gene expression values [170]. This technique involves placing small DNA fragments, known as probes, on a solid surface like a glass slide. These probes can specifically bind to complementary RNA molecules from a sample [171]. When the RNA molecules bind to their matching probes, they produce a fluorescent signal. By measuring the intensity of these signals, researchers can determine the abundance of specific genes in the sample [171]. Microarray technology has been a valuable tool in genomics research for many years to investigate gene expression patterns and easy understanding of various biological processes and diseases [170]. Although microarrays have been widely used in genomics research for decades, they have some limitations, such as reduced sensitivity, probe design issues, and background noise [172]. Despite this, microarrays still have a place in genomics research, particularly in applications where cost and data analysis simplicity are important factors [173].

However, newer approaches like RNA-Seq are gradually becoming more popular due to their increased sensitivity and ability to detect a wider range of transcripts [174]. RNA-Seq technologies are a group of methods used in genomics to investigate gene expression values. These technologies analyse the complete collection of RNA molecules in a cell or organism at a specific time, known as transcriptome [174, 175]. With the help of advanced high-throughput sequencing techniques, RNA-Seq offers detailed insights into the transcriptome, providing valuable information about which genes are active and to what extent, in order to understand complex biological processes [176]. By producing a large amount of data, these technologies present a novel approach for learning about the molecular mechanisms behind various biological processes, such as growth, change, and diseases [170]. One of the main advantages of RNA-Seq over microarray technology is its ability to detect a broader range of transcripts, including low-abundance and novel transcripts [177]. This increased sensitivity is particularly important when studying genes that are expressed at low levels, which can be challenging to detect using microarrays [178]. Additionally, RNA-Seq is not dependent on prior knowledge of the genome, allowing for the discovery of novel transcripts, alternative splicing events, and gene fusions [179].

Table 2.7 describes the Comparison between RNA-Seq and Microarray Technologies.

Table 2.7: Comparison between RNA-Seq and Microarray Technologies.

Aspects	**RNA-Seq**	**Microarray**
Detection Method	High-throughput sequencing technologies directly sequence RNA molecules and provide a digital representation of gene expression based on the number of reads mapped to each gene [180].	Hybridization between known complementary DNA probes on a solid surface and fluorescently labeled target cDNA or cRNA samples; fluorescence intensity corresponds to expression level [181].
Dynamic Range	Broader, allowing reliable detection of low and high abundance transcripts [182].	Limited by background noise and signal saturation at high expression levels [183].
Transcript Discovery	Can detect new transcripts, splice variants, non-coding RNAs, and fusion genes [184].	Relatively limited to measuring expression of known genes represented by array probes [183].
Quantification	Less cross-hybridization and background noise [185].	Depends on fluorescence intensity, which is affected by probe quality, dye biases, and sample labelling effectiveness [186].
Reproducibility	Better reproducibility between technical replicates and laboratories [187].	Depends on probe synthesis, array construction, and hybridization conditions [188].
Data Analysis	Needs advanced computational and bioinformatics pipelines [189].	Normalisation and differential expression analysis are utilised with relatively less computational processing required [190].
Cost	Relatively expensive [191].	Cost-effective, making it a popular choice for large-scale gene expression analysis [192].

2.5 Knowledge-based gene set enrichment analysis

Genomics describes the study of a person's genes, including the interactions of these genes and sets of genes with respect to diseases such as cancer. In addition, genomics helps scientists investigate why a patient develops certain diseases and why one disease recurs after cure while others do not. For example, many patients undergo cancer treatments such as chemotherapy and bone marrow transplant and have routine checkups but their cancer recurs. Other patients receive similar treatments but are cured. Genomics may hold the clues to explain these variations. Understanding the relationships between genes and gene sets through genomics helps scientists gain insight into diseases and patients.

Genomic data mining allows the estimation of gene expression levels of a large number of genes simultaneously. Many GSA approaches have been used to derive valuable information from genomic data, such as the discovery of differentially enriched genes associated with a particular biological function or disease phenotype. The integration of biological functions into gene expression data could be useful to explain the relationships between genomic data and disease and could help to understand how genomic data for specific patients relate to these functions [193]. In this section of the review, this thesis explore literature on knowledge-based GSEA that assesses the relationships between genomic data and disease.

2.5.1 Algorithms integrating biological knowledge

Stephan *et al*. [72] integrated pathway information (biological knowledge) and gene expressions to identify genes that are highly significant in cancer. Pathway data were collected from the GSEA [56] and sigPathway [194] databases. Both GSEA and sigPathway ranked genes based on differential expression and p-values, respectively. In this research, the random forest algorithm was used to select genes from gene expression data [195].

Chen *et al*. [196] combined GO and gene expression datasets and applied principal component analysis to separate relevant genes to predict survival consequences. The results showed that integrating biological knowledge (GO) outperformed prediction accuracy compared to approaches that only considered gene expression for prediction.

Miguel *et al*. [197] integrated gene sets with gene expression data to classify genes from the gene set provided by the expert. Four different classification algorithms (IBk, naive Bayes, support vector machine, C4.5) were used to select all the matched genes from a given gene set. The results showed that the combination of biological knowledge and gene expression data can outperform classifiers that rely solely on gene expression data.

Bandyopadhyay *et al.* [198] designed a pathway-centric approach to feature extraction by integrating pathway and gene expression data to reduce overfitting. The results showed that accuracy improved after biological knowledge was integrated.

Kim *et al.* [199] integrated pathway information with gene expression data to improve the efficiency of cancer subtype classifications.

Parraga Alava *et al.* [198] designed a clustering approach (NSGA-II [200], path-relinking [201] and Pareto local search [202]) by integrating biological knowledge and gene expression data to find gene clusters. First, the approach finds relationships between genes in the gene expression data to ensure there are duplicate genes in the collected gene expressions. Next, it searches for similar genes from the gene sets and clusters them if there is a match. The approach was applied to four different gene expression datasets.

Cui *et al.* [203] proposed a gene clustering approach using a deep learning autoencoder method [204]. The input was gene expression data and the biological information was gene networks. The novelty of this approach was to investigate whether the same gene was present in two networks using gene networks as biological information. The inclusion of gene network information improves the performance of gene clustering. The comparative result showed that the current method outperformed the traditional clustering of k-means and hierarchical clustering.

Bauer *et al.* [205] introduced a Bayesian inference approach to gene selection by integrating GO and gene expression data. Here, conditional probabilities are used to generate a score that identifies matching genes with the GO and categorises gene types. Here, genes were grouped based on gene expression data, and biological knowledge is used after analysis to perform a similarity check of the groups. However, this method has several shortcomings. In particular, grouping genes based on expression data only yields isolated co-expressed genes, but not significantly biologically coherent blocks [206, 207].

In a multivariate approach (MVA), Fagan *et al.* [208] and Busold *et al.* [209] integrated GO and gene expression data and grouped similar genes by measuring the distances between chromosomal positions between genes on both sides of the chromosomes.

Verbanck *et al.* [210] identified clusters of co-expressed genes (genes with similar expression levels) by integrating GO and gene expression data using an unsupervised clustering algorithm (k-means). The location of one co-expressed gene may be a gene regulatory network, while another could be from a specific biological response.

Brown *et al.* [211] applied a support vector machine (SVM) to classify genes from gene expression data by considering the known functions of gene sets.

Nepomuceno *et al.* [212] proposed an unsupervised bi-clustering approach that examines groups of genes with similar expressions according to a subset of preconditions. A scatter search, a metaheuristic optimization technique [213], helps identify gene patterns. The term bi-clustering refers to the identification of two distinct gene clusters from two separate input files. One input file contains gene expression data and the other file contains gene names associated with their GO annotations. The most fundamental difference of bi-clustering from conventional clustering is that bi-clustering aims to cluster genes and conditions (gene annotations) simultaneously. However, it has been shown that a scatter search leads to low accuracy. Fyad *et al.* [74] proposed a bisecting k-means leader clustering that forms multiple clusters of genes by integrating GO and gene expression. Finally, they identified the genes with shorter distances between these clusters.

Imoto *et al.* [214] proposed a gene network by combining biological knowledge and gene expression data with a Bayesian network that identifies genes from GO.

Gan *et al.* [215] proposed projection on convex sets (POCS) to find missing values, i.e., expression values, in gene expression data. Local least squares regression was used to identify genes by integrating gene expression and GO [216].

Kong *et al.* [217] presented a multivariate approach for clustering genes by integrating gene expression and gene sets. The approach identifies genes from multivariate samples by considering their phenotype in each gene set. Hotelling's T2 T^2 [218–222] measured the covariances between genes and gene sets for similar clustering genes.

Gene regulatory networks, representing either protein-protein interactions, cellular signalling, or precise molecular control, integrate GO and gene expression data to form a network [223–227]. Some popular approaches for gene regulatory networks are Boolean networks, differential equations, and Bayesian networks (BNs) [228–230].

Three different approaches have been used to summarise gene regulatory channels: Boolean networks, differential equations, and Bayesian networks (BNs). The strategy that has received increased attention involves an inference mechanism for causal probabilistic networks in particular BNs and dynamic Bayesian networks (DBNs) [228–238]. Finally, some other approaches that integrate both gene expression and gene ontology are presented in Table 2.8 with respect to R1-R3.

Table 2.8: Literature review of the research publications on existing approaches to knowledge-driven GSEA approaches.

Technique	R1	R2	R3
A statistical prior biological knowledge [239]	×	✓	×
Classification of cancers to compare the clinical markers [240]	×	×	×
An integration of biological knowledge [241]	×	×	×
A self-sufficient prediction for relevant pathways for functional enrichment [242]	×	×	×
Integrating gene signatures [243]	×	×	×
Microarray experiments to define signatures [244]	✓	×	×
Integrated genes association network [245]	×	×	×
Bi-clustering unsupervised method to search for patterns in gene expression [246]	×	×	×
Network construction for series of data [247]	×	×	×
Multivariate statistical procedure [248]	×	×	×
Multi-objective clustering algorithm [249]	×	×	×
A gene expression analysis to find the transcriptional activity of a cell [250]	×	×	×
Method for functionally classifying genes [251]	×	×	×

2.6 Data-driven GSA

2.6.1 Classification methods and visualization tools

A data-driven approach in GSA allows the analysis of genomic information to find patterns in genes or gene sets, or to determine insights into diseases associated with genes or gene sets. Gene expression contains a large amount of information about a patient profile, which are characteristics of patient diseases or biological insights. Over the years, many data-driven approaches have been developed to investigate the gene expression profile that explains a patient's condition [123, 252]. In genomic data analysis, high-throughput predictions and optimal visualisations are challenging for researchers. Therefore, many types of genomic research has been conducted to classify and visualise genomic or microarray data. One such approach is gene set enrichment analysis [56]. In addition, there are other approaches that identify genes while others visualise genes [253].

Yuanyuan *et al.* [254] designed a data-driven approach to classify a patient with respect to a particular cancer type using gene expression data. Similarly, Maisa *et al.* [255] applied neural networks (NNs) to investigate whether a patient has cancer or not. Shuguang *et al.* [256] developed an approach to classify the survival status of cancer patients using a supervised learning algorithm.

Ciaramella *et al.* [257] introduced a data-driven approach, FH-Clust, to discriminate patient subgroups from various omics data (e.g., miRNA expression, methylation, gene expression). Another data-driven approach is GeneSetCluster, which can identify gene clusters and visualise gene networks. GeneSetCluster uses an R package to identify a gene cluster of common genes [99] by calculating distances between overlapping genes. K-means and hierarchical clustering algorithms were applied to cluster genes from gene sets. GeneSetCluster represents gene networks in the visualisation phase to show the gene clusters using a dendrogram and a heatmap.

In addition, Wang *et al.* developed [100] GOMCL, a Python-based tool that captures terms (genes or gene sets) from Gene Ontology (GO) to reduce the redundancy of overrepresented genes. Markov clustering (MCL) is used to cluster GO terms from overlapping genes. GOMCL represents GO terms as heatmaps and hierarchical structures for cluster information.

A vaccine investigation ontology (VIO) used the LIMMA statistical model to reanalyze one of the Gaucher and Quebec datasets and found different gene lists to previous studies. The researchers also found that the gene and pathway lists in this study were significantly different from previous studies.

BEAVR is a tool [102] developed with the R language and uses DESeq2 as the engine for differential gene expression analysis. DESeq2 uses statistical tests and displays the results using a heat map, lists of graphs, and plots.

Another web application, GENAVi (Gene Expression Normalisation Analysis and Visualisation), is used to normalise RNA-seq data and display gene clusters in the form of gene expression correlation, principal component analysis, and gene enrichment scores [103].

Yousif *et al*. [104] developed an open-source platform, NASQAR (Nucleic Acid Sequence Analysis Resource) which enables the interactive visualisation of gene clusters. The NASQAR toolbox analyses and visualises transcription from metagenomics and RNA-seq data.

Zhu *et al*. [105] presented new perceptual methods to promote the exploration and exploitation of GO data using a directed acyclic graph. They also presented open-source software for connecting data visualisations. This tool is used for visualisations to identify gene sets based on gene expression.

Natanely *et al*. [106] proposed the Profiler of Multi-Omic data (PROMO) to investigate genomics datasets on cancer and their associated clinical data. PROMO uses different algorithms for prepossessing, visualising and clustering cancer data. PCA and t-SNE are applied for visualisation, k-means, hierarchical clustering, Click for data clustering and GO enrichment analysis for feature selections. Many similar data-driven approaches have been studied and are summarised in Table 2.9.

Table 2.9: Literature review of the research publications on classification methods and visualization tools for genome study analysis (GSA).

Technique	R1	R2	R3
Multi-view integration procedure for distinguishing patient subgroups [257]	×	×	×
Omics data similarity presentation using dendrogram for hierarchical clustering [99]	✓	×	×
Network-based gene weights [258]	✓	✓	×
Novel network-weighted gene-set clustering [259]	✓	×	×
Functional gene networks [260]	×	×	×
Engine for gene regulation [261]	×	✓	×
Scale-able and versatile factor [262]	×	×	×
Overdispersion analysis of gene set [263]	✓	×	×
Transcriptional profiling of microglia [264]	×	×	×
Repesent non-functional overlapped cluster genes [100]	×	×	×
A vaccine investigation ontology (VIO) visualization [101]	×	×	×
Visualize RNA sequence data [102]	×	✓	×
Gene annotation and visualization to normalize RNA sequences [103]	×	✓	×
An interactive visualization to normalize transcription data [104]	×	×	×
Graph theory for the exploration and utilization of GO data [105]	×	×	×
Profiler for examining huge genomics cancer data sets [106]	×	✓	×
Visualization of gene set over-representation analysis [265]	✓	×	×
PathDIP [266]	✓	×	×
Cluster profiler for over-representation gene sets [267]	✓	×	×
topGO [268]	✓	×	×

2.7 Enricher and related tools

Enricher is a web application that provides gene set enrichment analysis for annotated gene sets, i.e., functions of genes. Gene set enrichment finds annotated gene sets that are similar to a given query gene set [269]. Enrichment similarity measures evaluate a set of intersections. The significance of an intersection is evaluated using Fisher's exact test, a statistical significance test used to measure a contingency table which is a matrix that represents frequency distributions of several variables [270]. Fisher's exact test is a statistical test used to determine if there is a significant association between two categorical variables in a 2x2 contingency table [271]. It is often used when sample sizes are small and the assumptions of the chi-square test are not met.

Enricher contains a collection of 400000 annotated gene sets and 300 gene set libraries. Some popular libraries are KEGG (Kyoto Encyclopedia of Gene and Genomes) [6], GO Biological Process , ChEA, Wiki Pathways, Reactome, BioCarta [272], and many more. The annotated genes within Enricher provide insight into specific genes and gene sets associated with diseases, drugs, and biological processes (Table 2.7).

Table 2.10: Literature review of the research publications on gene set Enricher tools.

Technique	R1	R2	R3
Integrated method to visualize enrichment results [269]	×	×	×
Tool to recognize functional advancement of qualities [273]	×	×	×
MutEnricher tool used for researching somatic mutation enhancement [92]	×	×	×
Gene ontology analysis visualisation system [93]	×	×	×
Visually combining approach for expression data [94]	×	×	×
A graphical gene-set enrichment tool [95]	✓	×	×
An open-source gene set analysis tool [96]	✓	×	×
Gene sets annotation [97]	✓	×	×
Functional gene sets application [98]	✓	×	×
A novel web-based tool to compare gene lists [274]	✓	×	×

Zhang *et al*. [273] proposed AllEnricher, a GSEA tool to measure the functional enrichment of gene sets with respect to custom gene set libraries such as KEGG, ChEA, etc. Two statistical tests, namely Fisher's hypergeometric test and Fisher's precise test were used for ranking the gene sets with respect to the enrichment score. Moreover, AllEnricher used R to visualise the ranked gene sets.

MutEnricher, another GSEA enricher tool, was used to identify somatic mutation enrichments in both coding and non-coding genomic regions, which are defined next [92]. The coding genomes are the DNA that encodes proteins and the non-coding genomes do not encode proteins. MutEnricher investigates the transformation of coding and non-coding genes. A software package implemented in Python explores the enrichment of somatic mutation genes.

2.8 Predictive modelling

Predictive modelling is a process of transforming data into actionable future insights. Predictive analytics is a category of data analysis that aims to predict future outcomes based on historical data and analytical techniques such as statistical methods and machine learning. In genomics, advanced tools and models can be used to analyze past and present data, allowing researchers to make reliable predictions about future trends and behaviors in gene expression [275]. Predictive analytics plays a crucial role in this process. For genomics scientists and researchers, this approach can help detect unusual patterns in genomic data. For example, they can identify genes or sets of genes associated with diseases such as cancer [276].

Predictive approaches such as deep learning, neural networks, and decision trees are used to apply gene regulatory networks to identify genes from a large amount of gene expression data.

Kong *et al.* introduced a forest deep neural network (fDNN) to extract genes from the gene set [87]. In the initial stage, the random forest algorithm extracts genes from gene sets. Then, these genes are fitted into the deep network and overfitting genes are identified. Similarly, Arloth *et al.* proposed [88] DeepAWS to investigate single nucleotide polymorphisms (SNPs) and multinucleotide polymorphisms, i.e., variations of DNA nucleotides in an individual (A, T, C, G), using deep learning-based LASSO (Least Absolute Shrinkage and Selection Operator). DeepAWS identifies genes associated with disease-related SNPs from tissues. Roy *et al.* [67] used a supervised classifier to predict genes in invasive ductal carcinoma (IDC). They investigated IDC in two different stages, early IDC and late IDC.

Allahyar *et al.* [89] proposed a synergistic network (SyNet) using a linear regression model to aggregate and rank cancer genes. The authors also used the Spearman correlation coefficient to identify highly correlated breast cancer genes. Another network analysis identifies prognostic cancer genes using the Pearson correlation coefficient [90] between every two genes. In addition, Cytoscape [277] is used to visualise the co-expression and module networks. Permutation of genes is applied to construct module networks and find high density networks.

Moreover, the gene ranking algorithm was used to rank genes and identify important gene networks for cancer. Tong *et al.* [91] proposed a predictive approach to identify colorectal cancer genes by considering high-dimensional data with unsupervised clustering. In addition, a considerable amount of similar research on gene prediction was reviewed and is presented in Table 2.11.

Table 2.11: Literature review of the research publications on predictive approaches to gene set enrichment analysis.

Techniques	R1	R2	R3
Forest deep neural network (fDNN) to extract the features gene sets [87]	✓	×	×
Genome-wide association studies (GWAS) to integrate single nucleotide polymorphism [88]	×	×	×
Supervised classifier for invasive ductal carcinoma (IDC) progression [67]	×	×	×
A gene network approach for gene expression [89]	×	×	×
A module-based network analysis to identify cancer prognostic modules [90]	×	×	×
Prognosis approach to prediction colon cancer [91]	×	×	×
Dirichlet process regression to predict gene expressions [278]	×	×	×
Reduced accuracy to predict genetic information [279]	×	×	×
Subset of genes prediction [280]	×	×	×
Predictive performance of PrediXcan [281]	×	×	×
Testing process for cross-population to predict gene expression [282]	×	×	×
Comparative study of the prediction accuracy [283]	×	×	×
Reconstructed genomes sequencing [284]	×	×	×
Improvements to imputation machinery of genotype [285]	×	×	×
New phasing algorithm for Haplotype phasing genetics [286]	×	×	×
Machine learning method for cardiovascular disease [287]	×	×	×
Machine learning method for lipid traits [288]	×	×	×
Machine learning method for lipid modification [289]	×	×	×
Statistical model for genome wide association [290]	×	×	×
Gene expression (cis expression quantitative trait loci (eQTLs)) [291]	×	×	×
Prioritize causal genes [292]	×	×	×

2.9 Summary of the literature review

According to the discussion in sections 2.4, 2.5, 2.6, 2.7, and 2.8, the literature are primarily focused on identifying genes, predicting diseases, and determining how genes interact with one another.

Section 2.4 discussed on GSEA. GSEA uses collections of genes that have been categorized according to their shared functions. GSEA ranks genes in gene sets based on how strongly they are associated with the presence of a phenotypic trait. So, GSEA can be used to understand the impact of variation in gene expression characteristics. GSEA incorporates gene sets as well as gene expression profiles. GSEA selects a limited number of gene sets with the assumption that the selected gene sets are useful to explain the expression levels of the two classes (experiment and control).

Section 2.5 discussed about knowledge-driven approaches where gene expression values and gene sets are integrated to identify genes. All approaches predicted genes that are related to specific cancer diseases using both machine learning and statistical approaches.

Section 2.6 discussed data-driven approaches where gene expression values are used to predict diseases and genes. According to the literature, many of the data-driven methods discussed in this section involve analyzing genes or diseases using gene expression values.

Section 2.7 discussed tools that used to identify genes or gene sets related to diseases. Similarly, section 2.8 discussed about different AI and machine learning approaches such as deep learning, neural network, support vector machine, k-nearest neighbor, and rule-based systems.

RQ1 provides an opportunity to examine existing research to see if integrating gene sets with gene expression data helps to evaluate abnormalities in complex disease studies. Instead of looking at a few genes in the gene set, all member genes of a gene set will participate in the anomaly analysis. RQ2 focuses on predictive modelling and instance-based learning to propose treatment planning considering imbalanced genomic data samples. Finally, RQ3 focuses on explainable AI. Explainable AI helps to understand why prediction, classification, and cross-validation models are correct or incorrect.

2.10 Research gaps in genomics studies

The above-mentioned scientific literature in the field of genomics focuses mainly on the identification of genes associated with diseases by analyzing gene expression values. To achieve this goal, researchers use one of two main strategies. The first approach, called the data-driven method, focuses primarily on examining gene expression levels. The second approach, called the data-and knowledge-driven method, involves both gene expression values and gene sets and provides a deeper understanding of the relationships between genes and disease.

The overall goal of these research effort is to identify specific genes or gene sets that have a correlation with a particular disease.

In contrast, this thesis argues that gene sets not only provide information about individual genes, but also offer insights into the relationships and interactions between genes within a group. Consequently, examining these gene sets can help researchers better understand the impact of anomalous gene expression patterns and their potential implications in disease development and progression.

In addition, this thesis contends that analysis of variations in gene expression values is crucial for uncovering anomalies that may exist within these values. Such anomalies can contribute to our understanding of the molecular mechanisms underlying various diseases such as leukaemia, colorectal cancer, and breast cancer. By investigating these variations and their potential association with specific diseases, researchers can gain valuable insights that may ultimately lead to the identification of novel treatment planning or improved diagnostic techniques.

Predictive modeling and datadriven approaches are used to predict a patient's condition and gene. But there are ways to look at differences in the way genes are identified through data mining methods and link a patient's disease. There are ways to identify similar cohorts of patients in terms of variations in gene expression levels.

2.11 Hypothesis and research questions on gene set analysis

The systematic literature review on GSEA conducted in this thesis facilitates the adoption of effective methods to address research questions related to anomaly analysis using cancer gene expression data. By carefully examining existing GSEA literature, this thesis identifies relevant statistical methods, algorithms, and techniques that are widely used in this domain. Assessing the strengths and weaknesses of these methods allows for the determination of successful approaches for detecting anomalies in gene expression data, particularly in the context of cancer.

However, much of the existing research primarily focus on identifying specific gene sets associated with specific diseases from limited numbers of differentially expressed genes. However, consideration of the totality of genes when grouped into gene sets for each patient was not considered. This observation has motivated this thesis to develop a novel approach for detecting anomalies in cancer gene expression data when they are considered within gene sets groupings. By building on the insights gained from the literature review and addressing the limitations of previous studies, this thesis aims to develop an approach for analyzing gene expression data in the context of cancer research. As a result, the insights and limitations identified in the literature review shape the methods used in this thesis for addressing the research questions ensuring a well-informed and effective approach.

Based on the literature review and research gaps, this thesis argues that knowledge inherent in gene sets keeps an opportunity for profile-based data analytics by integrating gene sets and gene expression values. In addition, this thesis proposes a hypothesis ***that carefully aggregating gene expression values into measures over gene sets creates a potential for gaining insights from the analysis of data pertaining to gene expression profiles***. Specifically, this thesis proposes that in comparison to using expression values directly, integration of gene sets and gene expression values would be useful to measure anomalies exist in each profile. This anomaly analysis might be effective to enhance visibility and detectability of patterns across distinct expression profile.

- **Research question 1**

  How can aggregation of gene expression values into their respective gene sets provide opportunities for building a knowledge-based classification?

- **Research question 2**

  How can a technique be developed to assess variation of gene expression in each gene set for each patient and build a model using this assessment?

- **Research question 3**

How can we use an approach to map patients with respect to the biology of the patient that determines the response to cancer treatment?

2.12 Conclusion for systematic review on genomics related literature

Upon review, it was not uncommon to see an apparent variation in the methods used to analyse genomic data, from gene prediction to gene set ranking. Surprisingly, this thesis found that there are many ways to understand gene expression values and gene sets for complex disease analysis in an AI-based environment. This study motivated several research opportunities for gene set enrichment analyses to understand the gaps. However, finding solutions requires a remarkable interdisciplinary exercise between biology (science), mathematics (science), and data analysis.

2.13 States-of-the-art methodologies

Raw gene expression data can be processed in a number of ways and several existing pre-processing approaches have been reviewed. In next sections, I provide an overview of some of these pre-processing methods that generate modified gene expression values from raw gene expression values.

2.13.1 Gene fuzzy score

Abha and colleagues [293] proposed the Gene Fuzzy Score (GFS), a modified gene expression score derived from raw gene expression values. The term “fuzzy” in GFS means that the gene expression values are converted to a range between zero and one using a mathematical equation other than the fuzzy membership function, which is detailed in (Equation 2.2). The processing steps of GFS are as follows.

1. The raw gene expression values for each patient are ranked in ascending order. The position of a gene does not change in this ranking. The expression values for each sample are simply ordered from lowest to highest values.

2. The ordered gene expression values are processed to generate the GFS. To process the ordered gene expression data, two thresholds are considered, which are θ_1 and θ_2. Gene expression values above θ_1 are considered 1, and gene expression values below θ_2 are considered 0. Equation 2.2 is used to convert gene expressions between zero and one that exist between θ_1 and θ_2.

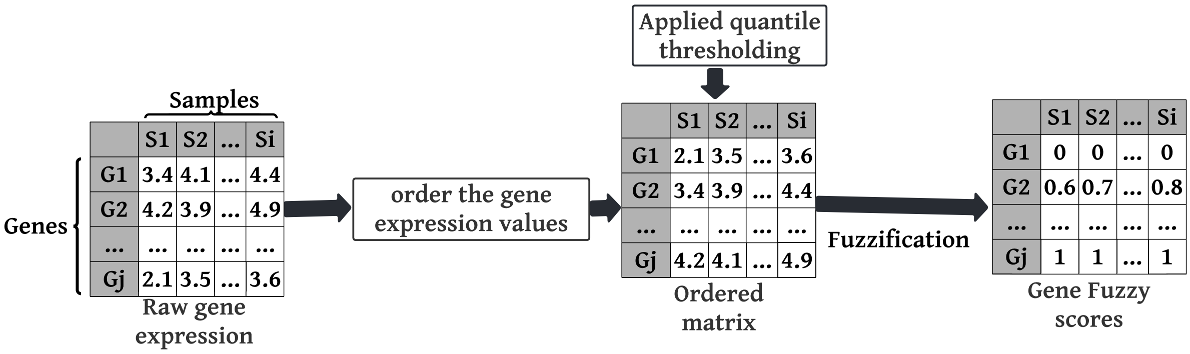
The processing steps of GFS are shown in Figure 2.8, beginning with ranking the raw gene expression values, splitting them into two thresholds of θ_1 and θ_2, and then converting the remaining gene expression values to 0 and 1 using equation 2.2.

$$f(g_i, p_j) = \begin{cases} 1,\ if\ q(p_j, \theta_1) < r(g_i, p_j) \\ \frac{r(g_i,p_j) - q(p_j,\theta_2)}{q(p_j,\theta_1) - q(p_j,\theta_2)},\ if\ q(p_j, \theta_1) > r(g_i, p_j) \geq q(p_j, \theta_2) \\ 0,\ Otherwise \end{cases} \tag{2.2}$$

where $f(g_i, p_j)$ is the fuzzy score for a gene g_i in patient p_j, and $r(g_i, p_j)$ is the rank of gene expression of a gene g_i in patient p_j, and $q(p_j, \theta_1)$ be the rank corresponding to the quantile thresholds of gene expression in patient p_j.

Objectives

- This pre-processing technique focused to find consistent results across different data sets for different phenotypes (e.g., healthy patients, cancer patients).
- To identify highly expressed genes from all given data sets.
- To ensure coherent groups of genes differ from other genes with respect to their GFS scores.



	S1	S2	...	Si
G1	3.4	4.1	...	4.4
G2	4.2	3.9	...	4.9
...	...	...	...	...
Gj	2.1	3.5	...	3.6

	S1	S2	...	Si
G1	2.1	3.5	...	3.6
G2	3.4	3.9	...	4.4
...	...	...	...	...
Gj	4.2	4.1	...	4.9

	S1	S2	...	Si
G1	0	0	...	0
G2	0.6	0.7	...	0.8
...	...	...	...	...
Gj	1	1	...	1

Figure 2.8: Processing steps of gene fuzzy scores.

2.13.1.1 Pros, cons and uses of GFS

Pros

- The GFS is a normalized gene expression score that scales raw gene expression data between 0 and 1. Top 15-20% of GFS scores were considered for cancer analysis.
- GFS ranked genes and embeds patients with respect to the highest-ranked genes.

Cons

- Only a limited portion of gene expression data was included in the GFS.

- Due to exclusion of a substantial proportion of gene expression values from the GFS, there is a possibility that some valuable gene expression data were overlooked.
- Patient embedding with respect to gene may not captured full biological functions as GFS excluded some genes.

Uses

- GFS has been used to show clustering of patients with respect to genes in two-dimensional plots.

2.13.2 Feature regression and classification (FRaC)

Keith and colleagues [294] developed a feature regression and classification (FRaC) approach which generates modified gene expression values from raw gene expression values. Initially, input data were divided into training and test samples and the model was trained using the training samples. Then, FRaC predicts the test gene expression values and calculates the prediction errors. The prediction errors are considered as pre-processed gene expression values. The steps performed by FRaC are as follows.

1. The input data is divided into training and target attributes (test gene expression values).
2. A regression and classification method is selected to predict target attributes.
3. Prediction errors are measured for each target gene expression value, i.e., the modified expression score for measuring changes in genes.

Figure 2.9 shows the processing steps of FRaC to identify prediction errors in raw gene expression values. First, FRaC sets some features as target features while other features are used to train the supervised prediction model to predict the target features. Then, the prediction errors for each target gene expression values are measured.

Consider a set of training samples, $X = \{X_1, X_2, \cdots, X_N\}$ with D number of features for each sample, $X_j = \{X_{j1}, X_{j2}, \cdots, X_{jD}\}$. First, FRaC selects the target features whose value is to be predicted and other features that will be used as training features. Separating the test sample from the training sample is crucial to avoid bias. If test and training data have similarities, it may result in higher accuracy than anticipated, which might not represent the model's actual performance on unseen data. By employing distinct test and training datasets, the model's performance can be evaluated more accurately, guaranteeing its effectiveness in generalizing to new, unseen cases.

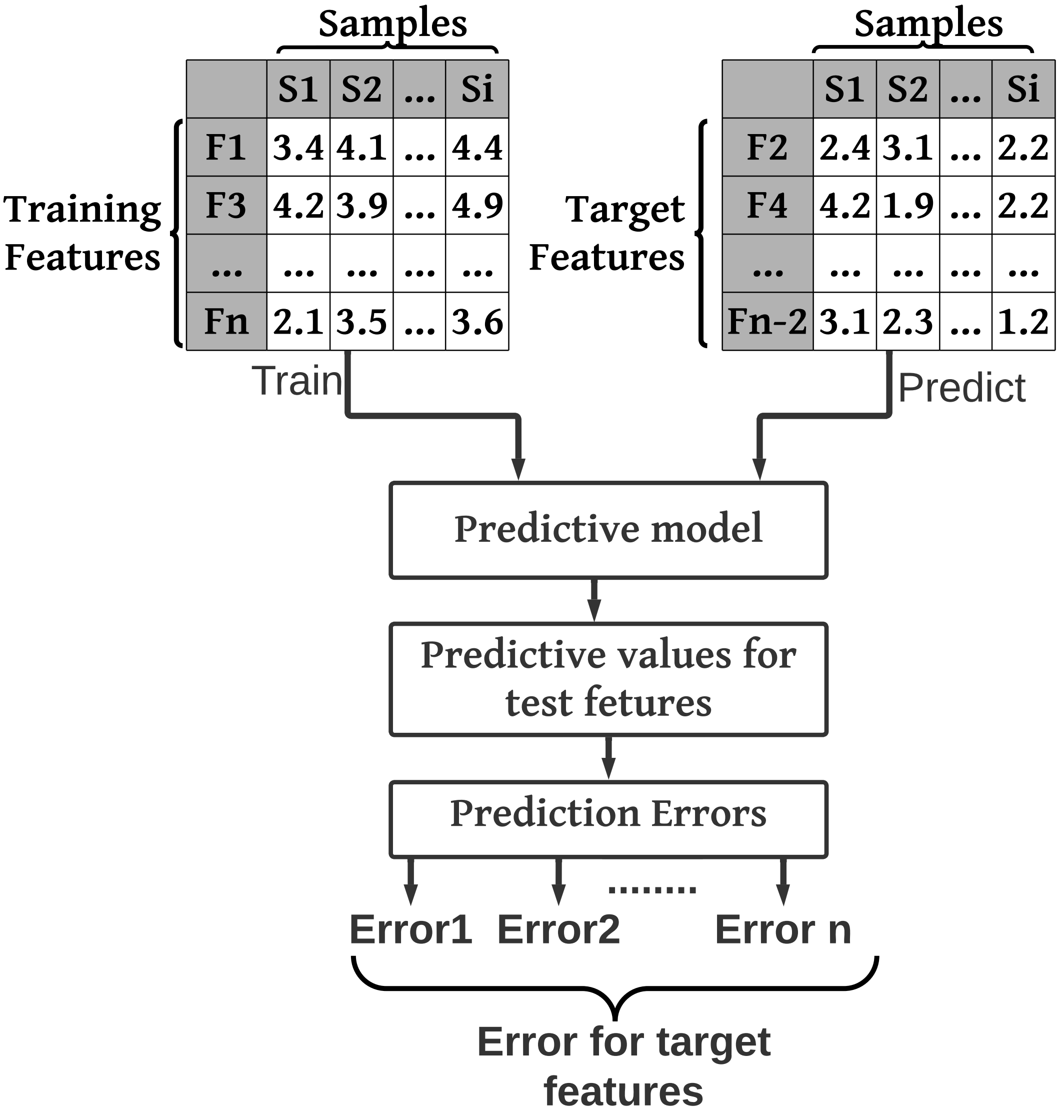
After separating the training and test features, FRaC selects supervised learning algorithms to predict the feature values. FRaC selects multiple predictor models for each feature prediction. Here, ρ_i refers to the set of predictor features used to predict the target feature i. Thus, the predictor feature for a sample X_j can be defined as follows,

$$\rho_i(X_j) = \langle x_{j1}, x_{j2}, \cdots, x_{j,i-1}, x_{j,i+1}, \cdots, x_{jD} \rangle \tag{2.3}$$

For each target gene, FRaC calculates a modified expression score based on a comparison of predicted values to actual values. The following equation is used to calculate the prediction error.

$$error(x_q) = \sum_{p=1}^{P}\sum_{i=1}^{D} distance(x_{qi}, C_{p,i}(\rho_i(x_q))) \tag{2.4}$$

where x_q is the target sample, $C_{p,i}$ is the classification predictor, P is the supervised learning model, and D is the features of the training sample.



Training Features:

	S1	S2	...	Si
F1	3.4	4.1	...	4.4
F3	4.2	3.9	...	4.9
...	...	...	...	...
Fn	2.1	3.5	...	3.6

Target Features:

	S1	S2	...	Si
F2	2.4	3.1	...	2.2
F4	4.2	1.9	...	2.2
...	...	...	...	...
Fn-2	3.1	2.3	...	1.2

Figure 2.9: Processing steps of FRaC on raw gene expression values.

Figure 2.10 shows an example of identifying prediction errors using raw gene expression values in leukemia with FRaC. The figure shows that three values of the target attributes DDR1, FIP1L1, and RFC2 are predicted with respect to four training attributes (SLC5A11, TAZ, ZNF133, and PPMIF). First, the model was trained with these four training samples and then the test samples were predicted with FRaC. Finally, the prediction errors were calculated by measuring the differences between the values of the original attributes and the predicted values of these target attributes.

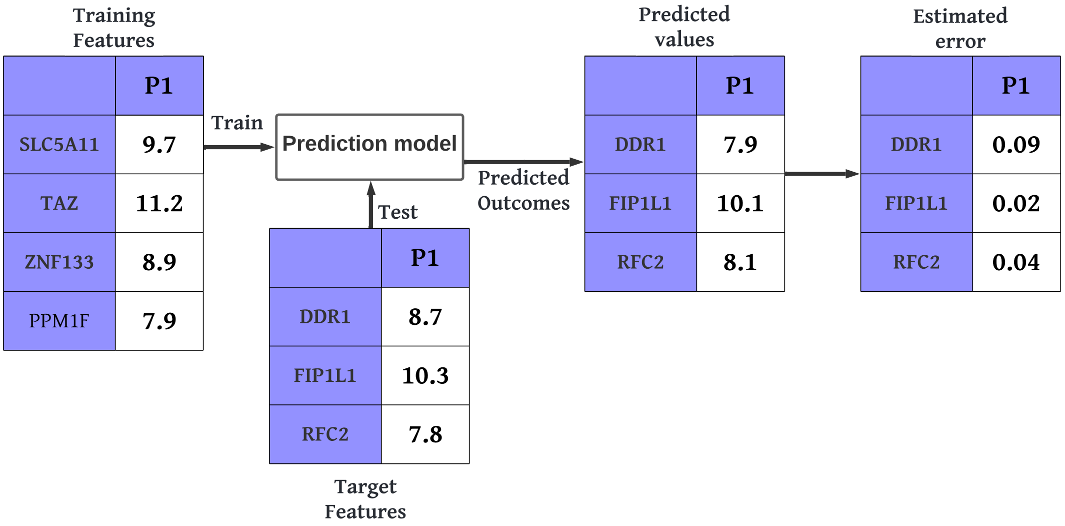


Figure 2.10: An example of prediction error identification for leukaemia gene expression values.

Objectives

- To establish a relationship between genes of two different distributions.
- To generate modified gene expression scores by considering noisy and irrelevant data.

2.13.2.1 Pros, cons and uses of FRaC

Pros

- FRaC predict gene using feature regression and classification approach and measures prediction errors.

Cons

- FRaC determines “variation scores” based on calculated prediction errors. However, it’s important to understand that these prediction errors might reflect limitations of the algorithm rather than actual variations in gene expression. In other words, just because an algorithm fails to predict accurately, it doesn't necessarily mean there is a variance in the gene expression values.

Uses

- A machine learning approach to predict genes with respect to training data sets.

2.13.3 Characterizing systematic anomalies in expression values (CSAX)

Keith and colleagues [295] introduced characterizing systematic anomalies in expression data (CSAX) to identify anomalous gene sets or pathways for a patient. The processing steps of CSAX are as follows.

1. The input data is divided into two groups: healthy and diseased. The healthy group is treated as a training sample and the diseased group is the test sample.

2. CSAX uses FRaC [294] to identify prediction errors in raw gene expression values for the test samples. This step produces modified gene expression values for each test attribute.

3. Using the modified gene expression values from the previous steps, CSAX uses GSEA [9] to select the number of gene sets that are at the top of the list.

Figure 2.11 shows the processing steps of CSAX. The figure shows that the input data are divided into two groups, training and test samples. The training data are all healthy gene expression values and the test data are a combination of both diseased and healthy gene expression values. CSAX uses FRaC to generate modified gene expression values by measuring the prediction errors for each test gene. After the modified gene expression values are obtained, CSAX applies GSEA. CSAX measures the gene set enrichment score to identify gene sets. The modified gene expression scores are calculated using the following equation.

$$anomaly\ score = \sum_{i=1}^{G} \gamma^{i-1} \times ES(M_i) \tag{2.5}$$

where G is the number of gene sets, γ controls how many gene sets are included in the calculation of the enrichment score, CSAX sets $\gamma = 0.95$, and M is the position of the gene sets in the ranking list.

This thesis applies the CSAX approach to acute lymphoblastic leukaemia gene expression scores (ALL) to generate modified raw gene expression scores. The input data were divided into two groups: normal and cancer patients. Then, CSAX calculates the prediction errors for the molecular function ontology gene sets (C5_ MF).

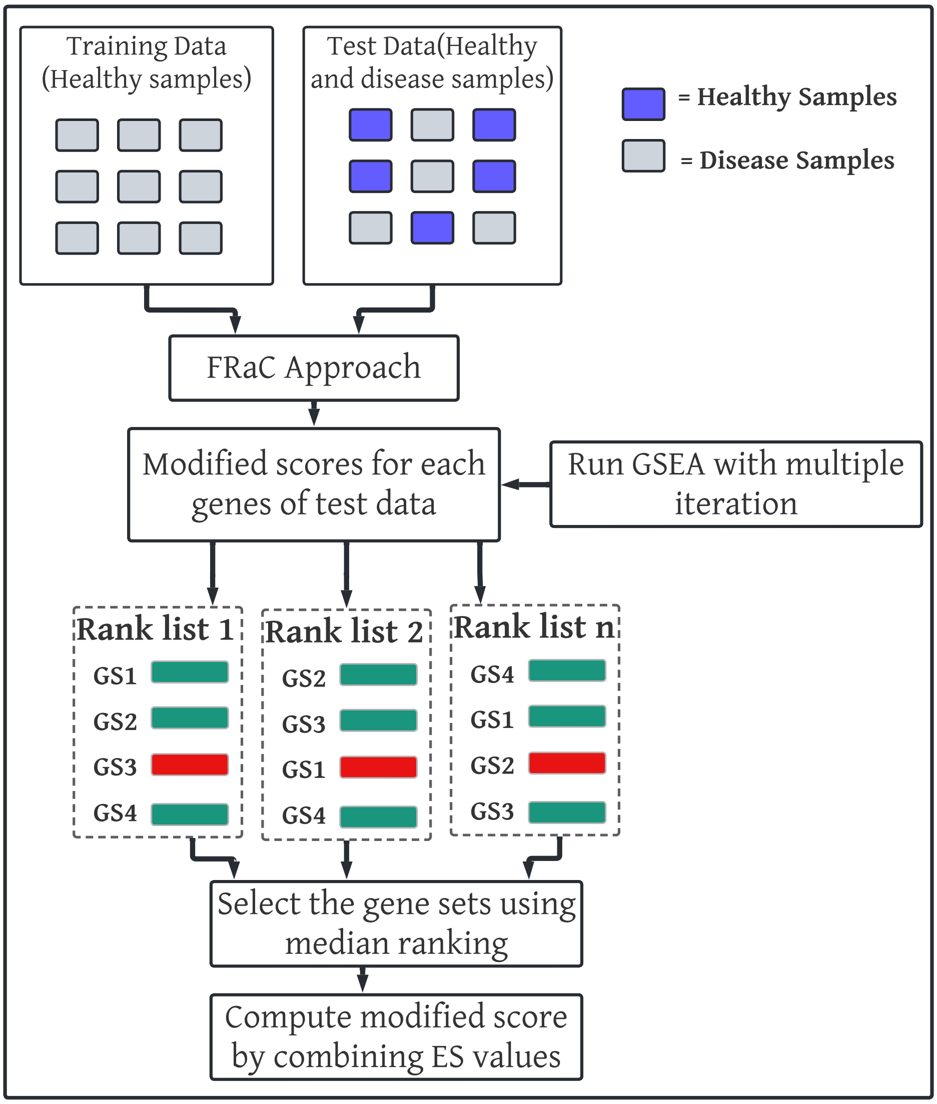

Training Data (Healthy samples)
Test Data(Healthy and disease samples)
= Healthy Samples
= Disease Samples
FRaC Approach
Modified scores for each genes of test data
Run GSEA with multiple iteration
Rank list 1
GS1
GS2
GS3
GS4
Rank list 2
GS2
GS3
GS1
GS4
Rank list n
GS4
GS1
GS2
GS3
Select the gene sets using median ranking
Compute modified score by combining ES values

Figure 2.11: Processing steps of CSAX.

Objective

- To identify anomalous gene sets or pathways.

2.13.3.1 Pros, cons and uses of CSAX

Pros

- CASX trained feature regression and classification approach using training data from healthy patients.
- CSAX uses a feature regression and classification to predict genes and then measures prediction errors of its predictions.

Cons

- lCSAX generates varied gene expression scores by computing prediction errors from the expected value. However, these prediction errors may actually indicate limitations of the algorithm itself. Therefore, these prediction errors shouldn't be interpreted as gene variation scores.

Uses

- A machine learning approach for predicting genes based on training data sets.

2.13.4 TEMPO: Detecting pathway-specific temporal dysregulation of gene expression in disease

Christopher and colleagues [296] proposed temporal modeling of pathway outliers (TEMPO) to identify changes in gene expression over time. This approach uses partial least squares regression (PLSR) [297], a feature classification technique, to measure prediction errors in the collected data sets. Collected gene expressions were divided into test and training samples, with healthy gene expressions considered as training samples and diseased expressions as test samples. PLSR was trained with a healthy training sample to predict the diseased sample and measure the prediction error. The processing steps of TEMPO are as follows.

1. The input data is separated into two phenotype groups: control and diseased.
2. Partial least squares regression (PLSR) is used to train the samples.
3. After the method is trained, the ages of the gene expressions (diseased) were predicted using the leave-one-out method and the prediction errors are measured.

Figure 2.12 shows the computational steps of TEMPO using PLSR. From the figure, it can be seen that the input data were divided into control (healthy) and diseased groups. TEMPO uses the gene expression data from the control group to train the PLSR model and predict the age of gene expression for the diseased samples. Multiple PLSRs were used to identify prediction errors. Finally, TEMPO calculates the prediction errors from the actual and predicted ages.

TEMPO generates a prediction error for each gene set G, which is measured using the differences between the predicted ages and actual ages. Here,$E_{G,s}$ is the prediction error for each sample,s under the gene set,G. TEMPO calculates a score using these prediction errors to find out which gene set, G is temporally dysregulated, which is calculated using the following equation.

$$score(G) = \frac{|C| \sum_{s \epsilon D} -log(\mathbb{N}(E_{G,s}))}{|D| \sum_{s \epsilon C} -log(\mathbb{N}(E_{G,s}))} \tag{2.6}$$

where the probability error $\mathbb{N}(E_{G,s})$ is calculated using the z-score with the normal distribution with mean μ_G and standard deviation, σ_G, D is the disease sample and C is the control sample.

Objectives

- To identify the temporal changes in raw gene expression for a particular disease.

- To identify the temporal dysregulation of biologically relevant gene sets for different diseases.

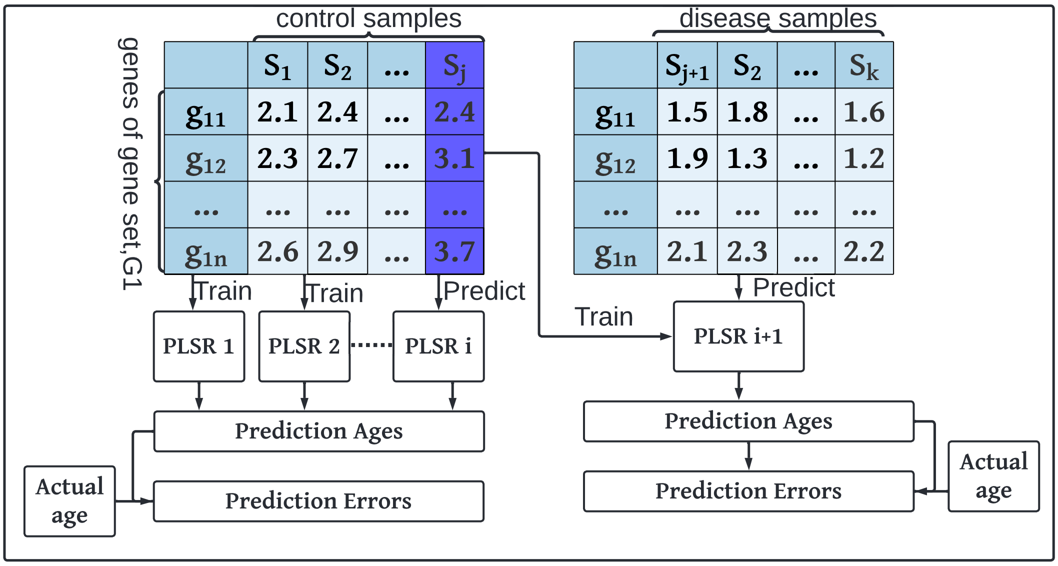


Figure 2.12: Computing prediction errors for groups of genes using PLSR.

2.13.4.1 Pros, cons and uses of TEMPO

Pros

- TEMPO identify prediction errors in gene expression using the feature classification approach, which is PLSR. TEMPO measures variations by predicting a gene expression values for a test gene and then measures the differences between test and taring values.

- TEMPO Incorporates both microarray and RNA-seq data sets to predict gene expression for a given test gene.

- TEMPO extract gene sets using GSEA tool.

Cons

- In order to get modified gene expression score, TEMPO computes variation of gene expression values from predicted value.

Uses

- A machine learning approach for predicting genes based on training data sets.

2.13.5 aTEMPO: anomaly temporal modeling of pathway outliers

Christopher and colleagues [298] introduced anomaly temporal modeling of pathway outliers (aTEMPO), an extension of TEMPO [296], to identify changes in gene expression values with respect to time. In contrast to TEMPO, aTEMPO uses microarray Significant Profiles (maSigPro) [299], a regression analysis method to find outliers genes. maSigPro is used to rank genes. After the genes are classified, the prediction errors are measured by predicting a test sample. The processing steps of aTEMPO are as follows.

- The input data is divided into two groups of phenotypes: healthy and diseased.
- Microarray Significant Profiles (maSigPro) [299], a regression analysis method, is applied to rank the gene expression values of normal samples.
- The gene sets are identified using GSEA [9] taking into consideration the ranks of genes determined in step 2.
- Fourth step is to train Partial Least Squares Regression (PLSR) [297] with training samples.
- The age of the gene expressions (diseases) is predicted with PLSR and cross-validated using the leave-one-out method and the prediction errors for the corresponding genes are measured.

Figure 2.13 shows the computational steps of aTEMPO where the blue square represents the training data (control samples) and the green square represents the test data (diseased samples). In both datasets, the rows contain the genes $(g_{11}, g_{12}, \cdots, g_{22})$ and the columns contain the samples $(S_1, S_2, \cdots, S_j, S_{j+1}, \cdots, S_k)$. In the next phase, i number of PLSR is trained on j training samples and the age $P_{G,c}$ is predicted. From the actual ages A_c and the predicted ages $P_{G,C}$, aTEMPO calculates the error $E_{G,C}$. Similarly, the predicted error $E_{G,D}$ is measured from the diseased data using Single PLSR, M_{test}. Finally, the error probability for the gene set G and sample s is calculated as follows.

$$L_{G,s} = -log(\mathbb{N}_G(E_{G,S})) \tag{2.7}$$

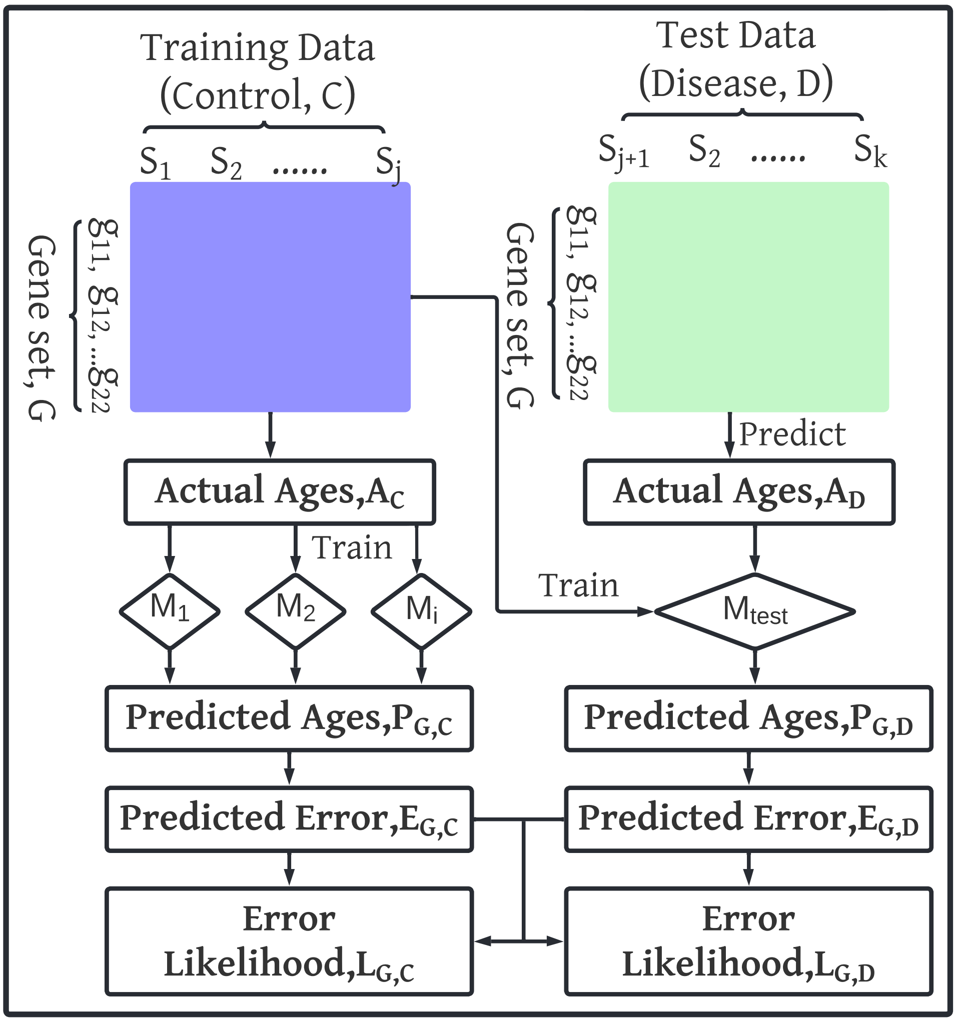


Figure 2.13: Processing steps of aTEMPO gene expression scores.

where $E_{G,S}$ is the prediction error calculated from the difference between the predicted and actual ages. $\mathbb{N}_G$ is the probability estimate calculated from the z-score using the mean and standard deviation of the prediction errors.

2.13.5.1 Pros, cons and uses of aTEMPO

Pros

- aTEMPO is a pathway-based outlier detection approach which used FRaC to detect prediction errors of test genes. aTEMPO predicted gene expression values with respect to age of training data (temporal data) and measures predicted errors.

- aTEMPO extracted gene sets using GSEA tool.

Cons

- In order to get modified gene expression score, aTEMPO computes variation of gene expression values from predicted value.

Uses

- aTEMPO identified gene sets from GSEA with respect to the prediction errors.

2.13.6 Outlier detection in gene expression data

Anindya and colleagues [300] describe an approach for identifying outliers in gene expression values, where outliers in gene expression are observations that quantify an abnormal distance between one sample and another, as opposed to the normal distance between two samples. The procedure for conducting this study is as follows.

1. The Pearson correlation between all genes is calculated.

2. A numerical value is assigned to each gene. This is due to the fact that after applying Pearson correlation, some gene-to-gene correlation values may be 1, resulting in an outlier situation. In this strategy, a numerical value is assigned to each gene to solve the problem.

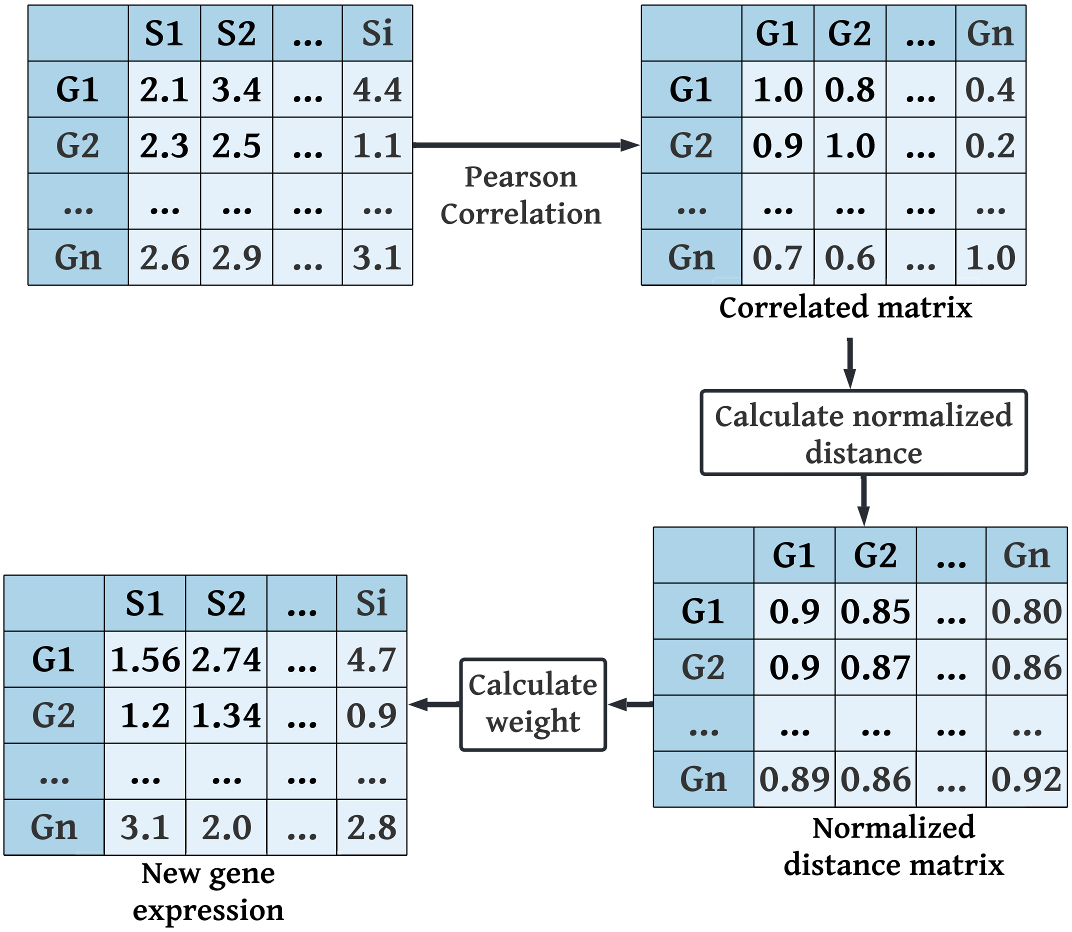
3. The normalized distance for each gene is measured. There is a relationship between the normalized distance and the assigned weight for each gene. If the normalized distance is smaller, the weight is larger, and if the distance is larger, the weight is smaller.

The steps of outlier detection approach are shown in figure 2.14. Gene expression values are shown in the first table in this figure, with rows representing genes and columns representing samples. Using Pearson correlation, this approach then provides a gene-to-gene relationship matrix indicating the degree of relationships between genes. Next, a normalized distance matrix is created from the correlation matrices and a weight is assigned to each gene based on the distance matrices. Finally, the new weight is multiplied by the gene expression values, resulting in new modified scores.

Let us consider a set of n genes $X = \{g_1, g_2, \cdots, g_n\}$ with $m-$dimensional gene expression. The similarity between the gene pair (g_i, g_j) is calculated using the Pearson correlation $corr(x_i, x_j)$, which is defined as follows:

$$corr(x_i, x_j) = \frac{\sum_{l=1}^{m}(x_{il} - \overline{x}_i)(x_{jl} - \overline{x}_j)}{\sqrt{\sum_{l=1}^{m}(x_{il} - \overline{x}_i)^2 \sum_{l=1}^{m}(x_{jl} - \overline{x}_j)^2}} \tag{2.8}$$

where x_{il} and x_{jl} are the gene expression values of i^th and j^th, respectively, and l is the sample. This similarity value is biased due to outliers. This similarity value should be close to 1 for coregulated genes, but due to outliers, the correlation value deviates greatly from 1. To solve this



	S1	S2	...	Si
G1	2.1	3.4	...	4.4
G2	2.3	2.5	...	1.1
...	...	...	...	...
Gn	2.6	2.9	...	3.1

	G1	G2	...	Gn
G1	1.0	0.8	...	0.4
G2	0.9	1.0	...	0.2
...	...	...	...	...
Gn	0.7	0.6	...	1.0

	G1	G2	...	Gn
G1	0.9	0.85	...	0.80
G2	0.9	0.87	...	0.86
...	...	...	...	...
Gn	0.89	0.86	...	0.92

	S1	S2	...	Si
G1	1.56	2.74	...	4.7
G2	1.2	1.34	...	0.9
...	...	...	...	...
Gn	3.1	2.0	...	2.8

Figure 2.14: Processing steps of outlier gene identification using raw gene expression values.

problem, the distance for each gene must be normalized. The normalized distance is calculated as follows.

$$D_{ijl} = \sqrt{(t_{il} - \bar{t}_i)^2 + (t_{jl} - \bar{t}_j)^2} \tag{2.9}$$

where D_{ijl} is the normalized distance, t_{il} and t_{jl} are the normalized gene expression values of expression x_{il} and x_{jl}. $\bar{t}_i)$ and $\bar{t}_j)$ are the mean of the normalized distance. Next, the weight is calculated for each pair of genes. When the normalized distance is smaller, the weight is higher and when the distance is larger, the weight is lower. The weight is defined for the gene pair (g_i, g_j) as follows.

$$w_{lij} = e^{-\alpha \times D_{ijl}} \tag{2.10}$$

where w_{lij} is the weight for the gene pair and $\alpha \geq 1$ is a constant. The new gene expression values are the product of the weight and the original gene expression, which is defined as follows:

$$
\begin{aligned}
x^w_{ijl} &= w_{lij} \times x_{il} \\
x^w_{jil} &= w_{lij} \times x_{jl}
\end{aligned}
\tag{2.11}
$$

where x^w_{ijl} and x^w_{jil} are the new gene expression for x_{il} and x_{jl}, respectively.

2.13.6.1 Pros, cons and uses of outlier detection

Pros

- Pearson correlation and Euclidean distance were used in this method to measure an amount of variation in gene expression data. According to the distances, the genes could be divided into two categories: those that were very similar to one another and those that were very different.

Cons

- Excluded a group of genes based on their similarities and differences. Employing the Pearson correlation coefficient in outlier detection might not be optimal due to its sensitivity to outliers. Future advancements in this field could consider alternative measures, such as the Spearman rank correlation [301]. This is a statistical technique that helps understand how closely two sets of data are linked. It works well even when the data isn't evenly distributed and is not as affected by outliers as the Pearson correlation [302].

Uses

- This approach identified a group of similar genes.

2.13.7 SNet: finding consistent disease sub-networks

Donny and colleagues [303] introduced SNet, which processes gene expression values by considering two thresholds. These thresholds generate a SNet gene expression score from raw gene expression values. This modified gene expression value is used to identify the relationship between genes in a gene network. The processing steps of SNet are as follows.

1. The gene expression values for each patient are ranked in ascending order. The position of a gene does not change in this ranking. The expression values for each sample are simply ranked from lowest to highest.
2. Ordered gene expression data is processed. Two threshold values are considered, $\alpha\%$ and $\beta\%$. SNet sets the value of $\alpha\%$ to 10% to extract the 10% best genes from the highest-ranked genes.
3. Specific gene expression values are selected. After selecting 10% of the genes in step 2, SNet assigns $\beta\%$ to 50%. This means that SNet selects the 50% best genes from the originally selected 10% best genes.

Figure 2.15 shows the processing steps of SNet. The first step in the figure is to rank the gene expression values. Then, thresholds are assigned to the tables of ranked gene expression values, and finally, the modified gene expression values are constructed for further processing. As shown in the figure, SNet ranks the raw gene expression values and extracts 10% of the genes from the top genes. Finally, 50% of the genes were selected from the top 10% of the genes.

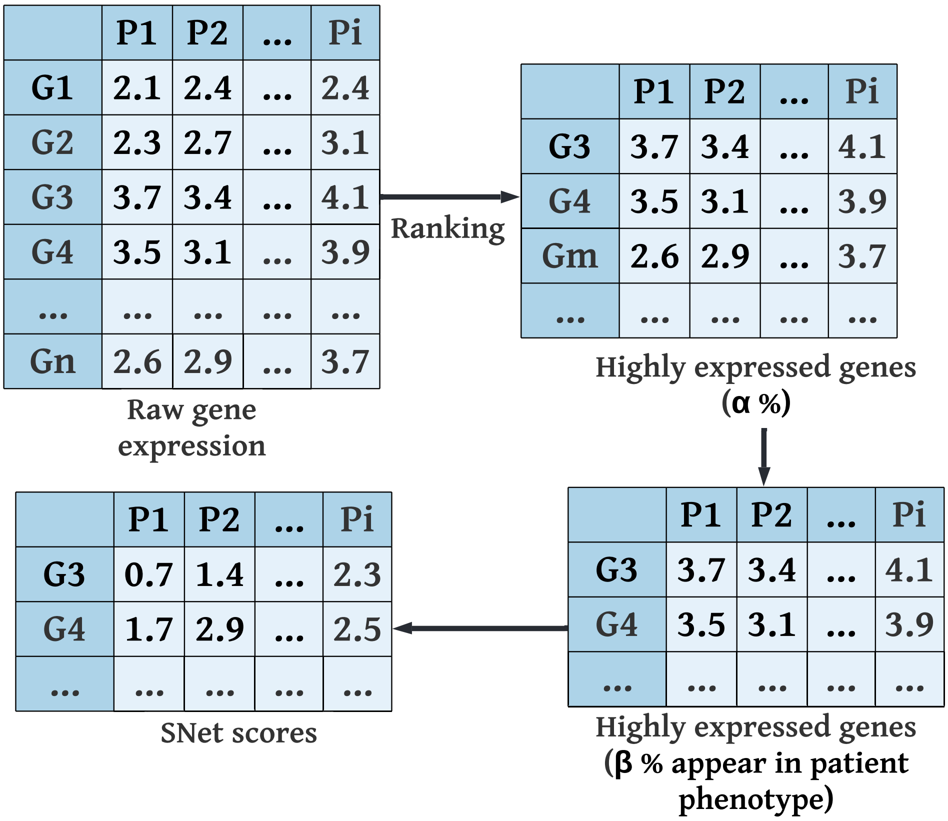
2.13.7.1 Pros, cons and uses of SNets

Pros

- SNets ranked genes according to their modified gene expression scores and included only top 10% of the total genes.

Cons

- The ranking list generated by SNets did not include 90% percent of the total genes.



	P1	P2	...	Pi
G1	2.1	2.4	...	2.4
G2	2.3	2.7	...	3.1
G3	3.7	3.4	...	4.1
G4	3.5	3.1	...	3.9
...	...	...	...	...
Gn	2.6	2.9	...	3.7

Raw gene expression
Ranking

	P1	P2	...	Pi
G3	3.7	3.4	...	4.1
G4	3.5	3.1	...	3.9
Gm	2.6	2.9	...	3.7
...	...	...	...	...

Highly expressed genes (α %)

	P1	P2	...	Pi
G3	3.7	3.4	...	4.1
G4	3.5	3.1	...	3.9
...	...	...	...	...

Highly expressed genes (β % appear in patient phenotype)

	P1	P2	...	Pi
G3	0.7	1.4	...	2.3
G4	1.7	2.9	...	2.5
...	...	...	...	...

Figure 2.15: Processing steps of SNet.

Uses

- SNets built sub-gene networks based on similarities between genes.

2.13.8 Finding consistent disease sub-networks using PFSNet

Kevin and colleagues [304] developed PFSNet, which processes raw gene expression values and generates PFSNet scores. PFSNet, unlike the SNet approach, is an extension that circumvents the limitations of the SNet approach. Two thresholds and a fuzzy membership function are applied to the gene expression value to generate new modified expression values. The processing steps of PFSNet are as follows.

1. The gene expression values for each patient are ranked in ascending order.
2. To generate modified gene expression values, PFSNet considers two thresholds, namely θ_1 and θ_1. Gene expression values above θ_1 are considered 1, and gene expression values below θ_2 are considered 0. For gene expression values between θ_1 and θ_2, the following fuzzy membership function is applied to convert them to a value between 0 and 1.

$$f_s(x) = \begin{cases} 0,\ x \leq a \\ \frac{x-a}{b-a},\ a < x \leq b \\ 1,\ x = b \\ \frac{c-x}{c-b},\ b < x \leq c \end{cases} \tag{2.12}$$

   where $f_s(x)$ is the fuzzy value for gene expression, and a,b,c are the minimum, mean and maximum gene expression values in the datasets, as shown in Figure 2.17.

Figure 2.16 shows the processing steps of PFSNet. The figure shows that gene expression values above θ_1 are converted to 1 and gene expression values below θ_2 are converted to 0. Gene expression values between θ_1 and θ_2 are converted to a value between 0 and 1 as shown in Figure 2.17.

2.13.8.1 Pros, cons and uses of PSNets

Pros

- PSNets is an updated version of SNets. PSNets converted raw gene expression values to zero and one using a fuzzy function.

- PSNets included a relatively small number of genes, each of which had a value of one, which enabled identification of highly expressed genes.

Cons

- The ranking list generated by SNets did not include a large section of genes.

Uses

- PSNets built sub-gene networks based on similarities between genes.

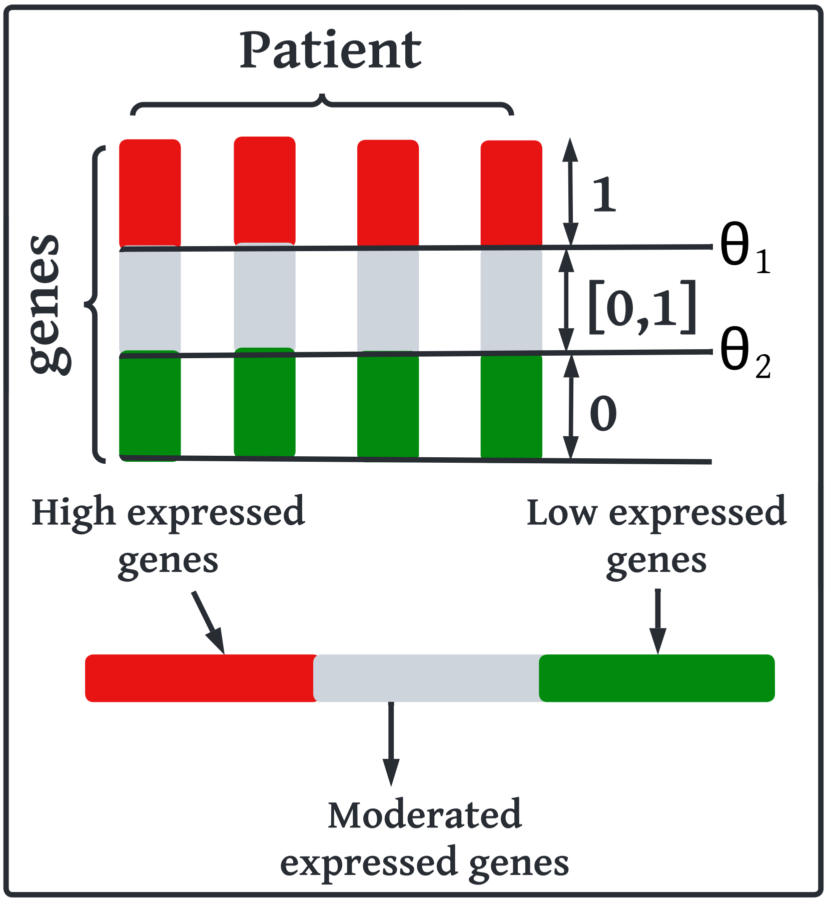


Figure 2.16: Fuzzification of gene expression values to between 0 and 1.

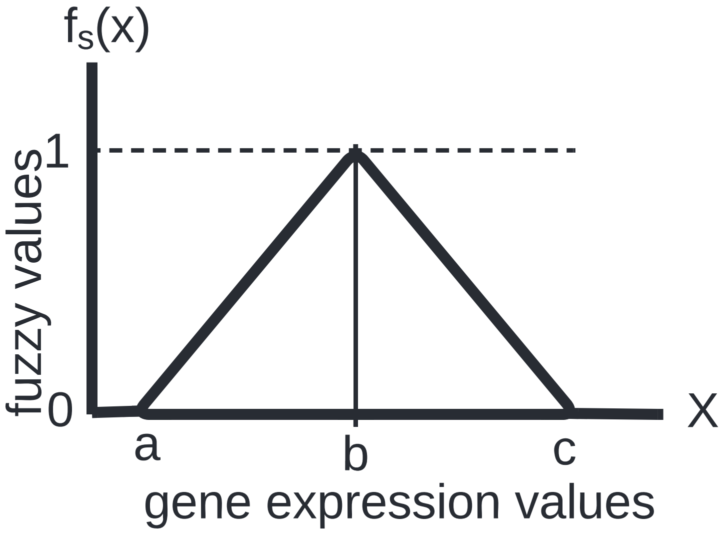


Figure 2.17: Triangular fuzzy values conversion.

2.13.9 qPSP: Quantitative proteomics signature profiling

Wilson and colleagues [305] proposed quantitative proteomic signature profiling (qPSP) that converts gene profiling data into fuzzy values. The steps of this method are as follows.

1. The gene expression values for each patient are ranked in ascending order based on the profiling values.
2. qPSP sets two thresholds, alpha1 and alpha2 and selects the 10% best genes considering the threshold alpha1. Gene expression values above alpha1 are considered as 1. qPSP then selects the next 10% of genes using the threshold alpha2.
3. qPSP determines the rank weight for the top 10-20% genes using four different ranges. For 10-12.5%, the weight is 0.80; for 12%-15%, 0.60; for 15%-17.5%, 0.40; and for 17.5% to 20%, 0.20. All other proteins outside alpha2 receive a weight of 0. All proteins above alpha1 receive a weight of 1 (Figure 2.18).

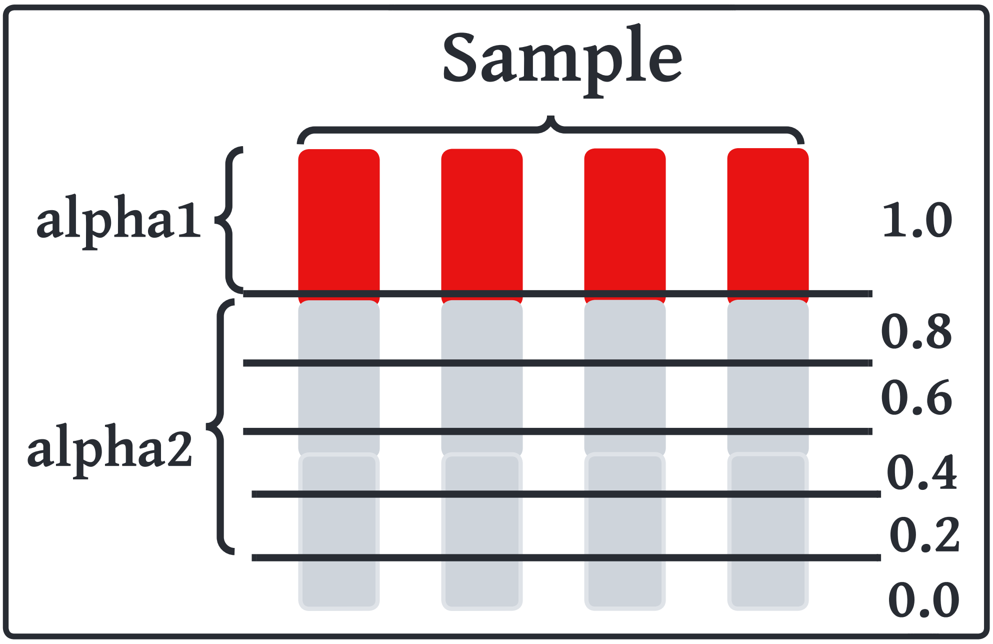


Figure 2.18: Fuzzification of raw gene expression values using qPSP.

2.13.9.1 Pros, cons and uses of qPSP

Pros

- qPSP produced modified gene expression scores of range between zero and one using a fuzzy function.

Cons

- qPSP did not use measure any variation in gene expression values.

Uses

- qPSP used to identify gene clusters.

2.13.10 Eigfusion: Outlier cancer gene detection with potential rearrangements

Alshalalfa and colleagues [306] proposed EigFusion, which rearranges genes by deletion, fusion, and overrepresentation to identify outlier genes from raw gene expression values. Eigfusion evaluates outlier genes based on the generated modified gene expression values. The steps performed by Eigfusion are as follows.

1. The raw gene expression values for the cancer samples are normalized by measuring the median values.

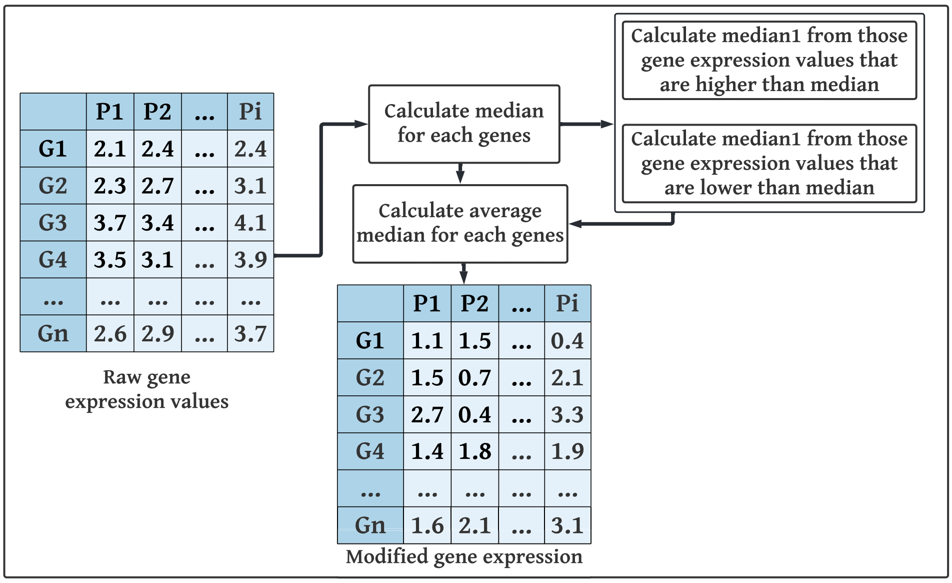
2. The cancer samples are divided into two groups according to the median values, where values that are greater than the median are group one and values that are less than the median value are group two. From these two groups, Eigfusion measures two median (one median from each group) values and measures the average median value of these two medians.

3. The modified gene expression scores were measured using the following equation.

$$\widehat{X}_{ij} = \frac{X_{ij} - AVG_{median}}{median(|X_{ij} - median_i|)} \tag{2.13}$$

   Here $\widehat{X}_{ij}$ is the modified gene expression values, $median_i$ is the median value of $gene_i$ for all the profiles.

Figure 2.19 shows the processing steps of Eigfusion. The first table shows the raw gene expression values. From this data, Eigfusion measures the median value. Then the samples are divided into two groups: gene expression values that are greater than the median value and gene expression values that are less than the median value. Finally, the modified scores are generated from the average median value.



	P1	P2	...	Pi
G1	2.1	2.4	...	2.4
G2	2.3	2.7	...	3.1
G3	3.7	3.4	...	4.1
G4	3.5	3.1	...	3.9
...	...	...	...	...
Gn	2.6	2.9	...	3.7

Raw gene expression values

Calculate median for each genes

Calculate median1 from those gene expression values that are higher than median

Calculate median1 from those gene expression values that are lower than median

Calculate average median for each genes

	P1	P2	...	Pi
G1	1.1	1.5	...	0.4
G2	1.5	0.7	...	2.1
G3	2.7	0.4	...	3.3
G4	1.4	1.8	...	1.9
...	...	...	...	...
Gn	1.6	2.1	...	3.1

Modified gene expression

Figure 2.19: Processing steps of Eigfusion to detect outlier genes.

2.13.10.1 Pros, cons and uses of Eigfusion

Pros

- Eigfusion is a way to find genes that are over-expressed or under-expressed. It did this by measuring the median values of raw gene expressions.

Cons

- There is a possibility to missing out group of genes as they focused one group at a time.

Uses

- Eigfusion used to recognise gene clusters.

2.14 Explainable AI for gene expression data analysis

Existing artificial intelligence (AI) approaches for DNA sequencing, gene expression analysis, drug prediction, personalised medicine, and next-generation sequencing allow users to observe processed results with greater accuracy. However, the internal data processing of these AI approaches is too complicated to be understood by humans without prior technical knowledge. Artificial algorithms are well suited to perform this task. However, many AI algorithms are often so complex that they are a black box, offering users few clues about internal data processing. Explainable AI (XAI) provides alternative analyses that are more understandable and technically equivalent to complex black-box AI approaches. In most cases, XAI and IML make it clear how features are linked together to produce the final predictions and analyses. The aim of this review is to examine the current XAI approaches in the areas of disease prediction, health systems, and gene expression analysis. In addition, a taxonomy of XAI is also discussed in this review.

2.14.1 Introduction

“Omics” datasets, such as genomics, proteomics, and metabolomics generate large-scale gene expressions that allow researchers to gain insights for cancer treatment planning. However, gene expression are not immediately comprehensible to humans. While experts are able to accurately identify images of fruits, they are unable to recognise genome sequences in general-at least not without the help of advanced computational models. AI approaches are used to extract features for insights with small assumptions and lots of processing capacity. Due to their higher processing capacities, AI approaches are often used in genomic data processing. In addition, AI algorithms are widely used in genomic data mining, medical imaging, and disease prediction [307, 308]. AI is helping to enhance our perception of complex relationships in underlying gene expressions, personalised medicine, treatment planning, and drug development [309].

Despite all their strengths and capabilities, many AI approaches present a number of challenges, particularly in the biomedical context, most notably, in terms of explainability, interpretability, and trustworthiness. Deep neural networks, for example, consist of layers of interconnected variables that are adapted by training the network on multiple instances [310]. As neural networks become more complex, it becomes more difficult to understand how number of parameters interact to make decisions. Even when AI developers have access to these parameters, the neural network's decisions cannot be accurately deconstructed [311]. Inputs may go through a series of iterative nonlinearities involving thousands of features before a decision is made. How can even the most accurate black-box approaches improve user understanding of biomedical data processing? How can users trust what they do not understand?

Explainable AI works between AI and humans to interpret the predictions of black-box AI approaches [312, 313]. Suppose x is a gene set in a person who has cancer. Predictions provide knowledge about what this gene set, x, represents. Explainability offers understanding as to why this collection of gene set, x, exists. Explainability thus brings value to making trustworthy and understandable decisions. The European Union's latest General Data Protection Regulation (GDPR) emphasises the importance of citizens understanding how AI systems make decisions [314]. Additionally, the Australian Federal Government has established eight ethical criteria for artificial intelligence, including those relating to explainability and transparency [315].

2.14.2 Explainable artificial intelligence and terminologies

Explainable AI refers to a set of approaches that operate between AI algorithms and users to increase the trustworthiness of the results produced by AI algorithms [316, 317]. The term explainable was used to focus on human understanding of the decisions of current AI approaches, the main interest being the human psychology of explaining. For more clarity, there are similar terms such as explainability that can help to reconcile the ideas with explainable AI [317–320].

- **Understandability:** A model's understandability refers to its ability to perform its purpose without requiring an explanation of the model's underlying structure or computations [319].

- **Comprehensibility:** Comprehensibility is the ability of a model to determine whether users understand the message conveyed by the model. Since this concept is difficult to quantify, the assessment of a model's complexity is related to its comprehensibility [321–323].

- **Interpretability:** The ability of a model to determine the extent to which a cause and effect may be detected in a system is referred to as interpretability (also known as interpretability index). In other words, it refers to the ability to forecast what will happen if the input or computational parameters are altered [318].

- **Explainability:** This term describes how well the internal decision-making mechanisms of an AI approach can be explained to humans in terms of understandability [318]. Explainability is a way to describe how a method gets to a result by doing a step-by-step calculation of given parameters without making any assumptions.

- **Transparency:** Transparency is the ability of a model to determine whether a model is explainable and whether the message conveyed by that model is interpretable to users [324].

2.14.3 Objectives of explainable artificial intelligence

There are many reasons to pursue AI explainability as follows:

- **Trustworthiness:** Trustworthiness refers to the confidence in or reliability of a decision made by an AI. Because of their simple and understandable analysis, XAI approaches strive for trustworthiness. Although trustworthiness should be a component of any explainable model, this does not imply that every trustable model is explainable [325].

- **Causality:** Causality refers to cause and effect and helps to understand the actions of an approach and the consequences of those actions within the data being studied. Prior knowledge is required to prove that observable effects are causal. An XAI approach can reflect causality by showing how an outcome is produced in terms of input data and their associations [326, 327].

- **Transferability:** Transferability is the ability to transfer qualitative data into meaningful associations [328].

- **Confidence:** Confidence is related to the reliability of a model. If the model produces understandable outcomes, it is reliable. An XAI model leads to a decision by showing how a decision was made, which increases user confidence [329, 330].

- **Informativeness:** The purpose of using machine learning models is to facilitate accurate decision making from a large volume of data. An XAI approach provides a list of features or variables in its decision-making process that provides users with a wealth of information about the data and processes [320].

- **Fairness:** Explainable AI approaches provide rationales for their decision-making process in relation to given inputs, ensuring the intelligibility of decisions and outcomes [331–334].

- **Accessibility:** It is easier for users to work with models when they understand them better. In explainable AI approaches, the decision-making processes are transparent to users, which can increase the accessibility of models [335].

2.14.4 Importance of explainable artificial intelligence

For legal, ethical, and security reasons, the requirement to declare the outcome of ML is critical when AI algorithms are used in healthcare, credit scoring, lending, and more [336]. Although there are many reasons why XAI is essential, the analysis shows that there are three concerns: (1) reliability, (2) clarity, and (3) trustworthiness of AI algorithms. XAI methods improve all three of the above concerns because some of the internal processes of ML use extremely complex algorithms with thousands of factors. XAI improves the understandability and fairness of judgments by creating humane reasoning and, when used correctly, can identify and eliminate prejudicial cases.

XAI improves the intelligibility and fairness of judgments by creating reasonable, humane reasoning and, when properly applied, can identify and eliminate prejudicial cases. The accuracy of predictions and the prevention of adverse cases is critical to clarity. An adverse case could prevent a classifier from making appropriate decisions if the classifier believes that an incorrect image is true. As autonomous methods become increasingly necessary to assist people in their daily lives, the quality of AI algorithms should be given the highest priority with respect to model interpretation. The reliability of deep learning methods is important to end users because it is a measure of confidence that a developed model will work in dynamic reality. Decisions and judgments depend largely on the knowledge and explanations of circumstances that people have access to and trust. Scientific or rational justifications for suboptimal decisions are preferable to very safe decisions for which there is no explanation. When it comes to building trust with end users, including professionals, developers, and scientists, it is critical to be able to answer why a specific decision was made.

On top of that, it is essential to build reliability on the path towards an AI-based socio-economic sector, and this is something that stakeholders and government agencies need to focus on. Bias in neural networks refers to the excessive weighting, bias, preference, or tendency of the learned model over subsets of the data caused by intrinsic biases in data collection and errors in the classification model. XAI improves fairness and helps mitigate biases caused by input data sets or an inadequate neural network design in AI decision making. Using XAI approaches to learn the model behaviour for different distributions of the input data could help us better understand the biases and skewness in the data. This could lead to a strong AI model.

2.14.5 Taxonomy of XAI methods

Taxonomies of XAI have been constructed to classify explainable mechanisms, but there is no generic taxonomy for XAI approaches. XAI methods have been classified into overlapping and non-overlapping classes. The taxonomy of XAI methods is shown in Figure 2.20.

- **Local Method:** Local interpretable methods are designed to explain the results of a prediction by focusing on a particular event and attempting to determine how the model arrived at its prediction. This can be achieved by approximating the relevant features in a black box system to use an interpretable model [337–339]. The prediction may simply depend linearly or monotonically on certain characteristics rather than having a complicated relationship. For example, the value of a house may not be proportional to its size. However, if we consider only a 200 square foot house, the prediction of the XAI model for this data may be linearly proportional to its size. To test this, it is possible to simulate how the expected price changes when the size increases or decreases by 30 square feet. As a result, local explanations may be more appropriate than global explanations.

- **Global method:** The models that consider the entire analysis of data processing to arrive at a decision are called global models [340–342]. The application of global models is an attempt to describe the nature of the model. By evaluating the importance of features, it is possible to determine which features are responsible for increasing the performance of a proposed model.

- **Model-specific**. Model-specific interpretations were developed from the parameters of specific models [337, 343]. Gradient-class activation mapping (Grad- CAM) allows the visualization of features, e.g., in convolutional neural networks (CNNs), but this method does not function with long short-term memory (LSTM). A model architecture is frequently used in model-specific techniques, such as feature maps, created by graph convolution. They are determined by the type and functionality of the particular model, such as tree interpreters. Graph neural networks explainer (GNNExplainer) is a unique feature of model-specific interpretability when complex metadata requires GNNs.

- **Model Agnostic:** Model agnostic methods are explanatory approaches that clarify how a model reaches its decisions [344, 345]. These methods, typically employed in post hoc analysis, can be applied to any model without relying on its internal details such as weights or structure. Consequently, every machine learning model can benefit from these methods for enhanced clarity. A typical way these methods function is by modifying inputs and observing the resulting changes in output. This provides insight into the inputs that have the most and least impact on the output. A notable example of such a method is LIME (Local Interpretable Model-agnostic Explanations).

- **Data modality-specific:** The term data modality refers to variables that are exclusively applicable to a particular type of information [346]. For instance, some approaches work exclusively with magnetic resonance imaging, while others work exclusively with tabulated clinical data. Grad- CAM applies only to images and not to other types of data such as text or tabular data. [347]. Model-specific explanatory methods are often used in conjunction with data-modality-specific explanatory methods. For instance, convolutional feature maps can be used to calculate what information a model needs to make a prediction in some explanatory procedures.

- **Data modality agnostic:** The approaches that are able to explain any data type are called modality agnostic. A good example is LIME, which can explain images, tabular data, and text [338]. These systems can handle a wide variety of data, making them valuable for clinical use. Perturbation-based methods are commonly used to extend the current approach to explaining models.

- **Post hoc:** Post hoc interpretability refers to an explanatory method used after a model has been trained. In particular, there are post hoc approaches that can be used for models that are inherently interpretable, since post hoc methods are often detached from the primary model [322]. Adding interpretability to the current approach through post hoc explanatory methods means an increase in understandability and confidence. Since they are model-independent, most post hoc XAI algorithms can be used in any network design. For instance, a neural network result that has already been trained and tested can be communicated without affecting the validity of the model.

- **Surrogate methods:** A specific example of supervised machine learning used in active learning to expand training datasets as training progresses, increasing training accuracy and effectiveness [348]. To evaluate alternative black-box models, surrogate techniques use an ensemble of different models. The decisions of the surrogate model can be better understood when compared to the decisions of the black box model. An example of the use of surrogate methods is the decision tree.

- **Visualization methods:** By using visual approaches, such as activation maps, some elements of the models can be easily understood. For example, patterns, lost information, and outliers can be discovered in a large data set [349]. Once all related information are available, data visualizations can be used to explain and display significant relationships in charts and graphs in a way which is more intuitive.

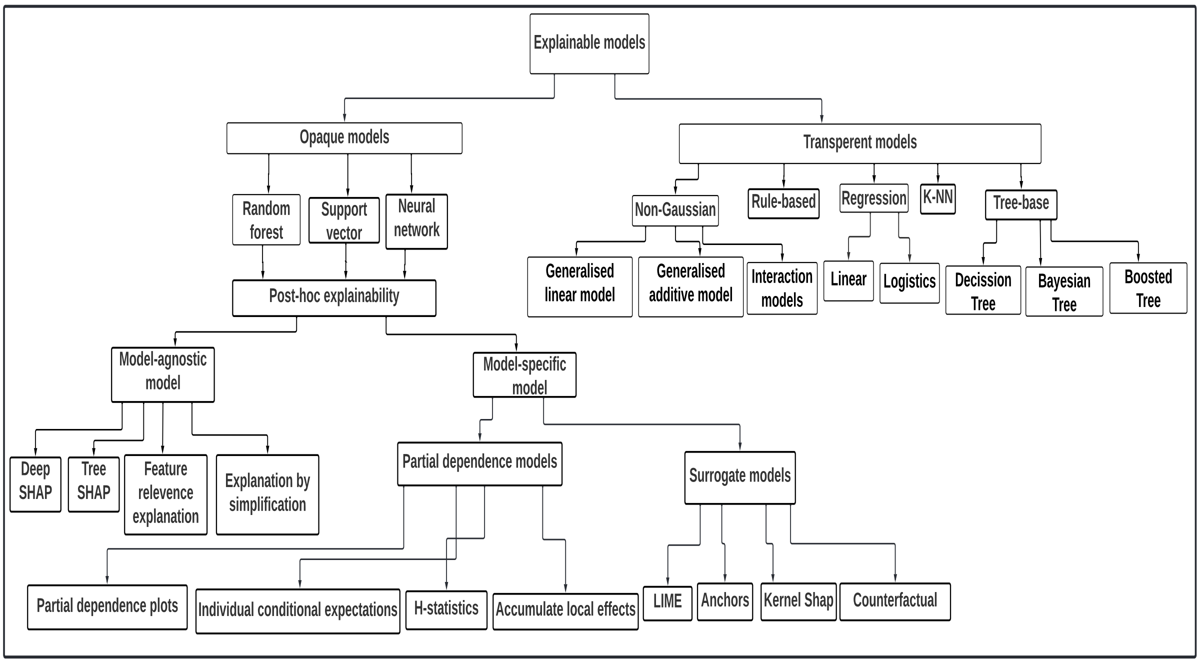


Figure 2.20: Taxonomy of the explainable artificial intelligence approaches.

2.14.6 Aims and research questions

The aim of this literature review is to describe the impact and importance of interpretable and explainable methods in health systems, gene expression analysis and clinical decision support systems. To achieve the goal, this thesis follow the following requirements.

Requirement 1 (R1): This thesis focuses to understand the impact of XAI methods in explaining different methods for predicting diseases (cancer or acute diseases) in health systems. The main focus here is on how Explainable AI (XAI) methods assist in interpreting the diagnoses of different diseases, notably cancer. In addition, this thesis focuses on how XAI methods can explain the analysis of patient clusters.

Requirement 2 (R2): This thesis focuses to understand how XAI methods create an understandable environment for clinical support systems. Here, XAI explains the properties of disease predictors for analysing the impact of clinical support systems.

Requirement 3 (R3): This thesis is centered around demonstrating how Explainable AI (XAI) methods can examine gene expression data to support disease prediction. Here, XAI methods mainly work to clarify the often complex and not easily understood outcomes from black box and graph-based disease forecasts.

This thesis propose the following research questions to understand the analysis of gene expression data in terms of explainability, interpretability, and causality in XAI:

Ref	review research questions
RRQ2.1.	How can XAI help understand the importance of traits in predicting disease?
RRQ2.2.	How can XAI explain black-box models and graph-based prediction results?
RRQ2.3.	How can XAI help understand biology when a predictive model makes a decision when analyzing gene expression data?
RRQ2.4.	How can XAI visualize model results so users can better understand them?

2.15 Literature search

It was necessary to search through academic databases to identify related research publications. Several well-known academic databases, namely Scopus, IEEE Xplore, ACM digital library, and Web of Science (WoS), were used to find relevant research papers, all of which are well known in the field of computer science. A wide range of AI and machine learning topics are covered by these databases. They have a user-friendly interface and have few access restrictions. Our search query for scholarly publications on explainability, interpretability, and causality in artificial intelligence, health systems, and genetic information included the following words: explainability, interpretability, and causality.

(AI or ML) (healthcare systems or genetic information) (XAI or explan or inter*).* From the databases, this thesis obtained publication titles, abstracts, keywords and year of publication using this query. During the initial search, the following records were discovered: WoS, IEEE Xplore, and Scopus.? Duplicate records are removed as well as records with blank or incomplete information. Finally, this thesis reduced search to three databases: IEEE Xplore, Scopus, and WoS. Figure 2.21 shows the PRISMA flowchart of our strategy.

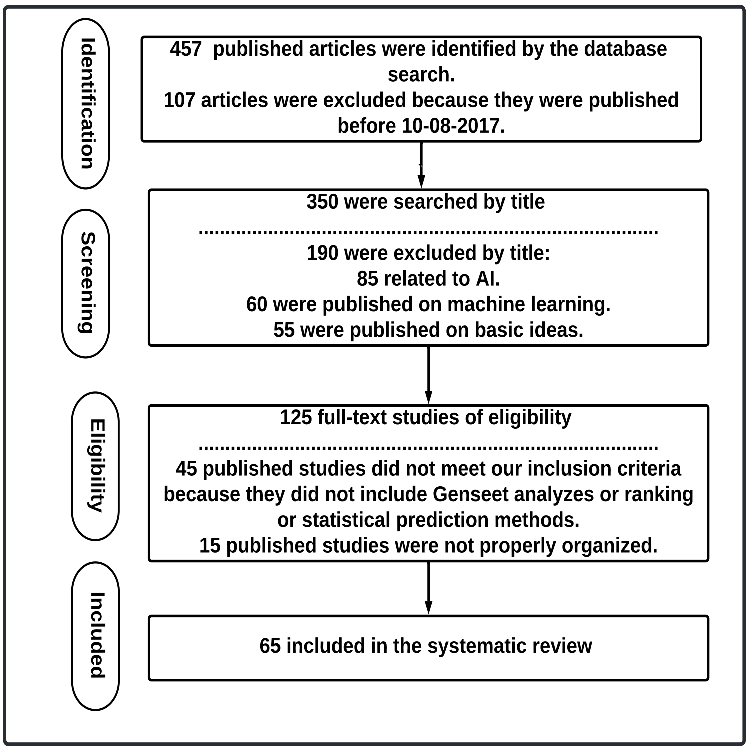

Identification
457 published articles were identified by the database search.
107 articles were excluded because they were published before 10-08-2017.
Screening
350 were searched by title
190 were excluded by title:
85 related to AI.
60 were published on machine learning.
55 were published on basic ideas.
Eligibility
125 full-text studies of eligibility
45 published studies did not meet our inclusion criteria because they did not include Genseet analyzes or ranking or statistical prediction methods.
15 published studies were not properly organized.
Included
65 included in the systematic review

Figure 2.21: Steps to identify articles or research papers related to explainable artificial intelligence using the PRISMA model.

2.15.1 Keyword search

In this SLR, this thesis aim to gain a deeper understanding of explainability, interpretability, and how current methods can help in this endeavour. To gain this insight, this thesis first looked at the terms used in the research publications in response to our search query. Using the search results from the IEEE, Scopus and WoS databases, this thesis identified the keywords in the titles, abstracts and main texts of the retrieved articles. Several keywords, including explainable AI, interpretation, and causality, were used to further narrow the results.

Figure 2.22 shows keywords found in literature for both XAI approaches, XAI applications and AI approaches. The arrow in the figure indicate keyword are related to an XAI approaches or XAI applicaiton or AI approches. Most of the keywords were from XAI-based research, and most had been recently published. This thesis also included terms from a variety of disciplines, such as health systems, gene expression, cancer prediction, deep learning, and other related topics. This thesis selected the research publications that primarily deal with XAI algorithms and their applications.

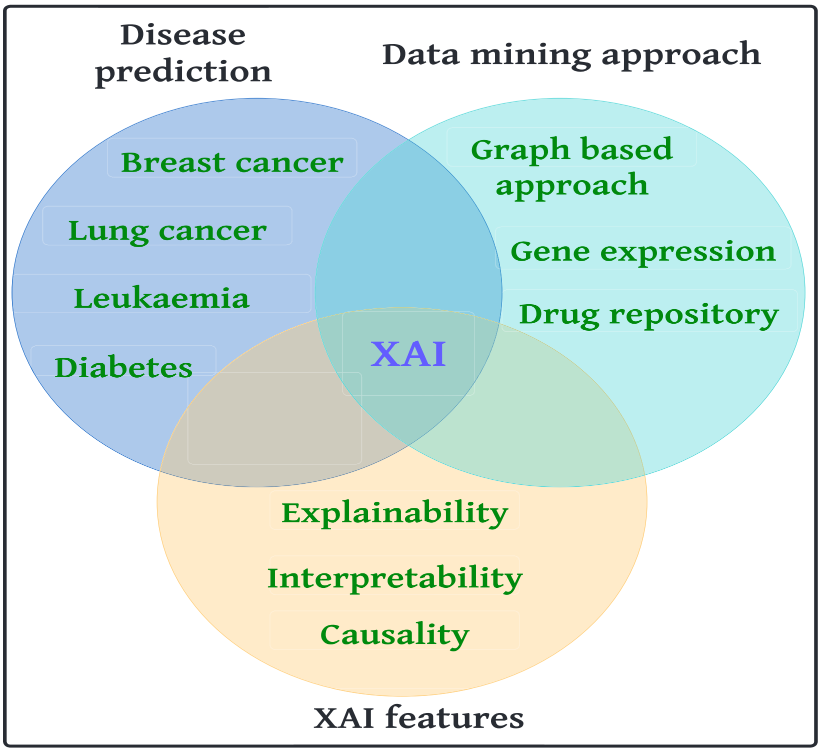


Figure 2.22: Keywords included in literature of XAI approaches, application and AI approaches.

2.15.2 Explainable artificial intelligence approaches

Several studies which examine the implications of explainable AI approaches in medical and biological data processing have been published. This thesis will now describe the XAI approaches from the literature that have been used to explain the results of the black-box model.

2.15.2.1 Local interpretable model-agnostic explanations (LIME)

LIME explains the predicted outcomes of a black-box model. First, a black-box model predicts an outcome, and then LIME explains how the black-box model arrived at a decision.

LIME is a local explainer that analyzes the responsible aspects of the input data [350]. LIME explains the predictive process of each classifier and establishes the relationship between the features that influence the process.

The word local refers only to those features that are strongly associated with a particular prediction. For example, in an input dataset, there are 1000 features, but only 8 features are closely related to the results predicted by the classifier. LIME explains these 8 features, which is why it is called a local explainer.

Working principles of LIME

- First, LIME permutes all input data with a normal distribution and generates a new set of features.
- These new data sets (features) are used by interpretable models (linear regression, logistic regressions, XBoost, decision trees, naive Bayes, K-NN).
- Then, the deviation between the transformed features and the original features is measured.
- Finally, the best features are selected by measuring the important scores of the features or using techniques such as Lasso [351].

LIME identifies the properties which are most strongly associated with a prediction when an approach makes a judgement. LIME provides a trustworthy explanation using the following formula (Equation 2.14).

Let P denote the space of features and p denote the instance of features in the dataset. Explainer (e) and black-box model (b) are two important components of LIME. To explain the process locally, LIME uses an interpretable function defined as follows.

$$exp(p) = argmin_{e \epsilon E} \theta(b, e, \lambda_p) + \Omega(e) \tag{2.14}$$

From the above equation, it is seen that LIME measures over a feature p to show how it comes to a decision for the feature where

b is a black box model, e is an interpretable model, λ_p is a distance between the permuted features and the original feature, $\theta(b, e, \lambda_p)$ is a loss function, Ω measures the complexity of the interpretable model. Since not every model is guaranteed to be interpretable, if the value of Ω is high, the interpretable model is hard to understand.

Limitations

- LIME works better for small datasets and linear models. When datasets become large, LIME suffers from higher feature extraction complexity.

- LIME is also not suitable for non-linear models.

2.15.2.2 Sub-modular pick local interpretable model-agonistic explanation (SP-LIME)

Although LIME gives the user some insight into the reliability of the classifier, it is not sufficient to analyze and evaluate the overall trustworthiness. By explaining a group of specific instances (SP-LIME) [352], a global understanding of the model is provided. This method is also model independent, but explains the results in a coherent way. SP-LIME calculates the overall importance of the features, $C(V, W, I)$, using the following equation.

$$C\left(V, W, I\right) = \sum_{j=1}^{d'} \left[\exists i \epsilon V : W_{ij} > 0\right] I_j \tag{2.15}$$

and

$$Pick(W, I) = \underset{v, |V| \leq B}{argmax}\, c(V, W, I) \tag{2.16}$$

where B is the budget, i.e., the number of explanations the user is willing to provide, $Pick(W, I)$ is the task of selecting an explanation from the total budget B. $W = n * d'$, W is the explanation matrix, n is the number of samples, and d' are the human-understandable features. $I(j)$ is the global importance of features j in the explanation space.

2.15.2.3 SHapley additive exPlanations (SHAP)

SHAP is an interpretable machine learning (IML) approach that analyses the predicted results of a predictor or classifier using trustworthy explanations [353, 354]. The key parameter of this approach is the Shapely value, which represents the average marginal value of a feature where all feasible combinations among all features are considered [355].

The following details how the Shapely value of a feature x can be calculated.

- Permute all the features entered.
- For each iteration, compute the average feature values of all features except feature x.
- Calculate the average feature values of all features, including x.
- Subtract the average feature value without x from the average feature value with x.
- The resulting value is the Shapley value of feature x.

To illustrate a situation where calculating the Shapley value is useful, assume there is a game consisting of n players who collectively receive a reward p that is to be distributed equally among each of the n players, taking into account their individual contributions.

The mechanism of SHAP is as follows:

Let us assume that
N represents the total number of executions.
d represents the total number of data points.
f denotes the characteristics of the data index
D denotes the data matrix.
L denotes the black-box model.

SHAP first chooses a sample of interest s, a feature f, and the number of iterations N. For each iteration, a random sample s is chosen from the data. Then, the feature order, $\omega_(0)$, is defined as follows: $\omega_0 = (\omega_1, ..., \omega_f, ..., \omega_p)$, then order the sample s, defined as: $s_0 = (s_1, ..., s_f, ..., s_s)$.
Construct two new instances: i. With features f: $\omega_{+f} = (\omega_1, ..., \omega_{f-1}, \omega_f, s_{f+1} ...., s_s)$. ii. Without features f: $\omega_{-f} = (\omega_1, ..., \omega_{f-1}, \omega_f, \omega_{f+1} ...., s_s)$.

Calculate the marginal contribution (Equation 2.17):

$$\varphi_f^n = \hat{f}(d_{+f}) - \hat{f}(\omega_{-f}) \tag{2.17}$$

Calculate the Shapley value as an average (Equation 2.18):

$$\varphi_f(\omega) = \frac{1}{N}\sum_{n=1}^{N}\varphi_f^n \tag{2.18}$$

To obtain all Shapley values, the technique was repeated for each of the features. To extract relevant features from a large number of features, the following equation was used to obtain the global significance values for each feature.

$$I_s = \sum_{i=1}^{n}\varphi_f^i \tag{2.19}$$

Features were ranked in order of importance from most important to least important.

2.15.2.4 Local Interpretation-Driven Abstract Bayesian Network (LINDA-BN)

LINDA-BN is a graphical XAI approach that displays the predicted results in a graph showing the associations between features in reaching a prediction. The LINDABN structure consists of three main steps: i) creation of permutations ii) computation of Bayesian network iii) f feature selections using a Markov blanket. The framework of LINDABN is illustrated in Figure 2.23.

LINDA-BN applies permutations, or rearrangements, to the input feature vector, represented as $F = \left\{F_1, F_2, ..., F_n\right\}$. This rearrangement is done using a uniform distribution, with a permutation variance ϵ that falls between 0 and 1. The permutations take place within the range $\left[F_i - \epsilon, F_i + \epsilon\right]$. By changing the arrangements of these features, the method assesses how different permutations influence the predictions made by the black box, also known as the classifier.

After rearranging the input features, LINDA-BN creates a graphical Bayesian network and makes a prediction on a class variable [356].

As depicted in Figure 2.23, once the permutations are done, the method employs the Markov

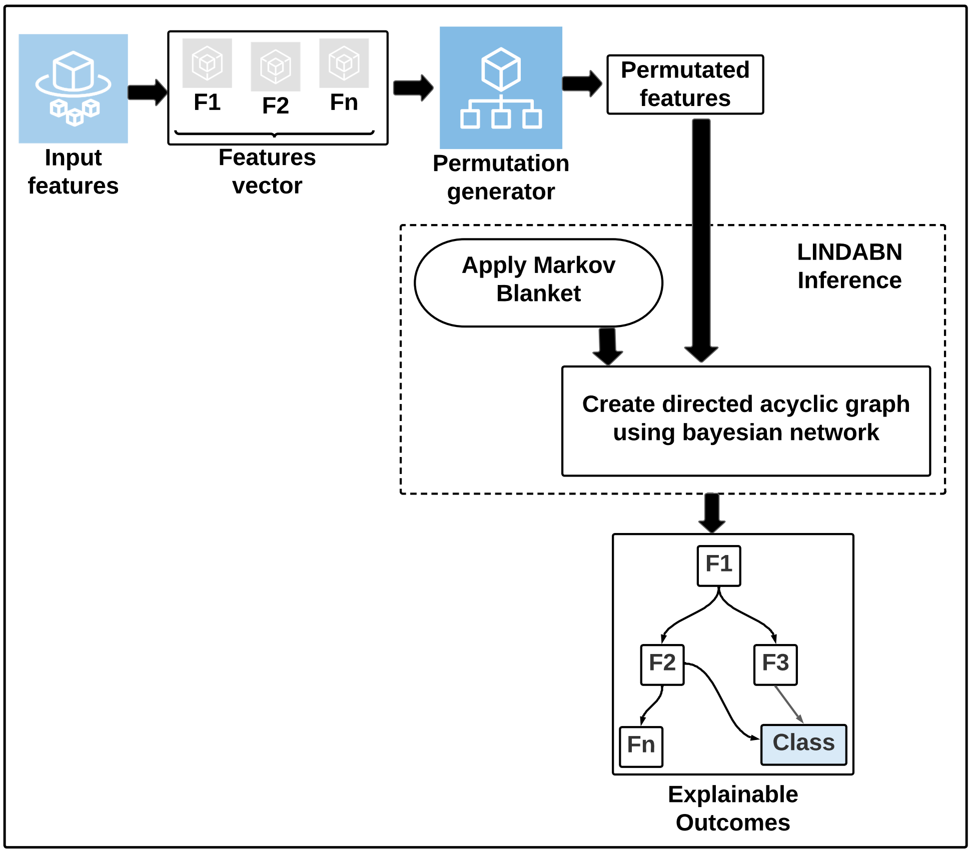


Figure 2.23: Schematic diagram showing the application LINDA-BN.

blanket and the Hill-climbing algorithm to make inferences and pinpoint the features responsible for a particular prediction. Here's how the graphical Bayesian network operates:

Suppose G represents a BN graph encompassing features $F_1, F_2, ..., F_n$. The probability P across a sample for graph G is described by the formula below [357]:

$$P(F_1, F_2, ..., F_n) = \prod_{i=1}^{n} P(F_i|S_{F_i}) \tag{2.20}$$

In this equation, S_{F_i} symbolizes the variables for all samples pertaining to the features F_i. The Bayesian network integrates all variables and makes use of the comprehensive joint probability theory for deduction. For specific events referred to as E and a recognized variable v, the inference of the Bayesian network can be computed through the following equation (equation 2.21) [357].

$$P(E|V = v) = \alpha P(E, v) = \alpha \sum_{w \varepsilon W} P(E, v, w), \quad with\ \alpha = \frac{1}{\sum_{e \varepsilon E} P(e, v)} \tag{2.21}$$

In this equation, W stands for the set of random variables that aren't part of either the events or the evidence.

This Bayesian network can express conditional dependence through a graph, G, along with a set of conditional probability parameters symbolized as π. Taking into consideration a dataset d comprised of n observations, the equation $P(G, \pi|d)$ involves a two-phase process: the learning of the structure and the understanding of the parameters. Each of these stages is explored further in Equation 2.22 [358].

$$P(G, \pi|d) = p(G|d).P(\phi|G, d) \tag{2.22}$$

In this equation, $p(G|d)$ signifies the process of structure learning, while $P(\phi|G, d)$ denotes parameter learning. The objective of structure learning is to discover the directed acyclic graph (DAG), represented by G, with the goal of maximizing $P(G|d)$. On the other hand, parameter learning is concerned with the probability parameter, π, that comes out of the structure learning. If the parameter π is distributed independently, the learning process can be described as follows [359, 360].

$$P(\phi|G, d) = \prod_i P(\phi_{F_i}| \prod F_i, d) \tag{2.23}$$

However, the structure learning problem is described by the following equation:

$$P(G|d) \propto \ P(G)P(d|G) \tag{2.24}$$

$P(d|G)$ can be decomposed into:

$$P(d|G) = \int P(d|G, \phi)P(\phi|G)d\phi \prod_i \int P(F_i| \prod F_i, \phi_{F_i})P(\phi_{F_i}| \prod F_i)d\phi_{F_i} \tag{2.25}$$

Maximum score in structure learning is determined by the Bayesian Information Criterion (BIC). This can be articulated through the equation 2.26.

$$SCore(G, d) = BIC(G, \phi|d) = \sum_i logP(F_i| \prod F_i, \phi_{F_i}) - \frac{log(n)}{2}|F_i| \tag{2.26}$$

In this equation, the sum of the log probabilities of F_i, given the product of F_i and ϕ_{F_i}, is reduced by half the log of n times the absolute value of F_i. This formula gives us the BIC score, which helps us find the best structure for learning.

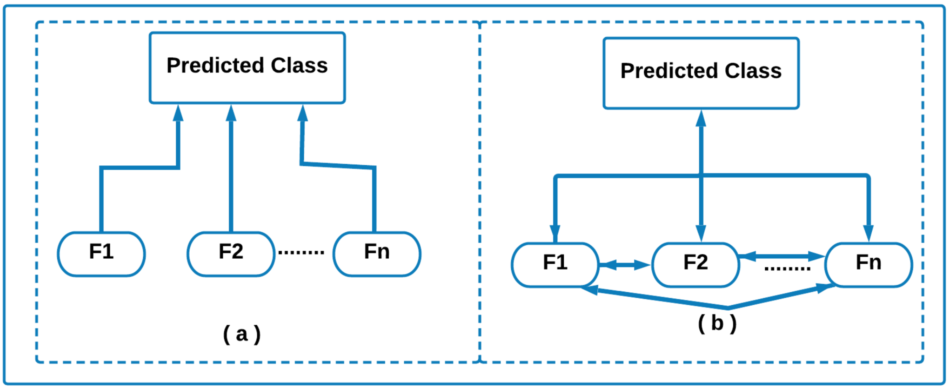


Figure 2.24: Feature selection on LINDA-BN for (a) conditionally independent and (b) conditionally dependent.

The reasoning for the Markov blanket is shown in Figure 2.24.

For example, in the prediction of a class label, $F_1, F_2, ..., F_n$ are independent, which means that knowing F_1 does not provide any further information for a judgment or prediction (Figure 2.24a). Figure 2.24b, on the other hand, shows the principle of linear regression, which states that the features $(F_1, F_2, ..., F_n)$ are conditionally independent of class only when the class variable is known. These features have a direct impact on the decision process for a target variable. A reliable interpretation of how a decision is made ensures that the user's entire reasoning is understood.

The relationship between the target variable (class variable) and the features can be inferred by abductive reasoning, i.e. human inference based on previously collected data [361]. When it comes to graphical structure, a user uses abductive reasoning to provide a reliable explanation. Abductive reasoning helps fit the Markov ceiling [362]), a strategy for selecting features for a given class. The union of conditionally independent features or variables of a target variable, such as parent, child, and co-parent (parent of a child), is called a Markov ceiling. Four conditions (high confidence, unreliable prediction, contrast effects, and uncertain prediction) determine which characteristics of LINDA-BN are used for a given target variable.

2.15.2.5 Algorithmic population descriptions

A Bayesian directed acyclic graph (DAG) is constructed to generate algorithmic population descriptors of given inputs (ALPODS) [363]. The decision network is constructed in a recursive manner as follows: first, a variable for the DAG's current node O is selected. The Simpson index is used to evaluate conditional dependencies in the selection of a variable (S). The estimated joint probability that sample components correspond to a similar or separate category is denoted by S. The S is a biologically inspired term that refers to predicted heterozygosity in population genetics or the probability of an interspecific encounter.

Second, the edges of DAGs are generated and linked with conditional dependencies to child nodes. Third, the generation of DAGs for all downstream nodes is stopped when the criterion is fully satisfied. ALPODS select a pair of data (X,Y) and calculate the probability differences for the pairs. For each pair of variables (X,Y), the probability differences of a class C are defined as follows:

$$prob\_Diff(X, Y, C) = P(X, Y|C) - P(X, Y|not\, C) \tag{2.27}$$

2.15.2.6 Partial dependency plot

Partial dependence plots are graphical representations that illustrate the relationship between certain input features and the predicted outcomes - these are often used in linear regression models [364, 365]. By calculating the partial dependence of the given input, it's possible to understand how a specific feature relates to a prediction. A partial dependence plot (PDP) lets us explore how alterations in the inputs influence the prediction outcomes of a complex model.

The partial dependence on an input feature at a specific point in time can be measured using the following formula:

$$pdp_f(v) = \frac{1}{N}\sum_{i}^{N} pred\,(x_i) \tag{2.28}$$

Here, $pred$ is the function that generates a probability of an outcome from an input row, and f represents the feature used to create the partial dependence representation. For each input row x_i, the formula computes an average result over all input rows, altering the value of the feature f to the input value v, while keeping the original input data unchanged. This way, it becomes possible to observe how the feature f influences the predicted probabilities.

2.15.2.7 DeepLIFT

Deep learning important features (DeepLIFT) is another explainable approach that assigns relevance values to input variables, similar to pixel-wise decomposition [358]. The basic premise of DeepLIFT is that it detects important aspects by comparing them to a reference state, determined based on the challenges. The reference input state is a neutral input with no special features. It can be set as the reference input activation value for each neuron in the network. The output of the network is calculated based on the reference input.

DeepLIFT tries to explain the difference between an output and a reference output by comparing an input with a reference input. Evaluation the changes in the activations of neurons in each intermediate layer with their reference state $(i.e., \Delta x_i)$ when the output of a neuron for a given input Δt is different from its reference output.

$$\sum_{i=1}^{n} C_{\Delta x \, \Delta t} = \Delta t \tag{2.29}$$

where C is the contribution scores.

2.15.2.8 Skater

Skater is a model-independent framework that allows model interpretation for all kinds of models so that one can develop an interpretable machine learning system for real-world use cases [366] . It is a free, open-source Python module that aims to clearly explain the learned structures of a black-box model both globally (based on an entire dataset) and locally (based on a single dataset).

Skater is a branch of LIME but has since evolved into a stand-alone framework with a variety of features and capabilities that enable the model-independent interpretation of any black-box model. Skater was developed as part of a research initiative to find ways to improve the human interpretability of predictive black boxes for both researchers and end users.

SKATER (Spatial 'K'luster Analysis by Tree Edge Removal) is an algorithm [367]. Specially, Skater is a Python library designed for interpretating and explaining machine learning models. While it shares its name with the spatial clustering algorithm, it is not directly related to that algorithm. The Skater library provides a comprehensive set of tools to interpret and

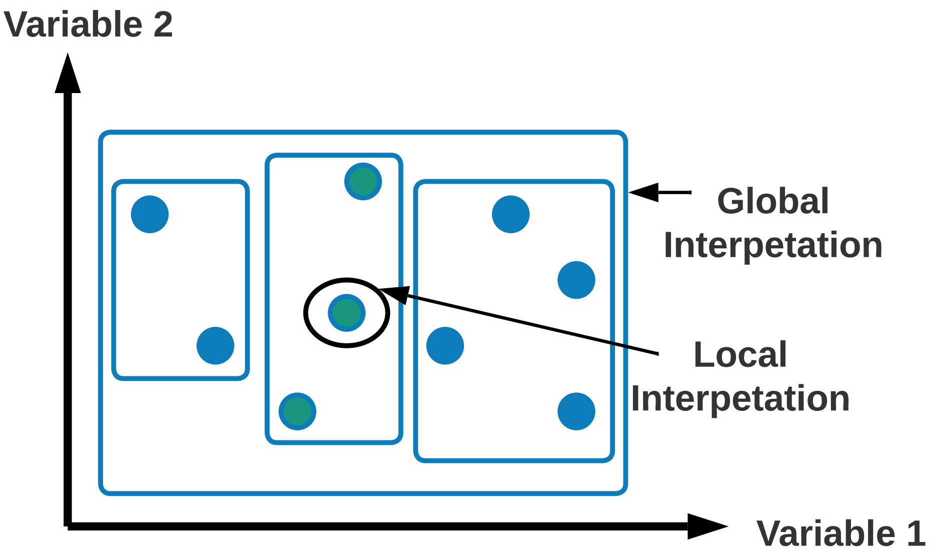


Figure 2.25: Skater Interpretive Overview.

explain the predictions of complex machine learning models. With Skater, it is possible to explore global feature importance, partial dependence plots, and local instance-level explanations. Skater has the following key features.

Global feature importance: Skater helps to assess the importance of each feature in a model by ranking them based on their contribution to the model's predictions.

Partial dependence plots: these plots visualize the relationship between a feature and the model's predictions. They help to understand how a specific feature affects the model's output while accounting for the average effect of all other features.

Local explanations: Skater also enables the generation of local explanations for individual predictions. Techniques such as LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations) can be used to understand the reasoning behind specific predictions and provide insight into the model's behavior for individual instances.

In the context of spatial analysis, the terms "an entire dataset" and "a single dataset" can be used to describe different scopes of analysis. However, the distinction between"global" and"local" approaches is more relevant here.

Global approaches: these methods focus on the entire dataset, analyzing and identifying patterns or relationships in the data as a whole. Global methods can help uncover general trends, structures, or features that are present throughout the entire dataset.

Local approaches: in contrast, local methods focus on smaller subsets or specific areas within the dataset. These techniques allow us to identify and analyze local patterns or variations that may be unique to certain regions or neighborhoods within the data.

In conclusion, global approaches address overall patterns or trends throughout the dataset, while local approaches focus on smaller areas or subsets within the dataset to identify localized patterns or relationships.

2.15.2.9 Class activation mapping

Class activation mMapping (CAM) is a widely used approach [344] to explain predictive outcomes in a trustworthy way. The CNN features responsible for image categorization decisions can be obtained using CAM. CAM uses a global average pooling layer after the convolutional layers and before the last fully concatenated layer.

Let $f_k(x, y)$ be the activation unit with weight w_c^k for each unit k. Then the input of the softmax layer for the corresponding class C is defined as follows:

$$S_c = \sum_{x,y} \sum_k w_c^k f_k(x, y) \tag{2.30}$$

The activation map M_c is computed for class c:

$$M_c(x, y) = \sum_k w_c^k f_k(x, y) \tag{2.31}$$

$M_c(x, y)$ shows the role of activation at the spatial point (x,y) in classifying its class c.

2.15.2.10 Layer-wise relevance propagation

Layer-wise relevance propagation (LRP) is another visual explanation technique [368]. LRP uses a decomposition technique that generates relevance values between the activation x_i of neuron i and its input. LRP generates the relevance values R_i^l of layer l with respect to layer $l + 1$ as follows:

$$R^l(i) = \sum_j \frac{x(i)w(i, j)}{\sum_i x(i)w(i, j)} R^{l+1}(j) \tag{2.32}$$

where $w(i, j)$ is the weight between neuron i and neuron j.

2.16 Interpretable and explainable methods in medical data analysis

Interpretable and explainable AI approaches increase trustworthiness by explaining the analytical decisions of AI approaches and explaining to the user how certain decisions were arrived at from given inputs [369, 370]. The differences between traditional AI and XAI approaches are illustrated in Figure 2.26. From the figure, it can be seen that XAI decisions are understandable to the user as it explains in detail how the decisions are made based on given input data. Traditional methods, on the other hand, only provide the results but no explanations, so the user cannot understand the decisions. This section is divided into two subsections.

This section is split in two parts. The first part talks about how XAI methods can predict diseases. The second part talks about how to explain data from genomics.

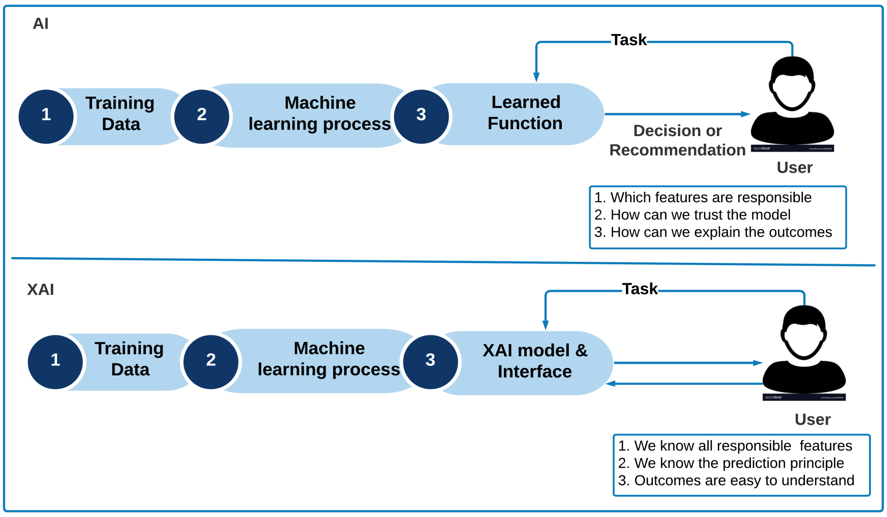

AI
Task
1
Training Data
2
Machine learning process
3
Learned Function
Decision or Recommendation
User
1. Which features are responsible
2. How can we trust the model
3. How can we explain the outcomes
XAI
Task
1
Training Data
2
Machine learning process
3
XAI model & Interface
User
1. We know all responsible features
2. We know the prediction principle
3. Outcomes are easy to understand

Figure 2.26: A framework showing how XAI adds value for users or clinicians.

2.16.1 Cancer Disease Prediction and Survival Analysis

Recently, XAI applications have been used to improve the understanding and confidence in the classification of various cancers such as breast cancer, colon cancer, etc. Table 2.12 details the studies on current XAI approaches to health care, preventive medicine, and disease prediction.

Chakraborty *et al.* investigated the data-driven relationship between features and the tumour environment in breast cancer [371]. The authors developed a data-driven XAI model using XBoost and SHAP to increase the explainability of breast cancer patient survival using RNA-seq data. They used local SHAP analysis to identify the inflection points for the tumour microenvironment (TME). They applied SHAP to interpret the XBoost results and explain the relationships between TME cell characteristics and cancer patients surviving more than five years. The authors found that T and B cells were the leading elements for TME factors in cancer patients surviving five years. XAI suggests that the number of T and B cells is the most influential factor in increasing survival rates in breast cancer patients.

Arturo *et al.* used breast cancer data to examine which patient features are important for patient survival [372]. The authors used publicly available data from the Netherlands Cancer Registry (NCR), where patients' tumour and treatment data are available. They used Cox proportional hazards (CPH) and various machine learning methods to estimate breast cancer survival.

CPH [373] is a semiparametric technique that calculates the influence of patient features on the probability of a recurrent event. It has been used to predict survival in breast cancer patients and to categorize them based on risk scores. Three machine learning techniques were also used, including Support Vector Machines, Gradient Boosting Machines, and XBoost. These methods were applied to further analyze and predict patient outcomes. The authors used SHAP to understand how these models perform their prediction. SHAP computes the shapely values for nine patient characteristics (i.e., age, tumour size, ratio of positive to removed lymph nodes, etc.) for CPH and ML classifiers. SHAP interprets the impact of these nine characteristics to predict cancer survival. Higher SHAP values of the patient indicate a higher risk than patients whose SHAP value is lower.

Another study conducted with NCR data using used several ML methods: random forest, extreme gradient boosting (XGB), k-nearest neighbours (K-NN), neural networks, naive Bayes, and logistic regression to investigate breast cancer patient survival [374]. The authors used 10-fold cross-validation to optimise the parameters of the ML models. LIME and SHAP explain the model results in this paper, LIME and SHAP showed more than 98% and 74% consistency in explaining patient characteristics, respectively. Both LIME and SHAP show that patients between the ages of 65 and 68 have difficulty surviving after the onset of breast cancer.

Bichindaritz *et al.* present a transparent, optimal, case-based breast cancer survival framework that predicts new cases to determine survival rates [375]. In this framework, they used gene expression, DNA methylation, and a combination of both data sets. The authors contributed to four topics: the optimal number of retrieved cases, elaboration of cases at multiple levels, adaptation of new cases, and explainability. To find the optimal case, the authors used case-based reasoning with confidence (CBR-CONF) to determine the similarity for each test case. High-order clustering was utilized in the multistage case processing to find the similarity features using DNA methylation data. In a new case, the authors used a confidence matrix using Euclidean distance to predict the survival of each test sample. Finally, to increase explanatory power, multivariate Cox regression was used for feature selection. To find features for patients, the authors proposed a prognostic score for DNA methylation and gene expression.

Amoroso *et al.* [376] proposed an XAI framework with adaptive dimension reduction (ADR) to analyse breast cancer prediction results. Adaptive dimension reduction is an iterative approach that uses principal component analysis (PCA) and K-NN. These dimension reduction methods reduce the dimension into two appropriate dimensions for XAI transparency and embed the clinical features for XAI justification. ADR calculated the distance between molecular subtypes and created a hierarchical clustering. This clustering showed how important the molecular subtypes were to the characteristics. They evaluated the efficacy of current therapies and the selected new therapeutic guidelines for cancer patients. For this experiment, the authors used 267 breast cancers patients and ADR helps to explain the selection of important features.

Pellegrini *et al.* examined breast cancer survival after surgery [377]. It is important to choose the right adjuvant therapy after breast tumour surgery to prevent tumour recurrence. In this study, the author focused on guidelines for adjuvant therapy and applied a new machine learning approach named coherent voting networks (CVN), which is suitable for nonlinear problem solving. The CVN model is used for predicting the survival of breast cancer patients who get extra treatment after surgery. This model is effective to predict whether patients will survive more or less than five years after surgery. CVN is useful for personalising treatment therapy with respect to a patient's molecular prognosis. This method validates clinical outcomes and explains how molecular functionalities affect cancer therapy. The proposed method uses gene expression profiles from patient tumour biopsy samples and explains gene functionalities for individual therapy selection. A knowledge diagram for cancer prediction (breast cancer) is shown in Figure 2.28.

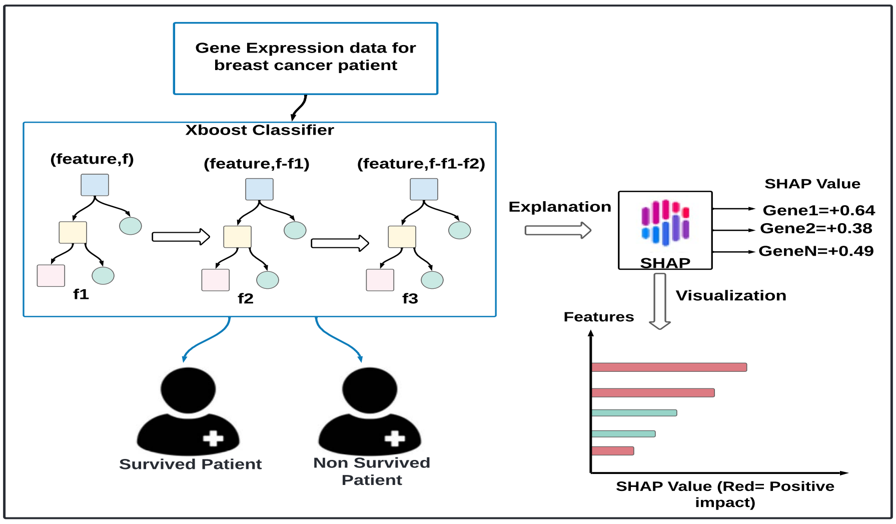


Figure 2.27: An example of interpretable outcomes of XGBoost classifier using SHAP to predict breast cancer survival.

Following the clustering method for predicting diseases, SHAP is also used to visualise the relationship between the characteristics from the data and the risk factors for corresponding diseases. SHAP is used to visualise the nonlinear interaction in predicting cancer survival posterior [378].

In a work by Li *et al.* a gradient-boosted tree model and SHAP values were used for model prediction. They used the national cancer database from America and considered the following characteristics such as age, Gleason score, percentage of positive cores (PPC), and prostate specific antigen (PSA). These characteristics have an impact on subsequent cancer risk and survival rates. Gradient boosting can efficiently predict these risk factors. In addition, SHAP helped visualise the nonlinear relationship between risk factors and patient survival rates. They used nonlinear relationships instead of linear relationships because they claimed that the latter were difficult to visualise and the results were not robust for different cut-off values. SHAP plotted a relationship between age and risk of death (or mortality rate) using SHAP dependency graphs. SHAP interpreted from this dependency diagram that patients over 70 years of age have a higher risk of mortality than patients aged 50 years. The SHAP dependency graph explained that patients have a high risk of developing cancer later in life if their PPC score is above 50%. An AI-based software project called Dr Answer for Prostate Cancer [379] was conducted. This tool explains the important variables for the tumour stage (T) of prostate cancer. The authors applied the random forest method to predict the important variables. The random forest method ranked the important variables based on characteristics with important scores. The authors also applied the synthetic minority oversampling technique (SMOTE) and the edited nearest neighbour (ENN) to account for imbalanced data. For this project, authors used 7,128 patient records after radical prostatectomy treatment.

Deep neural networks, both statistical and creative, are sometimes unable to extract features and interpretable patterns from enormous amounts of biological data. Arianna *et al.* proposed fuzzy logica with genetic algorithms to analyse the features of expression profiles in ovarian cancer [380]. To ensure interpretability, the authors used fuzzy inference systems as a rule-based method that uses if-then rules. These if-then rules extracted genes from gene expression values. Then, the authors applied a genetic algorithm that made the extracted genes more interpretable by human experts.

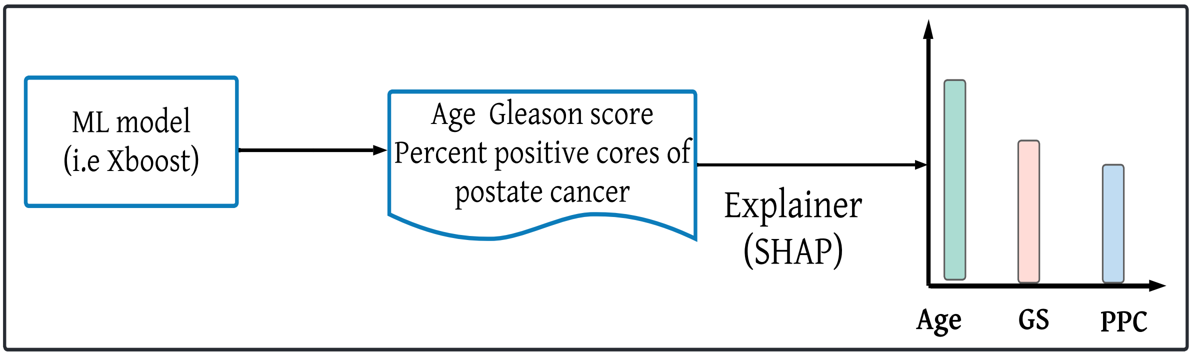


Figure 2.28: An example of interpretable outcomes of SHAP to predict prostate cancer prognosis.

2.16.2 XAI approaches in patient clustering

XAI approaches were used to explain the analysis of clustered patients and interpret the cluster results in relation to specific input data. A cluster-based explanatory approach for electronic health records was proposed [381]. Clementino *et al.* developed a multilevel clustering explainer (MCE) that can provide explanatory information to medical professionals at both local and global levels. The MCE approach is divided into three stages: preprocessing of data, creation of an explainer for local and global perspectives, and visualization of data. In the data preprocessing stage, the authors add patient identification numbers (ID) to their medical records. In the next phase, local and global explainers are introduced. In the local explainers, the authors describe the influence of clinical procedures. First, the importance scores were calculated using the learning function to represent the clinical procedures and normalize the importance influence scores. After calculating the influence scores, the scores were then sorted to rank the clinical procedures. This local procedure considers only a single cluster. From a global perspective, the entire process is considered for the entire cluster by calculating the average influence scores of each cluster, which represent the health data of the clustered patients with explainable characteristics for different traits.

Schulz *et al.* proposed clustering techniques in the explanatory space to derive disease subtypes [382]. The authors used two multiclass datasets: simple synthetic data for a proof-of-concept and data from the Cancer Genome Atlas. In this work, each statement represents how strongly and in which direction each feature contributes to prediction. PCA projected the data and derived distinct clusters that are easy to interpret. The authors used standard clustering quality indices (Davies-Bouldin index, Silhouette coefficient, and Calinski-Harabaz index) to identify structural differences between clusters. Agglomerative clustering was applied to find mutual information between subtypes clustering patients and actual outcomes.

Ultsch *et al.* [363] presented a new novel XAI method for disease classification based on clusters using high-dimensional data. Their new method is called algorithmic population descriptors (ALPODS) and is based on the Bayes decision network. ALPODS can explain its results to human experts in a trustworthy way. First, the Bayes decision network selected features and formed a directed acyclic graph (DAG). These features were selected based on conditional dependence using Simpson's index. The edges of DAG are created and assigned based on the dependencies of the successor nodes. This process was applied recursively for all successor nodes. They calculated the probability difference (ProbDiff) between the nodes. The highest ProbDiff values of the nodes are plotted over the clusters. The authors claim that this plot is easy to interpret and understand. ALPODS was applied to the iris dataset, peripheral blood (PB) and bone marrow (BM) datasets.

2.16.3 XAI approaches on precision medicine

XAI approaches have been used to explain the predictive decisions of personalized medicine. For example, SHAP is used in precision medicine for inflammatory bowel diseases (IBDs) using demographic, multiatomic (genomic and transcriptomic), and medical data [383]. The authors proposed the random forest (RF) and K-NN methods, which extracted features from genomic and demographic data. The applied SHAP framework interpreted the prediction of the models in terms of the important features for drug responses. The authors selected five drugs/doses and extracted the key features which are highly associated with drug responses from demographic, medical, and single nucleotide polymorphism (SNP) data. SHAP presents a ranking of the top twenty traits based on SHAP values for K-NN from demographic, SNP, and medical data. SHAP presents the genes most sensitive to the five selected drugs.

Al-Taie *et al*. proposed an XAI approach to analyse drug response outcomes for specific patient groups [384] by dividing patients into subgroups. Both genotypic and phenotypic data are used as input for patient stratification and drug responses. The drug repositioning knowledge base (DR-KB) is used to establish the relationship between genotype and phenotype information. The framework proposed by the authors is divided into two modules: subpopulation discovery and drug candidate evaluation. Subpopulation consists of three sub-modules. The first sub-module uses the stratification of patients using a patient network graph with the pathway expansion technique. The second sub-group is contrast to evaluate high contrast and significant patients by adding or removing nodes (patients) from the patient network. The contrast value is calculated to identify the network differentiation using the contrast pattern mining method for each candidate subgroup. Subgroup prioritization is the next submodule that generates the subpopulation contract score to rank the candidate subgroups. The final module is the candidate drug scoring module. For each drug in each patient group, this module calculates an overall drug score. This drug score can be used by clinicians to measure the efficacy of a medicine based on its molecular profile and gene pattern. These explainable results motivate clinicians to recommend drugs and analyze risk factors.

2.16.4 Clinical decision support systems

Recent developments in artificial intelligence (AI), XAI, have enabled the analysis of automated decisions through machine learning in clinical decision support systems (CDSS) [387].

Antoniadi *et al*. [387] used two scoring methods: McGill Quality of Life (QoL) (MQoL) and Single Item Score (SIS) to understand the factors that influence quality of life. The authors used the Irish dataset collected on the neurodegenerative disease, amyotrophic lateral sclerosis (ALS). The XBoost method was used to predict the QoL factors for this disease. Here, SHAP applied MQoL5 and MQoL3 to explain the features that are highly responsible in the prediction. Using the SHAP results, the authors found that the patient's age at disease onset is the most

Table 2.12: Literature review of the research publications on explainable and interpretable AI approaches for disease prediction.

Techniques	R1	R2	R3	R4
Data-driven XAI model using XBoost for breast cancer survival rates [371]	✓	×	✓	✓
Semi-parametric approach with SHAP to identify features for cancer survival [372]	✓	×	×	✓
Demographic feature extraction using LIME and SHAP to affect breast cancer survival [374]	✓	×	×	✓
Visualize non-linear relationship between risk factors and survival rates using SHAP [378]	✓	×	×	✓
SHAP framework for ranking genes for drug responses [383]	×	×	✓	✓
Local and global XAI methods (LIME, SHAP) and partial dependency plot for explaining critical features for hepatitis patients [385]	✓	×	×	✓
Bayesian rule lists (BRL) to explain the features of classified schizophrenia patients [386]	✓	×	×	✓
Case-based reasoning with confidence for breast cancer patients [375]	✓	×	×	✓
XAI framework based on an adaptive dimension for breast cancer therapies [376]	✓	×	×	✓
Coherent Voting Networks (CVN) for breast cancer prediction [377]	✓	✓	×	✓
A tool to explain the important variables for tumor (T)-stage prostate cancer [379]	✓	×	×	✓
A cluster-based explainability approach for health care records [381]	✓	×	×	✓
To explain mutual information between subtypes clustering patients [382]	✓	×	×	✓
A novel XAI method (ALPODS) to classify diseases from clusters [363]	✓	✓	✓	✓
XAI approach for patient stratification and drug repositioning [384]	✓	×	✓	×
XAI approach for patient stratification and drug repositioning [384]	✓	×	✓	×
XAI to develop a clinical decision support system (CDSS) to alert clinicians to improve patients' (QOL) [387]	✓	×	✓	✓
XAI method using an early warning score (EWS) system to predict acute critical illness from EHR [388]	✓	×	×	✓
Combined fuzzy rule systems and genetic algorithms to explain the features of ovarian cancer [380]	✓	×	×	✓

important characteristic for quality of life.

Peng *et al*. [385] used both LIME and SHAP to explain the predictive results and understand the characteristics that determine the deterioration of quality of life in hepatitis patients. Using the UCI database, the authors created a computer-aided design (CAD) for hepatitis patients. In this CAD system, several ML algorithms were used to predict this liver disease: decision tree, SVM, RF and XGBoost. Then the authors used LIME, SHAP and partial dependency diagrams (PDP) for model interpretation. SHAP explains the contribution of features using Shapely values as a global explainer. When Shapely values are greater than 0, it means a positive contribution and Shapely values less than 0 mean a negative contribution. SHAP outcome values are also plotted in the PDP to show the dependence of the characteristics bilirubin and alkphosphate. This dependency plot describes the effects of the characteristics bilirubin and alkphosphate for different bilirubin and alkphosphate values. In addition, LIME uses a local explanation plot in this paper where LIME explains the individual prediction results for hepatitis. LIME describes the critical characteristics responsible for survival and mortality rates. Instead of SHAP and LIME, the researchers used a rule-based method to explain models in health systems.

Mellem *et al*. [386] examined six weeks of data from a double-blind study for schizophrenia. The Personalized Advantage Index (PAI) algorithm was used to flag patients for treatment or nontreatment. PAI is a multiple linear regression approach that is more similar to the decision tree. Next, the authors applied Bayesian rule lists (BRL) to explain the PAI prediction results for those who were treated and those who were not treated. This BRL is the Boolean statement of if-then-else rules that explain the characteristics of patients classified as treatment-indicated and not treatment-indicated.

An XAI approach uses electronic health records (EHRs) to analyse early warning system (EWS) results and predict acute critical illness [388]. This XAI approach explains why it made the prediction it did. The authors examined secondary health data that included information from EHRs. The data also included information from the CROSS-TRACK cohort on biochemistry, medicine, microbiology, and procedure codes. During the study period, 163,050 inpatient admissions were available, of which 45.9% were male. Sepsis, AKI, and ALI were identified in 2.44 percent, 0.75 percent, and 1.68 percent of these admissions, respectively. Eighty percent of the data were training samples, the remaining twenty percent were test samples. The authors proposed temporal convolutional (TCN) to analyse the outcomes. The most common application of TCN is predictive modules. The Deep Taylor Decomposition (DTD) explanation module used for temporal explanations is another key module of the technique. The proposed XAI-EWS provides two perspectives for explanatory models: individual-based and population-based. The XAI-EWS module determines target variables related to clinical outcomes at a given point of an individual. Since the current clinical system often follows the patient without EWS parameters, knowing temporally precise clinical factors that are important for disease surveillance is critical, and XAI-EWS provides this pinpoint explanation for an individual time frame. Ultimately, XAI-EWS provides explanations for the model's results in a way that does not require deep knowledge of the model's mechanics.

For univariate and bivariate tester explanations, Marcin *et al*. [389] proposed a hybrid approach called Evolutionary Heterogeneous Decision Trees (EvoHDTree) in combination with Relative eXpression Analysis (RXA). EvoHDTree analyses the weight relationship between genes and searches the node structures for cancer and control samples. This evolutionary method represents the relationship between two genes, namely control genes and cancer samples.

2.16.5 XAI approaches on gene expression data

Genetic information or gene expression analysis has a significant impact on predicting disease survival, drug response, or gene prediction for clinical support. XAI approaches provide trustworthy explanations for the analysis of genes for various disease prediction and drug responses (Figure 2.29). From Figure 2.29 and 2.30, it is seen that XAI interpretation helps human experts understand disease-oriented genes or gene ontology by showing which genes are responsible for a particular decision.

Karim *et al*. [390] proposed a CNN-based VGG16 network for gene selection from the Pan-Cancer Atlas project data. They called the method OncoNetExplaine, which was used to classify 33 cancer types. The authors analysed cancer data with a heat map to show related genes. From all the cancer data, genes were ranked according to their values in the CNN and VGG16 network. The prediction accuracy for CNN is 89.75% and for the VGG16 network, the accuracy is 96.25%. They used gradient boosted trees and SHAP to identify top genes and cancer-specific driver genes for comparison. SHAP identified the top 20 genes responsible for

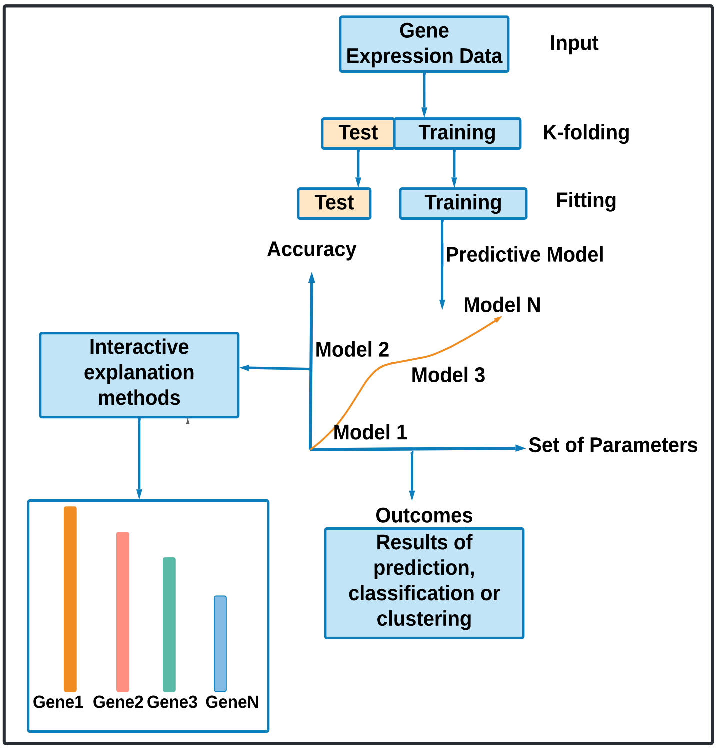


Figure 2.29: An example of explainable and interpretable features selection from gene expression values.

all tumour types. Of the listed genes, three genes are highly responsible for all types of tumours.

For non-small cell lung cancer, Kirienko *et al.* [391] proposed a method for evaluating radiological and genetic data to characterise the disease and predict outcomes. They selected 149 surgically treated patients who underwent fluorodeoxyglucose (FDG) positron emission tomography (FDG-PET / CT). The data included 74 tumour samples subjected to molecular analysis by the proposed method using a targeted RNAseq approach. They applied a logical learning algorithm (LLM) with Rulex (RULe eXtractor) to the datasets with output variables: squamous cell carcinoma, adenocarcinoma, or tumour recurrence. Rulex/LLM enables the role of radiological parameters in tumour recurrence to be explained. Rulex/LLM identified the distinguishing features from the images of PET and CT and determined the gene expression patterns associated with lung cancer.

Augusto *et al.* [392] proposed a unique rule-based XAI technique to find meaningful patterns from human gene expression data. They began by minimising the number of probes in their proposed technique, which simplifies the experimental difficulties and reduces the size of the input to the search for genes. Second, a new discretization technique converts the raw gene expression values into three discrete values. This discretization is used as a secondary dimension reduction of the data. The SRM method CMRules then generates sequence rules and ranks the genes in the discretized dataset according to a specific pattern. In the knowledge extraction stages, the authors recommend integrating the output gene rules with the functional annotation. Finally, they visualised the data in a common hierarchical model of gene patterns that allows them to quickly learn a lot about genes. Machine learning algorithms, which are widely used to improve decision making in healthcare and play an essential role in cancer diagnosis, are hampered by a number of problems, one of which is the black-box problem.

An alternative technique is proposed to understand lung cancer through gene expression data analysis using the XAI approach [393]. The authors analysed a large data set matrix to identify genes that play an important role [394]. They used publicly available data sets from the NCBI Gene Expression Omnibus. To capture the interactions between mRNA and ncRNA, the proposed mathematical model examined RNA functions in relation to gene expression values. The authors presented three different models: mRNA activity, ncRNA activity, and reciprocal activity of the species between cells in the tissue. Unless otherwise stated, this study uses a combination of Gaussian initial states to calculate population abundance and structure.

Olatunji *et al.* [350] proposed an explicable way to understand multimodal tumour types using multiplatform genetic data which increases trustworthiness for a black-box approach. They used RNA-seq (RNA-seq) transcriptome expression profiling and transcriptome expression profiling as the input data sets. Gene subtypes were selected using differential expression (DE) with Gaussian distribution and clustered gene filtering (CGF). Moreover, a deep neural network is used to predict the genes. Finally, LIME explains the deep neural network outcomes. Figure 2.30 shows the XAI analysis for gene selection using the XAI approach.

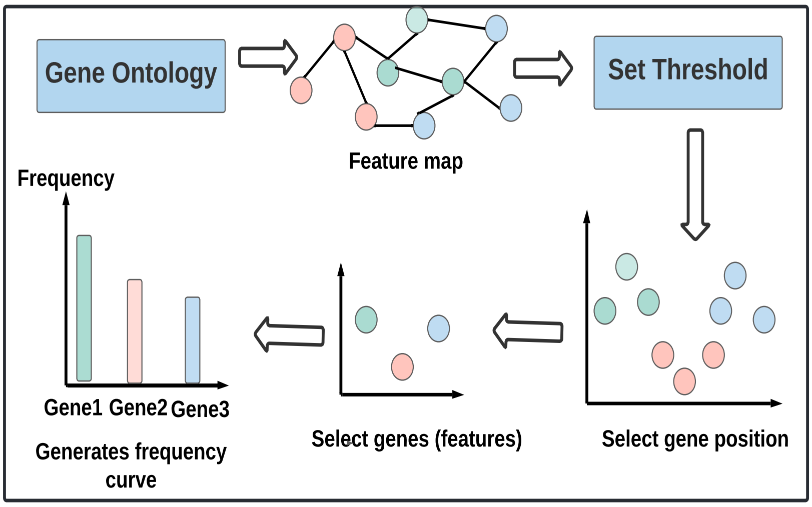


Figure 2.30: Interpretable feature selection steps from gene ontology.

Kaiwen *et al.* [395] proposed a graphical approach to aggregate gene expression data across sample and feature spaces by implementing a hierarchical graph convolutional network (HiGCN). The HiGCN is used for evaluating gene expression data in high-dimensional low sample size (HDLSS) environments. HiGCN was tested on four gene expression profiles in order to build gene interaction graphs and evaluate its overall effectiveness. HiGCN was trained with the Adam optimizer to reduce cross-entropy on labelled data. . AffinityNet and GEDFN as well as three classical algorithms (AdaBoost, Random Forest and Decision Tree) were compared with HiGCN for a more accurate comparison. HiGCN was able to produce more discriminating plots of gene expression values, even when the data were weakly labelled, and was thus able to improve classification accuracy. HiGCN also helps avoid over-smoothing by identifying significant features from noise in minimal time.

2.16.6 XAI approaches for gene ontology or networks

Bourgeais *et al.* [396] proposed Deep GONet for predicting cancer and identifying the genes responsible for cancer using gene ontology data sets. They used two data sets: the first is from an experimental study and contains 22309 samples of which 14749 are cancer and 7650 are non-cancer. The RNA-Seq dataset with 6464 samples is the second dataset. They preferred biological process ontology (GO-BP) as the hidden layers. Layerwise relevance propagation (LRP) was used in the Deep GONet approach to obtain average values for each neuron in the cancer samples. The neurons are ranked according to the relevance values of each layer, and

Table 2.13: Literature review of the research publications on explainable and interpretable AI approaches for gene expression analysis.

Techniques	R1	R2	R3	R4
SHAP methods on CNN-based VGG16 network to extract genes [390]	✓	✓	×	×
Logic learning machine (LLM) algorithm using Rulex (RULe eXtractor) for expression patterns associated with lung cancer. [391]	✓	×	✓	✓
A novel rule-based XAI strategy for identifying relevant sequential patterns from human gene expression data (GED).[392]	✓	×	×	✓
An alternative technique based on a coupled reaction-diffusion system for possible biomarkers signifying tumorigenesis [393]	✓	×	×	✓
LIME to extract a meaningful subset of genes [350]	✓	×	×	✓
A hierarchical graph convolution network for gene expression data [395]	✓	✓	×	✓
A self-explainable deep neural network (deep GONet) to integrate gene ontology [396]	✓	✓	×	✓
SHAP framework on CNN to predict tissue classification [397]	✓	×	×	✓
A multi-layer personalized network for gene regulatory tensor data [398]	✓	✓	×	✓

the ranking can be used to explain significant GO phrases.

Melvyn *et al.* [397] used SHAP for RNA-seq data to reflect gene functions. The authors designed a neural convolutional network to predict tissue classification based on genotype. SHAP was applied to CNN and evaluated silent genes to discriminate 47 tissue types. SHAP also identified protein-protein interactions. The gene ontology biological process (GO) described for SHAP-listed genes is enriched for tissue classification.

Heewon *et al.* [398] proposed DeepTensor to analyse a large-scale personalized network for the gene regulatory network. DeepTensor and Tensor Reconstruction-based Interpretable Prediction (TRIP) are two explainable AI approaches that were used to decompose the multilayer network. This individualized network was created for 762 cell lines under different settings of epithelial-mesenchymal transition (EMT). The explicable TRIP technique examines important genes to reveal the cognition process of the network.

2.16.7 Interpretable graph-based mapping algorithm

Graph-based predictions provide better visualization and analysis for disease prediction and data security. This section focuses on the explainable AI approaches of graph-based mapping algorithm.

Efficient graph-based decision graph classifiers such as decision trees (DTs) or binary decision graphs have been proposed [399]. The authors attempted to develop a polynomial-time algorithm for computing the explanation of these graphs. Explanation graphs (XpGs) form the basis for these novel algorithms. XpGs are a graph format that enables the efficient computation of explanations for decision graphs both in theory and in practice.

The graph CNN approach should be able to explain uninterpretable black-box models of contemporary deep learning methods [400]. Such a black-box model of machine learning is not capable of providing interpretable insights about the model. Moreover, decisions made by black-box approaches should undergo a trustable analysis before a conclusion is reached. The authors attempted to explain these neural network outcomes by considering the input variables in two sections: the first is a method of explaining the local analysis of a particular prediction using LIME and SHAP, the second are explanations by walking backwards through a computer network that outputs a prediction.

In other similar research [401], an attempt was made to explain GNN fidelity, stability, and fairness, and the first axiomatic framework for analyzing, evaluating, and comparing state-of-the-art GNN explanation techniques was proposed. These bounds only required knowledge of the generic form of messaging of GNNs and made no assumptions about the architecture of GNNs. The goal of this research was to show how these bounds can be applied to all current GNN explanation methods. The proposed bounds are quick to compute and are therefore ideal for empirical evaluations to determine which of several explanatory approaches provides the most trustworthy explanations. Several theorems have been applied, such as Random Explanations, GNN Gradients, Integrated Gradients, GraphLIME, PGMExplainer, GraphMASK, GNNExplainer, PGExplainer, etc. These theorems have been combined in various ways to identify the axes of fidelity, stability and fairness of GNNs.

Another XAI approach for multiple input modalities is multimodal causality with graph neural networks [402]. In this study, the authors used four different types of inputs: time series, histopathological images, knowledge bases, and textual data from patients. These inputs were linked by an interaction and correspondence graph (ICG) and positive/negative patterns were generated for this graph based on network or similarity structures. These positive/negative patterns are embedded in a low-dimensional space using the modality concept for interpretation.

Graph-based XAI systems were developed using a multi-scale convolutional neural network to train graph potentials (MS-CNN) [403]. The authors used fifty-eight medical images with hand rims to verify their results. MS-CNN showed better accuracy by successfully merging local and global texture information from the images.

A fully automatic graph-based XAI system [403] was proposed, where graph potentials were trained using a multiscale convolutional neural network (MS-CNN). For validation, the authors used fifty-eight medical photographs with hand boundaries. By successfully combining the local and global texture information of the images, MS-CNN showed that the segmentation accuracy of the proposed graph-based technique can be improved.

2.17 Research gaps in XAI

In the literature above, XAI approaches were used to predict specific gene and gene set for a wide range of diseases. A number of well-known XAI methods have been used in the literature, including LIME, SHAP, anchor, and SP-LIME. However, all of these XAI approaches adhere global learning (GL) mechanism. The GL techniques utilise the same feature values irrespective of the number of test instances used to predict a condition of an instance. For example, there were four test instances such as x, y, p, and q. For a test patient named x, GL will compare the feature values of x to the predefined feature values of all training samples. And then repeats the procedure for subsequent test patients such as y, p, and q. Predicting a condition of an instance would be useful if it were possible to use only the most relevant knowledge from the training data, which is incompatible with applying a predefined model to all test samples.

Instance-based learning (IBL) generates a separate set of knowledge for each test instance by comparing the feature values of the test instance with all feature values of the training instances, which is useful for predicting a condition of a patient. An IBL measures the distances between the feature values of a test instance and the feature values of all training instances. Then, it produces a matrix of values that approximate the feature values of the test instance. The IBL inference leads to putting a test instance to putting in the same group as its most common neighbour. IBL differs from other supervised learning approaches because it does not construct ambiguous abstractions like support vector machine, decision tree, and boosting. “Eager” or global learning, which works on a fixed model, performs the opposite of Instance-based Learning.

Moreover, an integrated approach of IBL and XAI would be useful to predict a patient medical condition with respect to the variation of gene expression values in each gene set. Because the prediction outcomes would allow individuals to identify genes or gene sets responsible for a particular disease for an instance.

2.18 Research questions

Regarding explainable AI in predictive analysis, the GSEA literature review provides insights into incorporating machine learning and AI models into gene expression data analysis. The review reveals that various AI approaches have been successful in generating accurate predictions while also highlighting their limitations with respect to explainability and interpretability. By evaluating the performance of existing AI models used in GSEA and their suitability for cancer research, this thesis applied suitable explainable AI approaches for prediction. This process ensures that the explainable AI approaches used to generate predictions outcomes and allowing for a clear understanding of the model’s decision-making process and promote confidence in the model’s outcomes.

Based on the literature, this thesis propose the following research question.

Research question 1

How can an explainable and interpretable method predict a patient status (healthy or cancerous or relapse or non-relapse) built on an individual instance in relation to variation of gene expression values in each gene set?

2.19 Conclusion

As machine learning techniques become more popular, the field of XAI is becoming increasingly important, and there are ethical, trust, transparency, and security issues to be addressed. In this review, this thesis explored existing literature on XAI in the context of machine learning and classifiers. This thesis have shown that research in this area is still in its infancy, but it is growing and has significant implications for explainable processes. Researchers are focusing on significant feature explanation, visualisation techniques, query-based explanations, and verification methods for disease and health system prediction, with a clear understanding and interpretation.

This thesis shifts discussion to the concept of AI and machine learning prediction, including transparency, fairness, and privacy. This thesis also talks about the application of XAI methods in the context of gene expression analysis, health systems, and disease prediction analysis. Various XAI working principles were also graphically described to explain the prediction process. The taxonomy and methods of XAI are also described in detail. The aim of this study is to show how the interpretability of models needs to be addressed in terms of XAI standards such as trustworthiness, fairness and transparency for various biological experiment analyses.

Chapter 3

Proposed method to understand variance in gene expression values in cancer

---

"Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less." – Marie Curie

3.1 Introduction

The literature review performed for this thesis reveals that differences in gene expression values could be the underlying cause of various life-threatening cancers, such as acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), colon cancer, breast cancer, and adrenal cancer. It is possible that there is a close relationship between various diseases and shifts in gene expression values. Therefore, identifying and explaining outliers in the normal distribution of gene expression values in cancer patients could be useful. Measuring abnormalities in gene expression values of cancer patients relative to biological functions or gene sets is one potential solution to this problem. This research investigates the potential relationships between patients and gene sets, especially with regard to variations in gene expression values. In particular, a prime purpose of this investigation is to identify which parts of a patient's biology are responsible for the onset of cancer or the relapse of cancer in a patient.

When building a model to infer the details of an individual's condition, either knowledge-driven or data-driven approaches can be applied. A knowledge-driven process uses information accumulated using the scientific method. A data-driven process uses labelled data, sampled from the domain, to build a model. Data-driven modelling ranges from simple techniques, like a logistic regression model, to more complex machine learning techniques, like support vector machines, Bayesian models, and artificial neural networks.

An approach in which a model is built based on sample data is referred to as a data-driven approach. A data-driven approach can be useful when one's understanding of the domain is weak and a large amount of sample data is available from the domain [50]. A knowledge-driven strategy involves developing one's understanding of the domain matter. Knowledge-driven modelling can be useful in situations where it is difficult to obtain sample data and when an individual possesses a strong understanding of the domain [51]. A combined data and knowledge-driven approach would be effective in handling small data sets [404]. Because small data sets are prone to having imbalanced groups, which occurs when one group has a disproportionately large number of observations compared to another group.

This thesis explores possibility of employing a data- and knowledge-driven approach to analyse gene expression data of cancer patients with leukaemia, colon, breast, and adrenal cancer. When using a model where each data point corresponds to a patient, there is a chance that the model may need to work with limited amounts of data. This can present challenges in accurately analyzing and drawing conclusions from the available information.

Ideally, a data analyst may like to combine both data and knowledge-driven ap- proaches to obtain the best possible model. However, in the analysis of gene expression data, it is not clear how to combine data analysis approaches with accumulating scientific knowledge.

In this thesis, knowledge is represented by the sets of genes identified within a specific gene ontology. A gene set in an ontology is characterized by a combination of genes and their products [405–407]. Gene sets are a collection of genes that share common properties and are united either by (1) a particular biological process, e.g. cell cycle, (2) a location, e.g. nucleus, (3) complications, e.g. leukaemia, or (4) a proximate in a pathway, e.g. genes associated in cell cycle pathway of the KEGG (Kyoto Encyclopedia of Genes and Genomes) [408–412]. Gene sets are members of a gene ontology and present a coherent arrangement of biological functions and genetic interconnections [413, 414]. Moreover, gene ontologies address which gene sets exist, how gene sets may be arranged within a hierarchy, and what are distinctive similarities/dissimilarities [415–419]. Additionally, gene ontologies indicate which biological processes are distinctive across gene sets [420, 421].

This thesis hypothesizes that carefully aggregating gene expression values into measures over gene sets provides opportunities to gain insights from the data analysis of gene expression profiles. In particular, it is posited that gene set anomaly scores increase the visibility and detectability of patterns across different expression profiles compared to using expression values directly. In this research, a family of methods is proposed for doing this and the approach is demonstrated from multiple analytic perspectives, capitalising on both the knowledge-driven analysis from using gene sets and a data-driven analysis from keeping profile data distinguishable.

In particular, the following questions are addressed:

1. How can gene expression variation be captured in gene sets, as anomaly scores, without prior assumptions about profile classes?
2. How can gene set anomaly scores offer insights into the biology of a cancer patient's response to treatment?
3. How do the distributions of gene set anomaly scores vary across different groups of patients?

3.2 Data preparation

This thesis used four datasets of gene expression values to investigate the proposed method. The first dataset is from patients with acute lymphoblastic leukaemia (ALL) with gene sets obtained from the Molecular Signatures Database (MSigDB) [8]. Leukaemia data was generated by The Children's Hospital at Westmead (CHW), Sydney, Australia and consisted of Affymetrix derived gene expression values identified in diagnostic bone marrow aspirates collected from 96 ALL patients, each patient provided more than twenty thousand gene subjects with gene expression values. Each patient received treatment involving chemotherapy and possibly bone marrow transplant based on their clinical risk stratification into standard, medium, high or very high, groups [422]. The dataset records the treatment strategy (low, medium and high) and the patient outcome (relapse or non-relapse). This dataset is publicly available [423].

The second dataset is from 54 colon cancer patients with 54676 genes measured [424, 425]. It also is in the public domain, available from National Centre for Biotechnology Information (NCBI, reference GSE4183).

The third dataset is from 64 adrenal cancer patients also available from the NCBI, (reference GSE4183). It covers the same 54676 genes [426, 427].

The fourth dataset focuses on breast cancer and comprises 189 patients' data, which is publicly accessible through NCBI (Reference GSE2990) [425]. This dataset includes microarray data for breast cancer (breast carcinomas) to evaluate histologic grade. Histologic grade characterizes the abnormal cancer tissues or cells within a tumor. Depending on cancer cell growth and spread, histologic grades are classified into grade 1, grade 2, and grade 3. Notably, histologic grade 2 tumors exhibited a high index of associated genes that were linked to recurrence.

For the colon and adrenal cancer dataset, we used six categories of gene sets: positional gene sets (C1), curated gene sets (C2), regulatory gene sets (C3), computational gene sets (C4), gene ontology gene sets (C5), oncogenic gene sets (C6), and immunological gene sets (C7) [8].

Additionally, the proposed approach could be applied to both RNA-Seq and microarray genomics data. The proposed method would be applicable to RNA-sequence gene expression data. The advancements in technology have allowed for high-throughput sequencing methods such as RNA-sequencing, which generate a large amount of gene expression data. This data is typically expressed numerically, representing the expression level of each gene. These numerical gene expressions form a table of data that is amenable to analysis using proposed method in this thesis.

The advantages of proposed approach in this thesis lies in its ability to analyze and interpret this high-dimensional data effectively. Specifically, it can generate a anomaly scores by integrating these gene expression values into gene sets. This approach provides an opportunity to gain in-depth insights from the data analysis of gene expression profiles, and it is particularly valuable when dealing with RNA-sequence gene expression data.

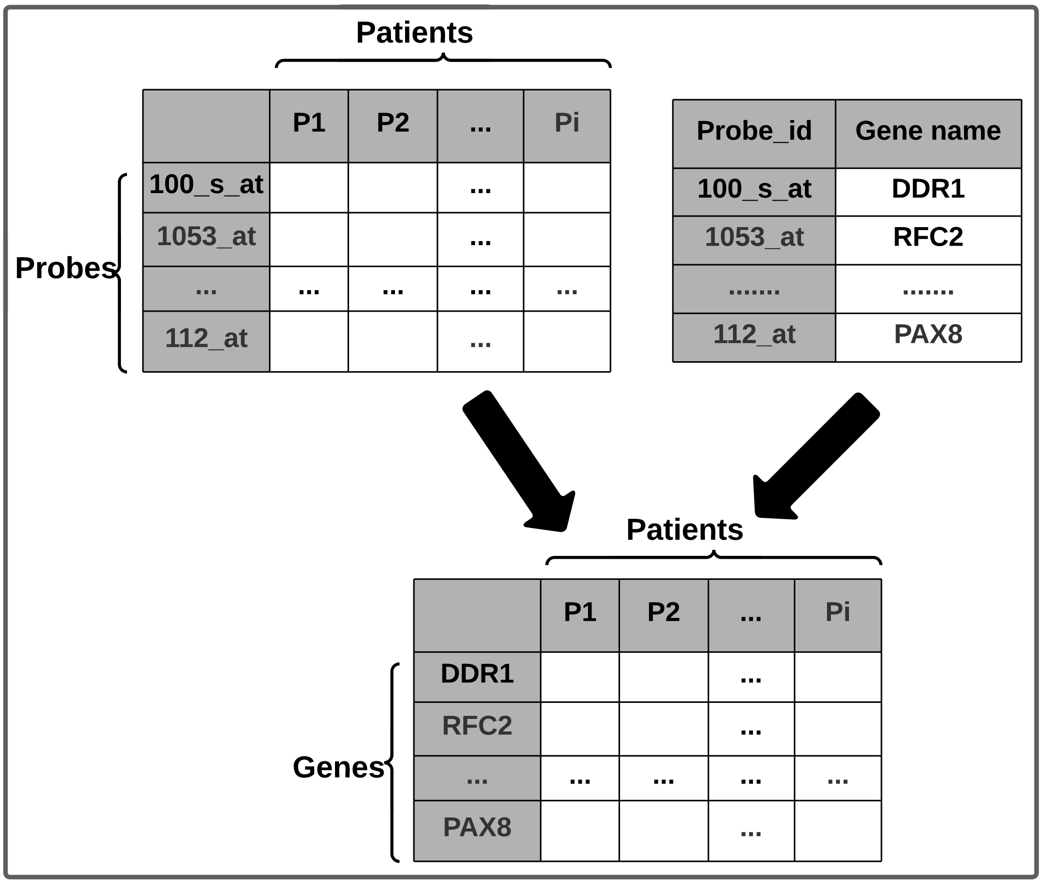
3.3 Method

To generate gene set anomaly score, the method operates in two steps: (1) pre-processing gene expression values, and (2) generating gene set anomaly score.

3.3.1 Pre-processing gene expression values and gene set

The data preparation step covers the collection, cleaning and organisation of the two input data sources, namely gene expression profiles and gene sets. Gene expression profiles are a collection of profiles where each has a gene expression value for a gene from the set of genes covering the data. Each profile can be thought of as relating to a single patient, but profiles can be more general than that. For example, each profile may be from the same patient but measured at multiple times. Gene set data is a collection of gene sets, i.e., each gene set has a name and a set of member genes. The number of gene sets may be quite large, for example all the gene sets from one or more ontologies. Unlike GSEA where gene sets are selected based on a hypothesis, the proposed method aims to identify important gene sets and exploit the knowledge inherent in them.

Figure 3.1 shows the initial phase of data preparation where one table contains gene expression values with probe ids and patients, and the other table contains genes with probe ids. These two tables are compared each other and the matched probe ids are identified, which results in the final table of gene expression values with gene names and patients. For example, the table contains the probe set name (e.g., 1007_s_at, 1053_at, etc.) with the patients (e.g., P_1, P_2, etc.). Another table contains same probe ids with gene names such as DDR_1, RFC_2, and etc. Finally, the probe set ids are matched between both tables and a table containing gene expression values with the gene names and patients is created, as shown in Figure 3.1.



	P1	P2	...	Pi
100_s_at			...	
1053_at			...	
...	...	...	...	...
112_at			...	

Probe_id	Gene name
100_s_at	DDR1
1053_at	RFC2
....	....
112_at	PAX8

	P1	P2	...	Pi
DDR1			...	
RFC2			...	
...	...	...	...	...
PAX8			...	

Figure 3.1: Pre-processing raw gene expressions values and probe IDs.

3.3.2 Generating gene set anomaly score

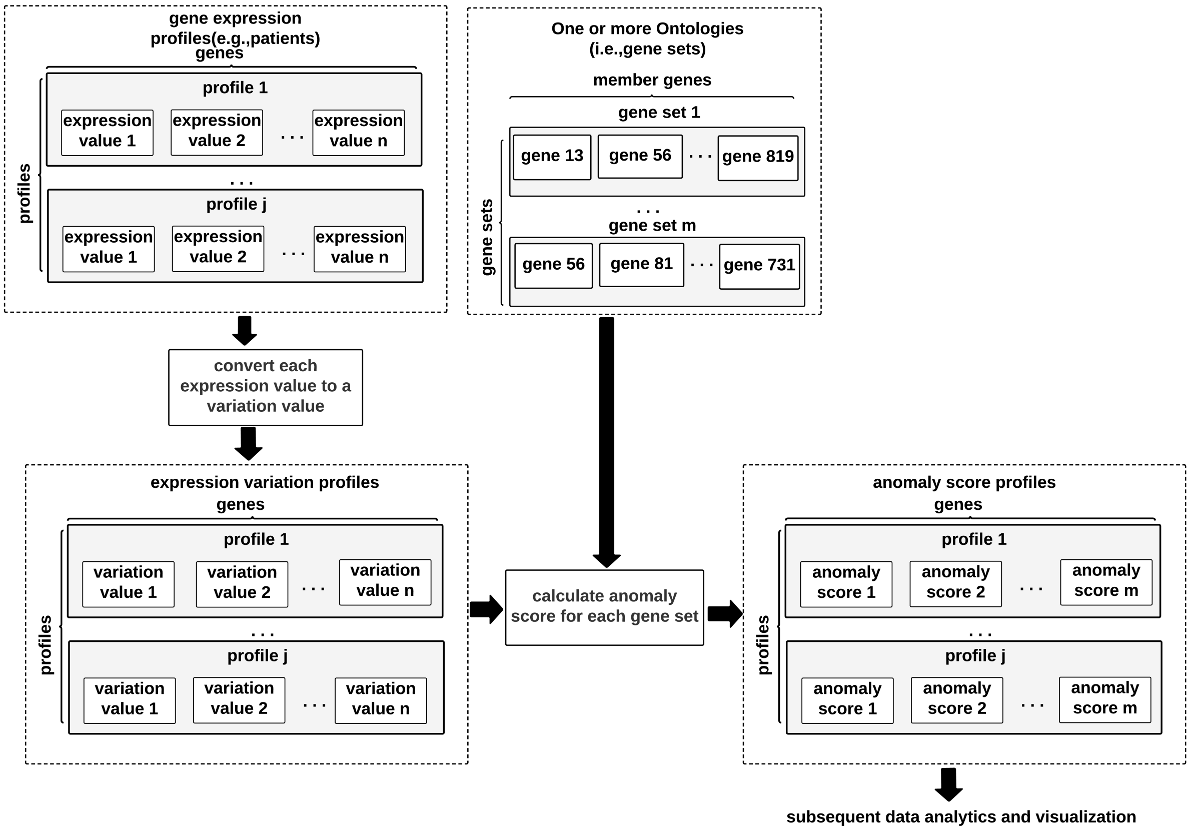
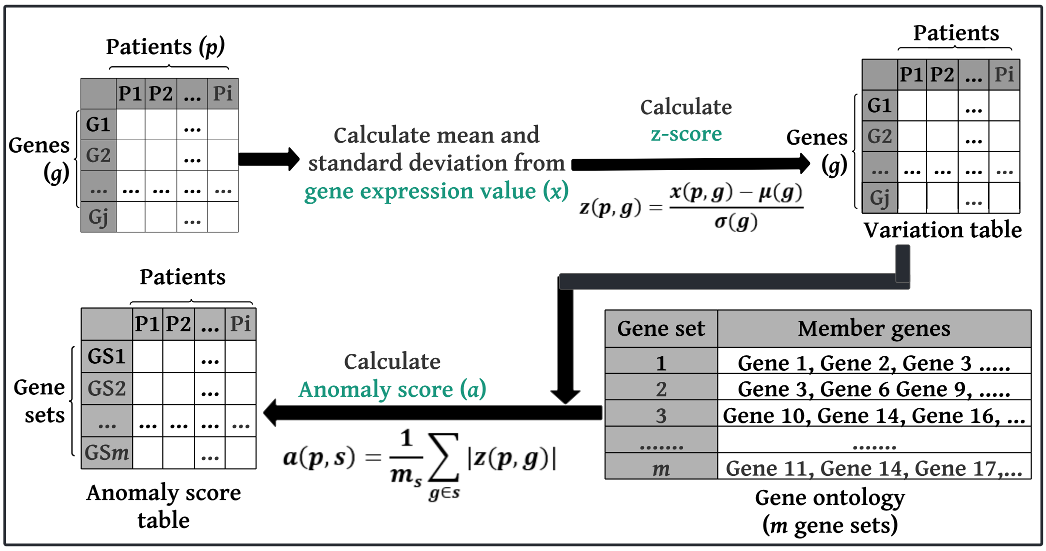


Figure 3.2: Schematic diagram showing an application of the proposed method representing genes (G), gene sets (S) and composite scores.

Proposed method (Figure 3.2) starts with gene expression profiles and gene sets. The proposed method is designed to work with large numbers of gene sets, for example all the gene sets from one or more ontologies. Figure 3.2 shows that generation of an anomaly score begins with measuring variations in gene expression data. This may be considered as a normalisation step that captures the researcher's intuitions about what it means for a gene expression to vary away from a typical value. In thesis thesis, z-scores were used but other measures are possible. Next, the variation measures of a profile and genes are aggregated to produce an anomaly score. This thesis reports on four methods: mean absolute score, root mean square score, cubic root mean cube score, and range mid-point. Thus, there are at least four variations of the anomaly score which we refer to as (1) z-absolute, (2) z-square, (3) z-cubic, and (4) z-mid-range. These are defined more formally below. This thesis reports on each of these and demonstrates that the proposed anomaly score method is a robust choice. The effect of the anomaly scoring phase is to re-map each profile from a space of gene expression values to a space of gene set anomaly scores.

Figure 3.3 shows the computational steps for the anomaly score, which are detailed as follows.

The proposed method generates an anomaly score for each gene set, for each gene expression profile. Thus, the entity for subsequent analysis are the profiles rather than the gene sets. The gene set anomaly scores are a way of re-representing each profile. This creates an opportunity for applying many data processing methods from statistical analytics and machine learning. This brings the advantage of knowledge inherent in gene sets but retains the opportunity for profile-based data analytics, including methods that do not require profiles to be placed in two classes (i.e., no classes or more than two classes).



$$z(p,g) = \frac{x(p,g) - \mu(g)}{\sigma(g)}$$

$$a(p,s) = \frac{1}{m_s}\sum_{g\in s} |z(p,g)|$$

Gene set	Member genes
1	Gene 1, Gene 2, Gene 3 .....
2	Gene 3, Gene 6 Gene 9, ...
3	Gene 10, Gene 14, Gene 16, ...
....	....
m	Gene 11, Gene 14, Gene 17,...

Figure 3.3: A data processing context for generating the anomaly score from raw gene expression values and gene sets.

- The first step of the process is to measure the z-score, or z-square score, or z-cube score for the raw gene expression values.
- The second step is to match the patient's genes to the gene sets and count the total number of matched genes from the gene sets.
- The third step is to divide the z-score values or the total z-squared values or the total z-cube values by the total number of matched genes.
- The final average z-score or z-square or z-cube value is the anomaly score, which represents the degree of anomaly with respect to the gene sets for each patient.

3.3.2.1 Z-score

The first step creates a measure of expression variation for each gene expression value. For profile p and gene g, the data set provides expression value $x(p, g)$. Z-score variation $z(p, g)$ is defined as

$$z(p,g) = \frac{x(p,g) - \mu(g)}{\sigma(g)} \tag{3.1}$$

where

$$\mu(g) = \frac{1}{n} \sum_{p} x(p,g) \tag{3.2}$$

$$\sigma(g) = \sqrt{\frac{1}{n} \left(\sum_{p} (x(p,g) - \mu(g))^2 \right)} \tag{3.3}$$

where n is the number of profiles (i.e., patients).

3.3.2.2 Z-absolute score

A z-absolute score is the mean absolute z-score variation value for profile p and s. For profile p and gene set s. Where s contains m_s number of genes, an anomaly score $a(p, s)$ defined on $z(p, g)$ as

$$a(p,s) = \frac{1}{m_s} \sum_{g \in s} |z(p,g)| \tag{3.4}$$

The intuition for this measure is to capture the average level of variation for a gene set, while disregarding the sign of a variation. Genes of a gene set may collectively operate to achieve some purpose, but when the gene set is not operating correctly, some genes may over express and some may under express.

3.3.2.3 Z-square score

The z-square score is the root mean square of z-score variation values for p and s. This is similar to z-absolute score but is more sensitive to variation. This may be considered a more traditional approach to aggregating anomalies.

In this case, an anomaly score $a(p, s)$ defined on $z(p, g)$ as

$$a(p, s) = \sqrt{\frac{1}{m_s} \sum_{g \in s} z(p, g)^2} \tag{3.5}$$

3.3.2.4 Z-cubic score

The z-cubic score is the cubic root mean cube of z-score variation values for p and s. Importantly, it is sensitive to the sign of variation. The cubic formulation also has the effect of emphasizing variation within a gene set when only a small number of genes show variation. The intuition is that many genes in a set may be operating normally, but if only a small number of genes are not operating normally, then the function of the whole gene set suffers.

The z-cubic score is defined as $a(p, s)$ on $z(p, g)$ as

$$a(p, s) = \sqrt[3]{\frac{1}{m_s} \sum_{g \in s} z(p, g)^3} \tag{3.6}$$

3.3.2.5 Z-mid-range score

The z-mid-range score is the min-max normalization of the z-score variation values for p and s. This is similar to the z-absolute score in that both approaches follow a certain scale of variation. z-absolute score considers unsigned variation, while z-mid-range considers the sign of variation. In the mid-range formulation, all genes are equally important in the calculation of the anomaly score. The intuitive goal of this measure is to capture the average level of variation present in a collection of genes in a gene set, considering the sign of a variation. All genes in a gene set may function together to make a desired effect. However, if the gene set is dysfunctional, some genes may be over expressed while others are under expressed.

The z-mid-range score is defined as $a(p, s)$ on $z(p, g)$ as

$$a(p, s) = \frac{1}{m_s} \sum_{g \epsilon s} \frac{z(p, g) - \max\limits_{x \epsilon z(p,g)}(x)}{\max\limits_{x \epsilon z(p,g)}(x) - \min\limits_{x \epsilon z(p,g)}(x)} \qquad (3.7)$$

3.3.3 Synthesizing anomaly score

An expression anomaly score of a gene set reflects the combined degree of deviation of the expression values from the expected values for a given patient. Gene expression data relates to 96 leukaemia patients in two categories, relapse and non-relapse. Expression anomaly scores were used to derive the similarities and dissimilarities between relapse and non-relapse patients. Gene sets reflect the biological functionalities (e.g., molecular functions, cell cycle, cellular locations, and biological processes) of a disease. A primary purpose is to map the patients by associating the anomaly score for each gene set with these biological functionalities to investigate the impact of the gene sets on a given disease.

This thesis contributes to the data processing context, as shown in Figure 3.4. The analysis and usage phases will explore how anomaly scores can provide insights into biology related to patients and diseases. In these phases, the re-mapped profiles are processed using standard data analytics. Figure 3.4 shows three analysis steps: (1) dimensional reduction to extract important gene sets or features, (2) analysis of anomaly score distribution to identify statistical variations, and (3) machine learning for classifiers.

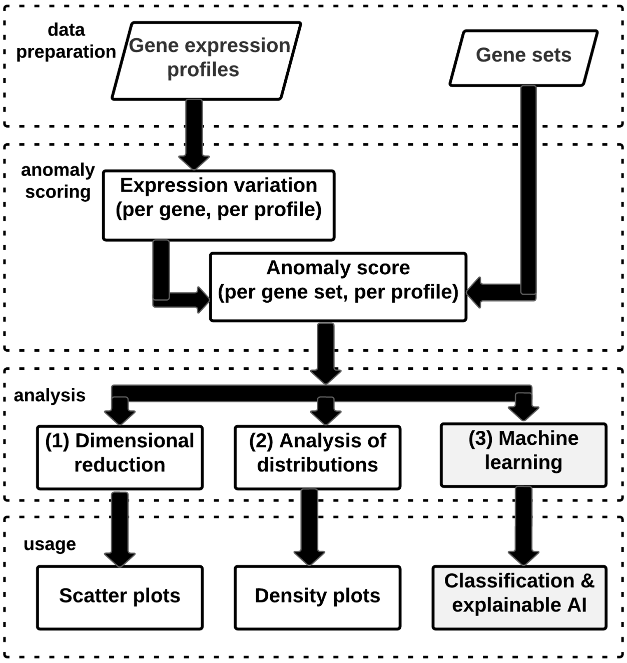

data preparation
Gene expression profiles
Gene sets
anomaly scoring
Expression variation (per gene, per profile)
Anomaly score (per gene set, per profile)
analysis
(1) Dimensional reduction
(2) Analysis of distributions
(3) Machine learning
usage
Scatter plots
Density plots
Classification & explainable AI

Figure 3.4: Architectural view of the proposed approach.

3.3.3.1 Dimension reduction

Different dimensional reduction approaches are explored. In fact, re-mapping to anomaly scores already provides significant dimensional reduction (from tens of thousands of genes to thousands of gene sets).

The re-mapped profiles still have a high dimension since there are a large number of gene sets. Not all gene sets are strongly associated with an identified class of profiles. Using principal component analysis (PCA) [428, 429], this work investigates how to reduce the large number of gene sets and identifies two or three strongly discriminating components (section 3.3.3.1).Maximum Relevance Minimum Redundancy (MRMR) [430] and Random Forest [431] are used as the feature selection techniques to identify the high-priority gene sets.

Twenty-seven thousand gene sets were collected to investigate the relationships between gene expression values and biological functions. Of these, 10,185 gene sets are linked to gene ontology (GO), which are divided into biological processes (BP, 7081), cellular components (CC, 996), and molecular functions (MF, 2108). The proposed approach provides a gene set anomaly score for gene sets with molecular functions (MF) by associating 22,000 patient genes and identifying 1644 gene sets that match the patient genes. Thus, each profile is represented in a 1644-dimensional space, which is already smaller than the 22000 dimensions of the input expression data. The dimensionality can be further reduced using standard techniques. This thesis considers PCA and MRMR as exemplar approaches.

3.3.3.2 Principal component analysis (PCA)

PCA is an unsupervised dimension reduction technique that reduces dimensions by transforming a large number of gene sets into a few principal components which capture most of the variation in the gene set anomaly scores [428, 429]. In this work, two principal components are chosen as being suitable for a scatter plot.

In short, PCA finds a transformation matrix U, such that $U \wedge U' = C$, where C is the covariance matrix of gene sets, i.e., covariance of $a(p, s)$ over p. The columns of U are ordered by descending eigenvalue so that the earlier columns transform data into principal components.

3.3.3.3 Maximum relevance minimum redundancy (MRMR)

MRMR is considered an alternative to PCA as it maintains the ability to identify gene sets after dimensional reduction. MRMR ranks gene sets by measuring mutual information, which is a measure of commonality in gene expression data [430]. The two highest-ranking gene sets were used to plot patients considering relapse, non-relapse, cancer, healthy patients outcomes.

MRMR is a feature selection technique that measures the reduction of uncertainty between features by considering associations of one given another [430]. Mutual information is used to identify the dependence between two or more features and calculate the resulting information gain. The value of mutual information can be zero or greater. If the mutual information is zero, it means that the variables are independent. MRMR is used to treat each gene set as a feature. MRMR ranks the gene sets and the top two are chosen to display in the scatter plots.

3.3.3.4 t-Stochastic neighbour embedding

T-SNE [432] simplifies high-dimensional data sets by converting them into low-dimensional data representations. This technique works by calculating the Euclidean distance between data points and using conditional probabilities to determine the relationships between them. By transforming complex data into a more manageable form, T-SNE enables easier analysis and visualization of the relationships within the data.

3.4 Evaluation methodologies and demonstrations

In this section, an evaluation of the proposed method is provided to comprehend values and implications of the anomaly score . This thesis explored three distinct methodologies to evaluate the proposed method, including (1) embedding patient profiles, (2) anomaly distributions, and (3) clustering patient profiles.

3.4.1 Methodology for embedding patients profiles

Patient embedding is an association between patients and gene sets with anomaly scores shown in Figure 3.5. The figure shows that a few patients were embedded over anomaly scores within two patient biology shown in x and y axis.

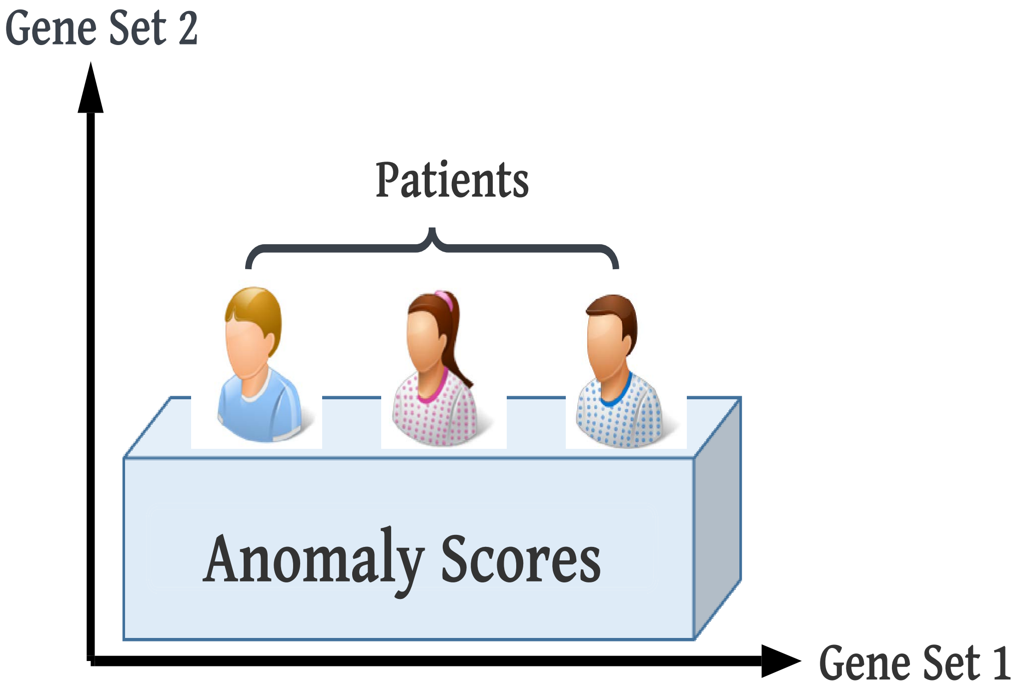


Figure 3.5: A schematic diagram for patient embedding with associated patients' biology.

After anomaly score calculation and dimension reduction, each profile (patient) is represented by two numbers, embedding the profiles in a 2D space. The results of these analyses are summarised in the embedding of patients in 2D and 3D plots in terms of PCA, MRMR, and t-SNE.

3.4.1.1 3D and 2D plots for leukaemia patient embedding with PCA

This section describes the embedding of leukaemia patients in 3D and 2D plots using z-absolute anomaly score for different groups of leukaemia patients in relation to their risk status and treatment plan. Of the 96 leukaemia patients, 13 were classified as high risk, 13 standard risk, 60 medium risk, and 10 low risk. Regarding treatment plans, 18 patients received chemotherapy and BMT, 68 patients received chemotherapy, and 10 healthy patients received no treatment plan. Of all the patients, 24 suffered a cancer relapse and 72 patients had no relapse.

Experimental setup

1. Technology: PCA, MRMR, random forest, t-SNE, and k-means clustering.
2. Python packages: NumPy, pandas, matplotlib , pyplot, math and matplotlib.patches.
3. Input data: Anomaly scores for leukaemia, colon, breast, adrenal cancer and scores for all state-of-the-art methodologies.

Figure 3.6 shows the leukaemia patients in a 3D plot, where each dimension represents a principal component from the z-absolute anomaly score. The figure shows four clusters that are in relatively similar locations for all patients. The figure 3.6 shows the data for all patients without distinguishing between those who relapsed and those who did not. This means that the visualization combines the outcomes for both groups, providing an overall view of the patient population rather than showing the differences between the relapse and non-relapse groups. Drawing from this result, the thesis postulates that clusters are associated with biologically significant differences, rooted in gene expression values.

Figure 3.7a shows leukaemia patients in a 2D plot in relation to treatment planning, where each dimension is a principal component from the z-absolute anomaly score. The treatment plans are chemotherapy and bone marrow transplantation (BMT). These two groups were divided into patients with and without relapse. The figure shows three clusters because both the patients with relapse and those without relapse tend to be in relatively similar locations. One cluster exclusively comprises non-relapse patients, while the other two clusters represent a mixture of patients with and without relapse. Similarly, figure 3.7b shows leukaemia patients in a 2D plot in terms of risk groups, where each dimension is a principal component from the z-absolute anomaly scores. The risk groups are BFM95 and Study 8, which are treatment

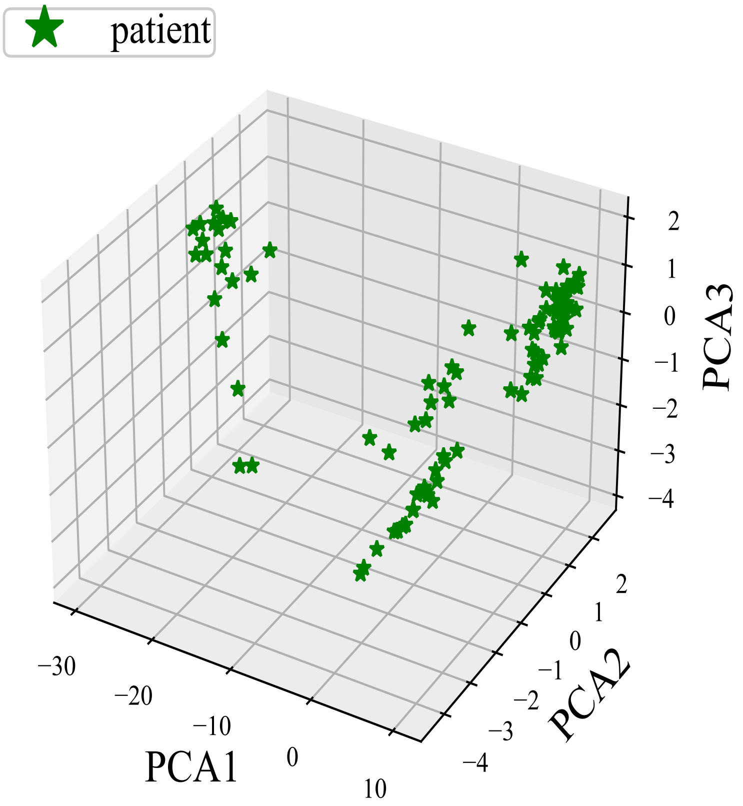
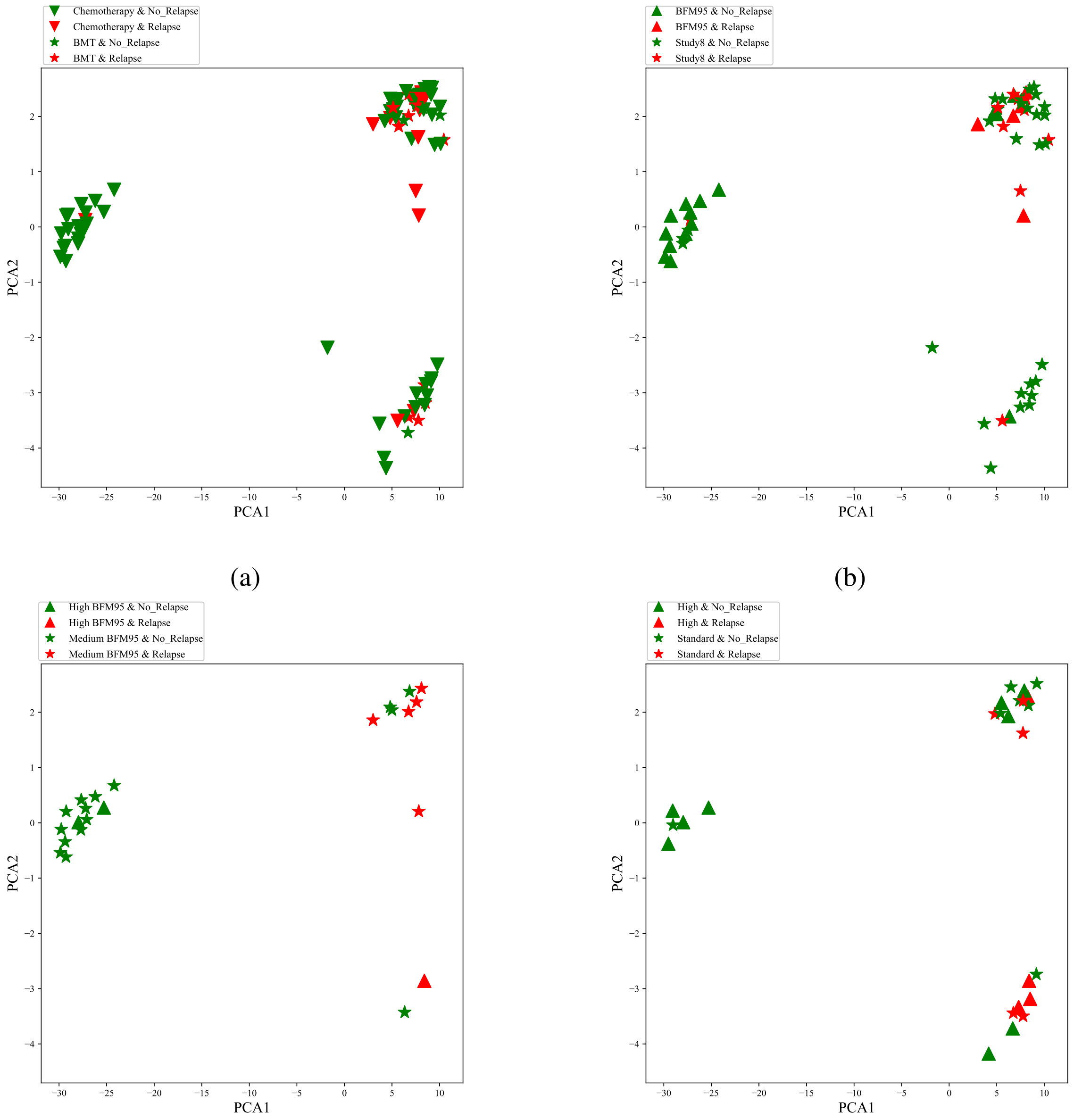


Figure 3.6: 3D projecting using a z-absolute anomaly score for leukaemia patients.

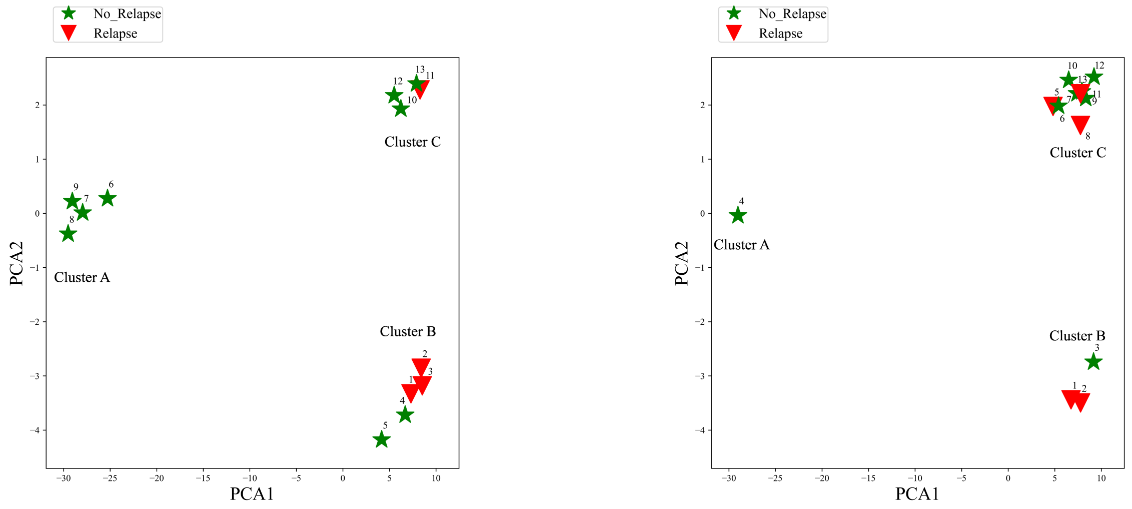
protocols, and these two groups are again divided into patients with relapse and patients without relapse. The figure shows three clusters because both patients with relapse and those without relapse tend to be in relatively similar locations. One cluster is exclusively comprised of patients without relapse, while the other two clusters represent a mixture of patients with and without relapse. Moreover, figure 3.7c shows leukaemia patients in a 2D plot relative to BFM95 for high-risk and medium-risk patients, where each dimension is a principal component of z-absolute anomaly score. The patient groups were divided into recurrences and non-recurrences for high and medium risk patients. The figure shows two clusters because both patients with relapse and patients without relapse tend to be in relatively similar locations. One cluster is exclusively comprised of patients with non-relapse, while the other cluster is a mixture of patients with and without relapse. Finally, figure 3.7d shows leukaemia patients in a 2D diagram for high-risk and standard-risk patients, where each dimension is a principal component from the z-absolute anomaly score. The patient groups were divided into relapse and non-relapse for



(a) (b) (c) (d)

Figure 3.7: Leukaemia patient embedding with z-absolute anomaly scores (a) with respect to treatment planning, (b) all medium patients with respect to BFM95 and study8, (c) for high and medium patients with respect to BFM95, and (d) high and standard risk patients.

both high-risk and standard-risk patients. The figure shows three clusters because both patients with relapse and patients without relapse tend to be in relatively similar locations. One cluster consists entirely of patients without relapse, while the other clusters are a mixture of patients with and without relapse. Drawing from these findings, the proposed hypothesis of this thesis implies that there is a relationship between the clusters and biologically meaningful differences, originate from gene expression measurements.



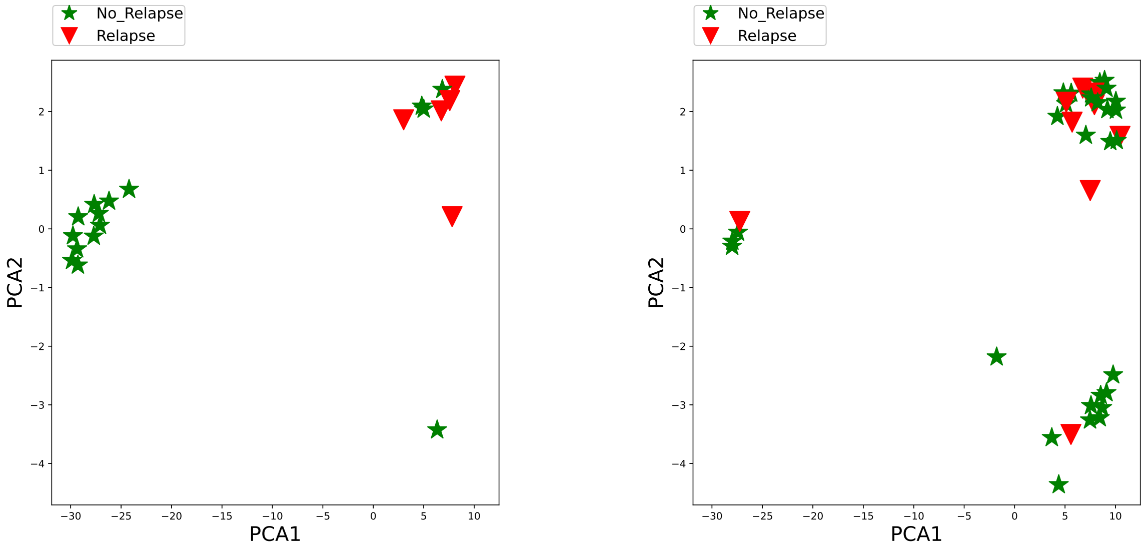
(a) (b)

Figure 3.8: Leukaemia patient embedding using z-absolute anomaly score (a) for high-risk patients and (b) standard-risk patients.

3.4.1.2 Relapse and non-relapse patients embedding with PCA

The results consist of all four types of anomaly scores (z-absolute anomaly score, z-square anomaly score, z-cube anomaly score, and z-mid-range anomaly score) when embedding patients in relation to patient strains, namely relapse, non-relapse, high risk, and medium-risk patients. Different cancer data, namely breast cancer, colorectal cancer, adrenal cancer and leukaemia, were used for the experiments.

Figure 3.8a shows high-risk leukaemia patients in a 2D plot where each dimension represents a principal component from the z-absolute score for abnormalities. The figure shows three clusters (labelled A, B, and C). Cluster A consists entirely of non-relapsed patients with four patients 6, 7, 8, and 9. Clusters B and C contain a mixture of relapsed and non-relapsed patients, but again it can be seen that the relapsed patients tend to be in relatively similar locations. Cluster B contains 3 patients who have relapsed, namely 1, 2, and 3. These three patients form a similar cohort to patients 4 and 5 who have not relapsed. Figure 3.8b shows standard-risk leukaemia patients in a 2D plot where each dimension rep- resents a principal component from the z-absolute anomaly score. The figure shows three clusters (labelled A, B, and C). Cluster A consists of only one patient who has not relapsed. Clusters B and C contain a mixture of relapsed and non-relapsed patients, but again it can be seen that the relapsed patients tend to be in relatively similar locations. Cluster B contains two patients who have relapsed, 1 and 2. These

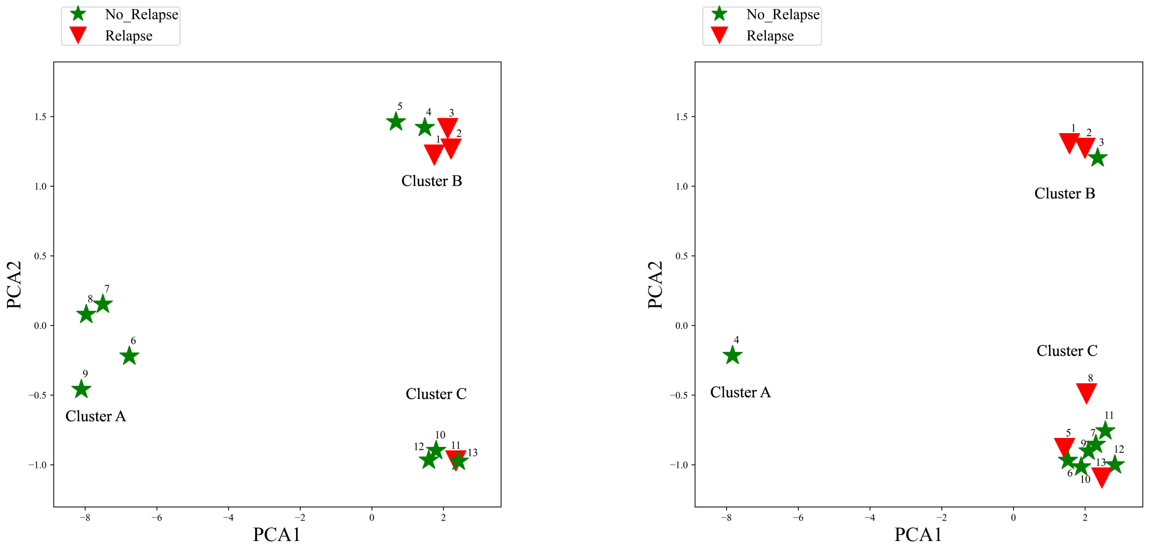


(a) (b)

Figure 3.9: Leukaemia patient embedding using z-absolute anomaly score (a) for medium-risk with BFM95 and (b) medium-risk with study8.

patients form a similar cohort to patient 3, who has not relapsed. Based on these outcomes, the proposed hypothesis of this thesis suggests that there is a correlation between the clusters and biologically significant variances, which originate from gene expression measurements.

Figure 3.9a shows medium-risk leukaemia patients in a 2D plot in relation to BFM95 treatment planning, where each dimension represents a principal component from the z-absolute anomaly score. The figure shows three clusters for patients with and without recurrence. One cluster consists entirely of patients without relapse. The other cluster contains a mixture of patients with and without relapse, but again it can be seen that the patients with relapse tend to reside in relatively similar locations. Again, figure 3.9b shows medium-risk leukaemia patients in a 2D plot in relation to study8 treatment planning, where each dimension represents a principal component from the z-absolute anomaly score. The figure shows three clusters for patients with and without recurrence. One cluster consists entirely of patients without relapse. The other clusters contain a mixture of patients with and without relapse, but again it can be seen that the patients with relapse tend to reside in relatively similar locations. The observation of the z-absolute anomaly score and PCA-based patient embedding results in this thesis indicates that the identified clusters are likely driven by notable dissimilarities in gene expression values. This suggests that the patient groupings are not just random or due to the method used, but have a relationship to biology.



(a)

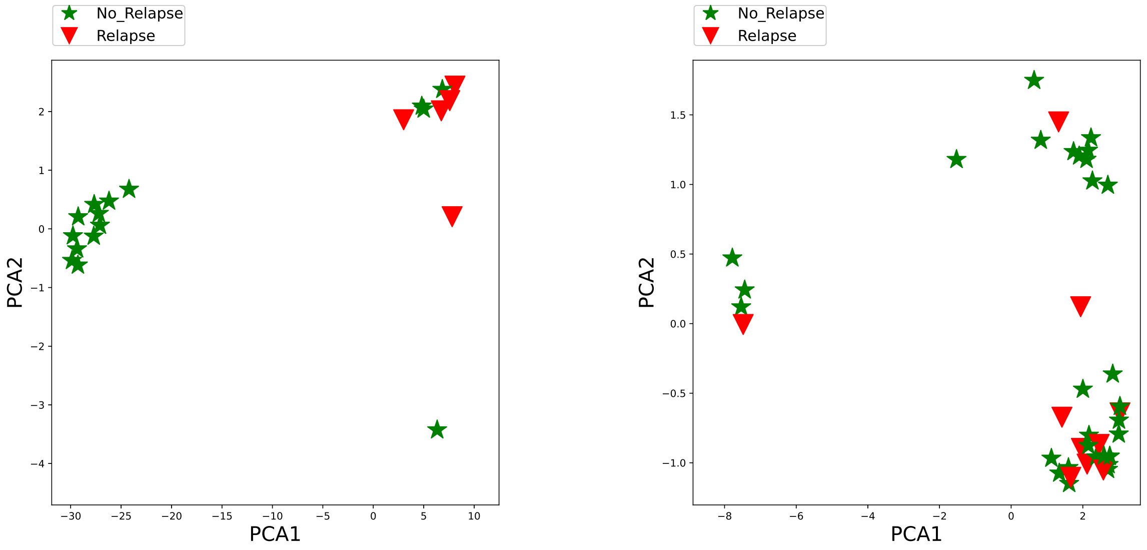
(b)

Figure 3.10: Leukaemia patient embedding using z-square anomaly score (a) for high-risk patients and (b) standard-risk patients.

Additional experiments were conducted using three alternative anomaly scores: the z-square anomaly score, the z-cube anomaly score, and the z-mid-range anomaly score. The outcomes for various patient groups are shown in Figures 3.10a, 3.10b, 3.11a, 3.11b, 3.12a, 3.12b, 3.13a, 3.13b, 3.14a, and 3.14b. These results demonstrate that the clusters generated by the z-square, z-cube and z-mid-range anomaly scores exhibit biologically meaningful differences concerning gene expression values. This finding suggests that the clusters are not merely random groupings, but instead represent significant differences between patient groups. Moreover, these results reinforce the application of multiple anomaly detection methods for identifying biologically relevant patterns across diverse patient groups.

Figure 3.10a shows high-risk leukaemia patients in a 2D plot, where each dimension represents a principal component from the z-square anomaly score. The figure shows three clusters (labelled A, B, and C) for patients with and without relapse. One cluster (cluster A) consists entirely of patients without relapse. The other clusters contain a mixture of patients with and without relapse, but again it can be seen that the patients with relapse tend to reside in relatively similar locations.

Figure 3.10b shows standard-risk leukaemia patients in a 2D plot, where each dimension represents a principal component from the z-square anomaly score. The figure shows three clusters (labelled A, B, and C) for patients with and without relapse. One cluster consists of



(a)

(b)

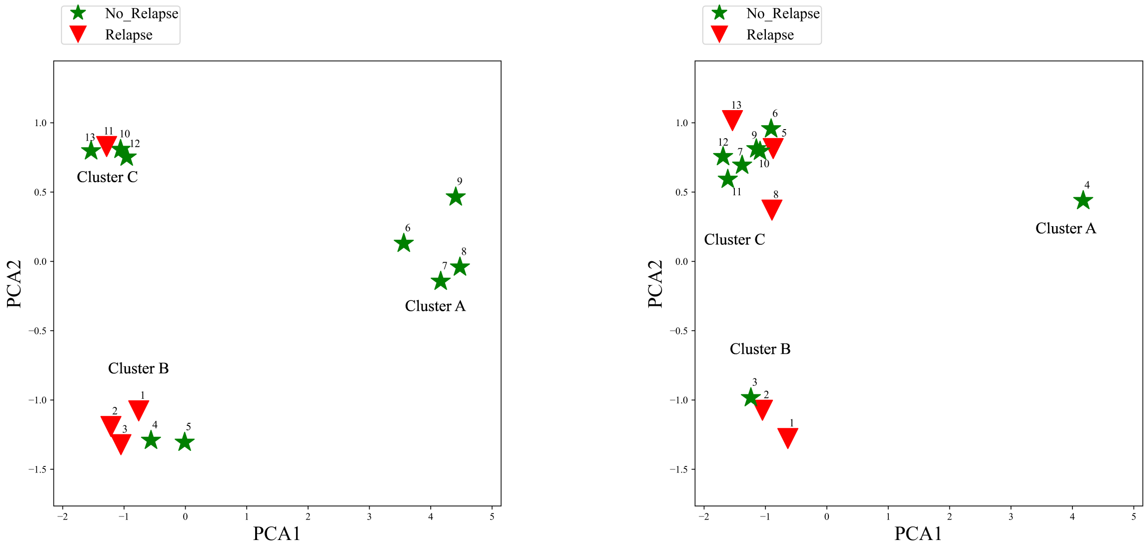
Figure 3.11: Leukaemia patient embedding using z-square anomaly score for (a) medium-risk patients with BFM95 and (b) medium-risk patients with study8.

only one patient (cluster A) without relapse. The other clusters contain a mixture of patients with and without relapse, but again it can be seen that the patients with relapse tend to reside in relatively similar locations.

Figure 3.11a shows medium-risk leukaemia patients in a 2D plot in relation to BFM95 treatment planning, where each dimension represents a principal component from the z-square anomaly score. The figure shows three clusters for patients with and without relapse. One cluster consists entirely of patients without relapse. The other cluster contains a mixture of patients with and without relapse, but again it can be seen that hypothesizes the patients with relapse tend to reside in relatively similar locations.

Figure 3.11b shows medium-risk leukaemia patients in a 2D plot in relation to study8 treatment planning, where each dimension represents a principal component from the z-square anomaly score. The figure shows three clusters for patients with and without relapse. Two clusters consist entirely of patients without relapse except for one relapse patient. The other cluster contains a mixture of patients with and without relapse, but again it can be seen that the patients with relapse tend to reside in relatively similar locations.

Figure 3.12a shows high-risk leukaemia patients in a 2D plot, where each dimension represents a principal component from the z-cube anomaly score. The figure shows three clusters for patients with and without relapse. One cluster consists entirely of patients without relapse. The



(a)

(b)

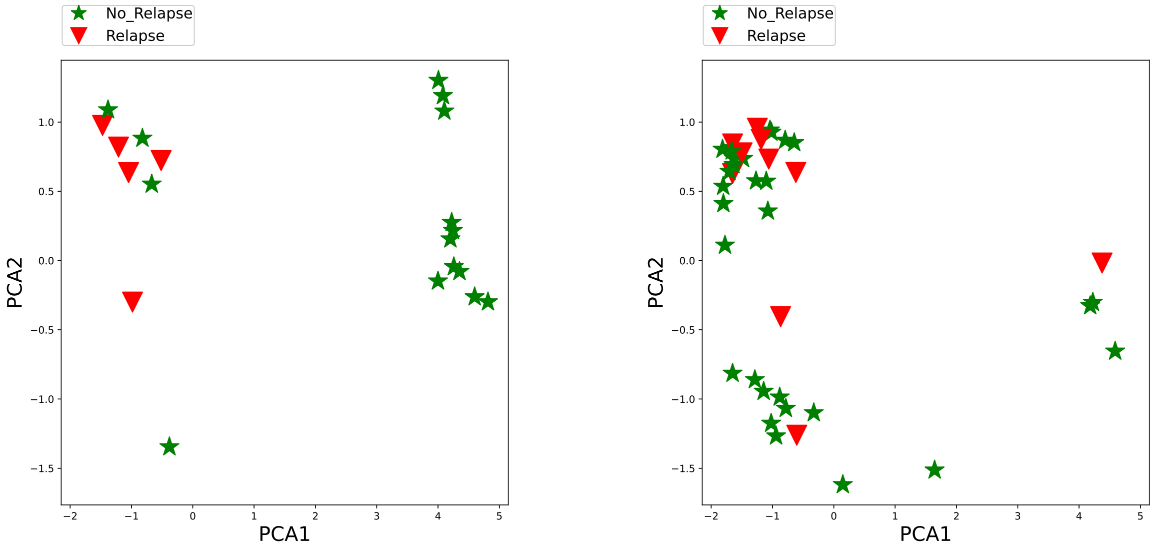
Figure 3.12: Leukaemia patient embedding using z-cube anomaly score (a) for high-risk patients and (b) standard-risk patients.

other cluster contains a mixture of patients with and without relapse, but again it can be seen that the patients with relapse tend to reside in relatively similar locations.

Figure 3.12b shows standard-risk leukaemia patients in a 2D plot, where each dimension represents a principal component from the z-cube anomaly score. The figure shows three clusters for patients with and without relapse. One cluster consists of only one patient without relapse. The other clusters contain a mixture of patients with and without relapse, but again it can be seen that the patients with relapse tend to reside in relatively similar locations.

Figure 3.13a shows medium-risk leukaemia patients in a 2D plot in relation to BFM95 treatment planning, where each dimension represents a principal component from the z-cube anomaly score. The figure shows three clusters for patients with and without relapse. Two clusters consist entirely of patients without relapse. The other cluster contains a mixture of patients with and without relapse, but again it can be seen that the patients with relapse tend to reside in relatively similar locations.

Figure 3.13b shows medium-risk leukaemia patients in a 2D plot in relation to study8 treat- ment planning, where each dimension represents a principal component from the z-cube anomaly score. The figure shows three clusters for patients with and without relapse. Two clusters consist entirely of patients without relapse except for one relapse patient. The other cluster contains a mixture of patients with and without relapse, but again it can be seen that the



(a)

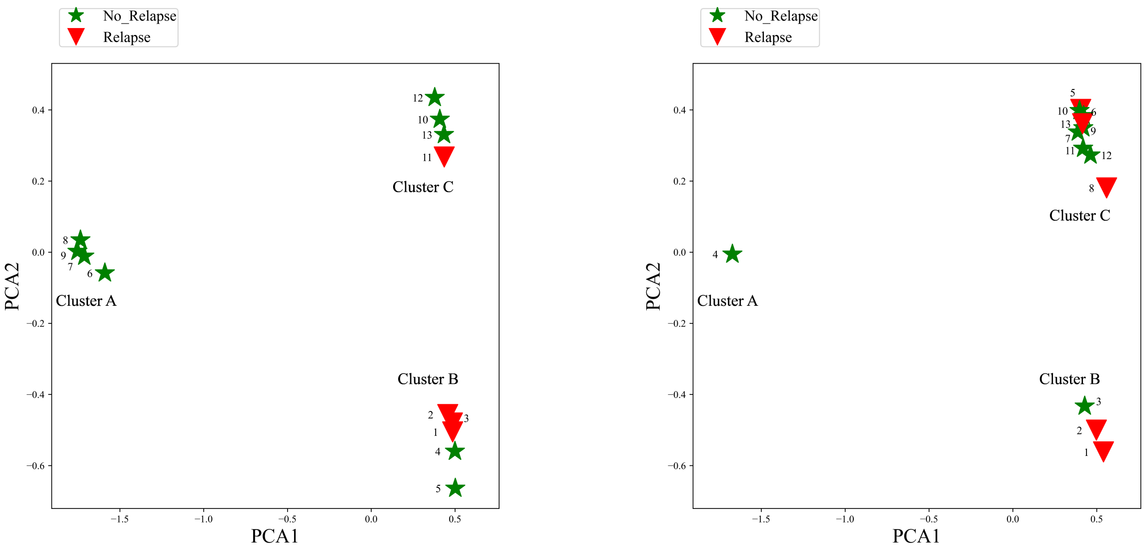
(b)

Figure 3.13: Leukaemia patient embedding using z-cube anomaly score for (a) medium-risk patients with BFM95 and (b) medium-risk patients with study8.

patients with relapse tend to reside in relatively similar locations.

Figure 3.14a shows high-risk leukaemia patients in a 2D plot, where each dimension represents a principal component from the z-mid-range anomaly score. The figure shows three clusters for patients with and without relapse. One cluster consists entirely of patients without relapse. The other clusters contain a mixture of patients with and without relapse, but again it can be seen that the patients with relapse tend to reside in relatively similar locations.

Figure 3.14b shows standard-risk leukaemia patients in a 2D plot, where each dimension represents a principal component from the z-mid-range anomaly score. The figure shows three clusters for patients with and without recurrence. One cluster consists of only one patient without relapse except. The other clusters contain a mixture of patients with and without relapse, but again it can be seen that the patients with relapse tend to reside in relatively similar locations.



(a)

(b)

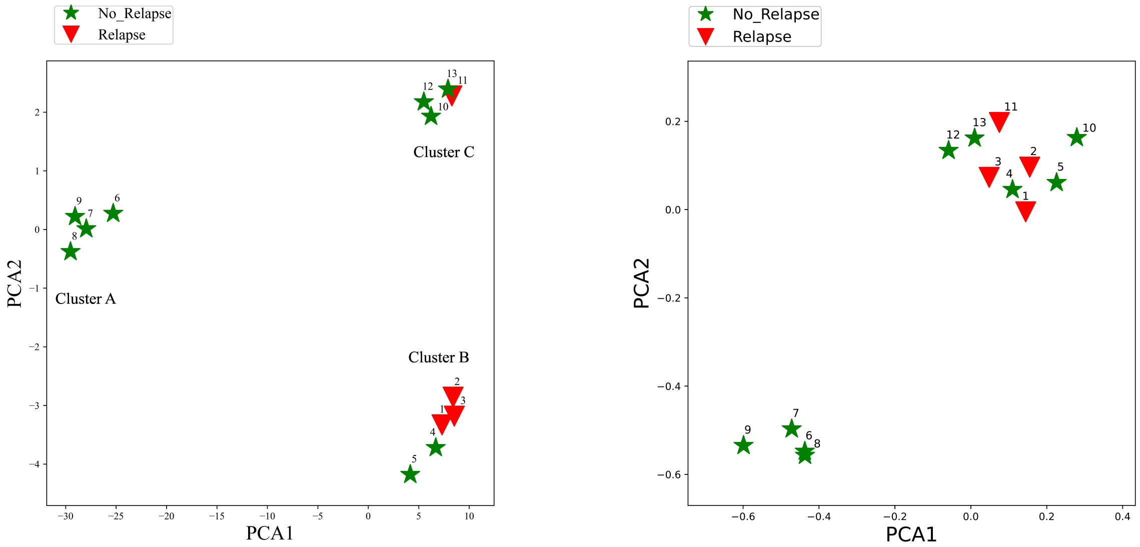
Figure 3.14: Leukaemia patient embedding using z-mid-range anomaly score for (a) high-risk patients and (b) standard risk-risk patients.

3.4.1.3 Comparative analysis between raw gene expression and anomaly score with PCA

In this thesis, a comparative analysis has been carried out between raw gene expression values and anomaly score patient embeddings. The aim of this analysis is to assess the effectiveness of anomaly score patient embeddings in supporting the hypothesis put forth in this thesis.

Two sets of visual representations have been generated for this comparison. The first set, comprising Figures 3.15a and 3.16a, illustrates the results derived from the anomaly score patient embeddings. The second set, consisting of Figures 3.15b and 3.16b, presents the data obtained from the raw gene expression values.

Upon careful evaluation of the outcomes derived from both sets of figures, it becomes evident that the anomaly score patient embeddings indeed provide substantial evidence in favor of the thesis hypothesis. The comparative analysis highlights the effectiveness of anomaly score patient embeddings over raw gene expression values, reinforcing the argument presented in the study.



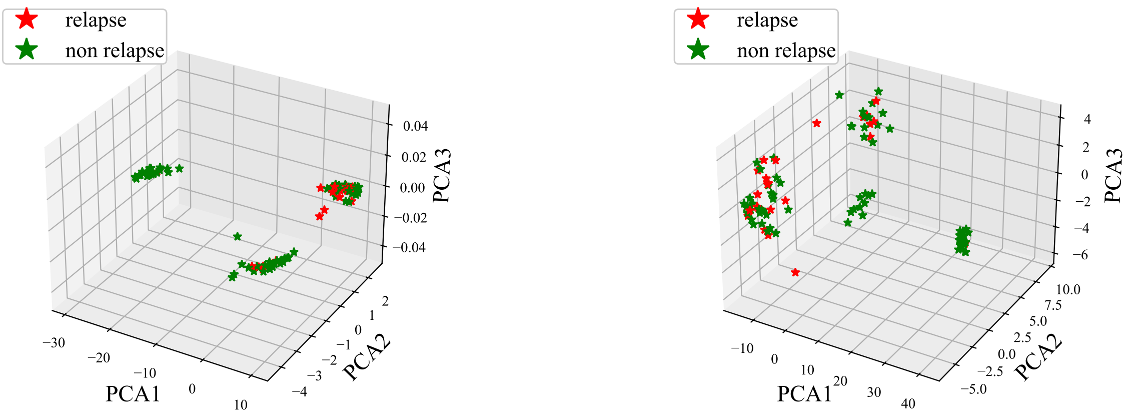
(a)

(b)

Figure 3.15: Leukaemia patient embedding (a) using z-absolute anomaly score and (b) raw gene expression values.

Figure 3.15a shows high-risk leukaemia patients in a 2D plot, where each dimension represents a principal component from the z-absolute anomaly score. The figure shows three clusters (labelled A, B, and C) for patients with and without relapse. Cluster A consist entirely of patients without relapse. The other clusters (cluster B and C) contain a mixture of patients with and without relapse, but again it can be seen that the patients with relapse tend to reside in relatively similar locations.

Figure 3.15b shows high-risk leukaemia patients in a 2D plot, where each dimension represents a principal component from the raw gene expression values. The figure shows that three patients are scattered in contrast with the anomaly score patient embedding.



(a)

(b)

Figure 3.16: Leukaemia patient embedding in a 3D plot (a) using z-absolute anomaly scores and (b) raw gene expression values.

Figure 3.16a shows high-risk leukaemia patients in a 3D plot where each dimension represents a principal component from the z-absolute anomaly score. The figure shows three clusters for patients with and without relapse. Two clusters consist entirely of patients without relapse. The other cluster contains a mixture of patients with and without relapse, but again it can be seen that the patients with relapse tend to reside in relatively similar locations.

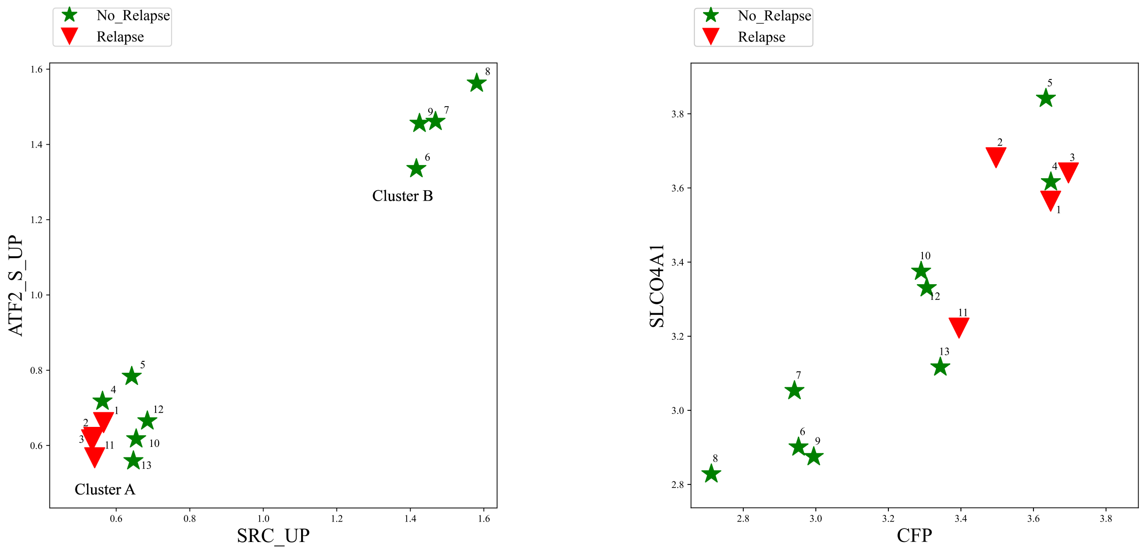
Figure 3.16b shows high-risk leukaemia patients in a 3D plot where each dimension represents a principal component from the raw gene expression values. The figure shows four clusters for patients with and without relapse. It can be seen that the patients were scattered in contrast with the anomaly score patient embedding for the raw gene expression embedding.

3.4.1.4 MRMR patient embedding

In the previous section, patient embedding results were discussed in relation to the principal components derived from PCA. Although PCA effectively identifies principal components, it does not revealing the specific gene sets involved as it does not provide gene set names. Capturing gene set names, rather than relying on principal components, could improve patient embedding and provide a more comprehensive understanding of the underlying biological processes.

This thesis presents an overview of patient embedding, focusing on gene sets that have a strong association with patients based on their anomaly scores. To effectively identify these gene sets, the thesis applies the MRMR, a feature selection approach. By measuring the mutual information of the gene sets, MRMR allows ranking these gene sets in terms of their associations.

By identifying and applying specific gene sets to patient groups, it may helps to determine which biological functions are closely associated with those groups. This improves patients embedding process and allows for a better understanding of the gene sets that play a role in different patient populations.



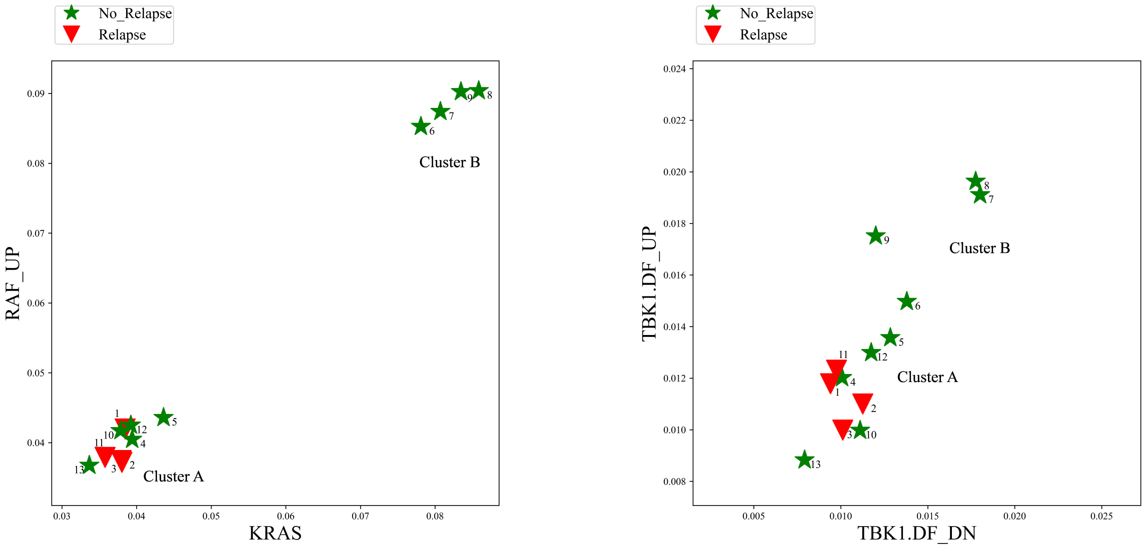
(a) (b)

Figure 3.17: MRMR high-risk leukaemia patient embedding (a) using z-absolute anomaly scores and (b) using raw gene expression values.

Figure 3.17a, 3.18a, and 3.18b demonstrate the MRMR patient embeddings with respect to z-absolute, z-square, and z-cube anomaly scores. In contrast, Figure 3.17b shows the MRMR patients' embedding using raw gene expression values. These visualizations support the thesis hypothesis, suggesting that utilizing MRMR anomaly scores for patient embeddings can provide relationship between the identified clusters and biologically significant distinctions. The clustering patterns observed when applying anomaly scores appear to be more meaningful than those based solely on raw gene expression values.

Figure 3.17a shows leukaemia patients at high risk in a 2D representation. Each dimension is a gene set ordered by MRMR based on z-absolute anomaly score. The figure shows two clusters (labelled A and B). Cluster B consists entirely of non-recurrent patients. Cluster A contains a mixture of relapsed and non-relapsed patients, but again we see that the relapsed patients tend to be in relatively similar locations.

Figure 3.17b shows the embedding of high-risk leukaemia patients with respect to raw gene expression values considering ranked genes using MRMR. Examining the raw gene expression data, the figure shows that the patients are scattered and there is no evidence of clustering.



(a)

(b)

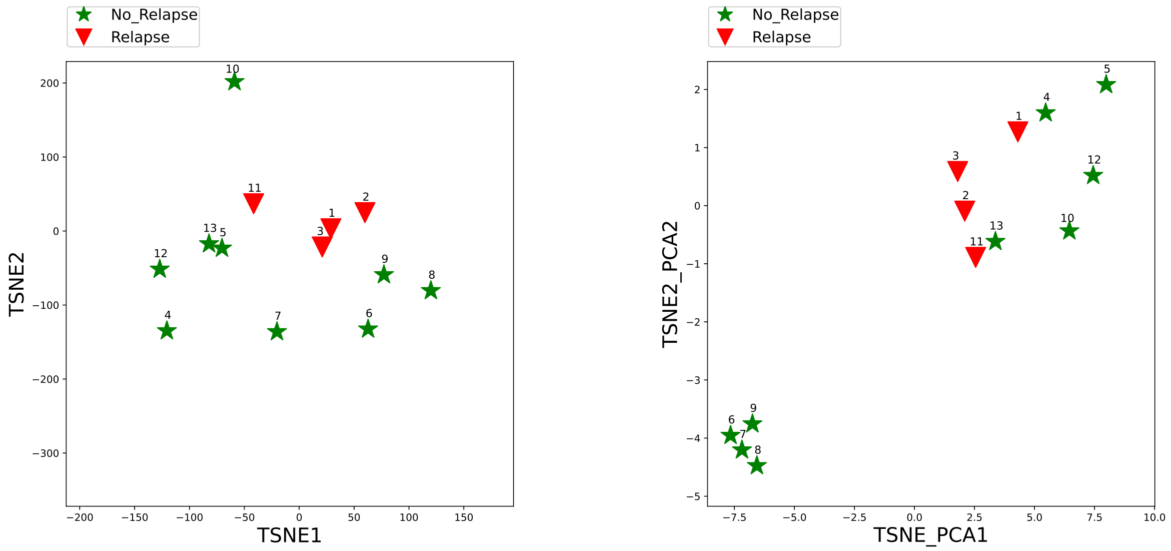
Figure 3.18: MRMR high-risk patient embedding (a) using z-square anomaly score and (b) using z-cube anomaly score.

Figure 3.18a shows leukaemia patients at high risk in a 2D representation. Each dimension is a gene set ordered by MRMR based on z-square anomaly score. The figure shows two clusters (labelled A and B). Cluster B consists entirely of non-relapse patients. Cluster A contains a mixture of relapsed and non-relapsed patients, but again it can be seen that the relapsed patients tend to be in relatively similar locations.

Figure 3.18b shows the embedding of high-risk leukaemia patients with respect to z-cube anomaly score considering gene sets ordered by MRMR. The figure shows two clusters (labelled A and B). Cluster B consists entirely of non-relapse patients. Cluster A contains a mixture of relapsed and non-relapsed patients, but again it can be seen that the relapsed patients tend to be in relatively similar locations.

3.4.1.5 Patient embedding with t-SNE

This section describes the results of patient mapping or embedding with t-SNE, which can reduce the dimensions of large gene sets to highly diversified 2D gene sets. Figure 3.19a and Figure 3.19b show patient embeddings using t-SNE based on anomaly scores. These results suggest that there is a correlation between the observed clusters and biologically relevant differences, which can be explained by meaningful variations in gene expression values.



(a)

(b)

Figure 3.19: Leukaemia high-risk patient embedding for z-absolute anomaly score (a) with t-SNE and (b) with PCA based t-SNE.

Figure 3.19a shows high-risk leukaemia patients in a 2D representation where each dimension represents a t-SNE from the z-absolute anomaly score. The figure shows three clusters for patients with and without relapse. All clusters consist only of identical patients, i.e., there are no mixed patients in any cluster.

Figure 3.19b shows high-risk leukaemia patients in a 2D plot where each dimension represents a PCA-based t-SNE from the z-absolute anomaly score. The figure shows three clusters for patients with and without relapse. All clusters consist only of identical patients, i.e., there are no mixed patients in any cluster.

3.4.2 Comparative analysis of a proposed method and state-of-the-art methodologies

Different modified gene expression scores of the existing methods were implemented on leukaemia gene expression data and the results of the proposed method were compared with these existing methods. This section describes the comparative results of the proposed methods and the existing different modified gene expression scores. The comparative analysis starts with the Gene Fuzzy Scores (GFS) and the proposed method.

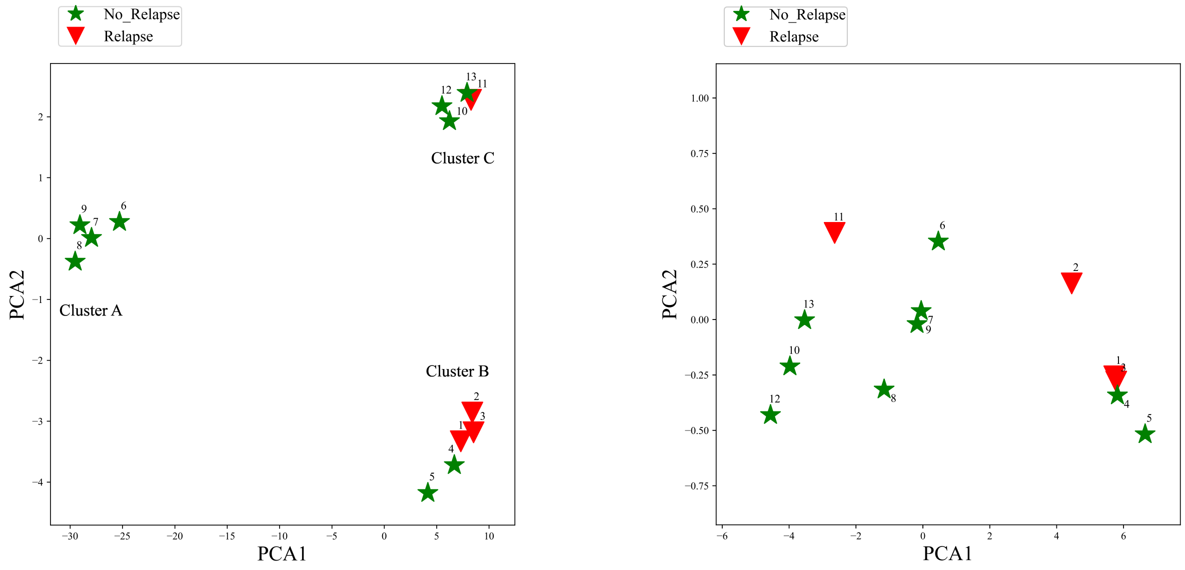
3.4.2.1 Comparative analysis between anomaly score and gene fuzzy score

Figure 3.20a shows patient embedding for high-risk leukaemia patients with z-absolute anomaly score using PCA. The figure shows that three different clusters were formed for both patients with relapse and patients without relapse. Cluster A consists of four patients who did not relapse, and clusters B and C show patients with their corresponding positions.

Figure 3.20b shows patient embedding for high-risk leukaemia patients with gene fuzzy scores using PCA. From the figure, it can be seen that the patients are more scattered than in Figure 3.20a.

Figure 3.21a shows patient embedding for high-risk leukemia patients with a z-score anomaly using MRMR. The figure shows that two clusters A and B clearly show relapse and non-relapse patients with respect to two strongly associated gene sets (SRC_ UP.V1_ DN and ATF2_S_ UP.V1_ DN) in the MRMR ranking.

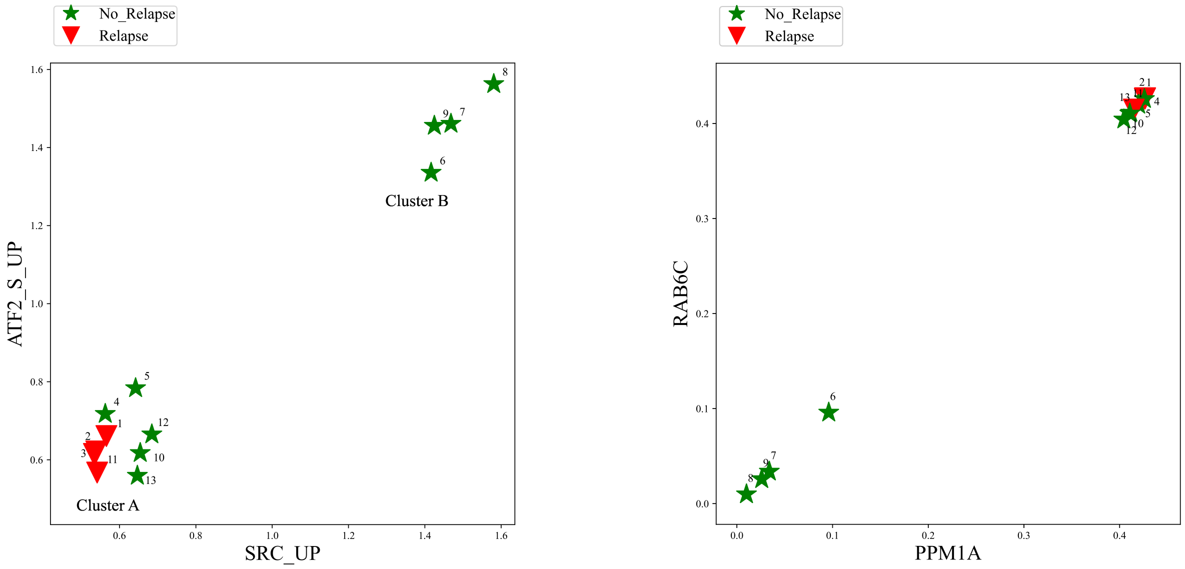
Figure 3.21b shows patient embedding for high-risk leukemia patients with the GFS using MRMR in relation to gene ranking. Here, PPM1A and RAB6C are the top two genes from the MRMR ranking. The figure shows that patients with and without relapse have a larger spread than in figure 3.21a.



(a)

(b)

Figure 3.20: Leukaemia patient embedding with PCA (a) using z-absolute anomaly score and (b) gene fuzzy score.



(a)

(b)

Figure 3.21: Leukaemia patient embedding with MRMR (a) using z-absolute anomaly score and (b) gene fuzzy score.

3.4.2.2 Anomaly score and feature regression and classification score

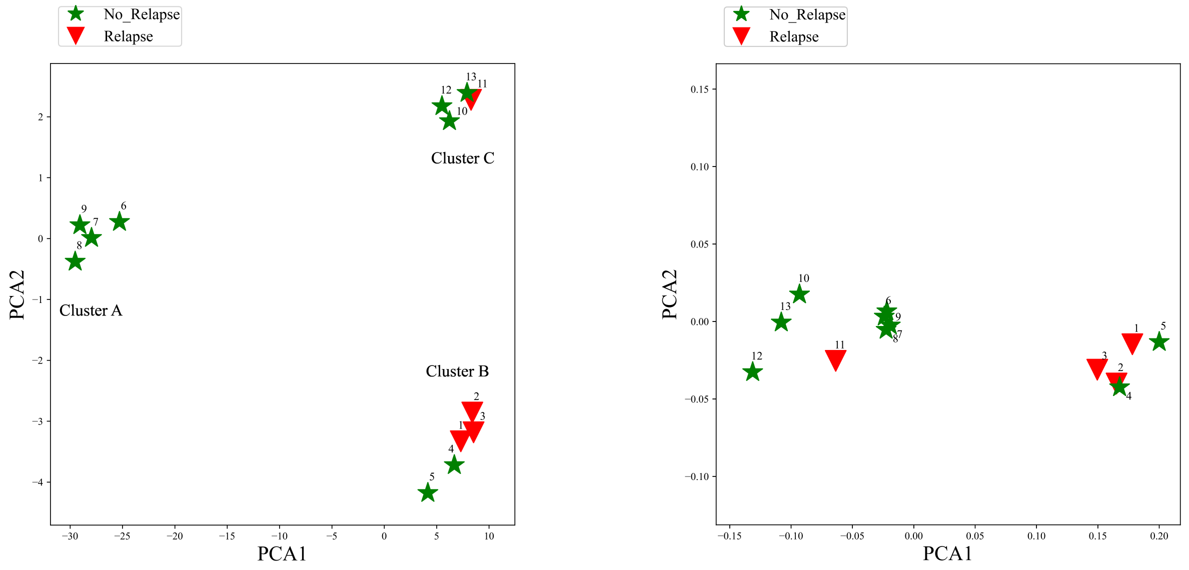
The following section compares the results between the anomaly scores of the proposed method and the Feature Regression and Classification (FRaC) scores for leukaemia gene expression data sets. The results show patient embedding with PCA and MRMR for high-risk leukaemia patients.

Figure 3.22a shows patient embedding for high-risk leukaemia patients with z-absolute anomaly score using PCA. The results are described in detail in Figure 3.20a.

Figure 3.22b shows embedding for high-risk leukaemia patients with feature regression and classi- fication using PCA. The figure shows that the patients are more scattered than in Figure 3.22a.

Figure 3.23a shows patient embedding for high-risk leukemia patients with z-absolute anomaly score using MRMR. The results are described in detail in Figure 3.21a.

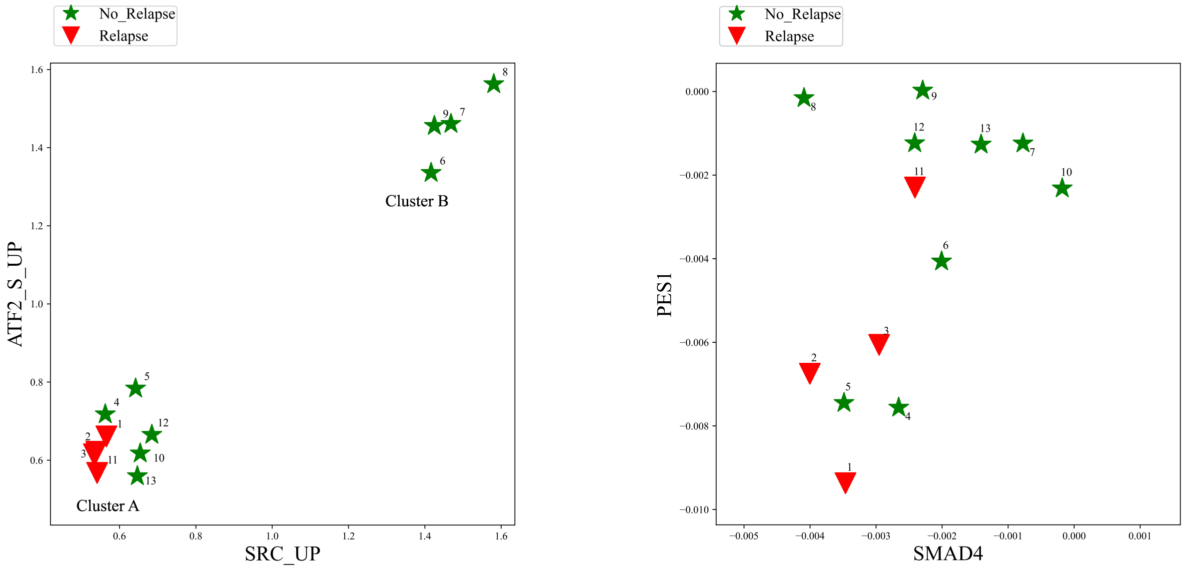
Figure 3.23b patient embedding for high-risk leukemia patients with feature regression and classification using MRMR in relation to gene ranking. Here, SMAD4 and PES1 are the top two genes from the MRMR ranking. The figure shows that the patients with and without relapse have a larger spread than in Figure 3.23a.



(a)

(b)

Figure 3.22: Leukaemia patient embedding with PCA (a) using z-absolute score anomaly and (b) FRaC.



(a)

(b)

Figure 3.23: Leukaemia patient embedding with MRMR (a) using z-absolute anomaly score and (b) FRaC.

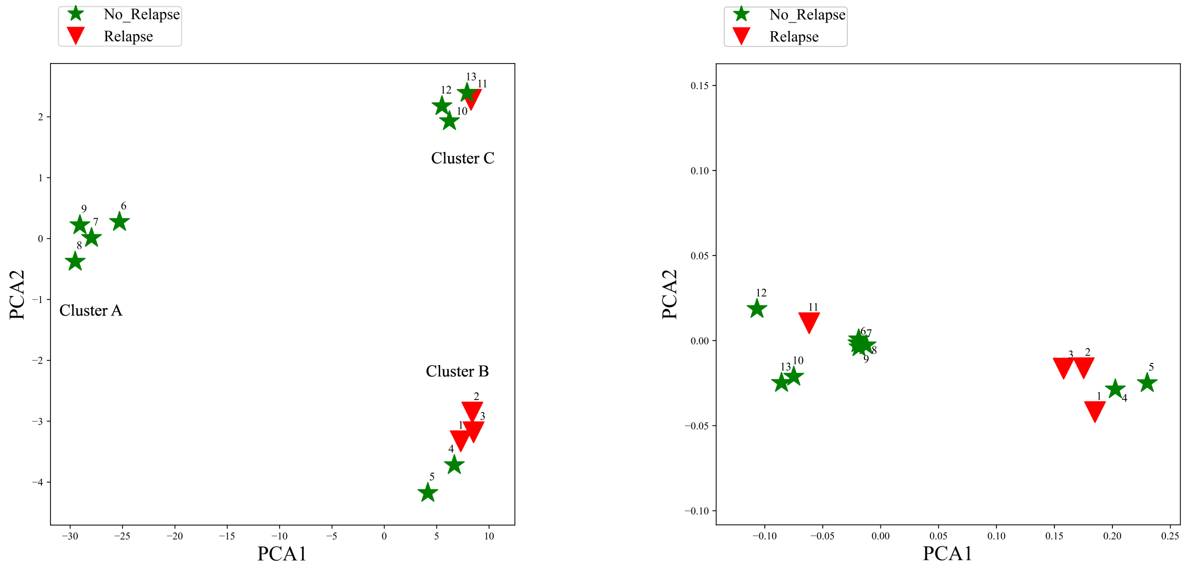
3.4.2.3 Comparative analysis between anomaly score and CSAX

Patient embedding was performed using anomaly scores and CSAX gene expression scores with PCA and MRMR. The results of this study are summarized in this section.

Figure 3.24a shows patient embedding for high-risk leukaemia patients with z-absolute anomaly score anomaly using PCA. Figure 3.24b shows patient embedding for high-risk leukaemia patients with CSAX using PCA. The figure shows that the patients are more scattered than in Figure 3.24a.

Figure 3.25a shows patient embedding for high-risk leukaemia patients with z-absolute anomaly scores using MRMR. The results are described in detail in Figure 3.21a.

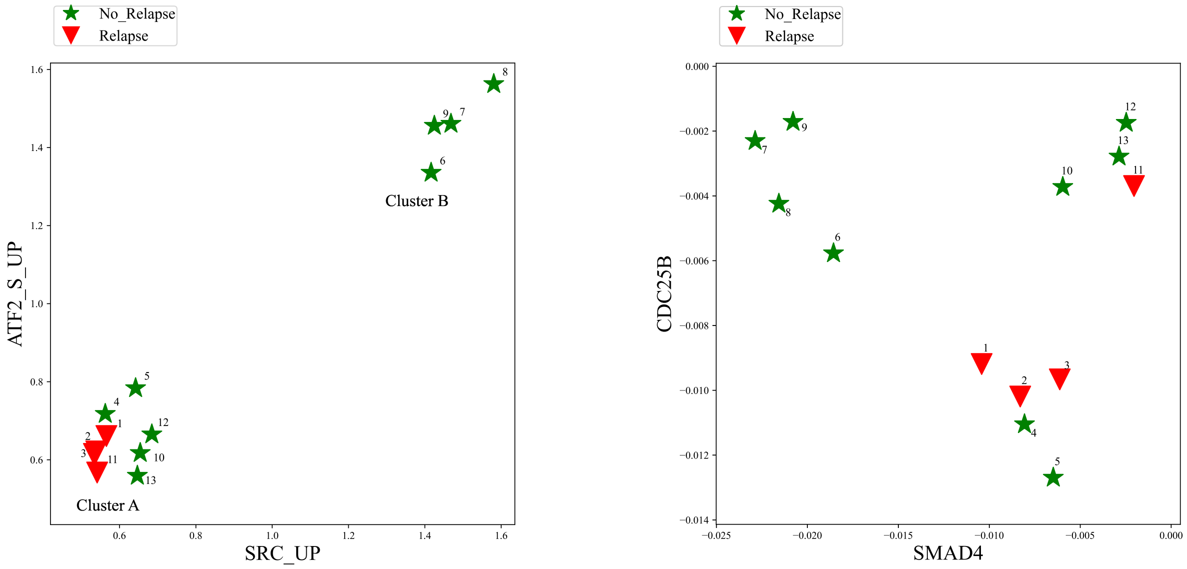
Figure 3.25b shows patient embedding for high-risk leukemia patients with feature regression and classification using MRMR in relation to gene ranking. Here, SMAD4 and CDC25B are the top two genes from the MRMR ranking. The figure shows that the patients with and without relapse have a larger spread than in Figur 3.25a.



(a)

(b)

Figure 3.24: Leukaemia patient embedding with PCA (a) using z-absolute anomaly score and (b) CSAX.



(a)

(b)

Figure 3.25: Leukaemia patient embedding with MRMR (a) using z-absolute anomaly score and (b) CSAX.

3.4.2.4 Comparative analysis between anomaly Score and TEMPO

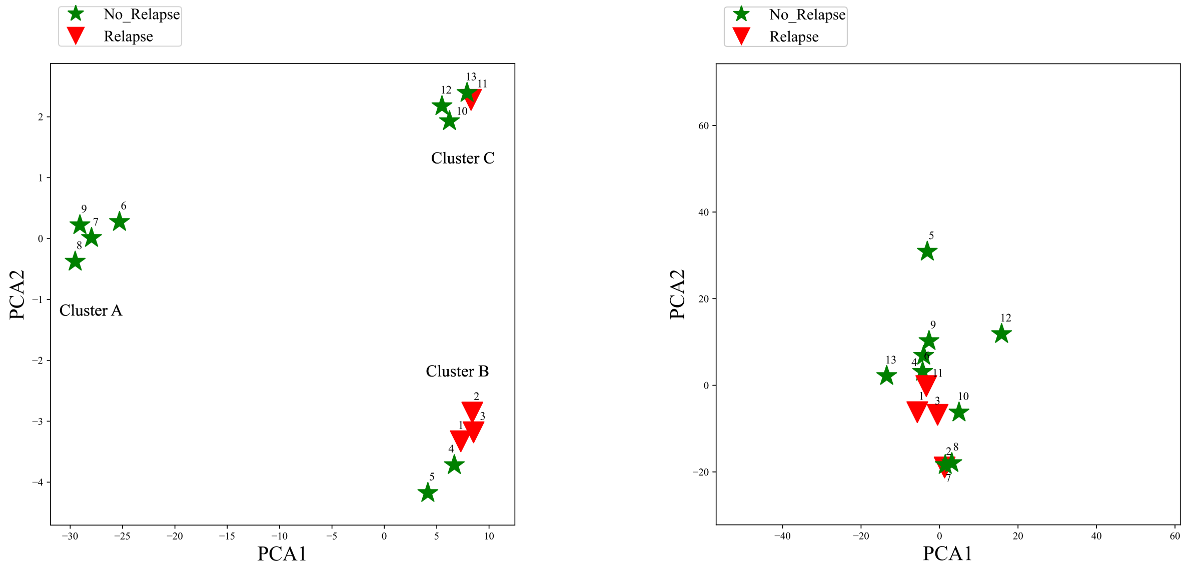
This section describes an experimental evaluation between anomaly scores and TEMPO gene expression scores. The evaluations were performed for patient embedding for high-risk leukemia with PCA and MRMR.

Figure 3.26a shows patient embedding for high-risk leukaemia patients with z-absolute anomaly score using PCA.

Figure 3.26b shows patient embedding for high-risk leukaemia patients with TEMPO using PCA. The figure shows that the patients are more scattered than in Figure 3.26a.

Figure 3.27a shows patient embedding for high-risk leukemia patients with z-absolute anomaly score using MRMR.

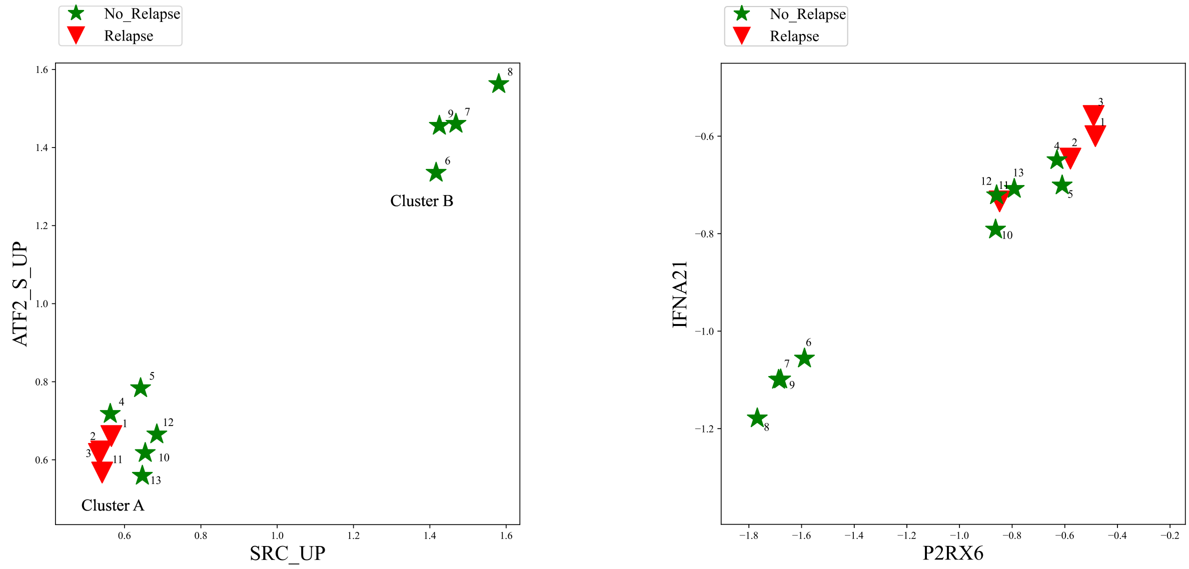
Figure 3.27b shows patient embedding for high-risk leukemia patients with TEMPO using MRMR in relation to gene ranking. Here, JAK3 and THPO are the top two genes from the MRMR ranking. The figure shows that patients with and without relapse have a larger spread than in Figure 3.27a.



(a)

(b)

Figure 3.26: Leukaemia patient embedding with PCA (a) using z-absolute anomaly score and (b) TEMPO.



(a)

(b)

Figure 3.27: Leukaemia patient embedding with MRMR (a) using z-absolute anomaly score and (b) TEMPO.

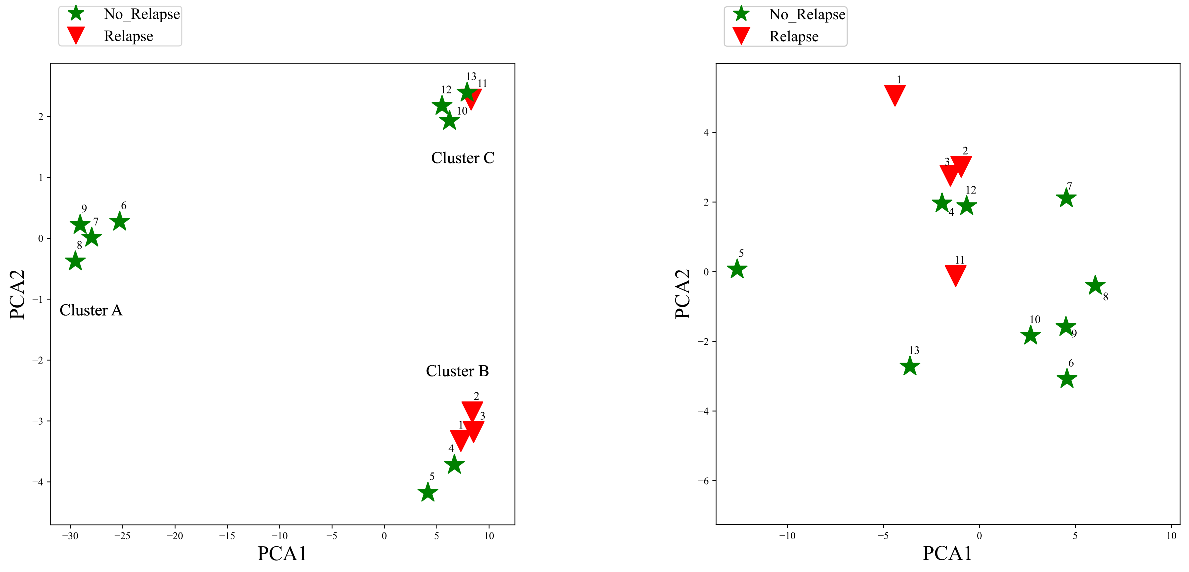
3.4.2.5 Comparative analysis between anomaly score and aTEMPO

This section compares the descriptions between anomaly scores and aTEMPO gene expression scores for embedding for high-risk leukaemia patients with PCA and MRMR.

Figure 3.28a shows embedding for high-risk leukaemia patients with z-absolute anomaly score using PCA. Figure 3.28b shows embedding for high-risk leukaemia patients using aTEMPO gene expression scores with PCA. The figure shows that the patients are more scattered than in Figure 3.28a.

Figure 3.29a shows embedding for high-risk leukemia patients using z-absolute anomaly score anomaly with MRMR.

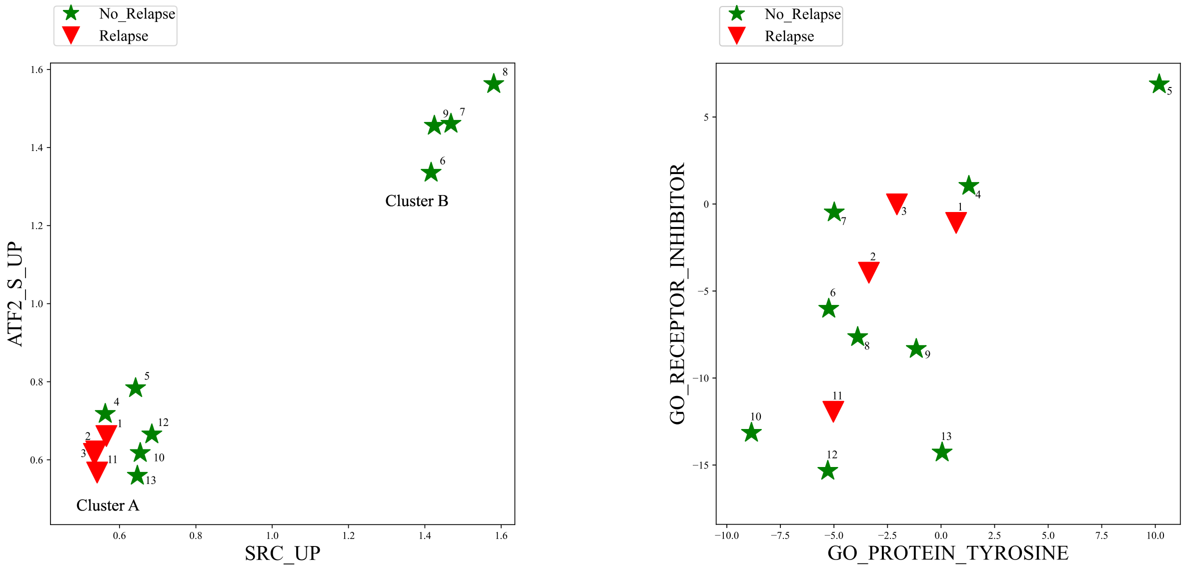
Figure 3.29b shows patient embedding for high-risk leukemia patients using aTEMPO gene expression scores with MRMR in relation to gene ranking. Here, GO_CYSTEINE and GO_ACTIN are the top two gene sets from the MRMR ranking. The figure shows that the patients with and without relapse have a larger spread than in Figure 3.29a.



(a)

(b)

Figure 3.28: Leukaemia patient embedding with PCA (a) using z-absolute anomaly score and (b) aTEMPO.



(a)

(b)

Figure 3.29: Leukaemia patient embedding with MRMR (a) using z-absolute anomaly score and (b) aTEMPO.

3.4.2.6 Comparative analysis between anomaly score and outlier analysis method

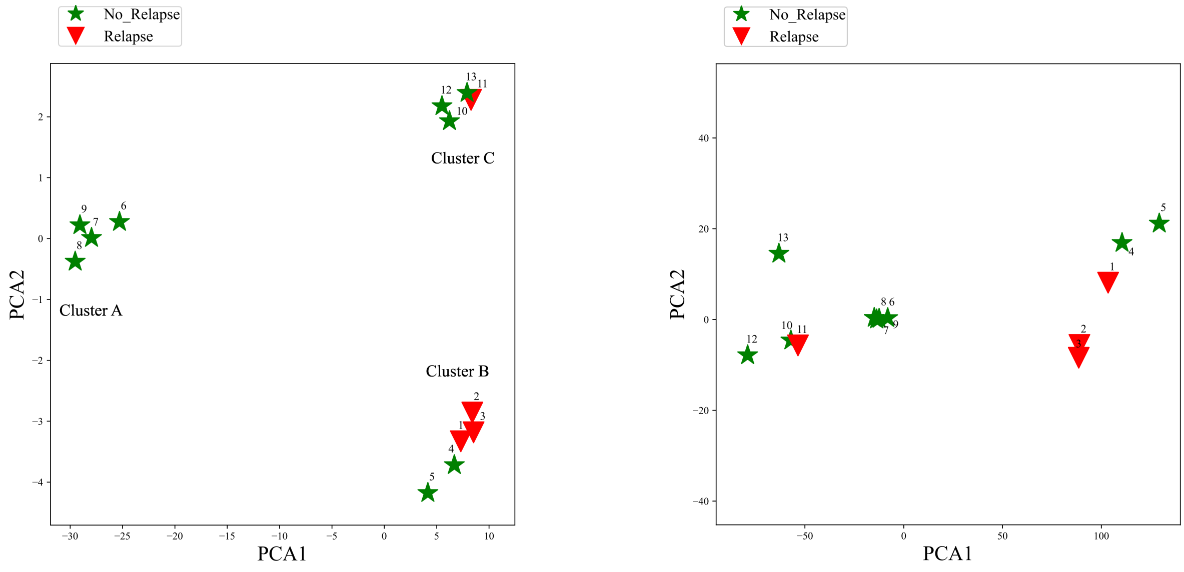
This section compares the descriptions between anomaly scores and outlier gene identification scores for embedding for high-risk leukaemia patients with PCA and MRMR.

Figure 3.30a patient embedding for high-risk leukaemia patients with z-absolute anomaly score using PCA.

Figure 3.30b shows patient embedding for high-risk leukaemia patients with outlier detection method using PCA. The figure shows that the patients are more scattered than in Figure 3.30a.

Figure 3.31a patient embedding for high-risk leukemia patients with z-absolute anomaly score using MRMR.

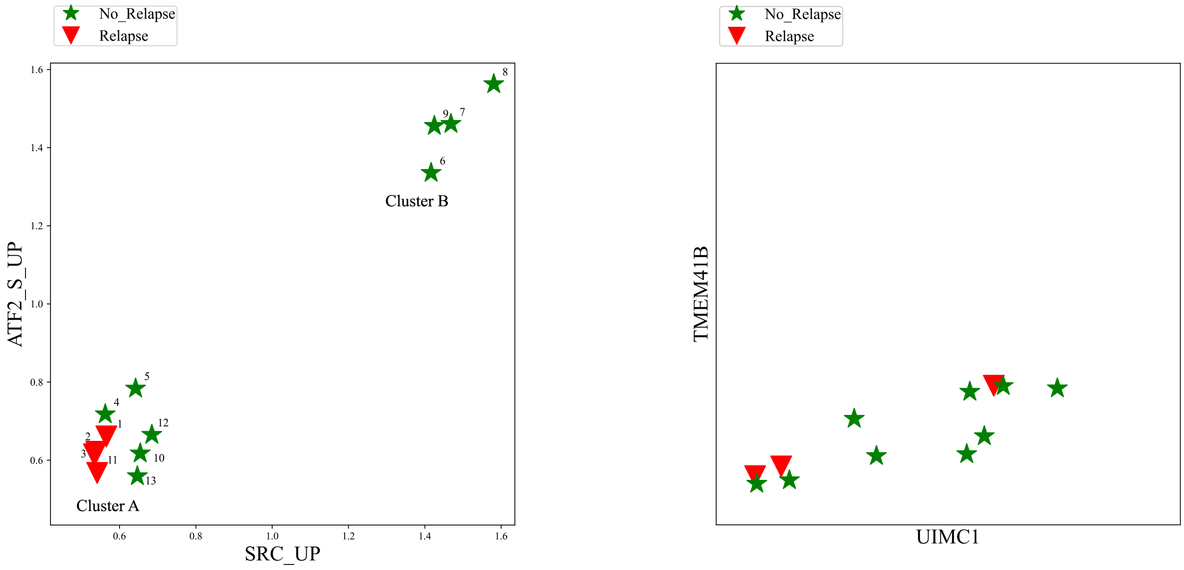
Figure 3.31b patient embedding for high-risk leukemia patients with aTEMPO using MRMR in relation to gene ranking. Here, UIMC1 and TMEM41B are the top two genes from the MRMR ranking. The figure shows that the patients with and without relapse have a larger spread than in Figure 3.31a.



(a)

(b)

Figure 3.30: Leukaemia patient embedding with MRMR (a) using z-absolute anomaly score and (b) Outliers detection approach.



(a)

(b)

Figure 3.31: Leukaemia patient embedding with MRMR (a) using z-absolute anomaly score and (b) Outliers detection approach.

3.4.2.7 Comparative analysis between anomaly score and SNet

This section compares the analysis with SNet and the proposed method.

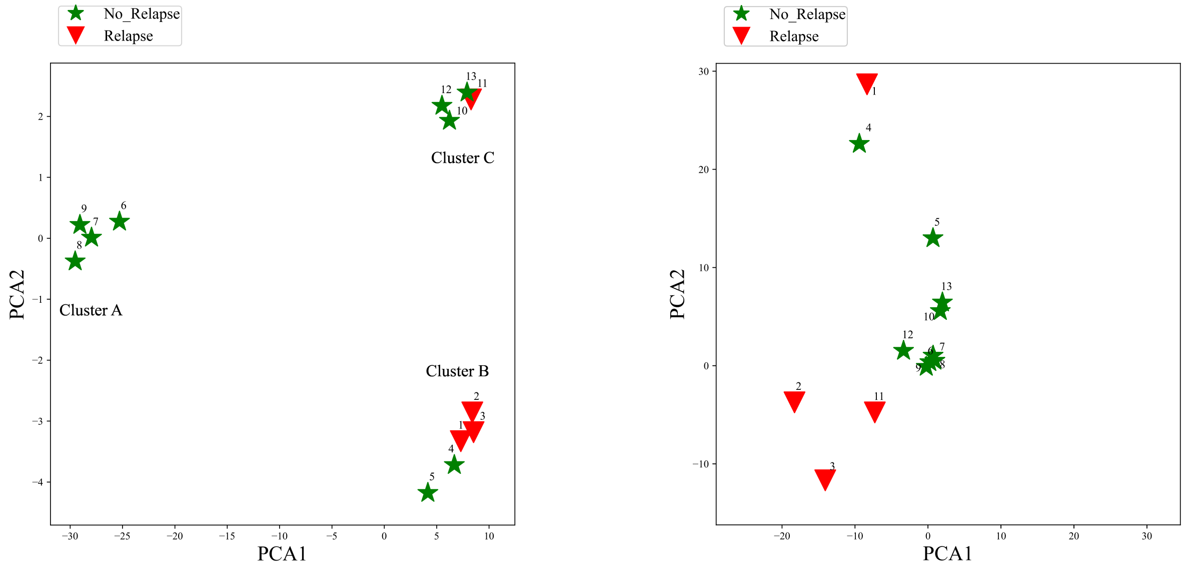
This section provides comparative descriptions between anomaly scores and SNet gene expression scores for embedding for high-risk leukaemia patients using PCA and MRMR.

Figure 3.32a patient embedding for high-risk leukaemia patients with z-absolute anomaly score using PCA.

Figure 3.32b shows patient embedding for high-risk leukaemia patients with SNet using PCA. The figure shows that the patients are more scattered than in Figure 3.32a.

Figure 3.33a shows patient embedding for high-risk leukemia patients with z-absolute anomaly score using MRMR.

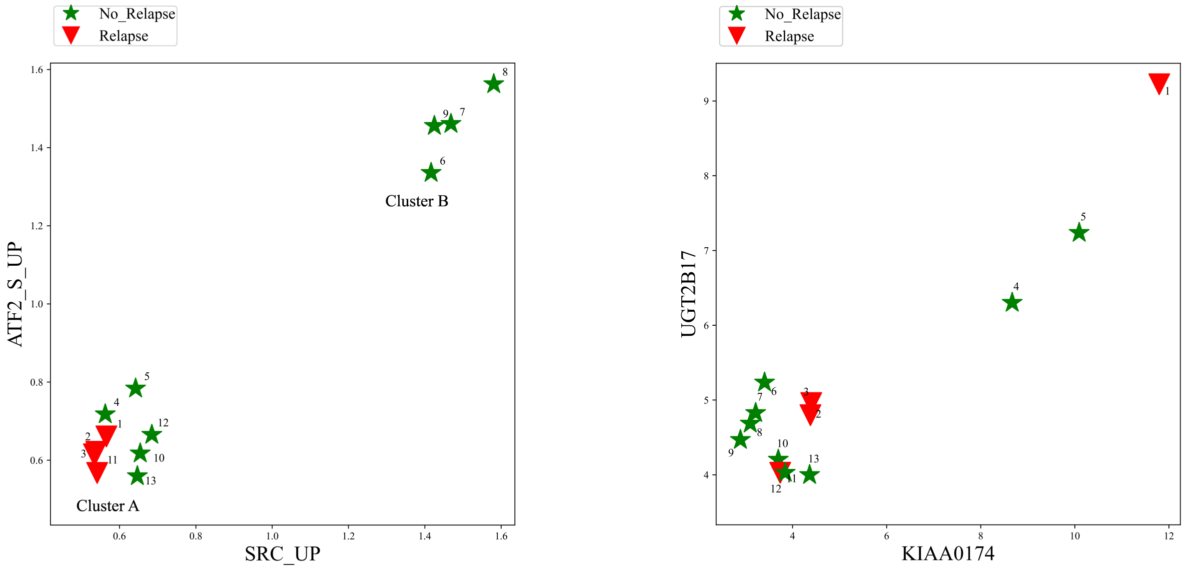
Figure 3.33b shows patient embedding for high-risk leukemia patients with SNet using MRMR in relation to gene ranking. KIAA0174 and UGT2B17 are the top two genes from the MRMR ranking. The figure shows that the patients with and without relapse have a larger spread than in Figure 3.33a.



(a)

(b)

Figure 3.32: Leukaemia patient embedding with PCA (a) using z-absolute anomaly score and (b) SNet.



(a)

(b)

Figure 3.33: Leukaemia patient embedding with MRMR (a) using z-absolute anomaly score and (b) SNet.

3.4.2.8 Comparative analysis between anomaly score and PFSNet

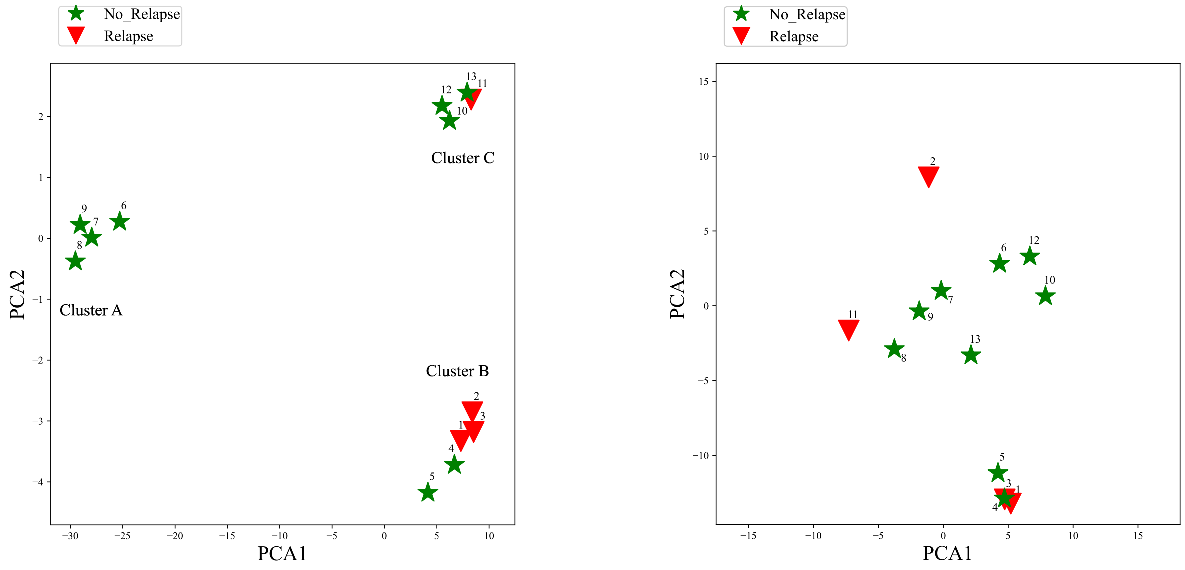
This section summarises and discusses the main findings of patient embedding using anomaly scores and PSFNet gene expression scores with PCA and MRMR.

Figure 3.34a shows patients embedding for high-risk leukaemia patients using z-absolute anomaly score with principal component analysis. The results are described in detail in Figure 3.20a.

Figure 3.34b shows patients embedding for high-risk leukaemia patients with PFSNet using PCA.

Figure 3.35a shows the patient embedding for high-risk leukemia patients using z-absolute anomaly score with MRMR.

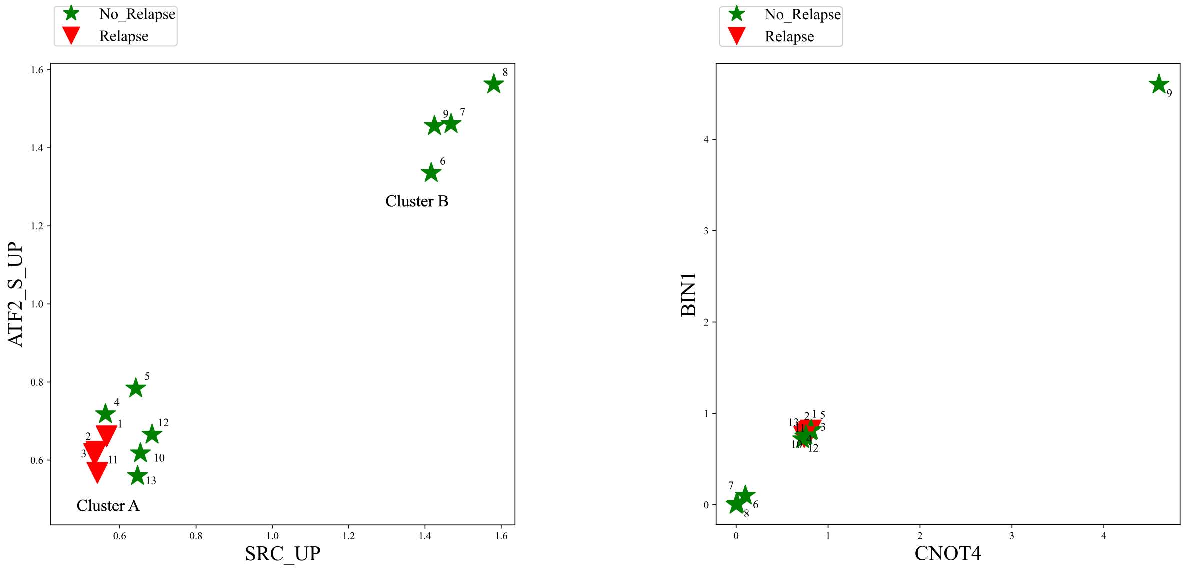
Figure 3.35b shows patient embedding for high-risk leukemia patients with PFSNet using MRMR in relation to gene ranking. CNOT4 and BIN1 are the top two genes from the MRMR ranking. The figure shows that the patients with and without relapse have no separate cluster in contrast to Figure 3.35a.



(a)

(b)

Figure 3.34: Leukaemia patient embedding with PCA (a) using z-absolute anomaly score and (b) PFSNet.



(a)

(b)

Figure 3.35: Leukaemia patient embedding with MRMR (a) using z-absolute anomaly score and (b) PFSNet.

3.4.2.9 Comparative analysis between anomaly score and qPSP

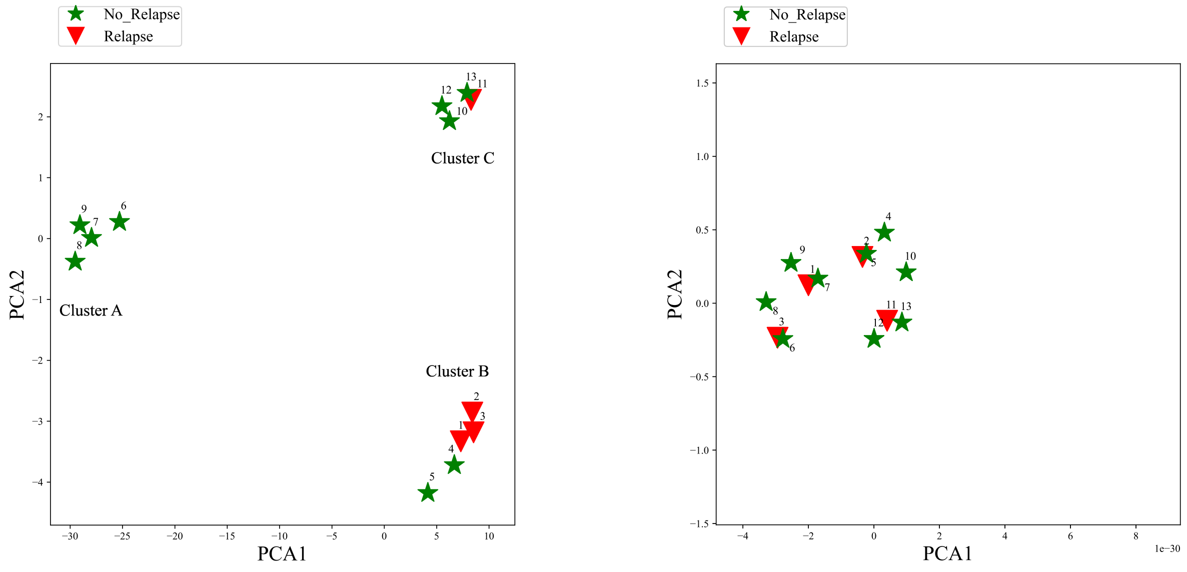
This section compares the descriptions between anomaly scores and qPSP gene expression scores for embedding high-risk leukaemia patients with PCA and MRM

Figure 3.36a patient embedding for high-risk leukaemia patients with z-absolute anomaly score using PCA.

Figure 3.36b shows patient embedding for high-risk leukaemia patients with qPSP using PCA. The figure shows that the patients are more scattered than in Figure 3.36a.

Figure 3.37a patient embedding for high-risk leukemia patients with z-absolute anomaly score using MRMR.

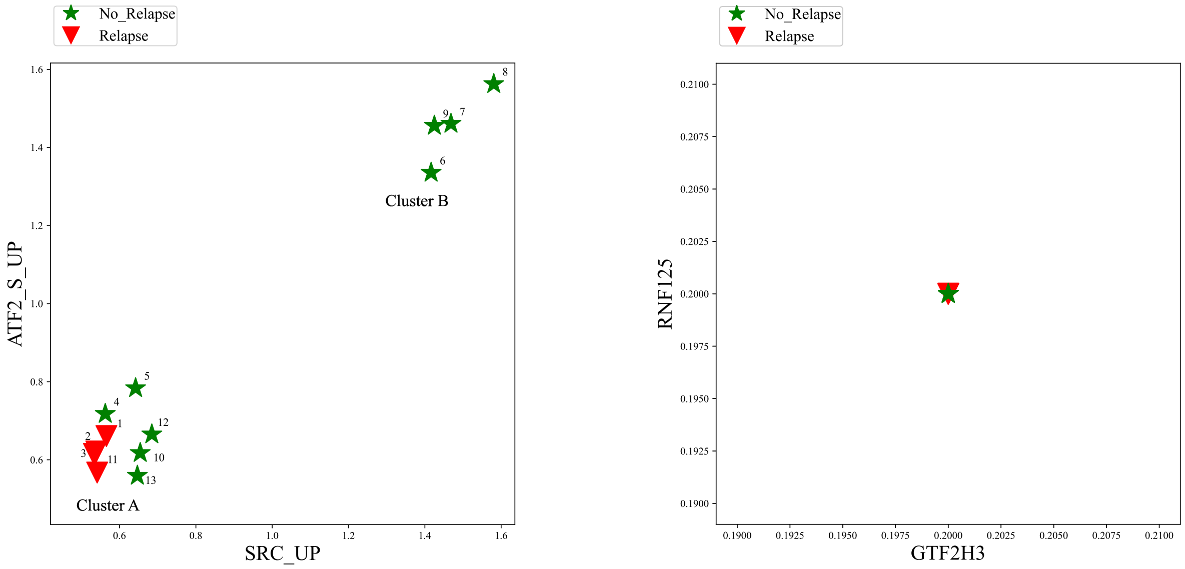
Figure 3.37b patient embedding for high-risk leukemia patients with qPSPt using MRMR in relation to gene ranking. KRTAP1-3 and PSMC6 are the top two genes from the MRMR ranking. The figure shows that the patients with and without relapse have no cluster.



(a)

(b)

Figure 3.36: Leukaemia patient embedding with PCA (a) using z-absolute anomaly score and (b) qPSP.



(a)

(b)

Figure 3.37: Leukaemia patient embedding with MRMR (a) using z-absolute anomaly score and (b) qPSP.

3.4.2.10 Comparative analysis between anomaly score and Eigfusion

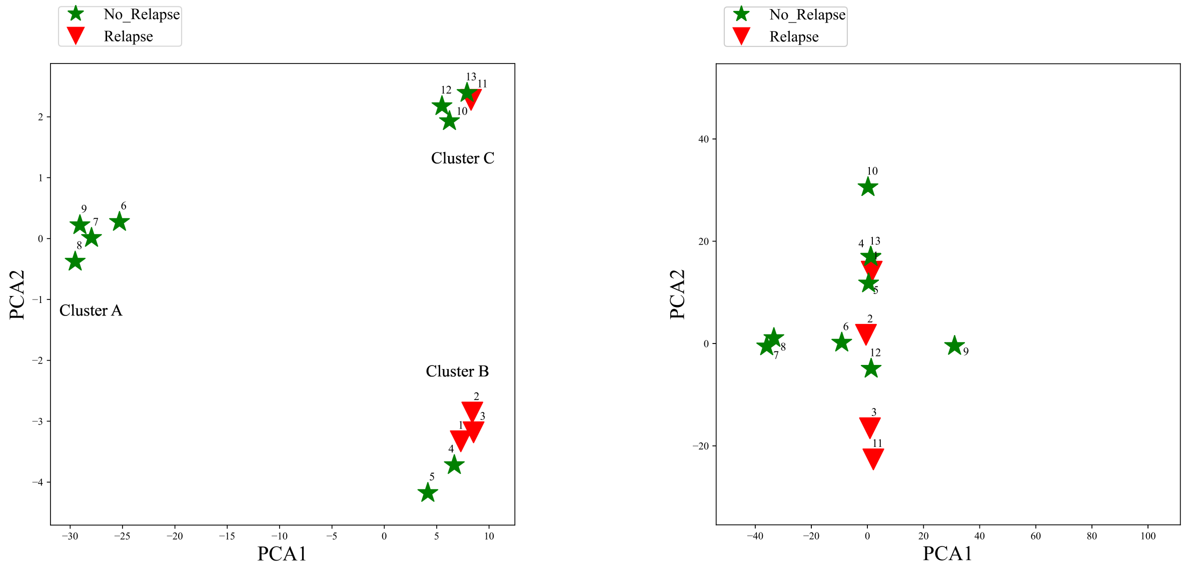
This section compares the descriptions between anomaly scores and Eigfusion gene expression scores for embedding high-risk leukaemia patients with PCA and MRM

Figure 3.38a shows patients embedding for high-risk leukaemia patients with z-absolute anomaly score using PCA.

Figure 3.38b shows patients embedding for high-risk leukaemia patients with Eigfusion using PCA. The figure shows that the patients are more scattered than in Figure 3.38a.

Figure 3.39a shows patient embedding for high-risk leukemia patients with z-absolute anomaly score using MRMR.

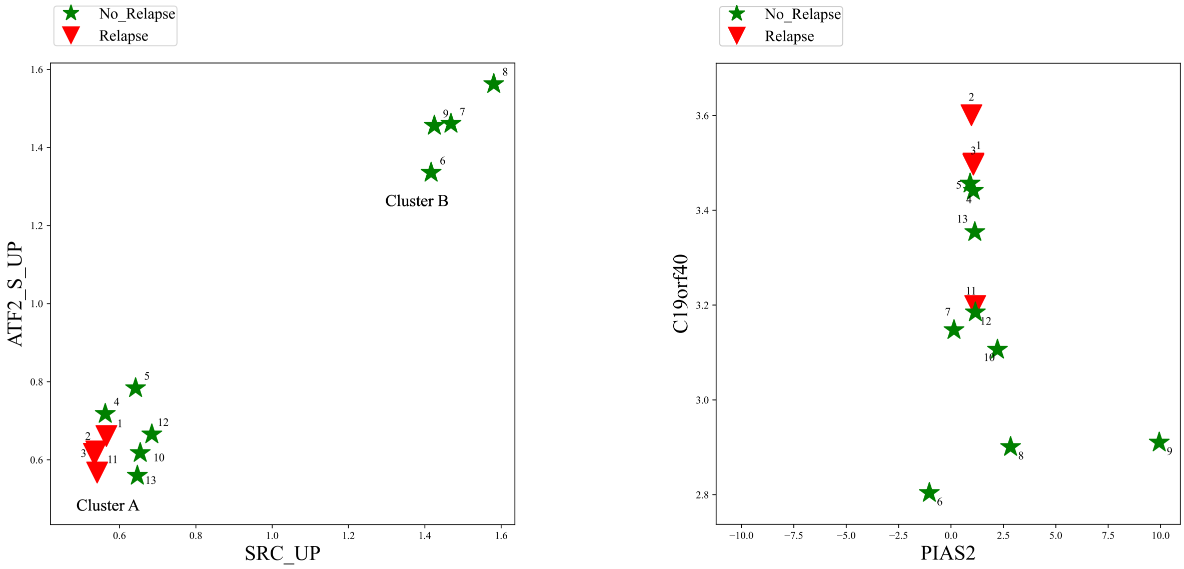
Figure 3.39b shows patient embedding for high-risk leukemia patients with Eigfusion using MRMR in relation to gene ranking. PIAS2 and C19orf40 are the top two genes from the MRMR ranking. The figure shows that the patients with and without relapse have no cluster.



(a)

(b)

Figure 3.38: Leukaemia patient embedding with PCA (a) using z-absolute anomaly score and (b) outlier detection approach for potential rearrangement.



(a) (b)

Figure 3.39: Leukaemia patient embedding with MRMR (a) using z-absolute anomaly score and (b) outlier detection approach for potential rearrangement.

3.4.3 Summary of the comparative outcomes

The results derived from the anomaly score patient embeddings provide more insightful clustering patterns than those obtained using state-of- the-art scores patient embeddings. These results reveal a relationship between the observed clusters and biologically significant changes that may be explained by meaningful variations in gene expression levels. This highlights the importance of the anomaly score approach in capturing the intricate relationships between patients. This will improve our understanding of the underlying biological processes and contribute to the progress of research in this area.

3.4.3.1 Breast cancer patients embedding

Figure 3.42a shows breast cancer patients in a 2D plot where each dimension represents a principal component from the z-absolute anomaly score. The figure shows three clusters including cancer and healthy patients. Two clusters consist entirely of cancer patients where another cluster includes both cancer and healthy patients.

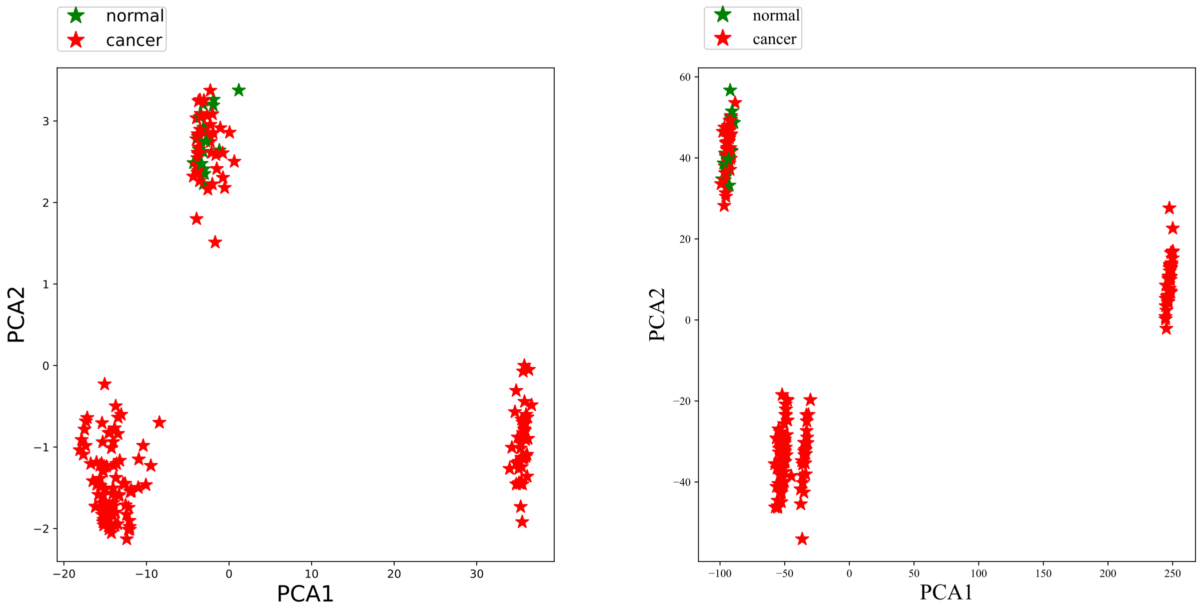
Figure 3.40b shows the embedding of breast cancer patients, where each dimension is a principal component of the raw gene expression values. The plot shows that fewer cancer patients are included in the clusters when raw gene expression values are used, in contrast to the plots based on the anomaly score (Figure 3.42a).

Figure 3.43a shows breast cancer patients in a 2D plot, and each dimension is a gene set raked by MRMR from z-absolute anomaly score. The two highest-ranking gene sets were used to plot patients considering cancer and healthy outcomes. The plot shows that most cancer patients are included in two clusters, whereas others include both cancer and healthy patients. MRMR is used on the breast cancer dataset and identified that HINATA_NFKB_IMMU_ INF ddependent signaling pathways relate to transcriptome factors NF-kappa B showing selective tissue effects. BCAT_GDS748_ DN reflects beta-catenin, which has a major impact on the canonical Wnt signaling pathway [433, 434].

Figure 3.40b shows breast cancer patients in a 2D plot, and each dimension is a gene raked by MRMR from raw gene expression values. The two highest-ranking genes were used to plot patients considering cancer and healthy outcomes. The plot shows that fewer cancer patients are included in two clusters, whereas others include both cancer and healthy patients.

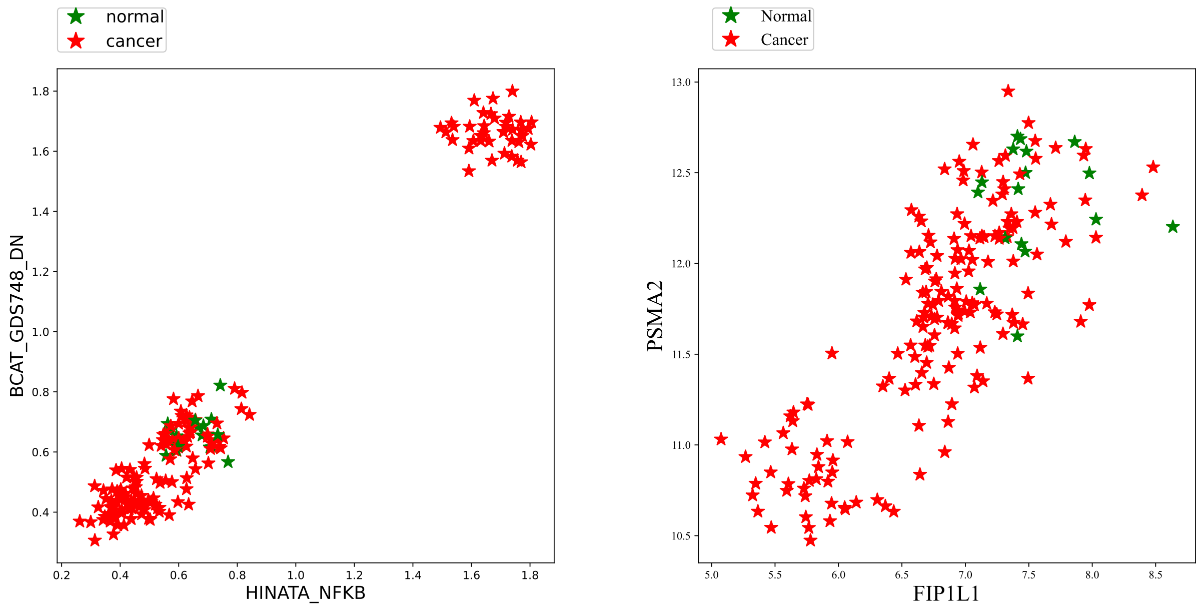
Figure 3.42b shows embedding of breast cancer patients, where each dimension is a principal component of the gene fuzzy score (GFS). The plot shows that both cancer and healthy patients are scattered when the GFS is used, in contrast to the plots based on the anomaly score (Figure 3.42a).

Figure 3.43b shows breast cancer patients in a 2D plot, and each dimension is a gene raked by MRMR from GFS. The two highest-ranking genes were used to plot patients considering cancer and healthy outcomes. The plot shows that both cancer and healthy patients are not properly clustered when the GFS values are used, in contrast to the plots based on the anomaly score (Figure 3.43a).



(a) (b)

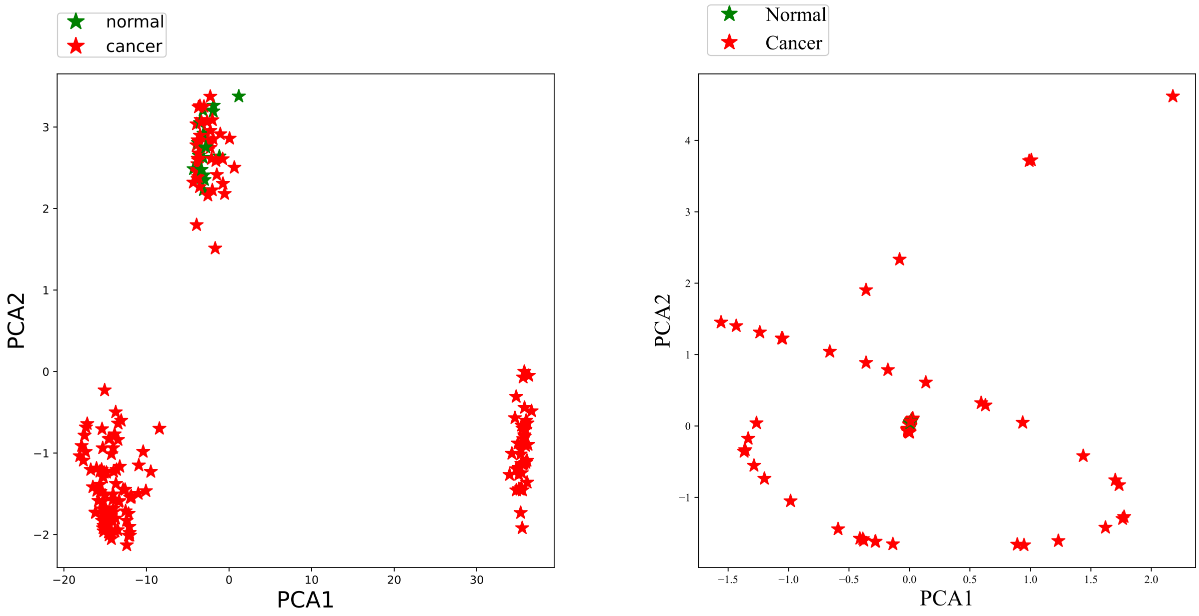
Figure 3.40: Breast cancer patient embedding with PCA (a) for cancer and healthy patients using z-absolute anomaly score and (b) using raw gene expression values.



(a) (b)

Figure 3.41: Breast cancer patient embedding with MRMR (a) for cancer and healthy patients using z-absolute anomaly score and (b) raw gene expression values.

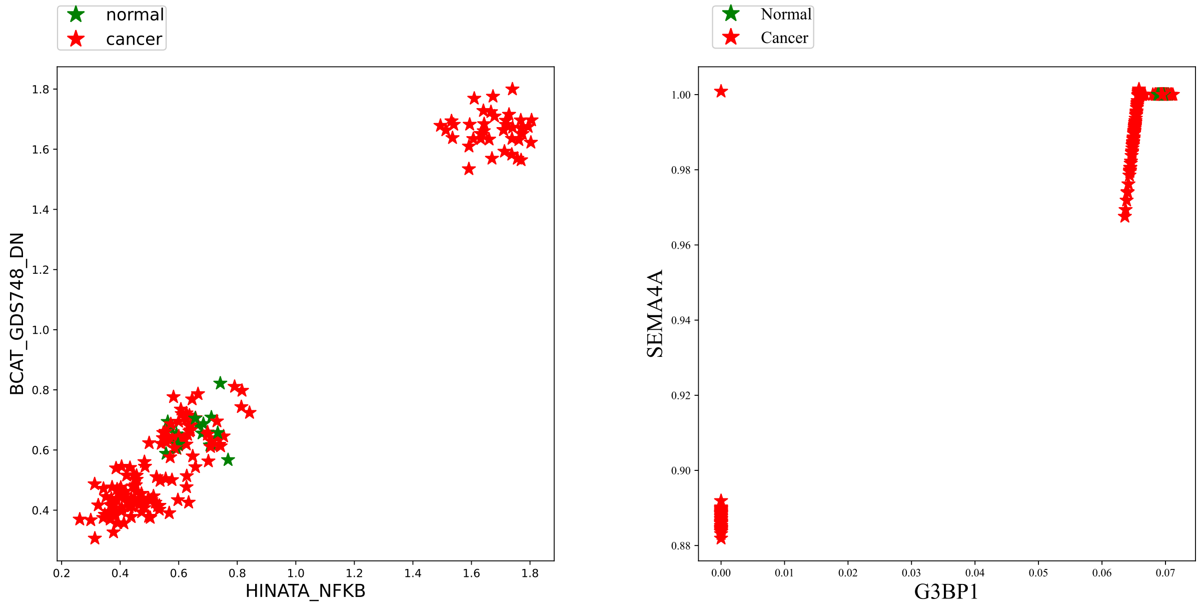
Breast cancer patient embedding for FRaC, CSAX, SNet, and PFSNet scores are shown in figures 3.44a, 3.44b, 3.44c, and 3.44d, respectively. The plots based on the gene expression scores for these approaches show that cancer and healthy patients are scattered in contrast to



(a)

(b)

Figure 3.42: Breast cancer patient embedding with PCA (a) for cancer and healthy patients using z-absolute anomaly score and (b) using GFS.

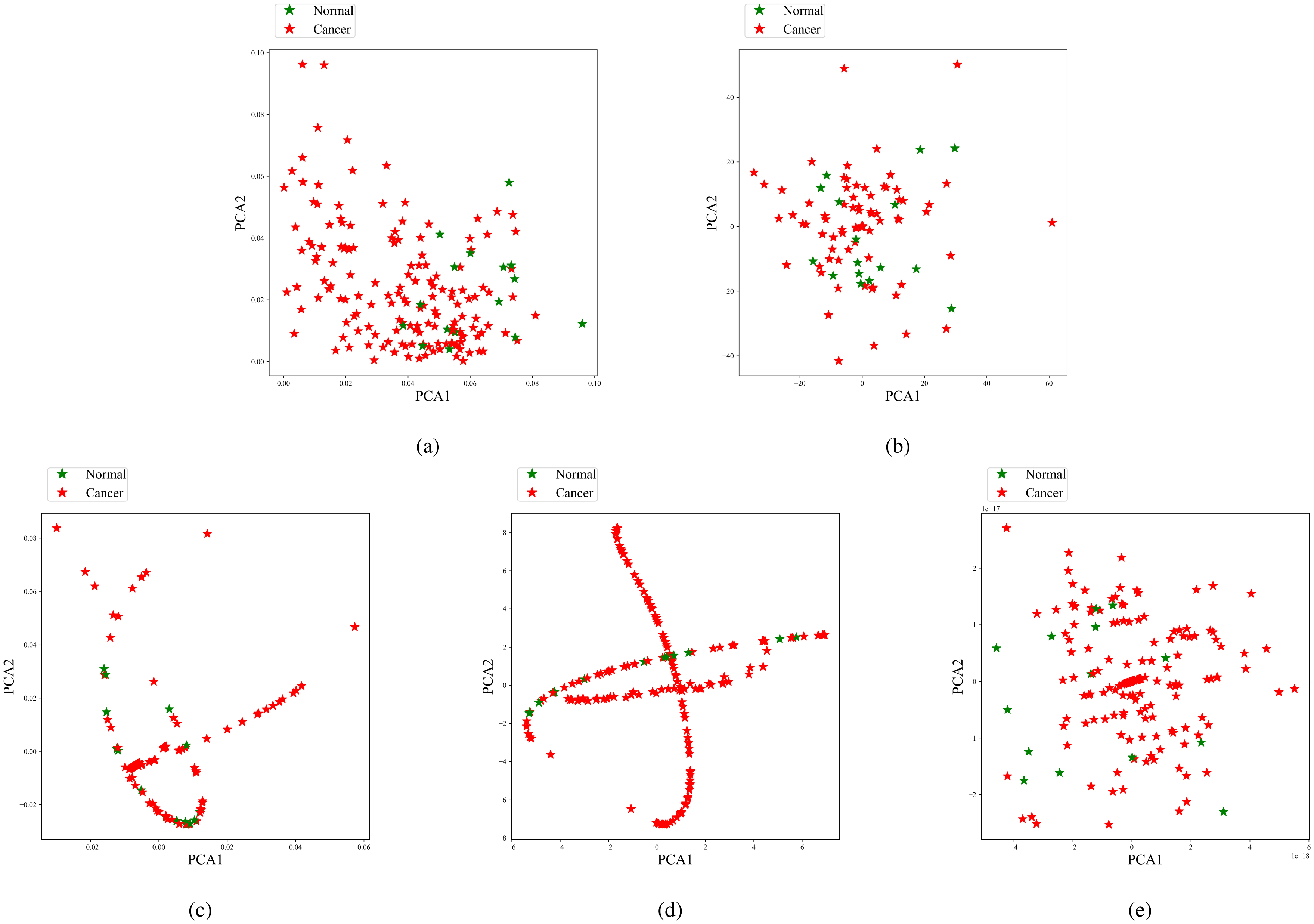


(a)

(b)

Figure 3.43: Breast cancer patient embedding with MRMR (a) for cancer and healthy patients using z-absolute anomaly score and (b) gene fuzzy scores.

the plots based on the anomaly score (Figure 3.42a). Similarly, Figures 3.44e, 3.45a, and 3.45b show the embedding of breast cancer patients for qPFS, outlier detection and EIgfusion. When the adjusted gene expression data from these methods are applied, the distributions of cancer and healthy patients were more scattered.



(a)

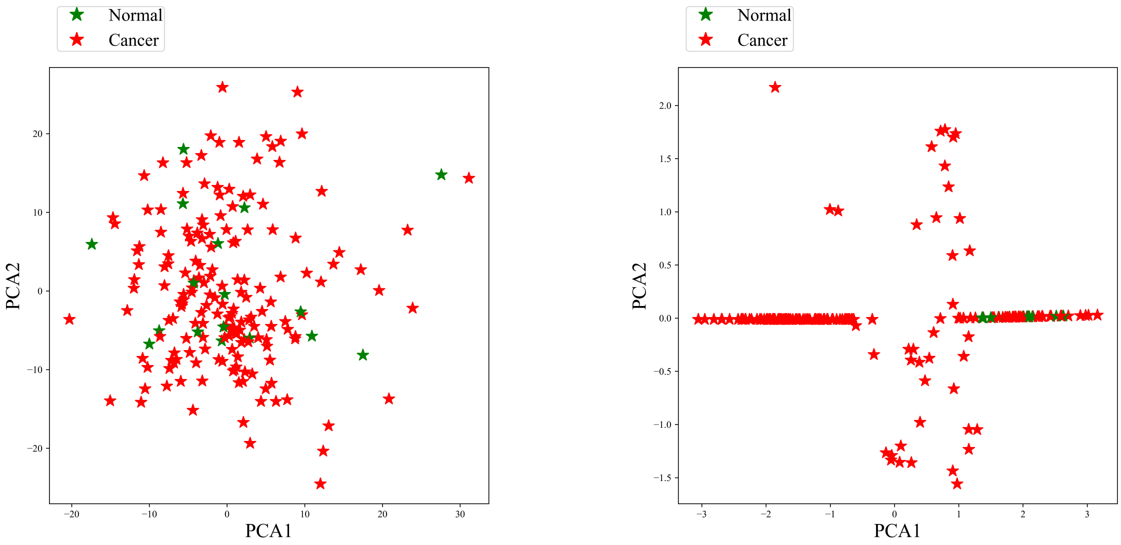
(b)

(c)

(d)

(e)

Figure 3.44: Breast cancer patient embedding for (a) FRaC, (b) CSAX, (c) SNet, (d) PFSNet, and (e) qPFS scores using PCA.



(a)

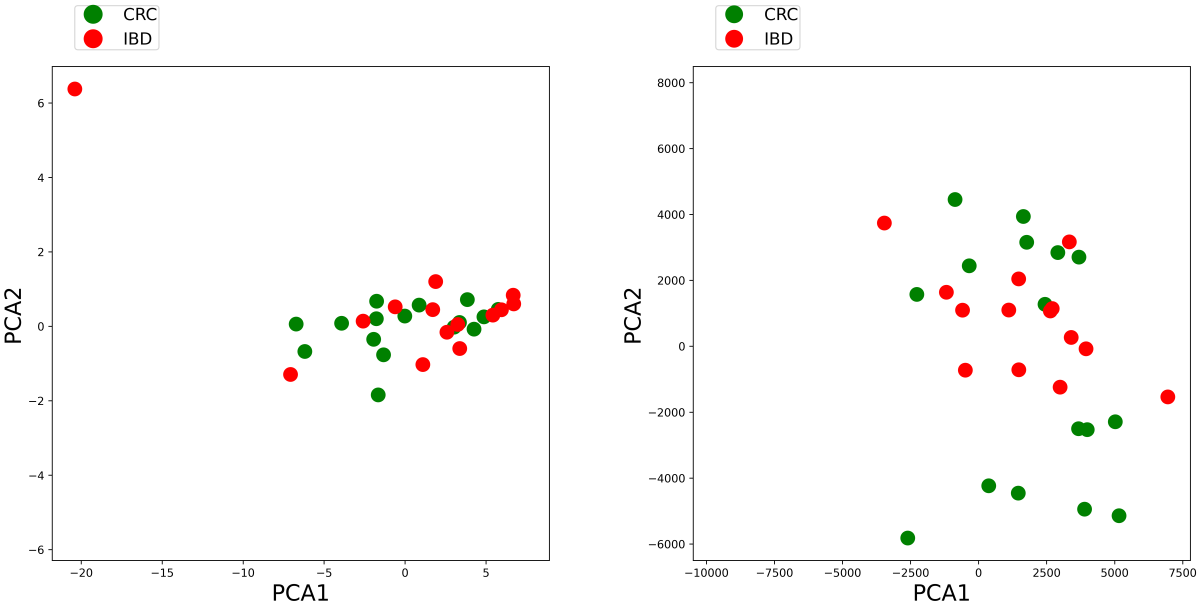
(b)

Figure 3.45: Breast cancer patient embedding for (a) outlier detection and (b) protein rearrangement using PCA.

3.4.3.2 Colon cancer patients embedding

Figure 3.48a shows colon cancer patients (CRC and IBD) in a 2D plot where each dimension represents a principal component from the z-absolute anomaly score. The figure shows that both IBD and CRC patients form relatively good clusters for anomaly scores.

Figure 3.46b shows colon cancer patients (CRC and IBD) in a 2D plot where each dimension represents a principal component from the raw gene expression values. The figure shows that both IBD and CRC patients form relatively poor clusters for raw gene expression values.



(a)

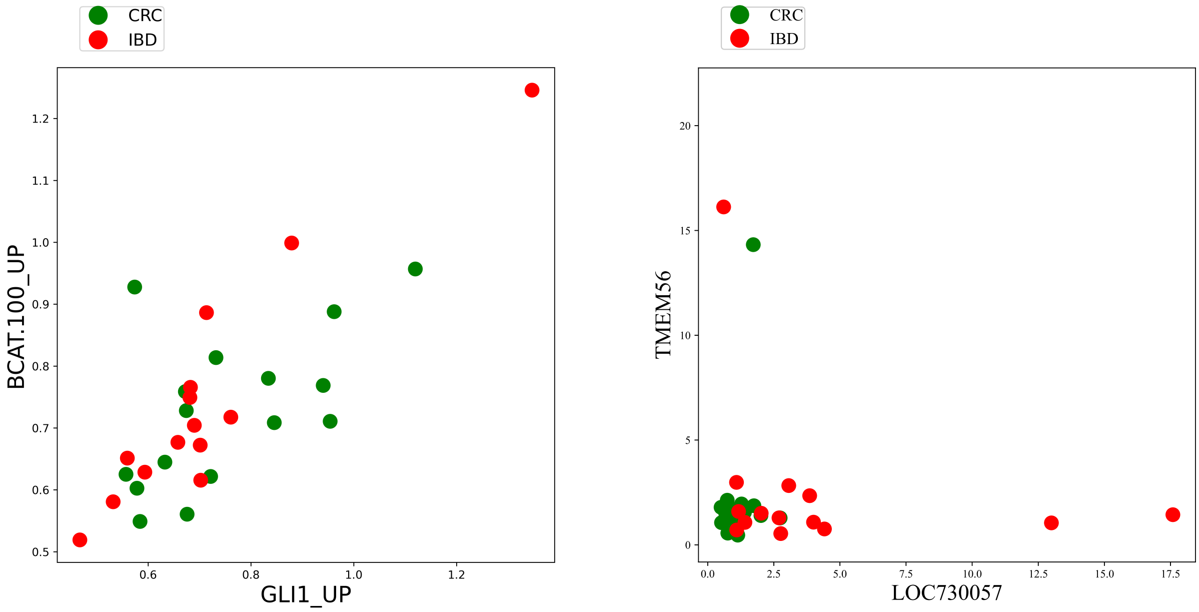
(b)

Figure 3.46: Colon cancer patient embedding with PCA (a) for IBD and CRC patients using z-absolute anomaly score and (b) raw gene expressions values.

Figure 3.49a shows colon cancer patients (CRC and IBD) in a 2D plot where each dimension represents a gene set ranked by MRRM using the z-absolute anomaly score. The figure shows that both IBD and CRC patients form relatively good clusters for anomaly scores.

Figure 3.47b shows colon cancer patients (CRC and IBD) in a 2D plot where each dimension represents a gene set ranked by MRRM using the raw gene expression values. The figure shows that both IBD and CRC patients form relatively poor clusters for raw gene expression values.

Figure 3.48b shows the embedding of colon cancer patients, where each dimension is a principal component of the gene fuzzy score (GFS) values. The plot shows that both CRC and IBD patients are scattered when raw gene expression values are used, in contrast to the plots



(a)

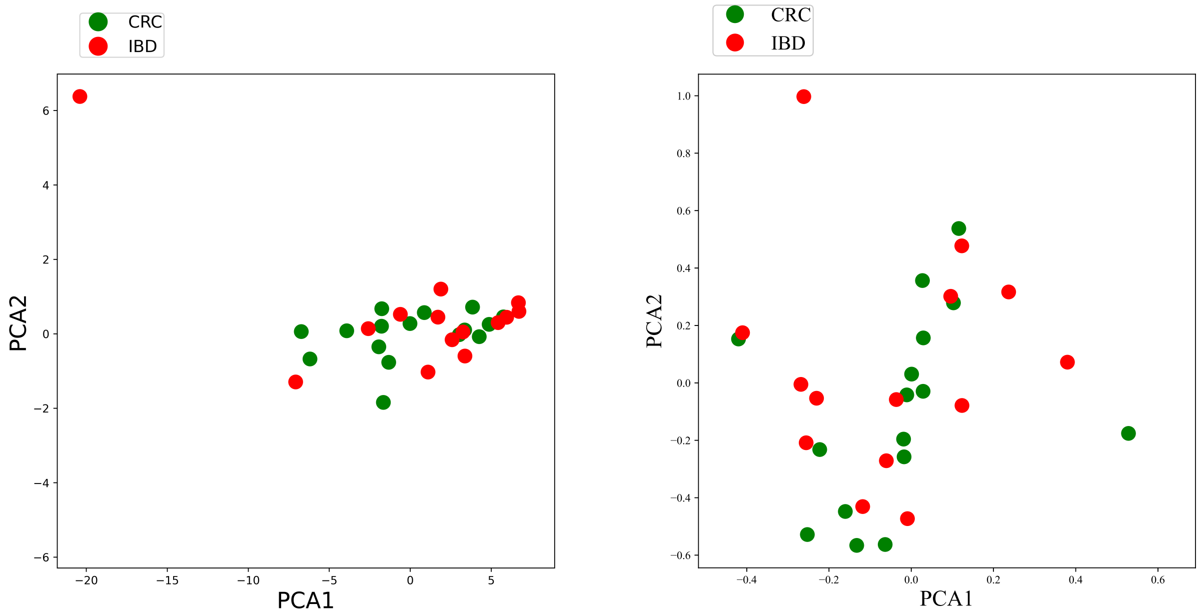
(b)

Figure 3.47: Colon cancer patient embedding with MRMR (a) for IBD and CRC patients using z-absolute anomaly score and (b) raw gene expressions values.

based on the anomaly score (Figure 3.48a).

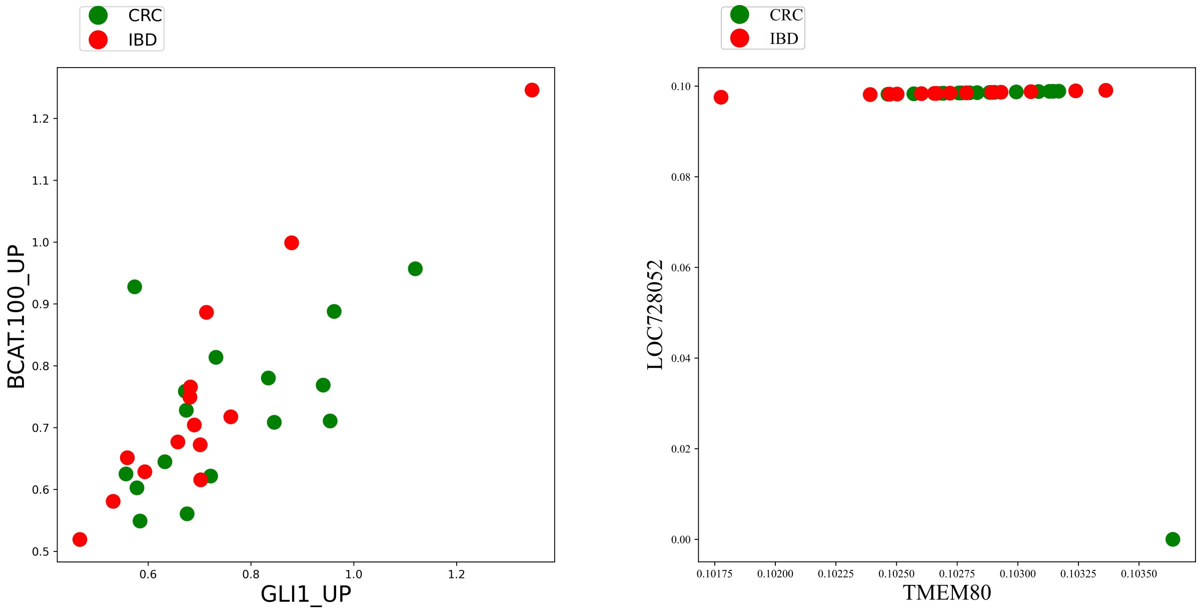
Figure 3.49b shows colon cancer patients in a 2D plot, and each dimension is a gene raked by MRMR from GFS values. The two highest-ranking genes are used to plot patients considering IBD and CRC outcomes. The plot shows that both IBD and CRC patients are scattered when the gene fuzzy scores are used, in contrast to the plots based on the anomaly score (Figure 3.49a).

Colon cancer patient embedding for FRaC, CSAX, SNet, PFSNet, qPFS, outlier detection and EIgfusion scores are shown the figures 3.44a, 3.44b, 3.44c, 3.44d, 3.44e, 3.45a, and 3.45b respectively. The plots based on the gene expression scores for these approaches show that both the CRC and adenoma patients are scattered in contrast to the plots based on the anomaly score(Figure 3.48a).



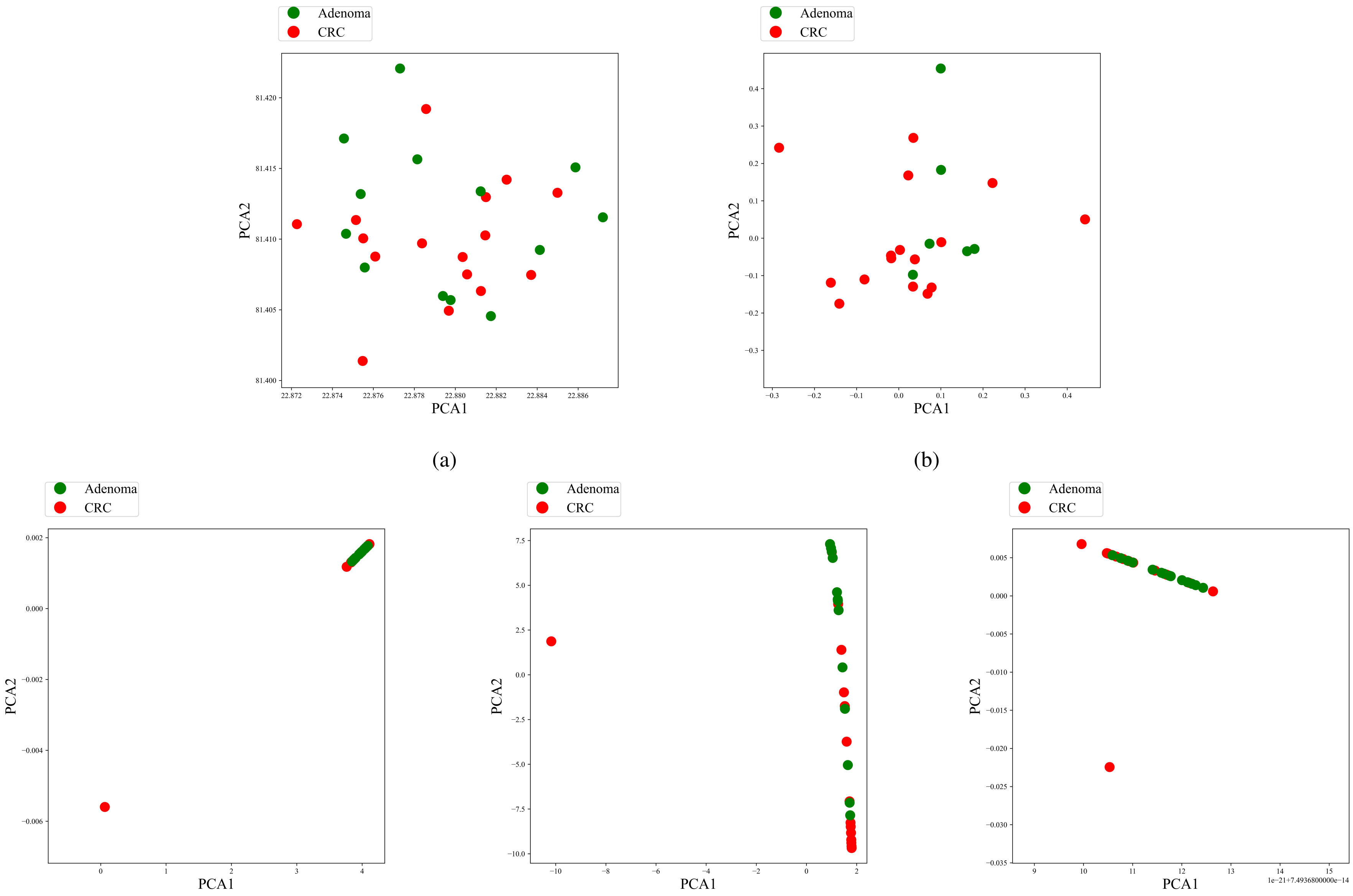
(a) (b)

Figure 3.48: Colon cancer patient embedding with PCA (a) for IBD and CRC patients using z-absolute anomaly score and (b) using GFS.



(a) (b)

Figure 3.49: Colon cancer patient embedding with MRMR (a) for IBD and CRC patients using z-absolute anomaly score and (b) using GFS.



(a)

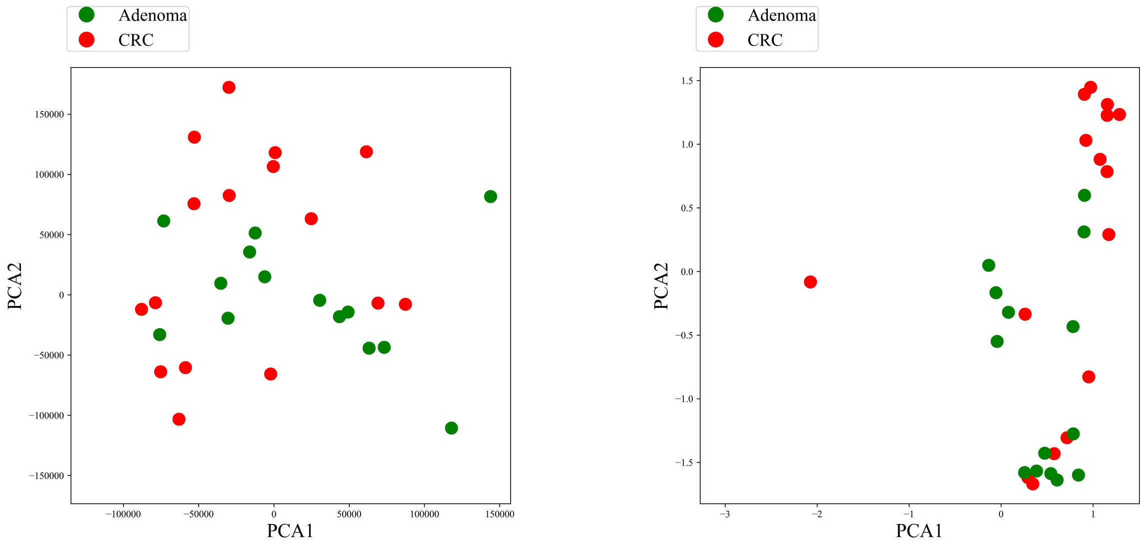
(b)

(c)

(d)

(e)

Figure 3.50: Colon cancer patient embedding for (a) FRaC, (b) CSAX, (c) SNet, (d) PFSNet, and (e) qPFS scores using PCA.



(a)

(b)

Figure 3.51: Colon cancer patient embedding for (a) outlier detection and (b) protein rearrangement using PCA.

3.4.3.3 Adrenal cancer patients embedding

Figure 3.54a shows adrenal cancer patients in a 2D plot where each dimension represents a principal component from the z-absolute anomaly score. The figure shows two clusters including adrenal carcinoma and adrenal adenoma patients. One cluster consists entirely of adenoma patients whereas the other cluster includes carcinoma patients.

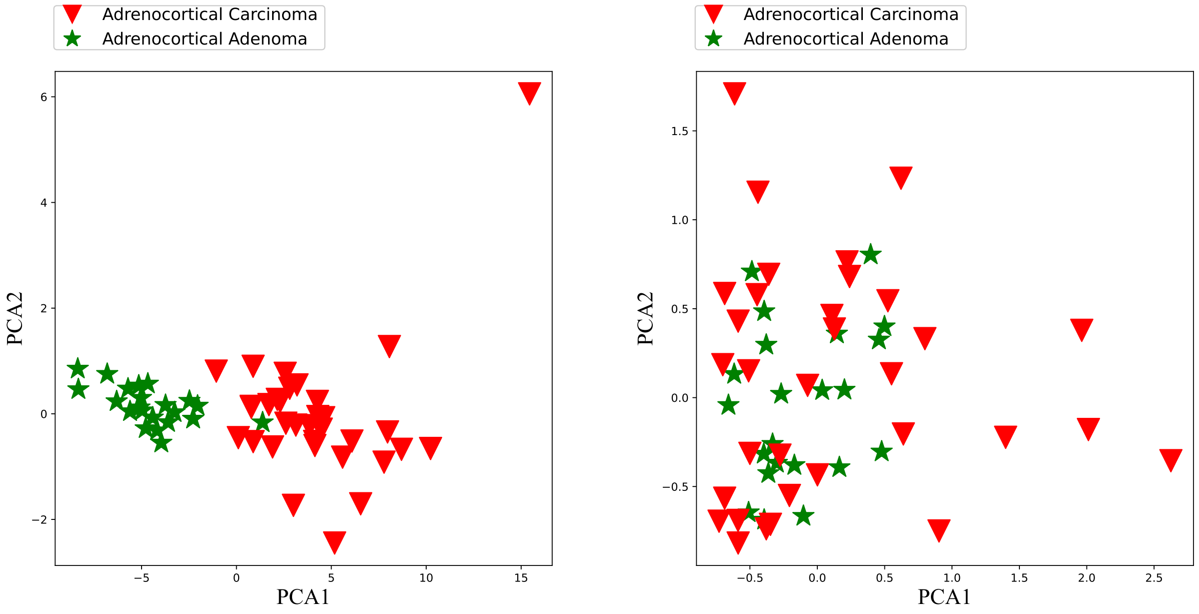
Figure 3.52b shows the embedding of adrenal cancer patients, where each dimension is a principal component of the raw gene expression values. The plot shows that both adenoma and carcinoma patients are scattered when the raw gene expression values are used, in contrast to the plots based on the anomaly score (Figure 3.54a).

Figure 3.55a shows adrenal cancer patients in a 2D plot, and each dimension is a gene set raked by MRMR from z-absolute anomaly score. The two highest-ranking gene sets are used to plot patients considering adenoma and carcinoma outcomes. The plot shows that adenoma and carcinoma patients are separated into two areas. This thesis hypothesizes that the clusters relate to biologically relevant distinctions that can be derived from gene expression values. MRMR is used on the adrenal cancer dataset and identified that CAHOY_ASTROCYTIC dependent signaling pathways relate to glial cells of the nervous system [435]. On the other hand, human malignancies with KRAS mutations tend to be aggressive and resistant to traditional therapies [436].

Figure 3.53b shows adrenal cancer patients in a 2D plot, and each dimension is a gene raked by MRMR from raw gene expression values. The two highest-ranking genes were used to plot patients considering cancer and healthy outcomes. The plot shows that both adenoma and carcinoma patients are scattered when the raw gene expression values are used, in contrast to the plots based on the anomaly score (Figure 3.55a).

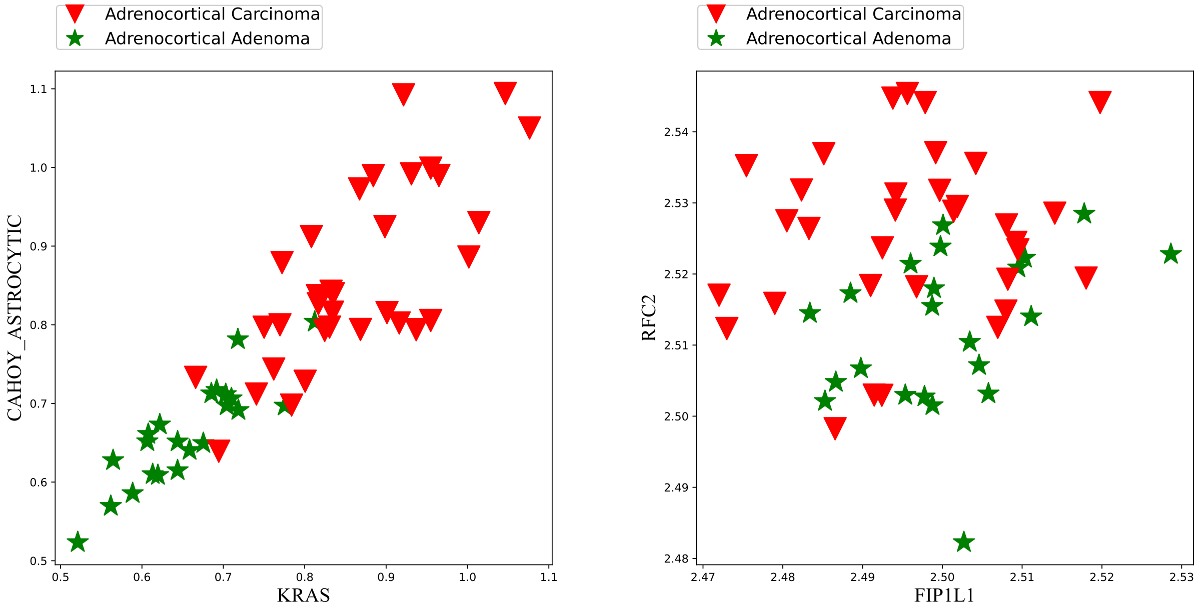
Figure 3.54b shows the embedding of adrenal cancer patients, where each dimension is a principal component of the GFS values. The plot shows that both adenoma and carcinoma patients are scattered when the GFS is used, in contrast to the plots based on the anomaly score (Figure 3.54a).

Figure 3.55b shows adrenal cancer patients in a 2D plot, and each dimension is a gene raked by MRMR from GFS values. The two highest-ranking genes are used to plot patients considering cancer and healthy outcomes. The plot shows that both adenoma and carcinoma



(a) (b)

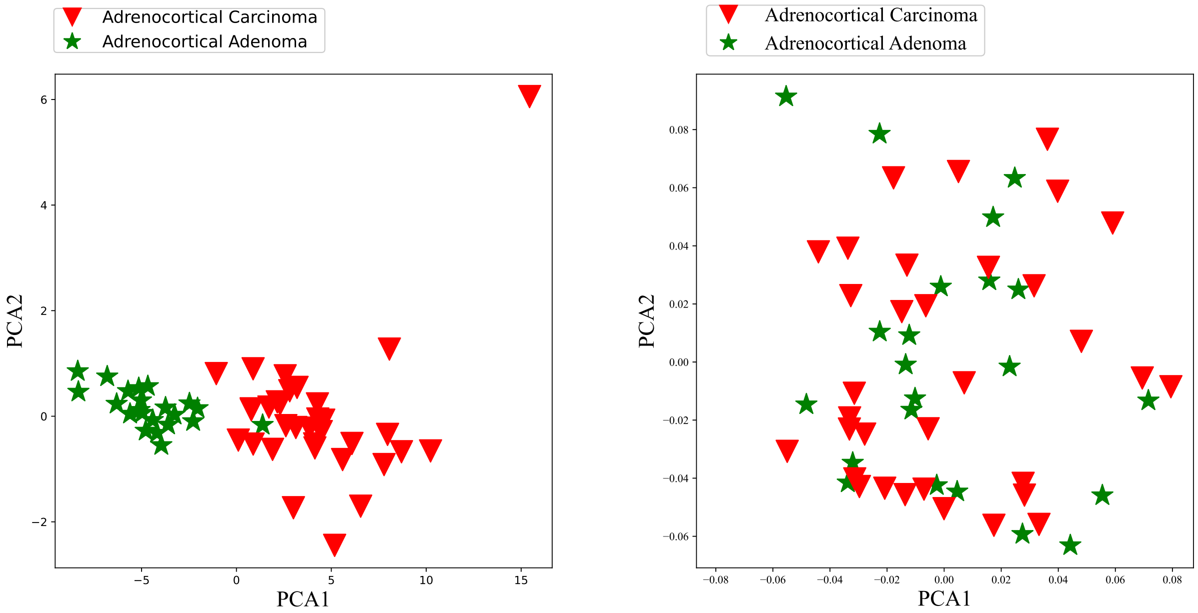
Figure 3.52: PCA Adrenal cancer patient embedding (a) for adrenal carcinoma and adrenal adenoma patients using z-absolute anomaly score (b) using raw gene expression values.



(a) (b)

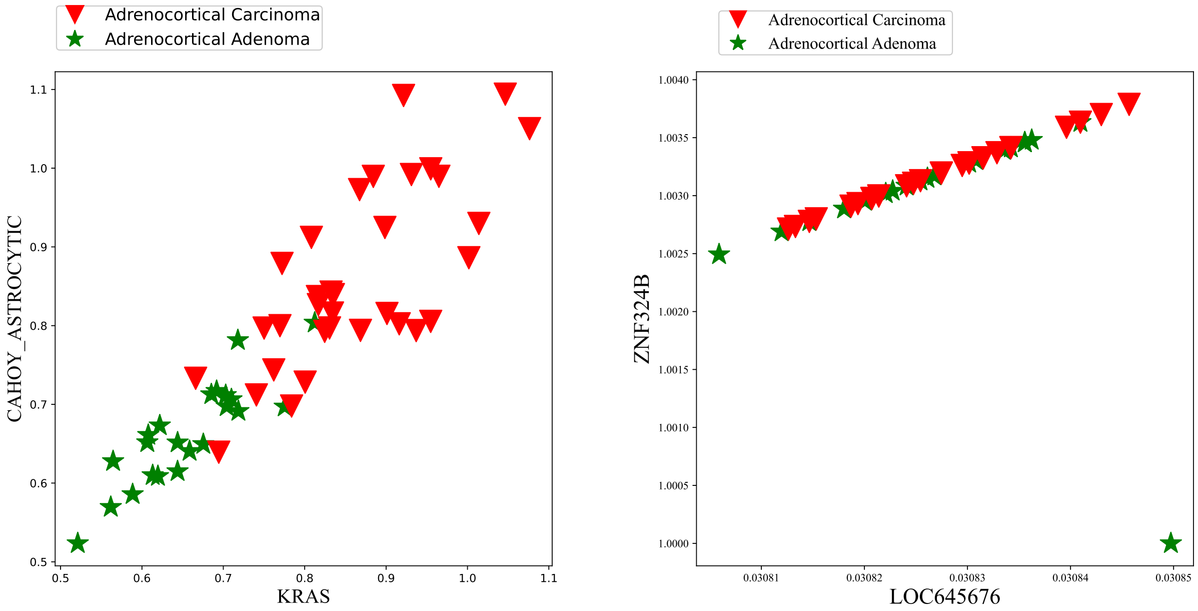
Figure 3.53: MRMR Adrenal cancer patient embedding (a) for adrenal carcinoma and adrenal adenoma patients using z-absolute anomaly score (b) using raw gene expression values.

patients are not clustered when the GFS values are used, in contrast to the plots based on the anomaly score (Figure 3.55a).



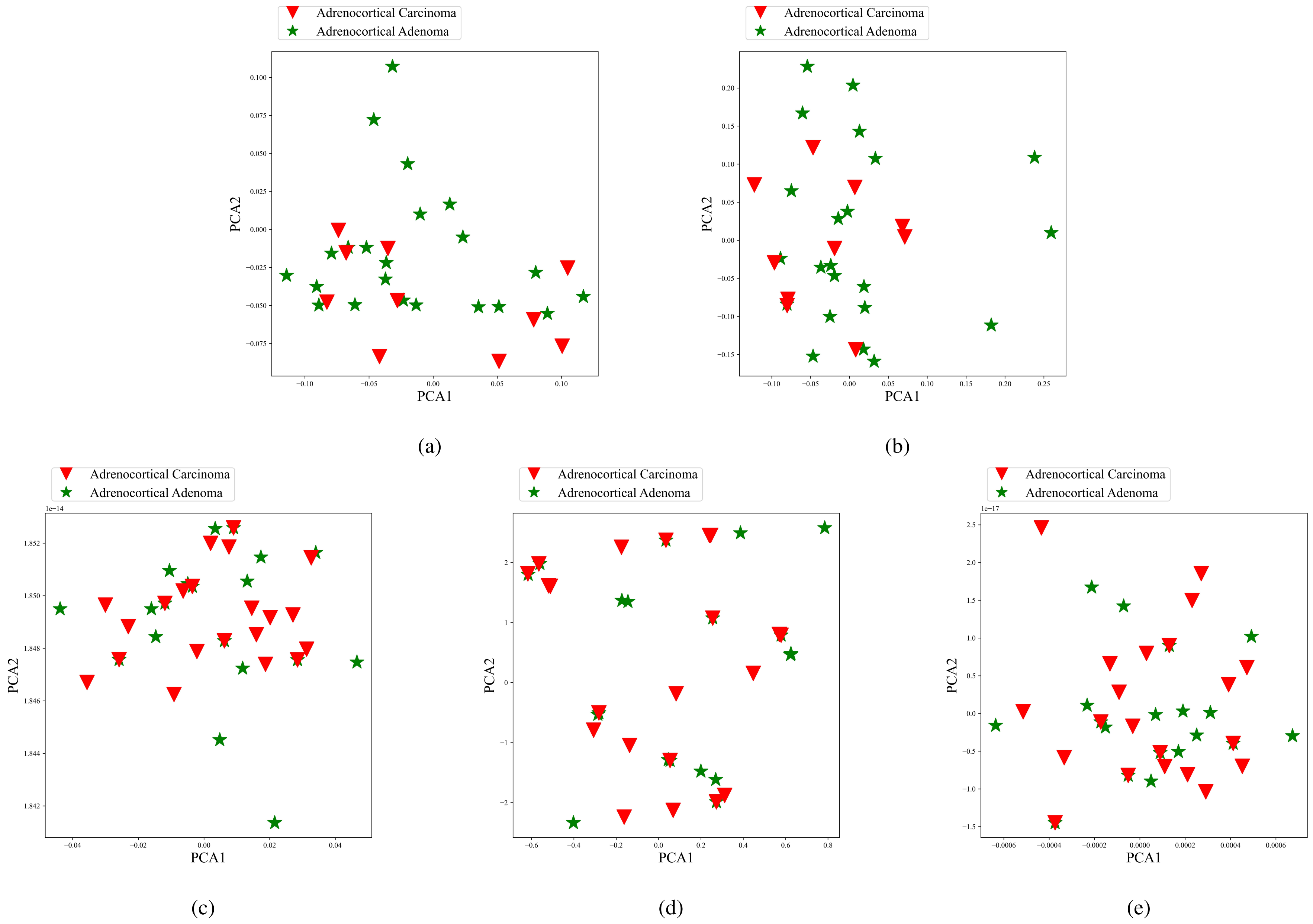
(a) (b)

Figure 3.54: PCA Adrenal cancer patient embedding (a) for adrenal carcinoma and adrenal adenoma patients using z-absolute anomaly score (b) GFS.



(a) (b)

Figure 3.55: MRMR Adrenal cancer patient embedding (a) for adrenal carcinoma and adrenal adenoma patients using z-absolute anomaly score (b) using GFS.



(a)

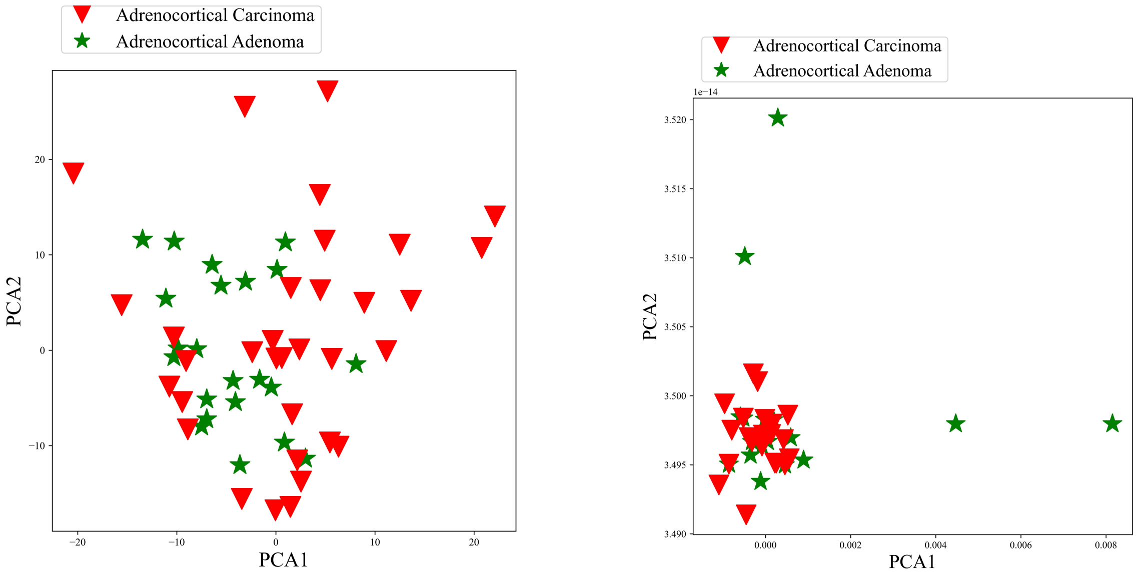
(b)

(c)

(d)

(e)

Figure 3.56: Adrenal cancer patient embedding for (a) FRaC, (b) CSAX, (c) SNet, (d) PFSNet, and (e) qPFS scores using PCA.



(a)

(b)

Figure 3.57: Adrenal cancer patient embedding for (a) outlier detection and (b) protein rearrangement using PCA.

Adrenal cancer patient embedding for FRaC, CSAX, SNet, PFSNet, qPFS, outlier detection and EIgfusion scores are shown in figures3.56a, 3.56b, 3.56c, 3.56e, 3.57a, and 3.57b respectively The plots based on the gene expression scores for these approaches show that both adenoma and carcinoma patients are scattered in contrast to the plots based on the anomaly score (Figure 3.54a).

3.4.3.4 Discussion for patients embedding

The thesis hypothesized that the patient clusters could be explained by differences in the levels of gene expression that are biologically significant. The outcomes by gene set anomaly score support that this thesis made proper alignment with its assumptions. The results that were presented above on patient embedding tested a hypothesis. This hypothesis posited that careful integration of gene expression values into gene set anomaly scores would make it possible to get useful insights from an investigation of gene expression data. Comparative outcomes between the proposed method and raw gene expression values and state-of-the-art methodologies show that gene set anomaly scores performed distinct clustering than state-of-the-art methodologies.

3.4.4 Expression anomaly distribution analysis

This thesis also used a histogram to evaluate anomaly distribution on various patient groups, including high-risk relapse patients, high-risk non-relapse patients, medium-risk relapse patients, medium-risk non-relapse patients, standard-risk relapse patients, and standard-risk non-relapse patients. Figure 3.58 shows data flow for anomaly distribution. The figure shows that anomaly scores were applied to a histogram, resulting in different distributions for different patient groups.

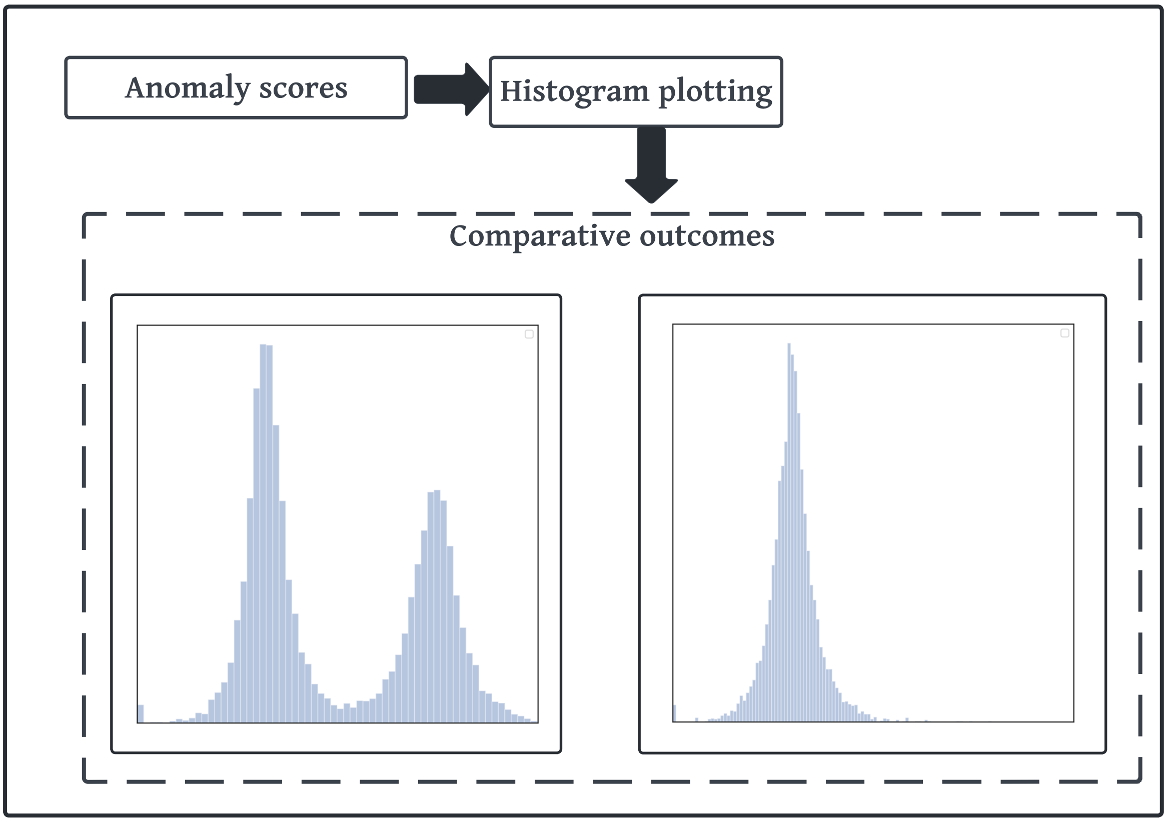


Figure 3.58: A data flow diagram for patient groups anomaly distribution.

The distribution of anomaly scores conditioned on three classes of profiles, (1) healthy, (2) relapse, and (3) non-relapse is investigated. The latter two classes relate to patients with an ALL diagnosis and their response to treatment.

The aim is to test whether anomaly scores potentially provide discriminating information using histogram plots.

Anomaly distribution analysis aims to determine if anomaly scores offer valuable insights through histogram plots by examining the distribution of scores based on three profile classes:

(1) healthy, (2) relapse, and (3) non-relapse. This stark difference between patients who had different clinical outcomes highlights possible novel insights about childhood leukaemia. The presence of a discrete gene sets and not a general transition for all gene sets to develop imbalanced gene expression activity indicated mechanisms are active in patients that may confer a survival signal as they associate with any patient who does not relapse during therapy regardless of risk stratification. This may reflect the presence of a unique clonal sub-population in these patients. Alternatively, these gene sets represent specific molecular mechanisms, such as apoptosis pathways that promote leukaemic cell sensitivity, that remain active at diagnosis and subsequently facilitate the effect of chemotherapy. This thesis considers that the examination of anomaly scores had drawn out different relationships in the data not previously identified, the prognostic significance of which warrants further exploration.

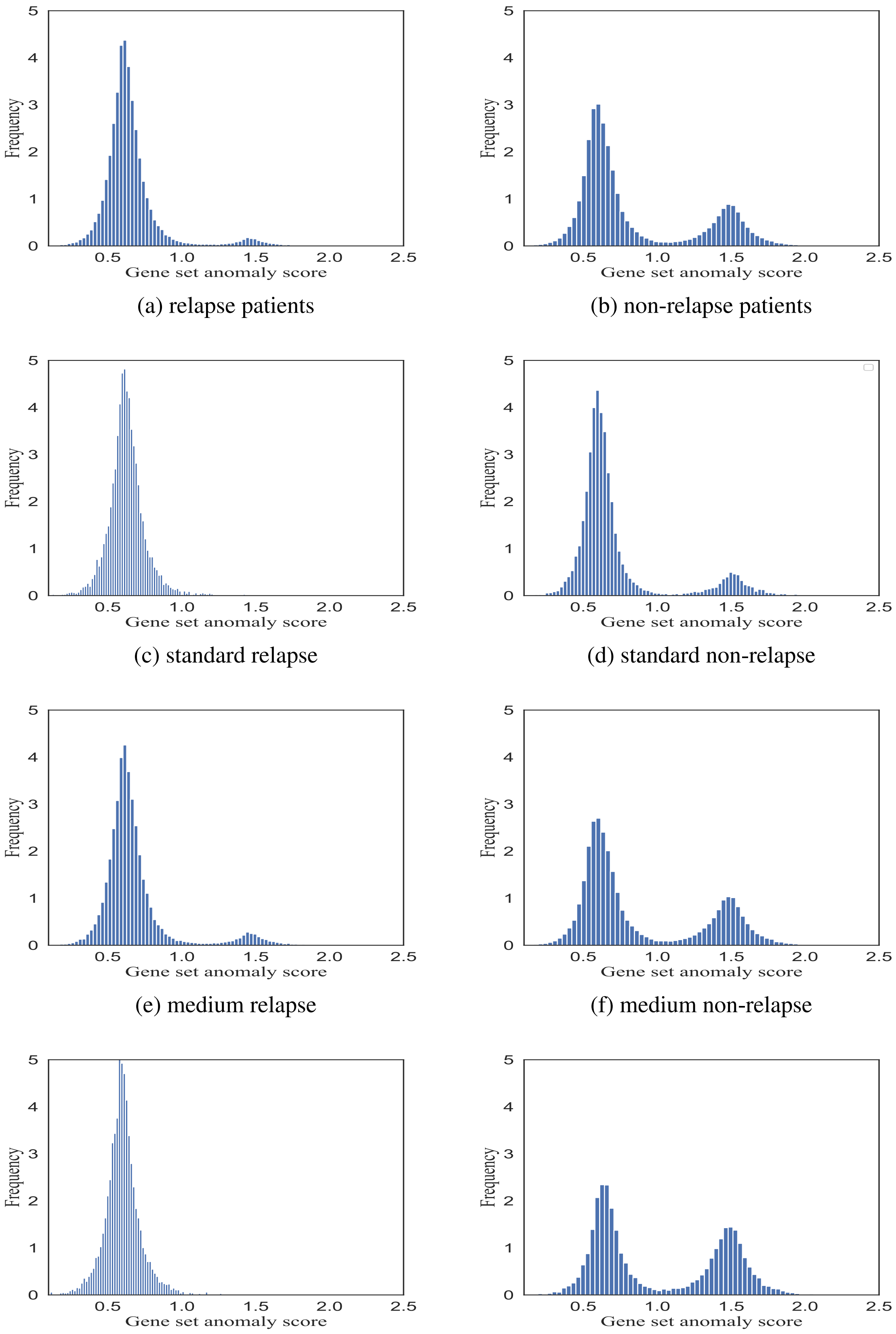
3.4.4.1 Results for anomaly distributions

Experimental setup:

1. Technology: Histogram.
2. Python packages: NumPy, pandas, matplotlib , and seaborn.
3. Input: Anomaly scores for leukaemia data sets.

This thesis seeks to understand how anomaly scores may change, conditioned on whether a cancer patient relapsed or not. Hence, a histogram of anomaly scores was plotted under the two conditions (Figure 3.59a and Figure 3.59b).

In general, it can be seen that the distribution of anomaly scores for relapsed patients is largely unimodal (Figure 3.59a), except for a much smaller secondary mode. Conditioning the distri- bution on the patients' risk stratification, it can be seen that the secondary mode comes from medium risk patients (Figure 3.59e). In contrast, it can be seen that for non-relapsed patients, the anomaly scores are distinctly bimodal.



(a) relapse patients

(b) non-relapse patients

(c) standard relapse

(d) standard non-relapse

(e) medium relapse

(f) medium non-relapse

(g) high relapse

(h) high non-relapse

Figure 3.59: Anomaly distribution on (a) relapse patients, (b) non-relapse, (c) standard relapse, (d) standard non-relapse, (e) medium relapse, (f) medium non-relapse, (g) high relapse, and (h) high non-relapse.

3.4.4.2 Discussion on anomaly distribution

This stark difference between patients who had different clinical outcomes suggests possible novel insights about childhood leukaemia. This thesis considers that the examination of anomaly scores indicates relationships in the data not previously identified, the prognostic significance of which warrants further exploration.

Multiple patient groups are evaluated in order to understand the variance of the anomaly distribution between them. These patient groups are relapse, non-relapse, high-risk relapse, high-risk non-relapse, medium-risk relapse, medium-risk non-relapse, standard-risk relapse, and standard-risk non-relapse. It can be seen from the distributions that non-relapse patients have two peaks in contrast to patients with relapse patients. This thesis hypothesises that more gene sets try to survive in non-relapse patients than in relapse patients.

3.4.5 k-means clustering

k-means [437] clustering algorithm is used to validate patient embedding, which is plotting patients into a two-dimensional space with respect to gene sets or principal components. The k-means algorithm works as follows.

1. Specify the number of clusters.
2. Initialise the centers from the patient's anomaly score and randomly select the centers for all patients.
3. Calculate the distance between the patient's anomaly score and all centers.
4. Assign each patient to the nearest cluster.

k-means cluster applies an expectation maximisation method which has an E-step and M-step, where the E-step assigns profiles to the nearest clusters and the M-step computes the centroid for each cluster. The objective function (F), which is the sum of squares of distances of each data point optimized by the k-means cluster is as follows.

$$F = \sum_{i=1}^{m} \sum_{j=1}^{n} w_{ij} \left\| d^i - c_j \right\|^2 \tag{3.8}$$

where d^i = each patient profile, n is number of clusters, and w_{ij} is an indicator variable which indicates whether a profile is exists in the cluster or not.

If profiles d^i belong to cluster n, then w_{ij} =1 else the value of w_{ij} is 0 and c_k is the centroids of d_i clusters.

The E-steps for selecting the closest clusters can be described by following equation.

$$\frac{\partial F}{\partial w_{ij}} = \sum_{i=1}^{m} \sum_{j=1}^{n} w_{ij\|d^i - c_j\|^2} \quad w_{ij} = \begin{cases} 1 & if\ k = argmin_j \left\| d^i - c_j \right\|^2 \\ 0 & otherwise \end{cases} \tag{3.9}$$

Patients d^i were assigned to the closest cluster as judged by the sum of the distance from the cluster's centroid. M-step is described as:

$$\frac{\partial F}{\partial c_j} = 2\sum_{i=1}^{m} w_{ij\|d^i - c_j\|=0} \tag{3.10}$$

M-step selects the new centroid for each cluster to reflect the new cluster operation. M-step helps to achieve a better cluster for all patients.

3.4.5.1 Results of k-means clustering

Experimental setup

1. Technology: k-means clustering algorithm, confusion_matrix,classification_report.
2. Python packages: NumPy, pandas, sklearn.cluster, and sklearn.metrics.
3. Input data: Anomaly scores for all four cancer data sets and scores for all state-of-the-art methodologies.

This section discusses the results of k-means clustering with anomaly scores and analyze how these results compare to other state-of-the-art methodologies.

The k-means clustering algorithm is used to analyze the accuracy of patient embedding provided by anomaly scores and state-of-the-art approaches for leukemia, colon, and adrenal cancer patients, respectively. Results of the clustering performed on these four data sets are presented in Table 3.1, together with the anomaly scores and state-of-the-art methods used. The performance of clustering is evaluated using precision, recall, accuracy, and f-measure evaluation metrics, as well as the following rates: true positive (TP), false positive (FP), true negative (TN), and false negative. It is said to be a true positive rate if all patients who relapse belong to the same cluster. Conversely, if patients who do not relapse are grouped together, this is considered a true negative result. When a cluster has a combination of different classes, either the rate of false positives or false negatives is increased.

Table 3.1: k-means clustering performance evaluation for proposed method and state-of-the-art approaches.

Methods		Leukaemia	Colon cancer	Breast cancer	Adrenal cancer
Anomaly score	z-absolute	75.10%	54.72%	69.84%	92.31%
	z-square	84.72%	98.89 %	88.37%	55.91%
	z-cube	84.03%	97.92%	88.37%	55.91%
GFS score		43.75%	50.94%	69.84%	55.38%
FRaC score		48.84%	35.56%	58.14%	56.25%
CSAX score		48.84%	35.56%	58.14%	56.25%
TEMPO score		49.67%	36.73%	59.32%	57.10%
aTEMPO score		57.32%	39.35%	60.42%	61.23%
Outlier score		43.75%	49.06%	30.16%	46.20%
SNet score		43.75%	30.19%	30.16%	46.15%
PFSNet score		43.75%	30.19%	69.84%	50.77%
qPFS score		43.75%	30.19%	30.16%	46.15%
Eigfusion score		43.75%	30.19%	30.16%	46.20%

3.4.5.2 Discussion on k-means clustering

According to the table, the clustering accuracy of the proposed method is 75.10% for leukaemia data sets with z-absolute anomaly score, which is a higher percentage than state-of-the-art methods. Accuracy is achieved by using state-of-the-art methods such as GFS (43.75%), FRaC (48.84%), CSAX (48.84%), TEMPO (49.67%), aTEMPO (57.32%), outlier detection (43.75%), SNet (43.75%), PFSNet (43.75%), qPSP (43.75%), and Eigfusion (43.75%). However, patient embedding for these state-of-the-art approaches shows far weaker clustering than their numerical precision would suggest.

In the colon cancer dataset, z-absolute anomaly scores improved clustering accuracy by 54.72%. In addition, the accuracy of the z-square and z-cube anomaly scores is 98.89% and 97.92%, respectively. In contrast, state-of-the-art methods such as GFS, FRaC, CSAX, TEMPO, aTEMPO, outlier detection, SNet, PFSNet, qPSP, and Eigfusion display comparatively lower accuracies. Specifically, GFS achieves 50.94%, while FRaC and CSAX both register a 35.56% accuracy. TEMPO delivers a slightly higher 36.73% accuracy, with aTEMPO at 39.35%. Outlier detection reaches 49.06%, while SNet, PFSNet, qPSP, and Eigfusion all achieved a 30.19% accuracy.

For the breast cancer datasets, the proposed z-absolute anomaly scores show an improvement in clustering accuracy and achieved a success rate of 69.84%, outperforming traditional state-of-the-art methods. Moreover, the z-square and z-cube anomaly scores achieved even higher accuracy, both reaching 88.37%. In comparison, state-of-the-art methods such as GFS, FRaC, CSAX, TEMPO, aTEMPO, outlier detection, SNet, PFSNet, qPSP, and Eigfusion achieved lower accuracy. GFS and PFSNet both achieved 69.84% accuracy, while FRaC and CSAX achieve 58.14% accuracy. TEMPO achieved a slightly higher accuracy of 59.32%, while aTEMPO achieved 60.42%. Outlier detection, SNet, qPSP and Eigfusion all had significantly lower accuracy of 30.16%. In conclusion, the proposed z-absolute, z-square, and z-cube anomaly scores represent an improvement in clustering accuracy for breast cancer datasets when compared with existing state-of-the-art methods.

In the adrenal cancer dataset, the proposed z-absolute anomaly scores achieved a clustering accuracy of 92.31%. However, the z-squared and z-cube anomaly scores have lower accuracy in this case, both reaching 55.91%. In comparison, state-of-the-art methods such as GFS, FRaC, CSAX, TEMPO, aTEMPO, outlier detection, SNet, PFSNet, qPSP, and Eigfusion achieved lower accuracy. GFS achieved 55.38% accuracy, while FRaC and CSAX both achieved 56.25% accuracy. TEMPO provided slightly higher accuracy of 57.10%, while aTEMPO achieved 61.23%. Outlier detection and Eigfusion achieved 46.20% accuracy, while SNet and qPSP both achieve 46.15% accuracy. PFSNet achieved an accuracy of 50.77%.

3.5 Complexity analysis

Complexity analysis is an approach used to measure the performance of an algorithm. It examines the amount of time and memory (or space) an algorithm requires to solve a problem [438]. This approach provides an overview of how the algorithm works in three distinct scenarios such as best, worst, and average cases [439]. Special symbols are used to express the time and space complexity of an algorithm. These symbols are Big $\mathcal{O}$ (pronounced "big oh"), Big Ω (pronounced "big omega"), and Big Θ (pronounced "big theta"). Big Ω (big omega) represents the best-case scenario, which is the minimum time that an algorithm might take, Big Θ (big Theta) represents the average case scenario, which indicates the average time required by an algorithm. Big $\mathcal{O}$ (big O) represents the worst-case scenario, which is the maximum time required by an algorithm [440]. Time complexity is measured with respect to processing time an algorithm requires, while space complexity is determined by the quantity of memory the algorithm needs [441].

Further more, run time refers to the time an algorithm takes to execute, usually measured by the number of basic operations it performs [442].

- Big $\mathcal{O}$: Big $\mathcal{O}$ notation is used to show the maximum growth rate of an algorithm's time or space complexity. The equation for Big $\mathcal{O}$ notation is:

  $$f\left(n\right) = \mathcal{O}\left(g\left(n\right)\right) \tag{3.11}$$

  Here $f(n)$ describes the time or space complexity of the algorithm based on the input size n. The $g(n)$ represents the maximum growth rate of the algorithm as the input size increases.

- Big Ω : Big Ω notation is used to show the minimum growth rate of an algorithm's time or space complexity. It is written as $f(n) = \Omega(g(n))$. Here, $f(n)$ describes the time or space complexity of the algorithm based on the input size n. The $g(n)$ represents the minimum growth rate of the algorithm as the input size increases.

- Big Θ: Big Θ notation is used to show the tight bound of the growth rate of an algorithm's time or space complexity. It is written as $f(n) = \Theta(g(n))$. This means that $f(n)$ grows at the same rate as $g(n)$ as n becomes large, up to a constant factor.

This thesis provides an overview of the time and space complexity of different approaches to generate anomaly scores such as z-absolute score, z-square and z-cube. In addition, this thesis analyzes the time and space complexity of PCA and MRMR for patient embedding.

3.5.1 Time complexity for z-absolute anomaly score

This thesis assesses the time and space complexity of the z-absolute score, z-square, and z-cube anomaly approaches. Time complexity of the z-absolute anomaly score is shown in table 3.2. The table shows a loop at the second step, line 2, that iterates over all patients (n) to calculate their z-score values. A 'loop' refers to a process that repeats itself until a certain condition is met. The time complexity of this loop is $O(n)$. Lines 4 and 5 contain nested loops, which are loops inside another loop. These nested loops iterate over the genes (r) and different gene sets (m) for each patient, aligning the patient's genes with the genes in the gene sets. The time complexity for this part of the process is $O(mr)$.

From the table it can be seen that the second step of the process, line 2, is a loop. A loop is like a cycle, going over and over again until it is completed with respect to a condition. In this case, it's going over all patients (n) to determine their z-score value. The time complexity for this loop is $O(n)$. This means that the more patients we have, the longer it may take. In lines 4 and 5 of the process involve two loops inside each other, which is called nested loops. These loops go through the genes (r) and the different gene sets (m) of each patient. These loops match the genes of each patient with the genes in the gene sets. The time complexity for this part is $O(mr)$. This means it might take longer the more genes and gene sets we have. Finally, line 8, is another loop that goes over each gene set (m) to calculate the anomaly score. The time complexity for this step is $O(m)$. So, the total time complexity the z-absolute anomaly score evaluation is $\mathcal{O}(mnr + mn)$ = $\mathcal{O}(mnr)$ (Equation 3.12)

$$\begin{gathered}
\textstyle\sum_1^n \left(\sum_1^m O(r)\right) + O(r) \\
n * (m * O(r)) + O(r) \\
n * (O(mr) + O(r) \\
O\left(mnr\right) + O(nr) \\
O\left(mnr + nr\right) \\
O\left(mnr\right)
\end{gathered} \tag{3.12}$$

Table 3.2: Time and space complexity analysis for z-absolute anomaly score.

Statements	Description	Complexity	Overall complexity
# Create a matrix of anomaly scores. 1. anomalies = np.zeros((num_patients, self.num_of_genesets),dtype=np.float) 2. for patient_index, (patient, row) in enumerate (z_scores.iterrows()): 3. counts = np.zeros(self.num_of_genesets, dtype=np.int)	Apply a loop on each patient to access the z-score value. We consider n number of patients over the z-score values.	O(n)	O(nmr+nr)
4. for gene_id, z_score in zip(gene_ids, row): 5. for geneset_index in self.geneset_map.get(gene_id, ()): 6 .anomalies[patient_index, geneset_index] += z_score 7. counts[geneset_index] += 1	Apply nested two loops to match the genes (m) with geneset's (r) genes to sum z-scores.	O(mr)	
8. for geneset_index in range(self.num_of_genesets): 9. count = counts[geneset_index] 10. if count >0: 11. anomalies[patient_index, geneset_index] /= count	Apply a loop to calculate anomaly scores for each genesets (r).	O(r)	

3.5.2 Space complexity analysis for z-absolute anomaly score

The space complexity of z-absolute anomaly score is: $\mathcal{O}(r)$, because memory space depends on the r-number of genes in micro-array data. If the number of genes in the micro-array data then more bits occupied for the data.

The time complexity of z-square and z-cube are similar as z-score approach. Because total number of nested loops of z-square and z-cube are similar as z-score. So the time and space complexity of both z-square and z-cube are $\mathcal{O}(mnr + mn)$ and $\mathcal{O}(r)$ respectively.

3.5.2.1 Time and space complexity for PCA patients embedding

The time complexity of PCA depends on the most common approach is to use a singular value decomposition (SVD) of the data matrix [443]. This thesis includes a data matrix of size n (number of patients) × d (number of gene sets). The time complexity of computing the SVD, and hence the PCA, is typically $O(n \cdot d^2)$ if $n > d$, or $O(d \cdot n^2)$ if $d > n$. Table 3.3 represents time complexity for PCA. Line 2, covariance matrix generated for the dimension (gene sets, m) over the data points (patients, n). Since the covariance between gene sets is checked, the time complexity is O(m^2) for the gene sets. Thus, the time complexity is O(m^2n). Line 4 uses the calculation of eigenvalues from the covariance matrix for the gene sets (m), and the time complexity is O(m^3). Thus, the time complexity of the PCA is $\mathcal{O}(m^2n + m^3)$.

The space complexity is $\mathcal{O}(m)$. The PCA operation is performed on the anomaly scores over the genesets and the memory requirement depends on the total number of genesets. In different geneset families (e.g., C5, C6, etc.), there is a different number of genesets. When the number of genesets increases, the memory required by the PCA method also increases.

Table 3.3: Time complexity for PCA.

Statements	**Description**	**Complexity**	**Overall Complexity**
1.dfmt=dfm.transpose(). 2.co1=np.cov(dfmt). 3.np.shape(co1).	Covariance matrix generated on genesets (m) and patients (n). Here m*m covariance matrix generated on n-number of patients.	O(m^2n)	O(m^2n+m^3)
4.eigen_vals, eigen_vecs=np.linalg.eig(co1) 5.tot=float(sum(eigen_vals)).	The PCA complexity also depends on eigen value generation for genesets (m).	O(m^3)	

3.5.2.2 Time and space complexity for MRMR patients embedding

This thesis analyzes the time and memory requirements of the MRMR for patients embedding with respect to their anomaly scores. The time complexity of MRMR is presented in Table 3.4.

Table 3.4 represents step by step descriptions to estimate the time required. First, the entropy, which is a measure of disorder or randomness, is calculated for each patient. In line 1, this process is repeated for each patient. Thus, if there are n patients, the time required is proportional to n (we call this $O(n)$).

in lines 4 and 6, there are two loops, one inside the other, that go through both the gene sets (m) and the patients (n) and perform another entropy calculation for each combination of gene set and patient. This part of the process takes time proportional to the product of m and n (which we refer to as $O(mn)$). Equation 3.13 shows how we calculate the total time required by MRMR.

In line 8, this thesis calculates the mutual information for each set of genes. This is done in a loop that passes through each gene set so that the time required is proportional to the number of gene sets, m (or $O(m)$).

Finally, in lines 10 and 12, nested loops are run to calculate the joint entropy, a kind of combined measure of randomness, over m times the number of gene sets and n times the number of patients. Again, the time required is proportional to m times n (or $O(mn)$).

When you add it all up, the total time required by MRMR is proportional to the product of the number of gene sets and the number of patients ($O(mn)$), and the total memory required is proportional to the number of gene sets ($O(m)$). This is because the memory required by MRMR is directly dependent on the total number of gene sets.

$$\begin{aligned} TC &= TC\,loop1 + TC\,loop2 \,\&\, TC\,loop3 + TC\,loop4 + TC\,loop5 \,\&\, TC\,loop6 \\ &= O(n) + O(mn) + O(m) + O(mn) + O(mn) \\ &= O(mn) \end{aligned} \tag{3.13}$$

Table 3.4: Time complexity analysis for MRMR.

Statements	Description	Complexity	Overall complexity
1.for i in range(n[1]): 2.entr_x_by_y = dft / py[i] 3.row_sum = []	Apply a loop to calculate entropy for patient (n) based on gene sets	O(n)	O(mn)
4.for i in range(n[0]): 5.s = 0 6.for j in range(n[1]): 7.s = s + dft.iloc[i, j]	Apply nested two loops to sum the entropy for each genesets (m) with corresponding all genesets (n).	O(mn)	
8.for i in range(n[0]): 9.e = -(row_sum[i] * math.log1p(abs(row_sum[i])))	Apply a loop to H-value for each genesests (m)	O(m)	
10.for i in range(n[0]): 11.e = 0 12.for j in range(n[1]): 13.p_x_y = entr_x_by_y.iloc[i, j] 14.p = dft.iloc[i, j] 15.if p ¿ 0: 16.e += -(p * math.log1p(abs(p_x_y)))	Apply two nested loops to calculate joint entropy of genesets (m) over the all patients (n).	O(mn)	

3.5.3 Run time memory usage

Run time memory usage refers to the amount of memory a computer program uses while it is running. This includes the memory used by the program itself, as well as any data structures and variables it creates and stores in memory during execution [444, 445].

Run time memory usage is a valuable consideration in the development of software applications, especially those that are memory-intensive or are intended to run on devices with limited memory resources. Run time memory usage controls the following two computing factors:

- Performance: The amount of memory used by an application at run time can have a significant impact on its performance. Applications that use too much memory may become slow and unresponsive, while those that use too little memory may not be able to handle large data sets efficiently. Optimizing run time memory usage can improve application performance and ensure that it runs smoothly.

- Resource management: Memory is a finite resource, and applications that use too much memory can cause other applications to run slowly or crash. Optimizing runtime memory usage helps to ensure that an application does not use more memory than it needs, freeing up resources for other applications.

This thesis measures the run time memory usages for generating anomaly scores as well as PCA and MRMR methods patients embedding. This thesis considers memory profiling, which is the process of analyzing a program's memory usage to identify any memory-related issues such as memory leaks, excessive memory consumption, and inefficient memory usage [446]. Memory profiling involves measuring the memory usage of a program at different points in its execution, and analyzing the data collected to identify potential problems.

The run time memory usage for z-score anomaly approach described in table 3.5, where it can be seen that the occupied memory allocation in mebibyte (MiB) unit. A Mebibyte (MiB) is a unit of digital information storage used to denote the size of data. It is equivalent to 1,048,576 bytes, or 1,024 Kibibytes (KiB) [447]. This term was defined by the International Electrotechnical Commission (IEC) to clarify the difference between the metric and binary interpretations of the units kilobyte, megabyte, and gigabyte [448]. Previously, the term "megabyte" (MB) was used ambiguously to refer to either 1,000,000 bytes (in accordance with the metric system) or 1,048,576 bytes (in accordance with the binary system), causing confusion.

Moreover, this thesis also represented the memory usages for PCA and MRMR patients embedding in table 3.6 and table 3.7.

Table 3.5: Run time memory usages for z-absolute anomaly anomaly.

Line Contents	Memory usage (MiB)	Increment (MiB)
self.mean = train_data.mean()	252.1 MiB	1.8 MiB
self.std = train_data.std()	252.5 MiB	0.4 MiB
num_patients = train_data.shape[0]	252.5 MiB	0.0 MiB
z_scores = (abs(train_data - self.mean) / self.std)	268.8 MiB	16.4 MiB
gene_ids = train_data.columns	268.8 MiB	0.0 MiB
anomalies = np.zeros((num_patients, self.num_of_genesets), dtype=np.float)	269.0 MiB	0.1 MiB
for patient_index, (patient, row) in enumerate(z_scores.iterrows()):	269.0 MiB	1.1 MiB
counts = np.zeros(self.num_of_genesets, dtype=np.int)	269.0 MiB	1.1 MiB
for gene_id, z_score in zip(gene_ids, row):	269.0 MiB	25016.1 MiB
for geneset_index in self.geneset_map.get (gene_id, ()):	269.0 MiB	71721.0 MiB
anomalies[patient_index, geneset_index] + = z_score	269.0 MiB	46706.0 MiB
· 1 Mebibyte (MiB) = 1,048,576 bytes		

Table 3.6: Run time memory usage for PCA.

Line Contents	Memory usage (MiB)	Increment (MiB)
l = np.shape(df_input)	199.3 MiB	0.0 MiB
dfm = df_input	199.3 MiB	0.0 MiB
for i in range(l[1]):	199.3 MiB	0.0 MiB
mean = df_input.iloc[:, i].mean()	199.3 MiB	0.0 MiB
dfm.iloc[:, i] -= mean	199.3 MiB	0.0 MiB
dfmt = dfm.transpose()	199.3 MiB	0.0 MiB
co1 = np.cov(dfmt)	200.1 MiB	0.8 MiB
np.shape(co1)	200.1 MiB	0.0 MiB
eigen_vals, eigen_vecs = np.linalg.eig(co1)	200.1 MiB	1.5 MiB
tot = float(sum(eigen_vals))	200.1 MiB	0.0 MiB
var_exp = [(float(i) / tot) * 100 for i in eigen_vals]	200.1 MiB	0.0 MiB
cum_var_exp = np.cumsum(var_exp)	200.1 MiB	0.0 MiB
eigen_pairs = [(np.abs(eigen_vals[i]), eigen_vecs[:, i])for i in range(len(eigen_vals))]	200.1 MiB	0.0 MiB
eigen_pairs.sort(key=lambda k: k[0], reverse=True)	200.1 MiB	0.0 MiB
w = np.hstack((eigen_pairs[0][1][:, np.newaxis], eigen_pairs[1][1][:, np.newaxis]))	200.1 MiB	0.0 MiB
emb = df_input.dot(w)	202.1 MiB	0.5 MiB
return emb	202.1 MiB	0.0 MiB
1 Mebibyte (MiB) =1,048,576 bytes		

Table 3.7: Run time memory usage for MRMR.

Line Contents	Memory usage (MiB)	Increment (MiB)
dft = X.transpose()	235.7 MiB	0.0 MiB
n = dft.shape	235.7 MiB	0.0 MiB
px = dft.sum(axis=1, skipna=True)	235.7 MiB	0.0 MiB
py = dft.sum(axis=0, skipna=True)	235.7 MiB	0.0 MiB
for i in range(n[1]):	236.1 MiB	0.0 MiB
entr_x_by_y = dft / py[i]	236.1 MiB	0.3 MiB
for j in range(n[1]):	236.1 MiB	0.0 MiB
s = s + dft.iloc[i, j]	236.1 MiB	0.1 MiB
for j in range(n[1]):	236.1 MiB	0.0 MiB
p_x_y = entr_x_by_y.iloc[i, j]	236.1 MiB	0.0 MiB
for j in range(n[0]):	236.1 MiB	0.0 MiB
s = s + dft.iloc[j, i]	236.1 MiB	0.0 MiB
Mi_h_x_s = h_x_s - joint_entp_mutual	236.1 MiB	0.0 MiB
avg_mi_xs = Mi_h_x_s.mean()	236.1 MiB	0.0 MiB
mrmr = Mi_h_x - avg_mi_xs	236.1 MiB	0.0 MiB
gene_frame = pd.concat([geneset, mrmr_frame], axis=1)	236.2 MiB	0.1 MiB
mrmr_id = gene_frame.index.values	236.3 MiB	0.1 MiB
GS1 = mrmr_id[0]	236.3 MiB	0.0 MiB
GS2 = mrmr_id[1]	236.3 MiB	0.0 MiB
return GS1,GS2	236.3 MiB	0.0 MiB

3.6 Conclusion

This thesis hypothesized that the careful aggregation of gene expression values into gene set anomaly scores would enable insights to be gleaned from the data analysis of gene expression profiles. A number of techniques for this purpose have been described and it has been shown that even simple methods such as using z-absolute anomaly score of gene expression values to measure variation and finding the arithmetic mean of the variations for each gene set are sufficient to provide an advantage over the direct processing of gene expression values.

The analysis of gene expression data from cancer patients was put through its paces using the proposed method, which was applied across four separate data sets. In particular, anomaly scores, followed by either PCA or MRMR, rendered clusters of cancer patient data visible in scatter plots. These clusters appear to be grouped with similar patient cohorts. Additionally, MRMR was able to identify prospective gene sets that have consequences that were important to biology. On the other hand, when the raw gene expression values and state-of-the-art methods were analyzed, biologically significant patterns could not be seen. The results of using k-means clustering compared to all existing state-of-the-art methods and my proposed approach also reflect the visualization of patient embedding.

When the patients who had relapsed were compared to those who had not, it was found that there was a distinct difference in the distribution of anomaly scores. Distributions for patients who had not experienced relapse displayed a substantial second mode. This raises the possibility of a fascinating difference that can be utilised in the quest for improved treatment.

This idea leads this thesis to conclude that gene set anomaly scores offers to gain insights from gene expression data. Utilising gene sets provides a tool that is knowledge-driven, which may then be paired with an analysis that is data-driven. This thesis proposes a new method of analysis as well as new directions for gaining an understanding of the genetic factors that contribute to disease.

Chapter 4

Interpretation of machine learning outcomes on disease analysis

---

"Negative results are just what I want. They're just as valuable to me as positive results. I can never find the thing that does the job best until I find the ones that don't." — Thomas A. Edison

4.1 Introduction

This chapter focuses on two goals. The first goal of this chapter is to explain how predictive models and anomaly scores were used to investigate a patient's health status. This thesis is interested in capturing information about genes or gene sets associated with predicting a medical condition of a patient. In general, a predictive model predicts a status of an instance and produces accuracy. However, when it comes to making medical decisions, predicting a medical status and being right about it may not be enough to give confidence to the user and medical professionals.

In addition to the accuracy and prediction of medical status, this thesis focused on interpretability and explainability to understand the prediction mechanisms and features associated with a prediction.

Therefore, this thesis investigates how XAI approaches can be used to predict a patient's medical condition and learn more about it. XAI approaches provide interpretability and explainability for the prediction decision process and identify the features associated with a prediction.

The second goal of this chapter is to present that the proposed anomaly score identifies novel gene sets in contrast to GSEA and state-of-the-art approaches.

4.2 Overview of the chapter

“Omics” data sets such as genomics, proteomics and metabolomics create large-scale gene expressions that help researchers learn more about diseases like cancer. In data-driven analysis, there has been growing interest in the use of predictive models and deep learning to enhance the accuracy of decision making in healthcare and medical data analysis. In addition, existing artificial intelligence (AI) approaches to DNA sequencing, gene expression analysis, drug prediction, personalised medicine, and next generation sequencing allow users to observe the results with more precision. AI approaches are used to extract features to gain insights with small assumptions and lots of processing capacity. Due to their higher processing capacities, AI approaches are often used in genomic data mining, gene prediction, and disease prediction [307, 308]. AI enhances our perception of complex relationships in underlying gene expressions, treatment planning, and patients biology [309].

Despite their usefulness and effectiveness, some AI approaches lacked certain characteristics, most notably explainability, interpretability, and trustworthiness. Deep neural networks, for example, consist of layers of interconnected variables that are adapted by training the network on multiple instances [310]. As neural networks become more complex, it becomes more difficult to understand how numbers of parameters interact to make decisions [311]. Inputs could go through hundreds of iterative nonlinear transformations involving a number of features before a decision is made. The internal data processing of these AI approaches is too complicated to be understood by humans without prior technical knowledge. Due to the intricate nature of the data processing stages involved in AI techniques, these algorithms are often referred to as black boxes. How can even the most accurate black-box approaches improve the user’s understanding of biomedical data processing?

Explainable AI (XAI) works between AI and humans to interpret decision making process of black-box AI approaches [312, 313]. For example, x is a medical condition of a patient such as cancer relapse or cancer non-relapse. Predictions define what x is, while explainability specifies why x exists. Explainability thus brings value to making trustworthy and understandable decisions. The most recent version of the General Data Protection Regulation, adopted by the European Union, focuses on need for citizens to know how AI systems arrive at their conclusions [314]. In addition, the ethical standards for artificial intelligence have been broken down into eight categories by the Federal Government of Australia. These standards include criteria pertaining to explainability and transparency [315].

The instance-based local interpretation-driven abstract Bayesian network (LINDA-BN) was utilized to determine an individual patient's medical state by using the gene set anomaly score. The purpose of this investigation is to identify which gene sets (patient biology) or genes are associated with predicting the condition of an individual patient. This provided insight into the biological factors related to a patient's health status, such as the presence or absence of cancer and the likelihood of a cancer relapse.

In addition, to specify the biological name of the gene set, gene ontology (GO) terms in publicly available databases were explored. These terms represent the biology of the gene set. In the end, a heatmap was used to represent the GO terms that were extracted using the proposed anomaly scores and state-of-the-art approaches. The purpose of the heatmap is to investigate whether or not the anomaly score could extract a new biology or similar biology that is comparable to the state-of-the-art approaches.

This chapter focuses two issues. First issue is an application of anomaly score using XAI. The objective of applying the XAI approach is to enable interpretable predictions of a patient's medical condition. A second issue involves the use of anomaly scores to identify cancer biology.

The following research question is addressed.

- How can an explainable and interpretable method predict a patient status (healthy or cancerous or relapse or non-relapse) built on an individual instance in relation to anomaly scores?

4.3 Method

A method consisting of the following two-step process is proposed to evaluate the usefulness of machine learning in disease analysis:

1. instance-based LINDA-BN.
2. representing new cancer biology.

4.3.1 Data sets

GSE14468 AML dataset The GSE14468 dataset comprises gene expression data from 524 acute myeloid leukemia (AML) patients, aged 16 to 60 years old. It covers 54,675 genes and was utilized by Warnat et al. [449] to identify CCAAT/enhancer binding protein alpha (CEBPA) mutations, comparing double and single CEBPA mutations using the gene expression data.

GSE12417-GPL570 CN-AML dataset The GSE12417-GPL570 dataset contains gene expression data for 163 cytogenetically normal acute myeloid leukemia (CN-AML) patients, covering 54,674 genes. Wang et al. [450] employed this data to develop a gene signature predicting overall survival (OS) in CN-AML and to identify a CN-AML patient cohort using supervised principal component analysis.

GSE24006 AML dataset The GSE24006 dataset includes gene expression data for 1,047 AML patients, encompassing 29,397 leukemia stem cell (LSC) genes. Gentles et al. [451] used this data to identify leukemic stem cell genes and patient cohorts for overall survival (OS), event-free survival (EFS), and relapse-free survival (RFS).

4.3.2 Instance-based learning (IBL)

Instance-based learning measures the distances between the attributes of a test instance and the attributes of all training samples to determine how similar or different the test instance is from the training samples [452]. The first step of IBL is to compare the feature values of a test instance with all the features values of all training instances using equation 4.1. The comparative results are stored in a matrix, which is a relevancy matrix or similarity matrix for a test patient (Figure 4.4).

$$IBL\,Matrix = \begin{cases} \text{if } |Training\,AS - Test\,AS| \leqslant 0.10, \text{ then, } AS = Training\,AS \\ |Training\,AS - Test\,AS| = 0, \text{ then, } AS = 1 \\ \text{otherwise, } AS = 0 \end{cases} \tag{4.1}$$

4.3.3 Local interpretation-driven abstract Bayesian network (LINDA-BN)

LINDA-BN is an XAI approach that graphically shows the data processing steps using the Bayesian network. The objective of LINDA- BN is to provide an interpretation of black-box predictions. LINDA-BN is capable of deriving an approximate Bayesian network, which serves as a simplified representation of a black-box model. This network corresponds to a specific prediction generated from any supplied input. For instance, LINDA-BN shows the relationships between gene sets in a given prediction using graphical inference [453].

4.3.3.1 LINDA-BN for interpreting prediction outcomes

Probabilistic graphical models (PGMs) are graphical approaches to represent probability distributions. Two useful PGMs are Bayesian networks (BNs) and hidden Markov models (HMMs). BNs use directed acyclic graphs (DAGs) to show the relationships between multiple random variables. They are typically used for processes that do not change over time. HMMs are used for Markov processes, which are systems that change over time, space, or other sets. HMMs have one hidden state variable and one observable variable that depends on the hidden state. Dynamic Bayesian networks (DBNs) are a PGM that combines the features of BNs and HMMs. DBNs can have multiple hidden random variables to represent processes that change over time.

The local interpretation-driven abstract Bayesian network (LINDA-BN) is a DBN that visually represents the relationships between different features. This thesis used LINDA-BN to aid the interpretation of the predictive outcomes for patients who relapse and those who do not, along with their associated gene sets. This approach provides an exaplainable decision-making process for determining the relevance of specific features or groups of features. The LINDA-BN process includes three distinct stages:

1. permutations of gene set
2. a graphical Bayesian network
3. analysing Markov blankets for feature associations.

These three steps are briefly described below.

(1) Permutations of gene sets

From the figure 4.1, it can be seen that input feature vectors were permuted using uniform distributions. For the gene set anomaly score, GS is an input vector as $GS = \left\{ GS_1, GS_2, ..., GS_n \right\}$, the variance of the permutations is ϵ, where $\epsilon \; \varepsilon \left[0, 1 \right]$, and the permutation interval for each gene set is $\left[GS_i - \epsilon, GS_i + \epsilon \right]$.

(2) Graphical Bayesian network

At this stage, LINDA-BN constructs a directed acyclic graph (DAG) as a Bayesian network for each gene set. As seen in Figure 4.1, a DAG is constructed after the permutations, providing structural graphical inference for gene sets regarding anomaly scores.

For the BN graphical inference from the DAG, let G be a BN graph over gene sets $GS_1, GS_2, ..., GS_n$, and the probability P of gene sets on G is defined as:

$$P(GS_1, GS_2, ..., GS_n) = \prod_{i=1}^{n} P(GS_i | Pa_{GS_i}) \tag{4.2}$$

Here, Pa_{GS_i} denotes the set of parent variables for gene sets GS_i. The joint probability of gene sets for specific events E with random variable v is calculated as follows:

$$P(E|V = v) = \alpha P(E, v) = \alpha \sum_{w \varepsilon W} P(E, v, w), with \alpha = \frac{1}{\sum_{e \varepsilon E} P(e, v)} \tag{4.3}$$

W represents the set of random gene sets not belonging to cancer prediction classes. To identify relationships between gene sets, conditional dependence in the DAG for all gene sets with respect to the anomaly scores. Let ϕ be the conditional dependence for gene set d with n observations, measuring the conditional dependence $P(G, \phi | d)$ in two phases: structure learning and parameter learning, represented by:

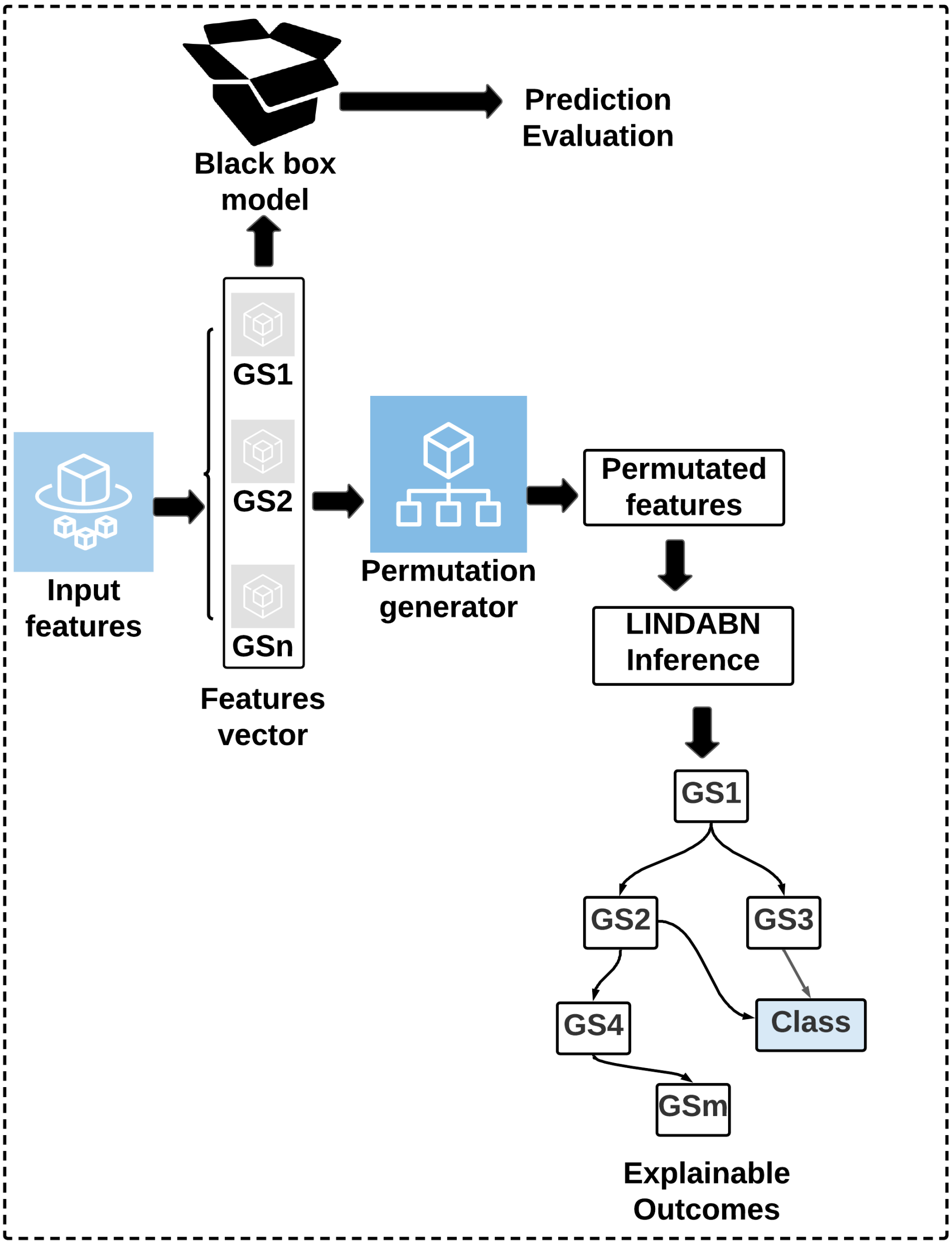


Figure 4.1: A framework for LINDA-BN.

$$P(G, \phi|d) = p(G|d)P(\phi|G, d) \tag{4.4}$$

In this context, $p(G|d)$ represents structure learning and $P(\phi|G, d)$ denotes parameter learning. The goal of structure learning is to identify the Directed Acyclic Graph (DAG) G by maximizing the value of $p(G|d)$ using the following equation:

$$P(\phi|G,d) = \prod_i P(\phi_{GS_i}|\prod GS_i, d) \tag{4.5}$$

Yet, it's important to note that structure learning is a problem classified as both NP-hard and NP-complete.

$$P(G|d) \propto P(G)P(d|G) \tag{4.6}$$

$P(d|G)$ can be decomposed into:

$$P(d|G) = \int P(d|G,\phi)P(\phi|G)d\phi \prod_i \int P(GS_i|\prod GS_i, \phi_{GS_i})P(\phi_{GS_i}|\prod GS_i)d\phi_{GS_i} \tag{4.7}$$

For structure learning, the optimal score is determined using the Bayesian information criterion (BIC), which is defined as follows:

$$SCore(G,d) = BIC(G,\phi|d) = \sum_i logP(GS_i|\prod GS_i, \phi_{GS_i}) - \frac{log(n)}{2}|GS_i| \tag{4.8}$$

LINDABN uses a hill-climbing algorithm to identify related features from the DAG.

(3) Applying Markov-Blanket for feature associations

Again from Figure 4.1, the Markov blanket is used on the DAG to identify features based on the predicted Bayesian network outcomes. The LINDA-BN graphically represents the Bayesian network and illustrates the relationships between the features and the predicted class (Figure 4.2a). According to the principle of the Naive Bayes classifier, the features are independent and uncorrelated. For example, $GL1_UP$, $E2F1_UP$, and $E2F1_DN$ are independent in predicting a class label, which means that knowing $GL1_UP$ does not provide any additional information for inference or prediction. Conversely, Figure 4.2b represents the principle of linear regression, meaning that features ($GL1_UP$, $E2F1_UP$, and $E2F1_DN$) are conditionally independent of a class only when the class variable is known. These features directly influence decision making for a target variable. A clear explanation of the decision-making process helps the user to fully understand the basic elements that are important in the process.

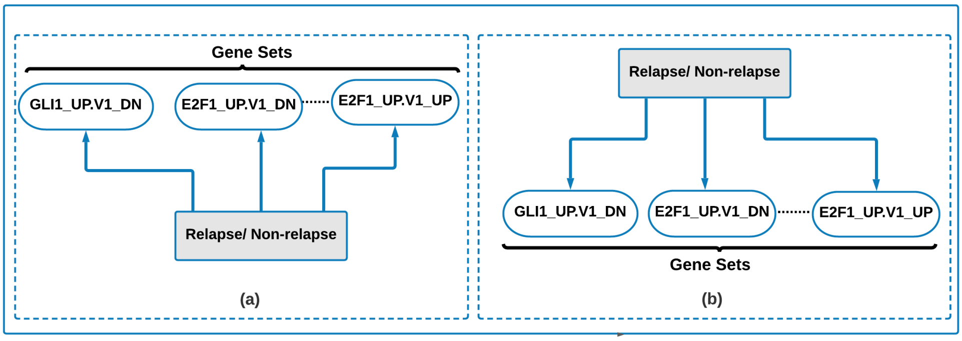


Figure 4.2: Different graph structure for reasoning.

The relationship between the target variable (class variable) and the features can be understood through an abductive reasoning process [361]. This involves human inference to draw conclusions from known information. Users apply abductive inference to generate reliable explanations related to the graphical structure. Abductive inference supports the target variable Markov blanket [362], an approach to feature selection for a specific class.

A Markov blanket of a target variable includes conditionally independent characteristics or variables that are parents, children, and co-parents (parents of a child) of that target variable (Figure 4.3). From Figure 4.3, this thesis observed that the Markov blanket selected three features: $GLI1_UP.V1_DN$, $EGFR_UP.V1_DN$, and $EGFR_UP.V1_UP$

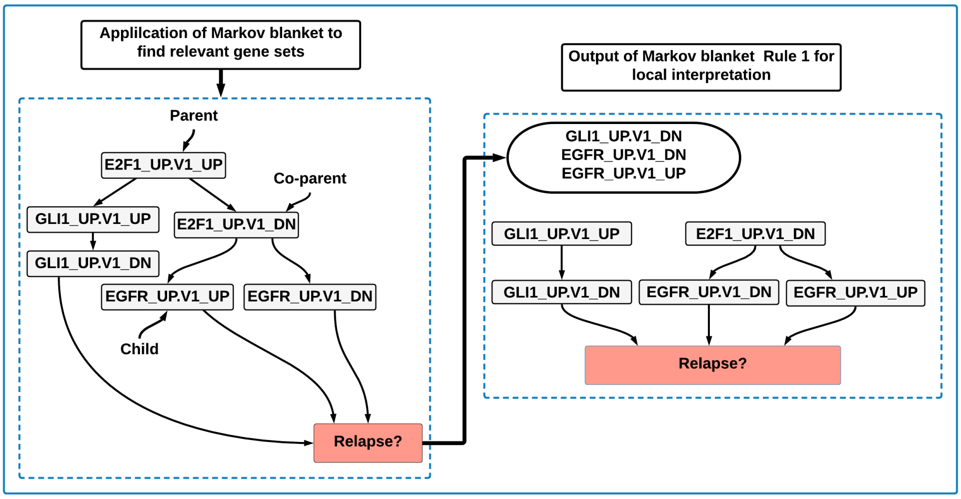
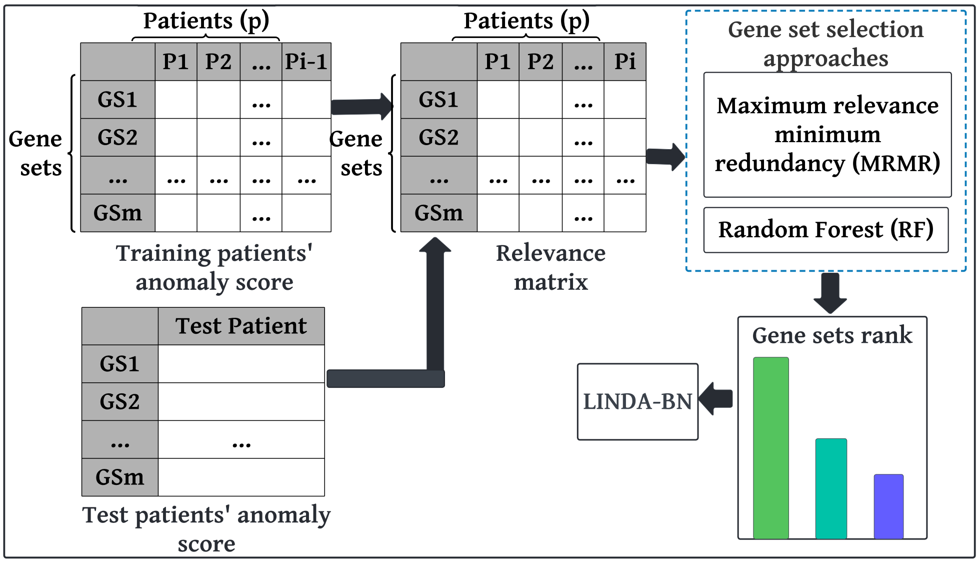


Figure 4.3: Features selection using Markov blanket.

4.3.3.2 Instance-based LINDA-BN

Instance-based LINDA-BN is placed in the data processing context as shown in Figure 4.4. The process begins by comparing a test instance with all the training samples and recording the results of this comparison into a matrix, referred to as a relevance matrix. Only those values that are comparable to the test instance with respect to a given condition are included in the relevance matrix. Following this step, feature selection approaches, namely maximum relevance and minimum redundancy (MRMR) and random forest (RF), are applied to the relevancy matrix to rank the gene sets. This is done to identify the gene sets that are of the utmost significance. In the final stage of the process, LINDA-BN [453] is applied to the ranked gene set list to determine which gene sets are responsible for predicting medical conditions affecting a patient.



Patients (p)
Gene sets
Training patients' anomaly score
Patients (p)
Gene sets
Relevance matrix
Gene set selection approaches
Maximum relevance minimum redundancy (MRMR)
Random Forest (RF)
Gene sets rank
LINDA-BN
Test patients' anomaly score

	P1	P2	...	Pi-1
GS1			...	
GS2			...	
...	...	...	...	...
GSm			...	

	P1	P2	...	Pi
GS1			...	
GS2			...	
...	...	...	...	...
GSm			...	

	Test Patient
GS1	
GS2	
...	...
GSm	

Figure 4.4: A schematic diagram showing processing steps of instance-based LINDA-BN.

4.4 Results

This section summarises and discusses the findings from the instance-based LINDA-BN. LINDA-BN identifies gene sets with respect to conditional dependence and independence. However, only conditionally dependent gene sets are associated with a prediction. Experimental results show that the method offers performance advantages in identifying gene sets or genes associated with a prediction. The gene sets that are associated with a prediction have been highlighted in blue in the figure in this chapter. The other gene sets are not associated with a prediction due to their conditional independence.

4.4.1 Instance-based LINDA-BN using anomaly score for leukaemia datasets

LINDA-BN identifies the gene sets that are associated with a disease prediction of a particular patient. Figure 4.5 shows the gene sets that are associated with the prediction of acute lymphoblastic leukemia in patient ALL10, based on the z-absolute anomaly score. The figure shows that two gene sets, CTIP DN and PRC2 EED, are associated with the likelihood of cancer relapse in patient ALL10. Based on the figure, this thesis hypothesizes that these gene sets (biological functionalities) are responsible for a cancer prognosis for the ALL10 patient. The CITP complex enhances the growth of breast tumors and the PRC2/EED complex is associated with the increased expression of lymph node metastases in breast cancer [454, 455].

Table 4.1: Gene set ranking for the ALL10 patient using z-absolute, z-square, and z-cubic anomaly scores.

Z-absolute	Z-square	Z-cube
BCAT.100_UP.V1_DN	BCAT.100_UP.V1_DN	BCAT.100_UP.V1_DN
KRAS.50_UP.V1_DN	KRAS.50_UP.V1_DN	KRAS.50_UP.V1_DN
KRAS.KIDNEY_UP.V1_DN	HINATA_NFKB_IMMU_INF	HINATA_NFKB_IMMU_INF
RELA_DN.V1_DN	HINATA_NFKB_MATRIX	HINATA_NFKB_MATRIX
PDGF_ERK_DN.V1_DN	GLI1_UP.V1_DN	GLI1_UP.V1_UP
ALK_DN.V1_UP	GLI1_UP.V1_UP	YAP1_DN
CTIP_DN.V1_DN	SINGH_KRAS_DEPENDENCY_SIGNATURE_	GLI1_UP.V1_DN
SNF5_DN.V1_UP	BCAT_GDS748_UP	SINGH_KRAS_DEPENDENCY_SIGNATURE_
PRC2_EED_UP.V1_DN	KRAS.50_UP.V1_UP	YAP1_UP
PRC2_EZH2_UP.V1_DN	YAP1_DN	TBK1.DN.48HRS_DN
BRCA1_DN.V1_DN	BCAT_BILD_ET_AL_DN	BCAT_BILD_ET_AL_UP
ESC_V6.5_UP_EARLY.V1_UP	TBK1.DN.48HRS_UP	KRAS.50_UP.V1_UP
PRC2_EZH2_UP.V1_UP	CAHOY_NEURONAL	TBK1.DN.48HRS_UP
KRAS.PROSTATE_UP.V1_DN	CAHOY_OLIGODENDROCUTIC	BCAT_GDS748_DN
PRC2_EED_UP.V1_UP	CAHOY_ASTROCYTIC	BCAT_GDS748_UP
NOTCH_DN.V1_UP	YAP1_UP	BCAT.100_UP.V1_UP
P53_DN.V1_UP	BCAT.100_UP.V1_UP	CAHOY_NEURONAL
KRAS.BREAST_UP.V1_DN	EIF4E_UP	BCAT_BILD_ET_AL_DN
CYCLIN_D1_KE_.V1_DN	CORDENONSI_YAP_CONSERVED_SIGNATURE	CORDENONSI_YAP_CONSERVED_SIGNATURE
STK33_NOMO_DN	CAHOY_ASTROGLIAL	EIF4E_DN

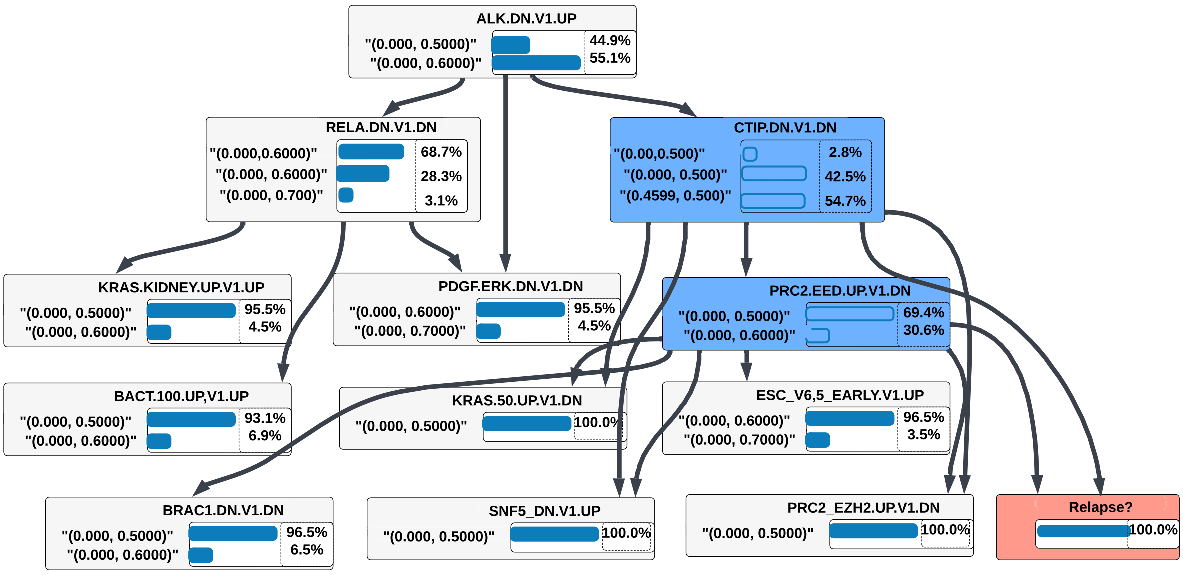


Figure 4.5: Gene sets associated with the disease prognosis in ALL10 patient using z-absolute anomaly scores.

Similarly, Figure 4.6 shows that the gene sets KRAS.50_UP, and BACT.100_UP are associated with the disease prognosis of ALL10 using z-square anomaly scores. KRAS plays a role in

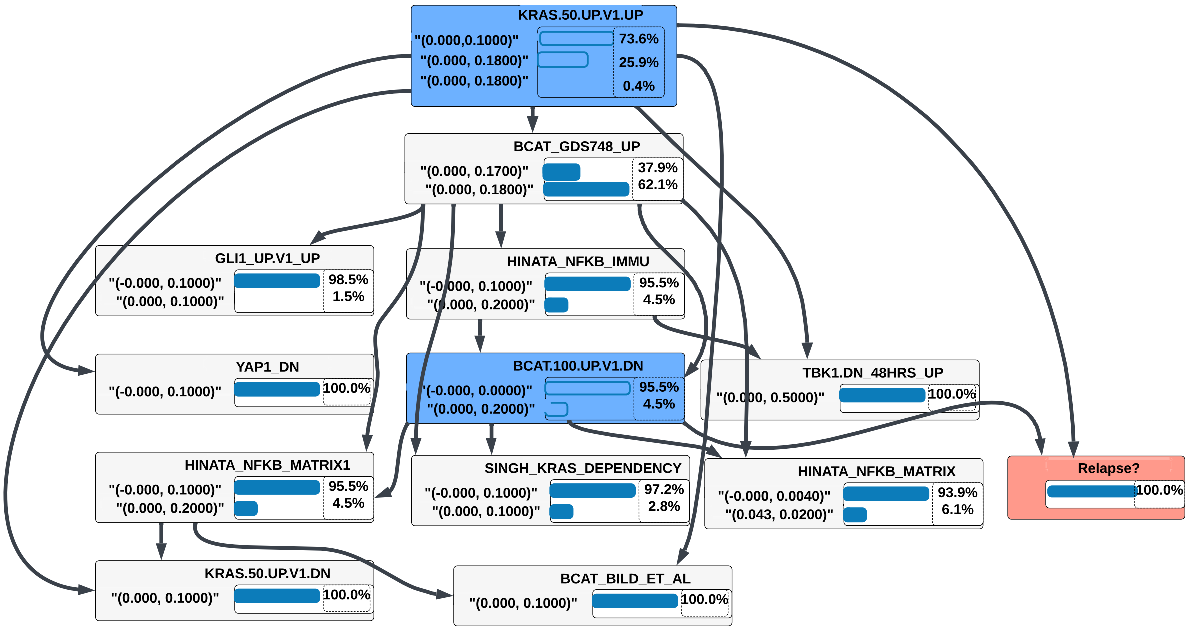


Figure 4.6: Gene sets associated with the disease prognosis in ALL10 patient using z-square anomaly scores.

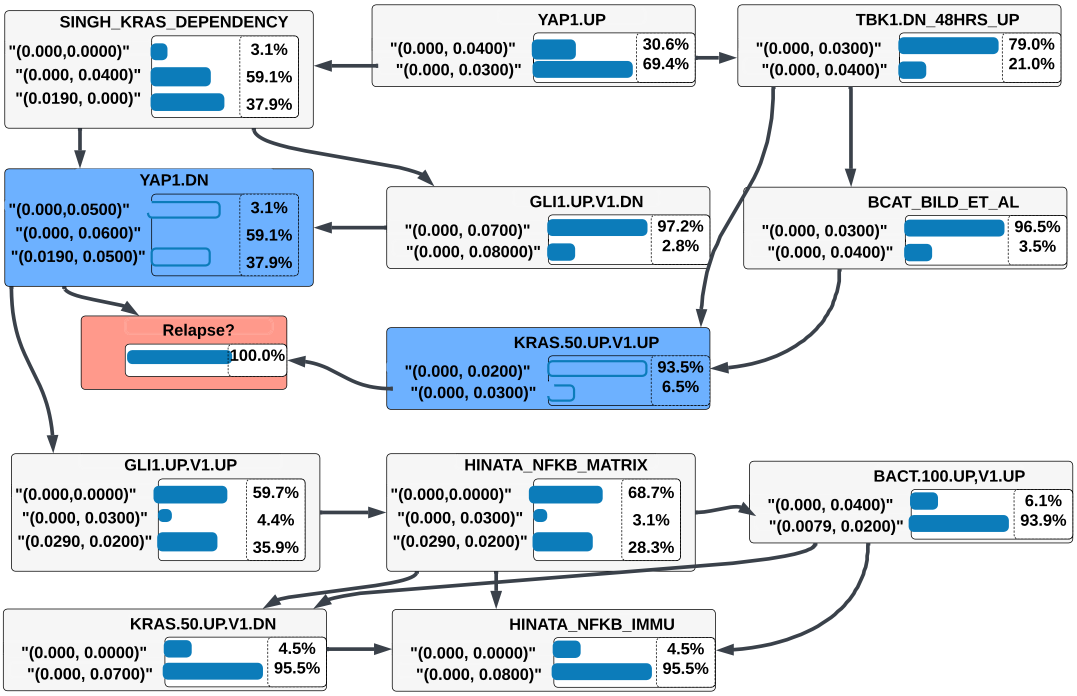


Figure 4.7: Gene sets associated with the disease prognosis in ALL10 patient using z-cube anomaly scores.

the development and progression of cancer, particularly in the colon, pancreas, lung, and blood plasma [456].

Figure 4.7 shows that the gene sets KRAS.50_UP, and YAP1_DN are associated with the disease prognosis of patient ALL10 using z-cubic anomaly score. YAP plays a role in the development of cell proliferation, apoptosis evasion, and stem cell proliferation by altering gene expression [457].

4.4.2 Instance-based LINDA-BN for ALL100 and ALL123 leukaemia relapse patients

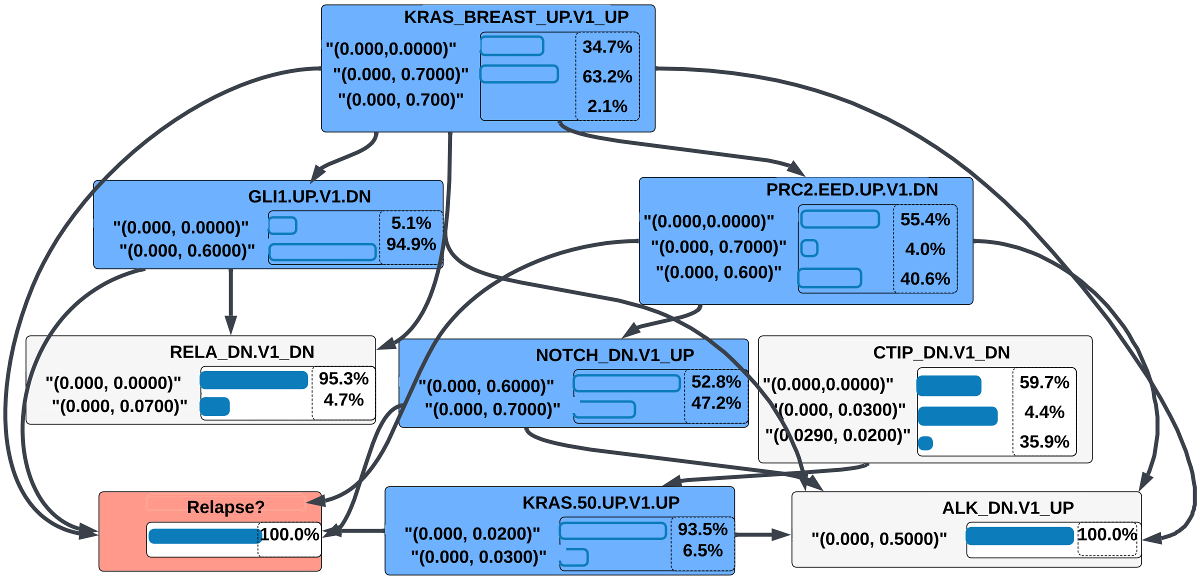


Figure 4.8: Gene sets associated with a disease prediction in ALL100 patient using z-absolute z-cube anomaly scores.

The gene sets KRAS, GLI, RELA, and NOTCH are correlated with a prediction of relapse in ALL100, as shown in Figure 4.8. This association was found using the z-absolute anomaly score. DNA sequence alterations in KRAS cause colorectal cancer [458].

Figure 4.9 shows the results of a prediction using the z-square anomaly score for ALL100 relapse patient. SING KRAS and GLI are associated with prediction of relapse. Changes in GLI gene expression cause cell cancer, medulloblastoma and sarcoma [459].

Figure 4.10 shows outcomes of a prediction using the z-cube anomaly score for ALL100 relapse patient. BACT and YAP these two gene sets associated with relapse prediction.

GLI and KRAS gene sets are common among all three anomaly scores.

Figure 4.11 shows outcomes of a prediction using the z-absolute anomaly score for ALL123 relapse patient. BACT and YAP1 these two gene sets associated with relapse prediction and BCAT gene set responsible for kidney disease [460].

SING, KRAS, and BCAT are associated with prediction of ALL123 relapse, as shown in Figure 4.12 using z-square anomaly score. SING causes lung and pancreatic cancer [461].

Figure 4.13 shows the associated gene sets for predicting ALL123 patients using z-cube anomaly scores. Here, SING, YAP, KRAS and GLI are associated gene sets for this relapse

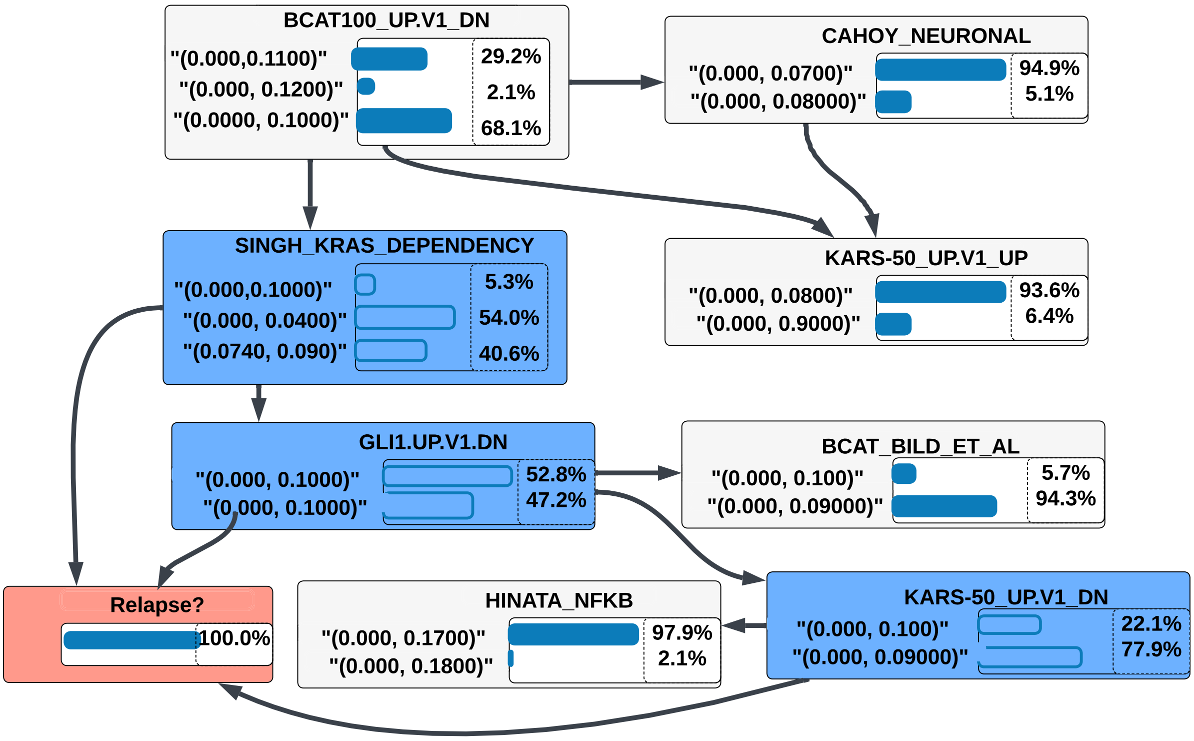


Figure 4.9: Gene sets associated with a disease prediction in ALL100 patient using z-square anomaly score.

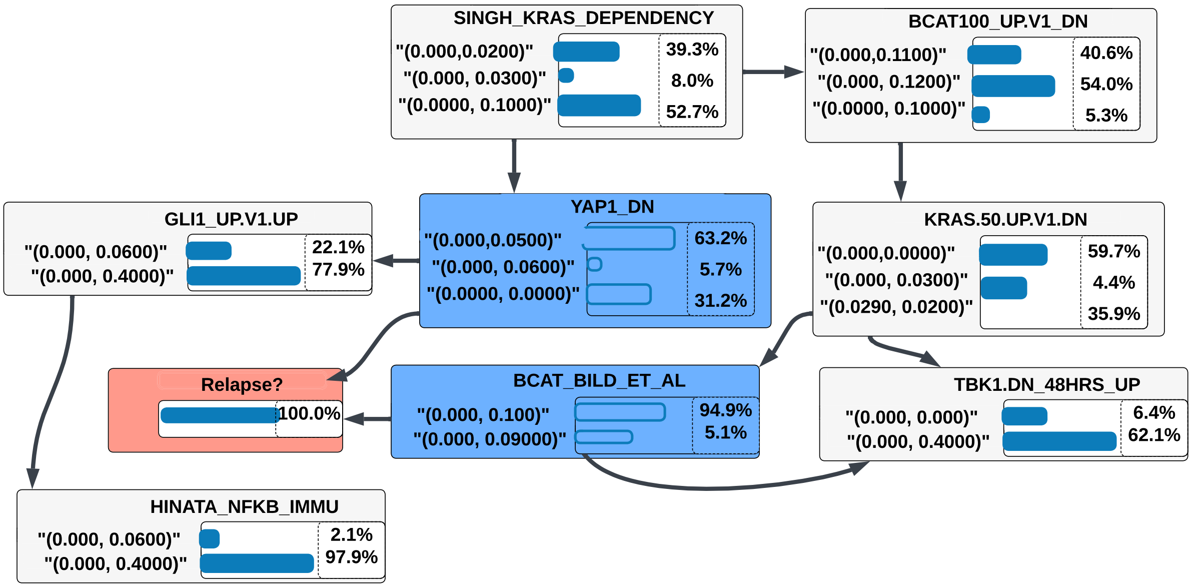


Figure 4.10: Gene sets associated with a disease prediction in ALL100 patient using z-cube anomaly scores.

prediction.

From the prediction of ALL123 patient, GLI, YAP, KRAS and SING are common gene sets for ALL123.

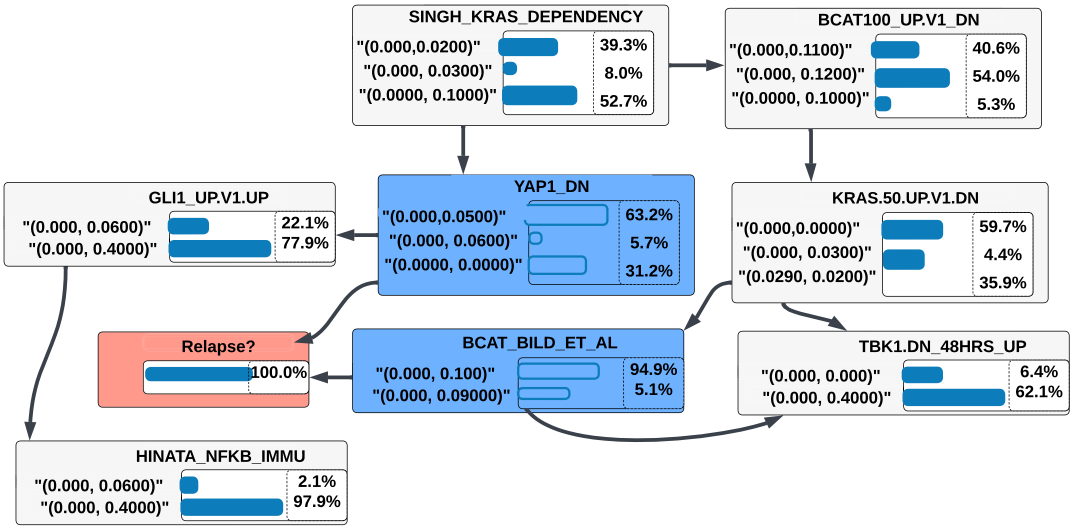


Figure 4.11: Gene sets associated with a disease prognosis in ALL123 patient using z-absolute anomaly scores.

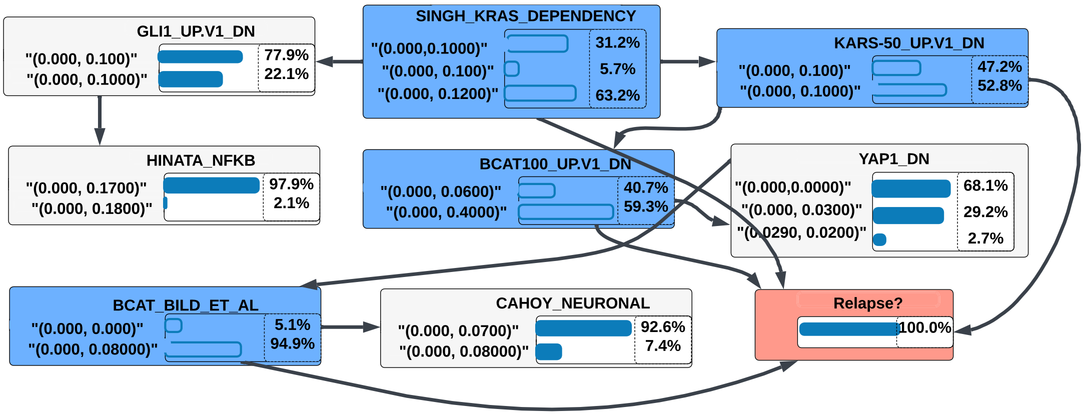


Figure 4.12: Gene sets associated with the disease prognosis in ALL123 patient using z-square anomaly scores.

4.4.3 Instance-based LINDA-BN using raw gene expression values

In this section, this thesis presents patient biology for disease prognosis using raw gene expression values. Table 4.2 the relevancy matrix for ALL10 patient with respect to raw gene expression values. MRMR and LINDA-BN are applied on the relevancy matrix to show the genes ranking and the genes associated with disease prognosis. Table 4.3 shows the MRMR gene ranking and Figure 4.14 shows the associated genes (SLC6A19 and PREP) in disease prognosis.

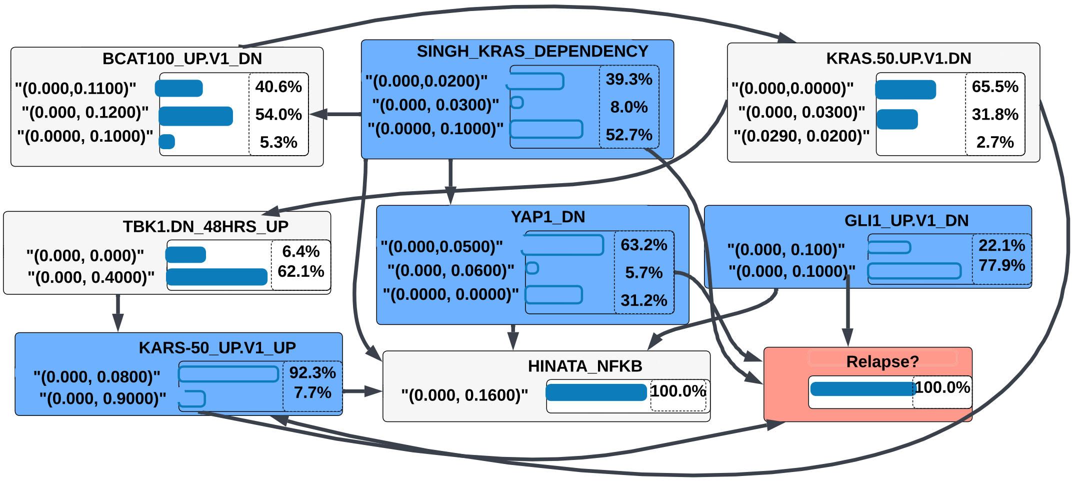


Figure 4.13: Gene sets associated with a disease prognosis in ALL123 patient using z-cube anomaly scores.

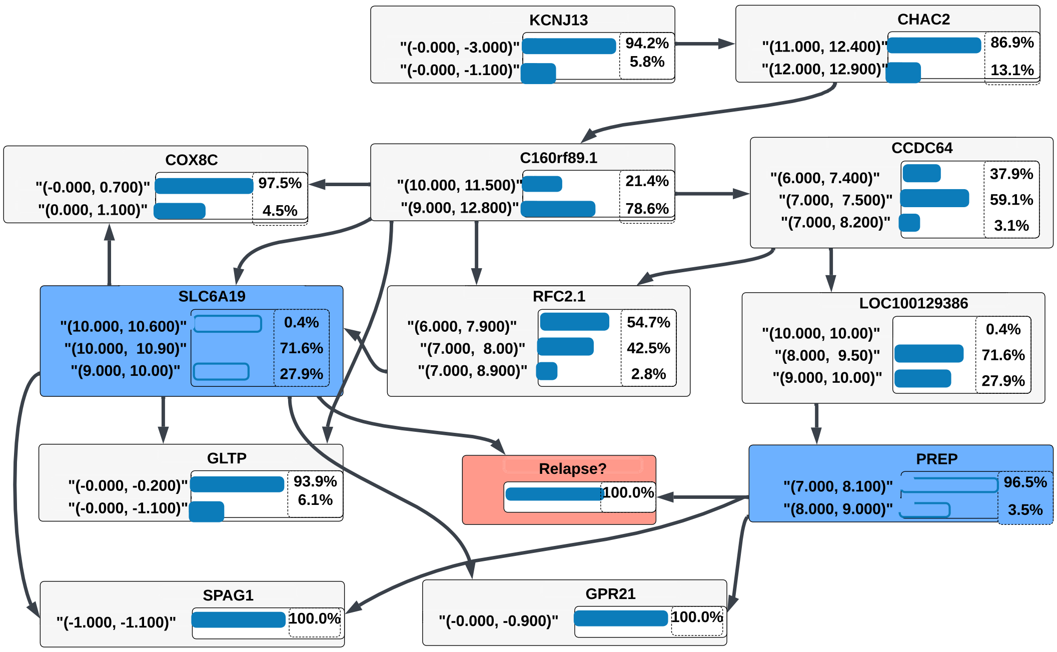


Figure 4.14: Gene associated with disease prognosis in ALL10 patient using raw gene expression values.

Table 4.2: Relevancy matrix for ALL10 patient using raw expression values.

Patient_id	CCDC62	CCDC64	HMCN2	KLHL9	RAG2	RBX1	ZNF385D	ZNF385D.1	LOC284912
ALL10	1	8.593171	7.76254	1	1	9.136168	1	11.4839	1
ALL11	7.015256	8.881025	7.885409	9.691786	10.21017	0	0	0	7.56703
ALL123	0	8.700832	7.806368	9.872934	9.562384	8.550011	9.934019	11.27889	8.185629
ALL13	0	8.674912	0	9.757159	0	0	10.41809	0	8.531556
ALL143	0	8.585872	0	9.811864	10.21606	9.128788	10.15997	11.68882	8.031329
....	....	....	....	....	....	....	....	....	....
ALL18	7.044321	8.83409	7.789628	0	10.03938	0	0	0	8.282644
ALL26	6.931048	8.470855	7.610796	0	0	8.505452	0	0	8.390844
Normal6	0	8.473055	7.895149	9.176048	9.935783	0	10.57319	11.58419	8.761325
Normal7	7.000854	8.491335	7.866489	9.016161	9.735301	0	10.43432	11.27442	0
Normal8	0	8.510356	7.715922	9.353065	0	0	0	11.05733	0

Table 4.3: MRMR top 20 genes for ALL10 patient using raw gene expression values.

KCNJ13,CHAC2,C16orf89.1,GLTP,COX8C,SLC6A19,
RFC2.1,CCDC64,LOC100129386, GPR21,SPAG1,PREP,
ZNF471, TBC1D23, KIAA1267, IRGQ.3, UBE2K, LOC388456,
CCDC101, MAPK8IP3, LOC391132, IP6K4, CHAC2.1, C19orf15.1

4.4.4 Instance-based LINDA-BN for state-of-the-art approaches

This section presented outcomes of instance-based LINDA-BN using modified gene expression scores from the state-of-the-art approaches. This is valuable for determining which genes are retrieved by by state-of-art approaches. These experiments illustrated two things: (1) showing genes and, (2) predicting a patient medical condition. The results of experiments show that state-of-the-art approaches can be used to find genes that can be used to predict a patient's medical condition. However, statistical analysis for the state-of-the-art don't support the hypothesis that an gene expression values and gene set can be integrated together to understand patient biology (gene set) responsible for anomalies in cancer. Overall, the outcomes of the state-of-the-art approaches to identifying genes and predicting a patient's condition are inconclusive.

4.4.4.1 Instance-based LINDA-BN using GFS

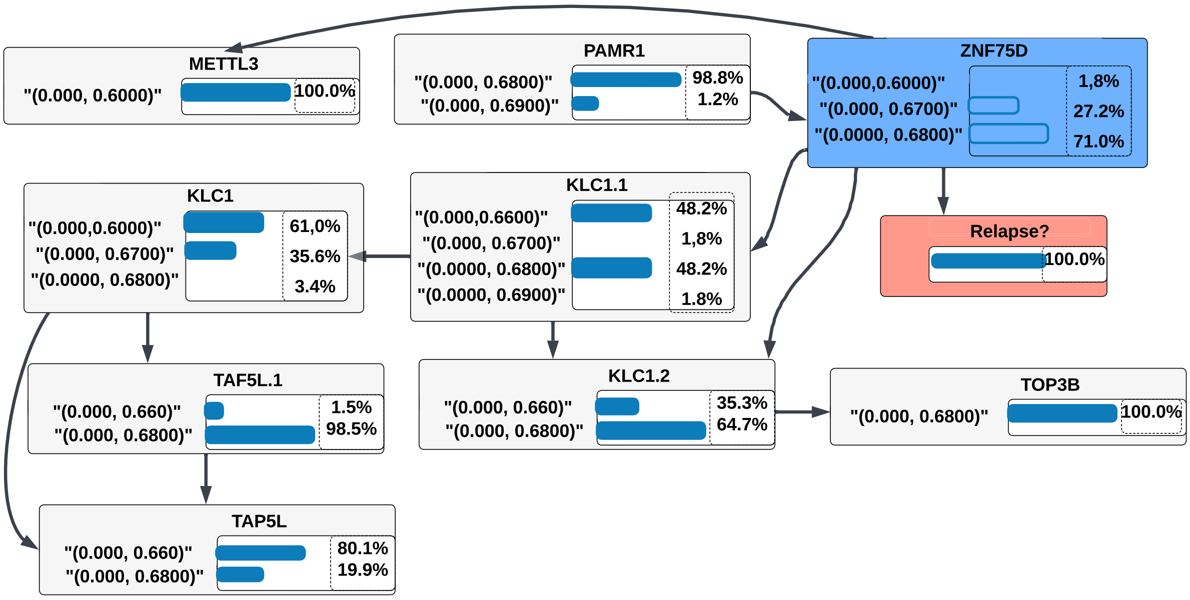


Figure 4.15: Instance-based LINDA-BN for leukaemia data sets using GFS scores.

Figure 4.15 shows an associated gene for predicting a medical condition of a test patient using GFS scores on leukaemia datasets. The plot shows that only one gene ZNF75D (zinc finger protein) is related for predicting cancer relapse in ALL10. This gene encodes a protein that is likely to have the function of a transcription factor. In addition, this gene is associated with GO annotations involving nucleic acid binding and DNA-binding transcription factor activity [462].

4.4.4.2 Instance-based LINDA-BN using FRaC

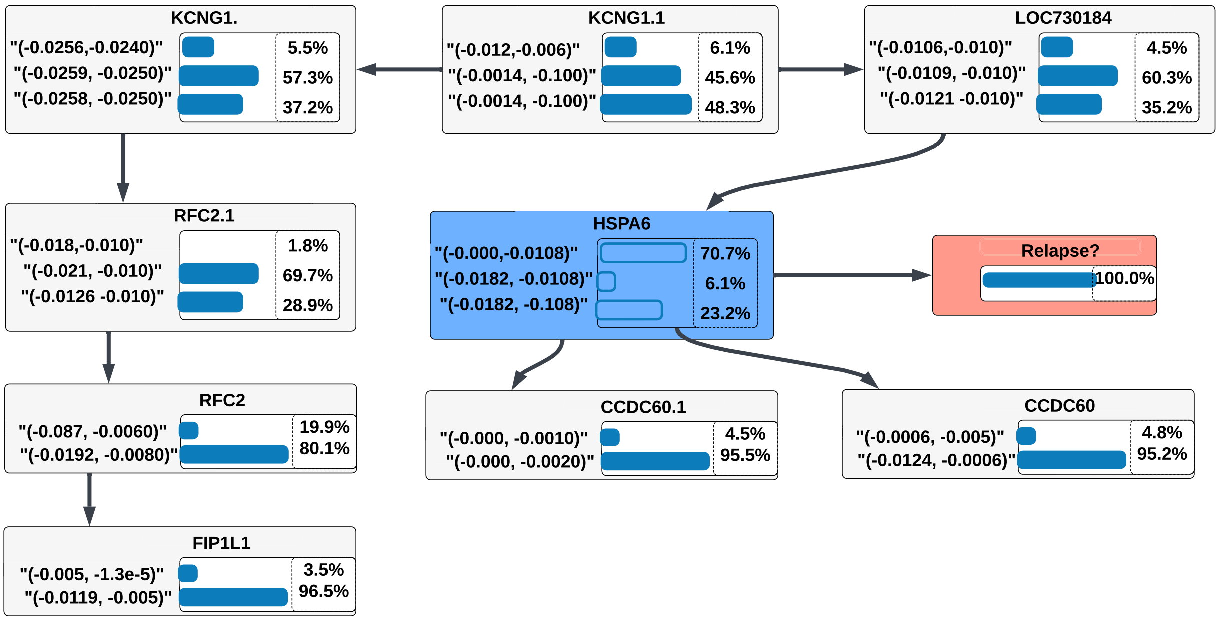


Figure 4.16: Instance-based LINDA-BN for leukaemia data sets using FRaC scores.

Figure 4.16 shows a gene called HSPA6 (heat shock protein family)) associated for predicting a medical condition for ALL10 using FRaC scores on leukaemia datasets. This gene is a protein coding gene and involved in cervical squamous cell carcinoma [463, 464].

4.4.4.3 Instance-based LINDA-BN using CSAX

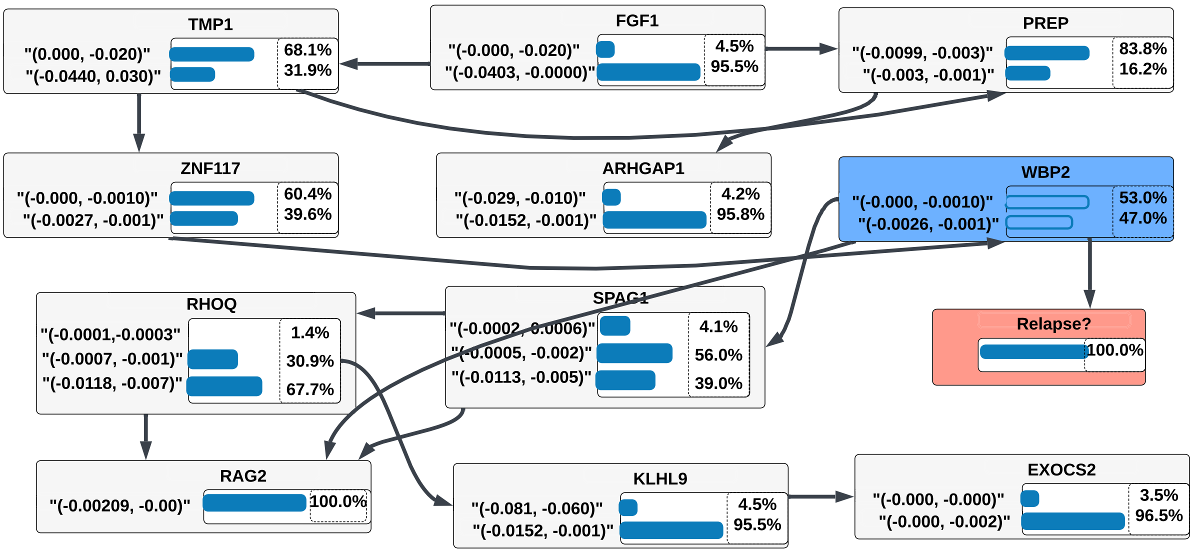


Figure 4.17: Instance-based LINDA-BN for leukaemia data sets using CSAX scores.

Figure 4.17 shows three associated genes (WBP2, SPAG1 and RHOQ) for predicting a medical condition of a test patient using CSAX scores on leukaemia datasets. WBP2, SPAG1, and RHOQ are all protein-encoding genes. WBP2 causes hearing loss, Dfnb, and autosomal recessive 107. SPAG1 is associated with diseases such as primary ciliary dyskinesia-28 and epilepsy. RHOQ influences colitis caused by clostridium difficile [465–470].

4.4.4.4 Instance-based LINDA-BN using TEMPO

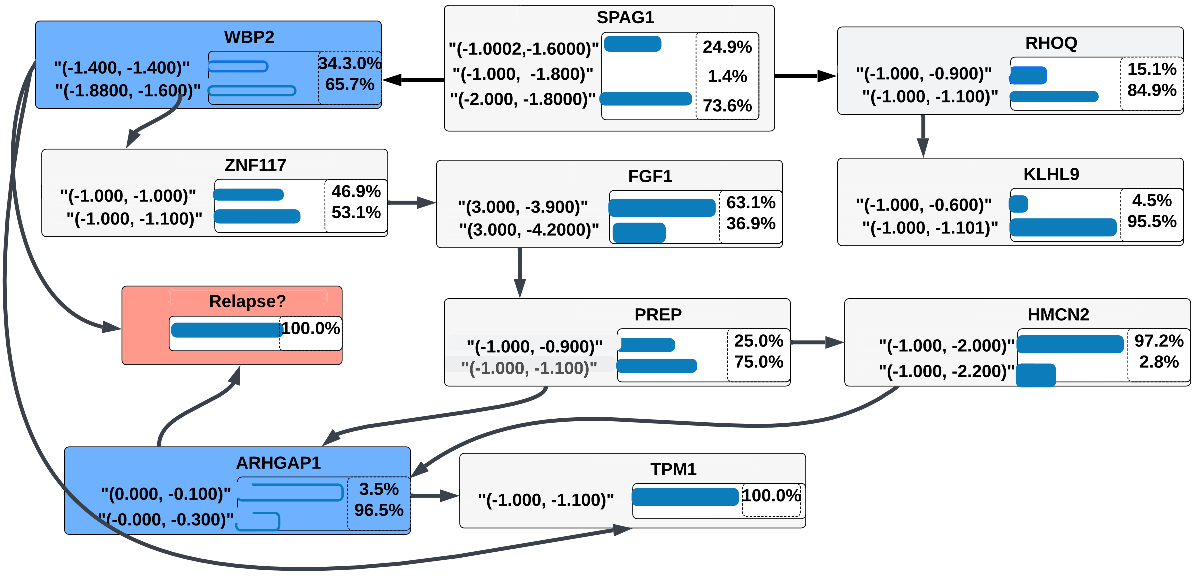


Figure 4.18: Instance-based LINDA-BN for leukaemia data sets using TEMPO scores.

Figure 4.18 shows two associated genes (WBP2 and ARHGAP1) for predicting a medical condition of a test patient using TEMPO's modified gene expression scores on leukaemia datasets. Both WBP2 and ARHGAP1 are protein-encoding genes. The gene ARHGAP1 is associated with Lowe Oculocerebrorenal Syndrome [471, 472].

4.4.4.5 Instance-based LINDA-BN using PFSNet

Figure 4.19 shows two associated genes (FETUB and ALDOB) for predicting a medical condition of a test patient using PFSNet's modified gene expression scores on leukaemia datasets. Both FETUB and ALDOB are protein-encoding genes. FETUB causes encephalitozoonosis and Caffey's disease [473, 474]. ALDOB is associated with diseases such as fructose intolerance, hereditary and fructosuria, and essential fructose [475, 476].

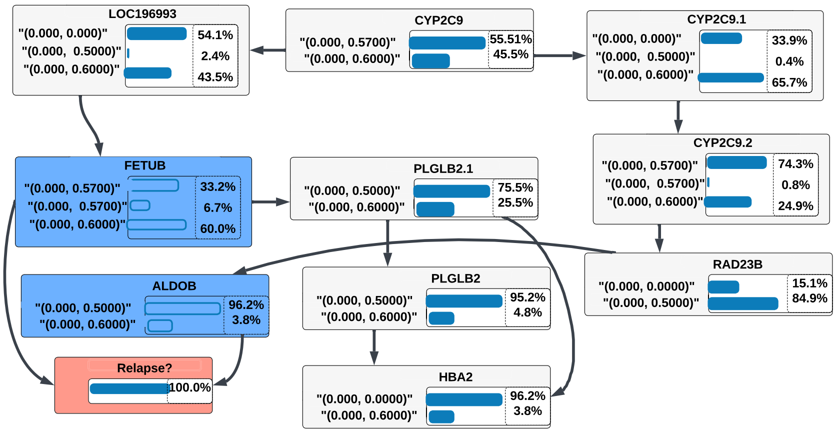


Figure 4.19: Instance-based LINDA-BN for leukaemia data sets using PFSNeT scores.

4.4.4.6 Instance-based LINDA-BN using SNets

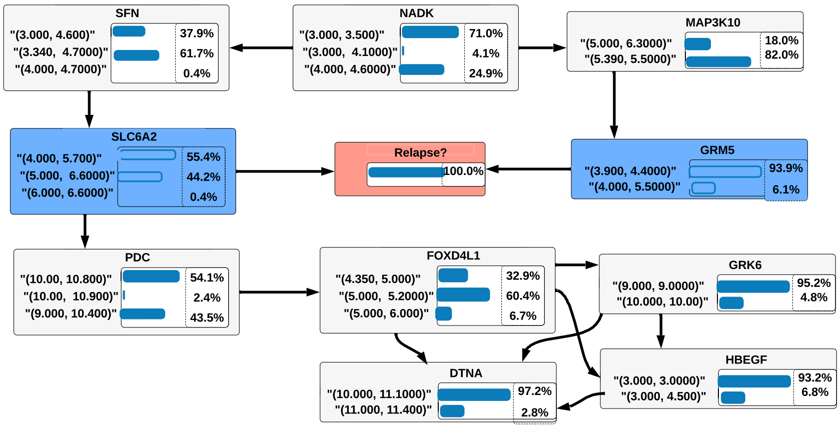


Figure 4.20: Instance-based LINDA-BN for leukaemia data sets using SNet scores.

Figure 4.20 shows two associated genes (GRM5 and SLC6A2) for predicting a medical condition of a test patient using SNets scores on leukaemia datasets. Both are protein-encoding genes. GRM5 is related to fragile X syndrome and central nervous system disorders [477, 478]. SLC6A2 is related to orthostatic intolerance and syncope [479, 480].

4.4.4.7 Instance-based LINDA-BN using qpFS

Figure 4.21 shows a gene (RAD1) is associated for predicting a medical condition of a test patient using qpFS scores on leukaemia datasets. The RAD1 gene is associated with mantle cell lymphoma disease [481, 482].

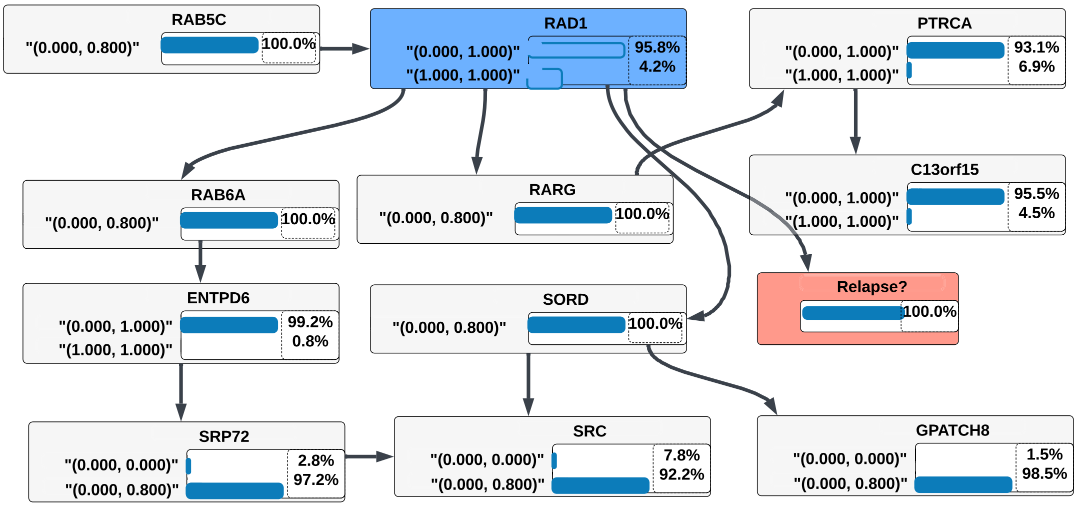


Figure 4.21: Instance-based LINDA-BN for leukaemia data sets using qpFS scores.

4.4.4.8 Instance-based LINDA-BN using outlier dection

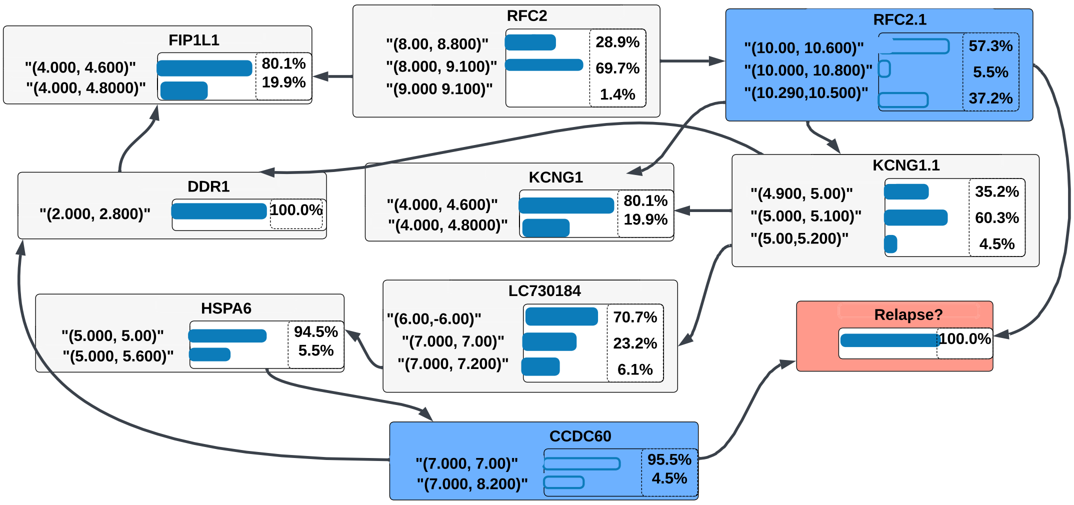


Figure 4.22: Instance-based LINDA-BN for leukaemia data sets using outlier detection scores.

Figure 4.22 shows two associated genes (RFC2 and CCDC60) for predicting a medical condition of a test patient using outlier scores on leukaemia datasets. Both are protein-encoding genes. RFC2 causes Williams-Beuren and Seckel syndromes [483, 484]. CCDC60 is related to both neuronitis and congenital muscular dystrophy-dystroglycanopathy type A6 disorder [485, 486].

4.4.5 Comparative analysis on multiple acute myeloid leukaemia (AML) datasets

The goal of this experiment was to use MRMR and LINDA-BN to investigate patient biology across multiple datasets. Specifically, this thesis investigated whether MRMR and LINDA- BN can generate similar results from multiple datasets of acute myeloid leukemia. The results confirmed that MRMR and LINDA-BN exhibit similar patient biologies across multiple datasets.

Three separate gene expression datasets were used for acute myeloid leukemia (AML), namely GSE14468, GSE12417-GPL570, and GSE24006. The GSE14468 gene expression data was from AML cancer patients, GSE12417-GPLS570 gene expression data was from healthy patients, and GSE24006 gene expression data was from both AML cancer patients and healthy patients. Salunkhe *et al.* [487] used two microarray data sets by considering acute myeloid leukaemia (AML) and healthy patients. These microarray data were publicly available in NCBI [488, 489] and were used to find gene pairs between AML and normal patients. There were 598 gene expression profiles for AML data (accession number GSE14468). Among the 598 samples, there are two mutations (CEBPAdouble-mut, CEBPAsingle-mut). There was cytogenetically normal acute myeloid leukaemia (CN-AML) to identify the gene pairs with AML data (accession number GSE12417-GPL570). These CN-AML microarray data had 163 gene expression profiles with gene expression values treated as healthy AML patients. The patient biology (gene sets) was determined from all three data sets. This thesis considered the top 5% gene sets.

Table 4.4 details the similar and dis-similar gene sets from the GSE14468 and GSE24006 data sets.

Table 4.4: Top 5% of MRMR ranked similar and dis-similar gene sets from the GSE14468 and GSE24006 data sets.

GSE14468	GSE24006
CAHOY_OLIGODENDROCUTIC	CAHOY_OLIGODENDROCUTIC
PDGF_ERK_DN.V1_UP	PDGF_ERK_DN.V1_UP
RB_P130_DN.V1_UP	RB_P130_DN.V1_UP
RB_DN.V1_UP	RB_DN.V1_UP
ALK_DN.V1_DN	ALK_DN.V1_DN
IL21_UP.V1_DN	IL21_UP.V1_DN
GLI1_UP.V1_DN	MEL18_DN.V1_DN
RPS14_DN.V1_DN	KRAS.600.LUNG.BREAST_UP.V1_UP
RB_DN.V1_UP	CRX_DN.V1_UP
GLI1_UP.V1_DN	PRC2_EZH2_UP.V1_UP
PDGF_ERK_DN.V1_DN	PRC2_SUZ12_UP.V1_UP
ESC_J1_UP_LATE.V1_UP	KRAS.PROSTATE_UP.V1_DN
PDGF_UP.V1_DN	CTIP_DN.V1_DN
RB_DN.V1_DN	KRAS.600.LUNG.BREAST_UP.V1_DN
NRL_DN.V1_DN	ATM_DN.V1_DN
ALK_DN.V1_DN	P53_DN.V2_UP

Table 4.5 details the similar and dis-similar gene sets from the GSE12417 and GSE24006 data sets.

Table 4.5: Top 5% of MRMR ranked similar and dis-similar gene sets from GSE12417 and GSE24006 datasets.

GSE12417	GSE24006
KRAS.600.LUNG.BREAST_UP.V1_UP	KRAS.600.LUNG.BREAST_UP.V1_UP
IL21_UP.V1_DN	IL21_UP.V1_DN
PRC2_EZH2_UP.V1_UP	PRC2_EZH2_UP.V1_UP
MEL18_DN.V1_DN	MEL18_DN.V1_DN
KRAS.PROSTATE_UP.V1_DN	KRAS.PROSTATE_UP.V1_DN
CTIP_DN.V1_DN	CTIP_DN.V1_DN
KRAS.600.LUNG.BREAST_UP.V1_DN	KRAS.600.LUNG.BREAST_UP.V1_DN
ATM_DN.V1_DN	ATM_DN.V1_DN
P53_DN.V2_UP	P53_DN.V2_UP
ATM_DN.V1_DN	CAHOY_OLIGODENDROCUTIC
PTEN_DN.V1_UP	PDGF_ERK_DN.V1_UP
DCA_UP.V1_DN	RB_P130_DN.V1_UP
KRAS.600_UP.V1_DN	RB_DN.V1_UP
PRC1_BMI_UP.V1_UP	ALK_DN.V1_DN
WNT_UP.V1_UP	IL21_UP.V1_DN
NOTCH_DN.V1_DN	MEL18_DN.V1_DN

Figure 4.23 shows the associated gene sets for a test patient using the AML data sets. The plot shows that two gene sets (RB_DN.V1_DN and CAHOY_OLIGODENDROCUTIC)) are responsible for an AML disease prognosis. The RB gene set plays a role in increased proliferation and abnormal differentiation [490]. Oligodendroglioma is a type of primary tumor that affects the central nervous system (CNS) [491].

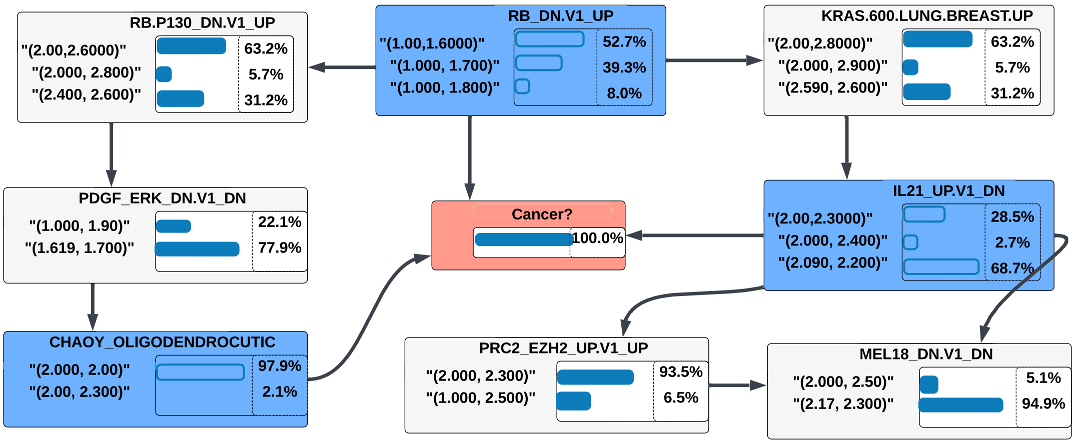


Figure 4.23: Genes associated with the disease prognosis in a test of an acute myeloid leukemia (AML) patient using the z-absolute anomaly score.

Figure 4.24 shows the associated gene sets for test patients using the AML datasets. The plot shows that two gene sets (RB_P130_DN.V1_UP and RB_DN.V1_UP) are responsible for a healthy AML disease prognosis.

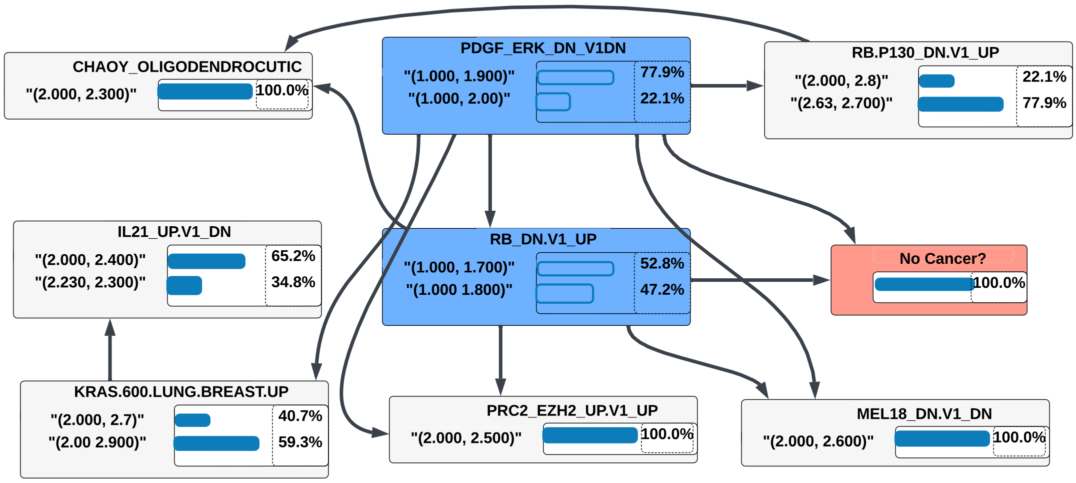


Figure 4.24: Genes associated with the disease prognosis in a test of a cytogenetically normal acute myeloid leukemia (CN-AML) patient using the z-absolute anomaly score.

4.4.6 Instance-based LINDA-BN for colon, adrenal and breast cancer

This section of the thesis provides a summary and discussion of the findings from the application of instance-based LINDA-BN to the colon, breast, and adrenal cancer datasets. Careful investigation shows that LINDA-BN provides graphical relationships between gene sets to predict a patient's medical condition. The experimental findings demonstrate that the approach is useful for identifying related gene sets for prediction goals. Only gene sets that are conditionally dependent on each other are used for prediction in LINDA-BN. In this thesis, the gene sets related to prediction have been highlighted with blue color. The remaining gene sets are not relevant for prediction due to their conditional independence.

Table 4.6 for colon, breast, and adrenal cancer datasets for there different patients such as GSM95473, GSM65316, and GSM277090

Table 4.6: MRMR gene set ranking for colon, breast and adrenal cancer for three test patients (colon cancer: GSM95473, breast cancer: GSM65316 and adrenal cancer: GSM277090).

Colon	Breast	Adrenal
YAP1_UP	YAP1_UP	GLI1_UP.V1_UP
BCAT_GDS748_DN	BCAT_GDS748_DN	GCNP_SHH_UP_EARLY.V1_DN
TBK1.DN.48HRS_DN	TBK1.DN.48HRS_DN	GCNP_SHH_UP_LATE.V1_DN
YAP1_DN	YAP1_DN	RAPA_EARLY_UP.V1_UP
GCNP_SHH_UP_LATE.V1_DN	GLI1_UP.V1_UP	HINATA_NFKB_MATRIX
CYCLIN_D1_KE_.V1_DN	RAPA_EARLY_UP.V1_UP	CYCLIN_D1_KE_.V1_DN
HINATA_NFKB_MATRIX	GCNP_SHH_UP_EARLY.V1_DN	GCNP_SHH_UP_LATE.V1_UP
AKT_UP.V1_DN	CAHOY_NEURONAL	RAPA_EARLY_UP.V1_DN
PKCA_DN.V1_DN	CRX_NRL_DN.V1_DN	HINATA_NFKB_IMMU_INF
CAHOY_ASTROGLIAL	KRAS.LUNG_UP.V1_DN	GCNP_SHH_UP_EARLY.V1_UP
BRCA1_DN.V1_UP	SIRNA_EIF4GI_UP	GLI1_UP.V1_DN
CTIP_DN.V1_UP	CRX_DN.V1_DN	E2F1_UP.V1_DN
ESC_V6.5_UP_EARLY.V1_DN	CAMP_UP.V1_DN	E2F1_UP.V1_UP
SIRNA_EIF4GI_DN	CORDENONSI_YAP_CONSERVED	EGFR_UP.V1_DN
NRL_DN.V1_DN	NFE2L2.V2	EGFR_UP.V1_UP
RB_DN.V1_DN	TBK1.DN.48HRS_UP	ERBB2_UP.V1_DN
CRX_NRL_DN.V1_UP	IL2_UP.V1_DN	ERBB2_UP.V1_UP
PTEN_DN.V1_DN	DCA_UP.V1_DN	CYCLIN_D1_KE_.V1_UP
AKT_UP.V1_UP	PRC2_SUZ12_UP.V1_UP	CYCLIN_D1_UP.V1_DN
MTOR_UP.V1_DN	ESC_J1_UP_LATE.V1_UP	CYCLIN_D1_UP.V1_UP

4.4.6.1 Instance-based LINDA-BN for colon cancer

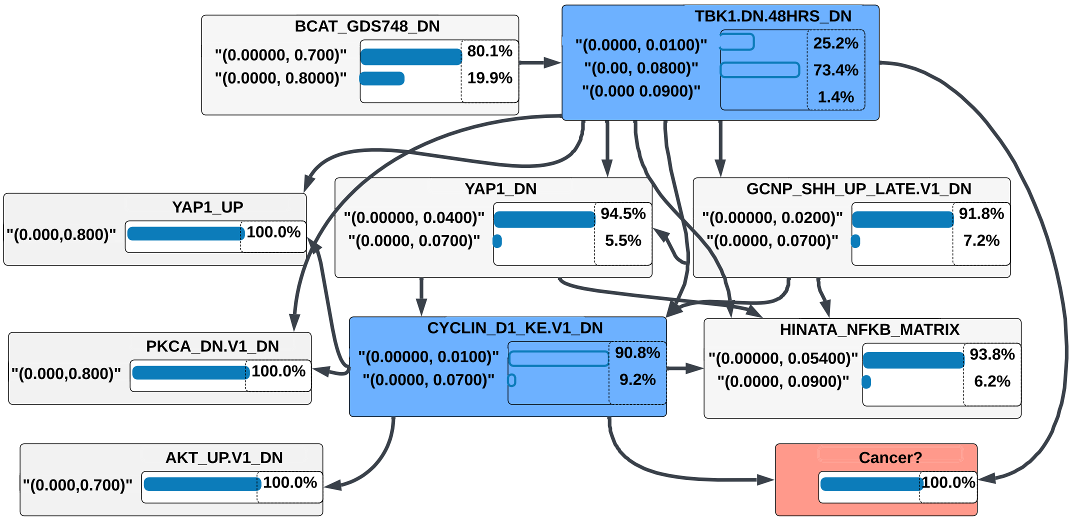


Figure 4.25: Gene sets associated with disease prognosis of colon cancer patient GSM95473 using z-absolute anomaly scores.

Figure 4.25 associated with the diagnosis of colon cancer in patient GSM95473 with respect to z-absolute anomaly scores. The results from LINDA-BN suggest that two gene sets (TBK1_DN, and CYCLIN_D1) are associated with the likelihood of cancer in patient GSM95473. Based on the figure, this thesis investigates the hypothesis that these gene sets (biological functionalities) are responsible for the cancer prognosis of colon cancer patient GM95473 and a similar analysis is conducted for each test patient.

4.4.6.2 Instance-based LINDA-BN for breast cancer

Figure 4.26 shows the gene sets associated with the breast cancer prognosis of patient GSM65316 with respect to z-absolute anomaly scores. The results from LINDA-BN suggest that three gene sets (GLI1_UP, GCNP_SHH and BACT_GDS748) are associated with the likelihood of cancer in patient GSM65316. Based on the figure, this thesis hypothesizes that these gene sets (biological functionalities) are responsible for the cancer prognosis of breast cancer patient GSM65316and a similar analysis is conducted for each test patient.

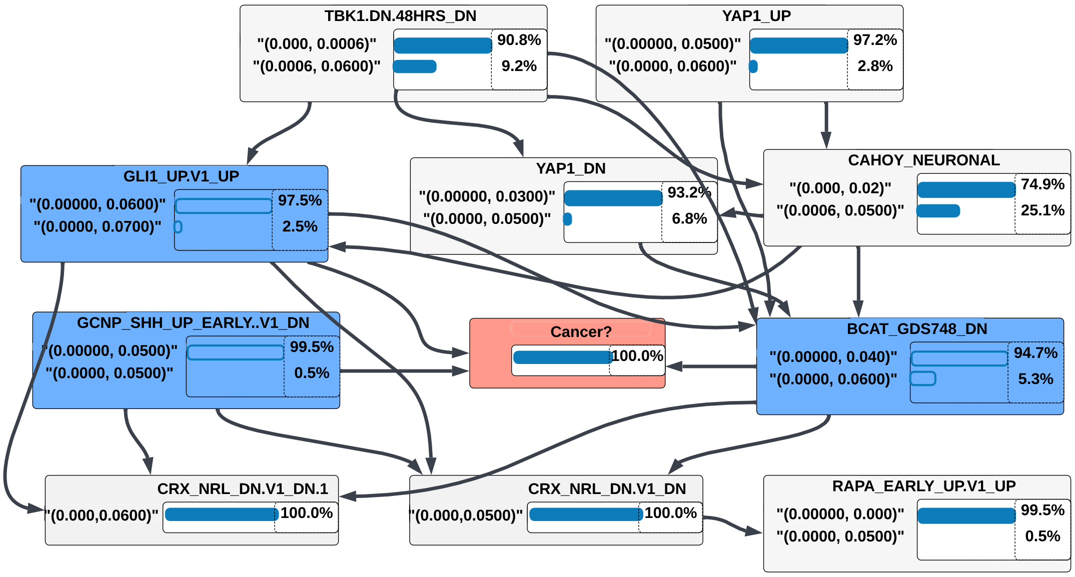


Figure 4.26: Gene sets associated with the disease prognosis of breast cancer patient GSM65316 using z-absolute anomaly scores.

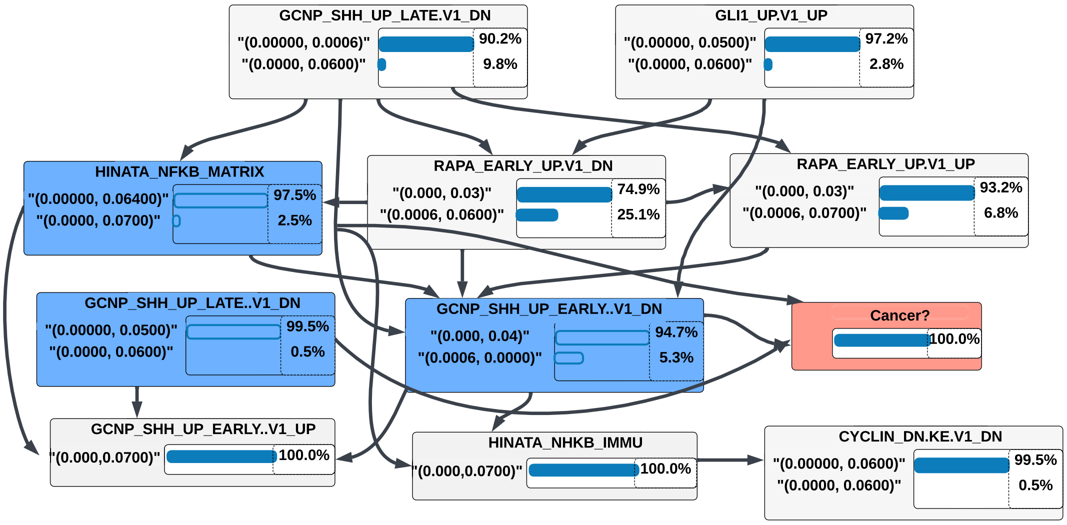


Figure 4.27: Gene sets associated with the disease prognosis of adrenal cancer patient GSM277090 using z-absolute anomaly scores.

4.4.6.3 Instance-based LINDA-BN for adrenal cancer

Figure 4.27 shows the gene sets associated with the prognosis of adrenal cancer for patient GSM277090 with respect to z-absolute anomaly scores. The results from LINDA-BN suggest that two gene sets (HINATA_NFKB, GCNP_SHH_UP.EARLY and GCNP_SHH_UP.LATE) are associated with the likelihood of cancer in patient GSM277090. Based on the figure, this thesis hypothesizes that these gene sets (biological functionalities) are responsible for the cancer prognosis of adrenal cancer patient GSM277090 and a similar analysis is conducted for each test patient.

4.4.7 Outcomes from AI approaches

Table 4.7 shows prediction accuracy of k-nearest neighbor (K-NN), support vector classifier (SVC), random forest (RF), AdaBoost, and XGBoost. All of these approaches performed well while predicting a test patient.

Table 4.7: Prediction accuracy of K-NN, SVC, RF, AdaBoost, and XGBoost.

Classifier	Accuracy	Precission	Recall	F-measure
K-NN	71%	37%	49%	42%
SVC	70%	57%	56%	57%
RF	72%	63%	54%	53%
AdaBoost	71%	67%	58%	58%
XGBoost	73%	63%	59%	60%

4.5 Discussion

Instance-based LINDA-BN is a probabilistic graphical model which is a graphical representation of probabilistic distributions. Bayesian networks and Hidden Markov Models (HMMs) are notable probabilistic graphical models. The structure of a Bayesian network, which is a directed acyclic graph (DAG), shows the conditional independence of a set of random variables. HMMs are used to represent Markov processes. Instance-based LINDA-BN is a combination of Bayesian network and HMMs.

This thesis explores the use of instance-based LINDA-BN, an approach that improves interpretability to understand the decision-making process behind predicting a patient's medical condition. The focus is on the method's ability to analyze the implications of various outcomes and provide a clear understanding of the process involved in making these predictions. The overall aim is to streamline the decision- making process making it easier to foresee the prediction of a patient's medical condition

The primary objective of this research is to understand the role of patient biology, particularly gene sets, in predicting a patient's medical condition. This aspect is crucial as it allows for the identification of specific gene sets that may be associated with a particular disease. By focusing on these gene sets, it is hoped that the biological mechanisms driving disease progression in individual patients, thereby improving the ability to predict and manage specific medical conditions.

This thesis bulit on the hypothesis that a careful integration of gene expression values into gene sets could facilitate understanding anomalies in cancer. In other words, the thesis argues that thoughtful and calculated synthesis of these gene expression values could help unravel the anomalies that characterize cancer. This process, guided by instance-based LINDA-BN, could help to identify the biology of the patient in the context of a disease.

However, due to the lack of wet lab validation capabilities, this research had to rely on analyzing the functions of these gene sets from the available literature. The results showed that the associated gene sets are responsible for various forms of cancer. This expands the understanding of the role of gene sets in disease prediction and provides valuable insight into the biological factors that may influence a patient's medical condition. This research therefore has the potential to make a valuable contribution to personalized medicine and disease analysis.

In contrast to the instance-based LINDA-BN, this thesis also experimented with existing methods such as K-NN, SVC, RF, AdaBoost, and XGBoost for predicting medical conditions. These methods are referred to as "black box" models due to their lack of interpretability and explainability. While these models are often very good at making correct predictions, they are inadequate when it comes to explaining the reasoning behind their decisions. In other words, they can predict the likely outcome, but their decision-making processes are complex so a clear explanation of the reasons or process that led to that particular outcome is difficult to understand. This is a major limitation, especially in the medical field, where understanding the rationale behind a diagnosis or prediction is as valuable as the prediction itself.

When we compare black-box models with our approach in this thesis, it becomes clear that black-box models have some limitations. We use a method called instance-based LINDA-BN, which not only gives us trustworthy predictions, but also allows to better understanding of the decision-making process. This is possible because it identifies and shows how it reached the decision to select gene sets associated with the medical condition.

If we put black-box models and instance-based LINDA-BN side by side, both might be able to predict accurately, but black-box models cannot provide the same level of understanding that LINDA-BN does. This makes LINDA-BN a better choice, especially in a medical field.

4.6 Conclusion

Black-box approaches predict medical condition of a patient with higher accuracy. However, there is limited scope to know which gene sets or gene are associated with the prediction process.

A patient could have thousands of genes. So, it would be more useful to know sets of genes that are associated with a patient's medical condition than to know just one or two genes. Comparative outcomes between anomaly score and state-of-the-art approaches show that instance-based LINDA-BN for proposed anomaly score shows gene sets for predicting a medical condition. On the other hand, state-of-the-art approaches show only one or two genes.

4.7 Presenting new cancer knowledge

Gene expression profiles can be analyzed by aggregating gene expression levels into gene sets. Chapter 3 described that gene set anomaly scores make patterns across expression profiles more visible and detectable than raw gene expression values alone. Gene set anomaly scores improve the extraction of insights from gene expression data. Using gene sets brings a knowledge-driven aspect that can be combined with data-driven analysis. In this research, multiple methods are applied to investigate new cancer biology using anomaly scores.

4.7.1 Gene set selection

To rank the gene sets, MRMR and RF are applied to the relevance matrix for four anomaly score variants, namely the z-absolute anomaly score, the z-square anomaly score, the z-cubic anomaly score, and the mid-range anomaly score. Table 4.1 shows the ranking of the gene sets for leukaemia patients in relation to z-absolute, z-square, and z-cubic anomaly scores. The color blue indicates gene sets that are similar across all three anomaly scores.

4.7.2 Heatmap for analyzing gene ontology terms

A heatmap is a specific kind of data visualization in which the frequency of an occurrence is shown as a color scale that extends over two dimensions [492]. A heatmap depicts the associations between the gene ontology words (biological names) of gene sets that have been found using the proposed methods as well as other state-of-the-art approaches.

The following considerations were taken into account when developing the heatmap for this experiment:

1. MRMR uses anomaly scores to determine the ranking of gene sets for leukaemia data sets in relation to cancer patients. MRMR is implemented by measuring the mutual information of cancer gene sets for leukaemia patients.

2. GSEA is used to identify gene sets by including 4,800 genes for each of the state-of-the-art approaches. Once the gene sets have been extracted using GSEA, the next step is to select the top 10 gene sets based on their enrichment values.

3. The GO terms are examined for both the proposed method and the current state-of-the-art for each gene set. The heatmap shows the GO corresponding to each gene set for the proposed method and the state-of-the-art.

4.8 Synthesis

A data processing framework is developed to identify GO terms regarding cancer by utilising anomaly scores, as shown in Figure 4.28. This process is broken down into three phases: the first phase involves selecting the top gene sets based on anomaly scores using MRMR and RF; the second phase involves generating the top gene sets using gene set enrichment analysis (GSEA) and state-of-the-art approaches; and the final phase involves producing heat maps based on anomaly scores and state-of-the-art approaches.

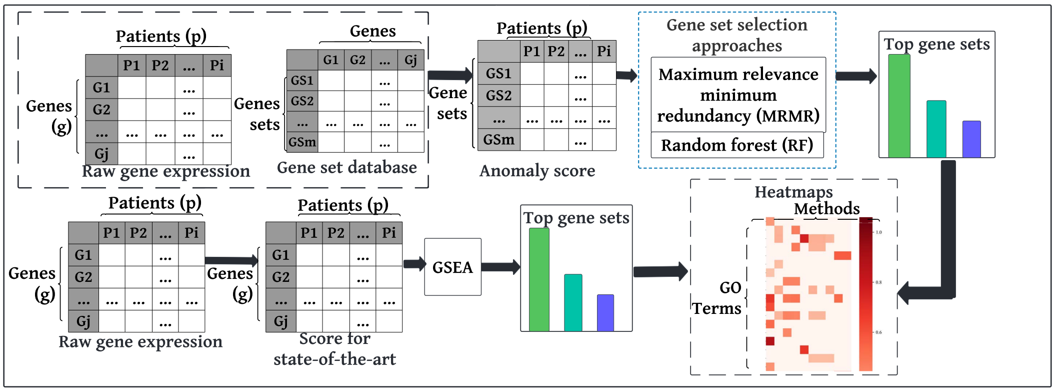


Figure 4.28: Schematic diagram showing methods representing new cancer biology.

4.9 Results

4.9.1 Gene sets ranking

The effectiveness of the anomaly scores is measured by conducting an experiment with colon cancer, breast cancer, and adrenal cancer data. Specifically, the gene sets are ranked for pre-cancerous and cancerous colon cancer lesions using random forest and MRMR assessment of anomaly scores and are compared to GSEA for selected genes.

Table 4.8 shows the top ten gene sets for the leukaemia datasets. The left column lists the gene sets and the right column lists the associated genes for a given gene set. From the table, it can be seen that gene sets identified by anomaly scores ranking are distinct from those generated by applying GSEA.

Table 4.9 shows the top ten gene sets for colon cancer. From the table, it can be seen that gene sets identified using anomaly scores ranking are different from those generated by GSEA.

Table 4.10 shows the top ten gene sets for breast cancer datasets. Again for this datasets, the table shows that the anomaly scores ranking identified gene sets that are different from those extracted using GSEA.

And finally, table 4.11 shows the top ten gene sets adrenal cancer datasets. From the table, it can be seen that that anomaly scores ranking identified gene sets that are different from those extracted using GSEA.

Table 4.8: Top 10 gene sets identifying leukaemia patients who relapsed using random forest, MRMR and GSEA.

(a) Gene set ranking using random forest

Gene Sets	Gene Names
GOMF_PROTEIN_ARGININE_OMEGA_N_MONOMETHYLTRANSFERASE_ACTIVITY URL	PRMT1, PRMT7, PRMT6, PRMT8, PRMT9
GOMF_QUATERNARY_AMMONIUM_GROUP_TRANSMEMBRANE_TRANSPORTER_ACTIVITY URL	SLC25A29, SLC25A48, SLC25A45, SLC25A47,SLC22A1, SLC22A3, SLC22A4, SLC22A5, SLC25A20,SLC22A16
GOMF_G_QUADRUPLEX_RNA_BINDING URL	MCRS1, DHX36, DHX30, FMR1, AFF2, XRN1, LIN28A
GOMF_HEPARAN_SULFATE_GLUCOSAMINE_3_SULFOTRANSFERASE_1_ACTIVITY URL	HS3ST5, HS3ST6, HS3ST4,HS3ST3B1, HS3ST3A1,HS3ST2, HS3ST1
GOMF_L_LYSINE_TRANSMEMBRANE_TRANSPORTER_ACTIVITY URL	SLC25A29, PQLC2, SLC7A1, SLC7A2, SLC7A3
GOMF_ARGININE_TRANSMEMBRANE_TRANSPORTER_ACTIVITY URL	SLC25A29, SLC38A9, PQLC2, SLC7A1, SLC7A2,SLC7A3
GOMF_NAADP_SENSITIVE_CALCIUM_RELEASE_CHANNEL_ACTIVITY URL	TPCN2, MCOLN2, TPCN1, MCOLN3, MCOLN1
GOMF_CROSSOVER_JUNCTION_ENDODEOXYRIBONUCLEASE_ACTIVITY URL	EME1, EME2, GEN1, SLX1A, RAD51C, XRCC3, SLX1B, MUS81, SLX4
GOMF_LIPID_KINASE_ACTIVITY URL	PIK3C2B, AGK, SPHK2, CERK, SPHK1
GOMF_CERAMIDE_1_PHOSPHATE_BINDING URL	COL4A3BP, GLTPD2, PLEKHA8P1, GLTP, PLEKHA3,CPTP, PLEKHA8

(b) Gene set ranking using MRMR

Gene Sets	Gene Names
GOMF_ACYL_CARNITINE_TRANSMEMBRANE_TRANSPORTER_ACTIVITY URL	SLC25A29, SLC25A48, SLC25A45, SLC25A47, SLC25A20
GOMF_ALPHA_AMYLASE_ACTIVITY_RELEASING_MALTOHEXAOSE_ URL	AMY1A, AMY1B, AMY1C,AMY2A, AMY2B
GOMF_PROTON_CHANNEL_ACTIVITY URL	OTOP1, OTOP3, ASIC5, NOX5, SLC4A11, HVCN1, OTOP2
GOMF_PHEROMONE_RECEPTOR_ACTIVITY URL	VN1R2, VN1R3, VN1R4, VN1R5, VN1R17P, VN1R1
GOMF_HISTONE_METHYLTRANSFERASE_ACTIVITY_H4_K20_SPECIFIC_ URL	KMT5A, KMT5B, NSD1, NSD2, KMT5C
GOMF_N_ACETYLGALACTOSAMINE_4_O_SULFOTRANSFERASE_ACTIVITY URL	CHST14, CHST13, CHST11, CHST8, CHST9
GOMF_ALPHA_N_ACETYLGALACTOSAMINIDE_ALPHA_2_6_SIALYLTRANSFERASE_ACTIVITY URL	ST6GALNAC2, ST6GALNAC3, ST6GALNAC6,ST6GALNAC1, ST6GALNAC5
GOMF_TRACE_AMINE_RECEPTOR_ACTIVITY URL	TAAR9, TAAR1, TAAR6, TAAR8 TAAR5 TAAR2 TAAR3P
GOMF_2_ACYLGLYCEROL_O_ACYLTRANSFERASE_ACTIVITY URL	MOGAT1, AWAT2, MOGAT3, MOGAT2, DGAT2, DGAT1
GOMF_PYRUVATE_TRANSMEMBRANE_TRANSPORTER_ACTIVITY URL	SLC16A11, MPC2, MPC1L, MPC1, SLC16A7

(c)Gene set ranking using GSEA

Gene Sets	Gene Names
GOMF_PHEROMONE_RECEPTOR_ACTIVITY URL	VN1R2, VN1R3, VN1R4, VN1R5, VN1R17P, VN1R1
GOMF_INORGANIC_PHOSPHATE_TRANSMEMBRANE_TRANSPORTER_ACTIVITY URL	SLC25A3, ANKH, SLC17A7, SLC20A1, SLC20A2
GOMF_BITTER_TASTE_RECEPTOR_ACTIVITY URL	TAS2R39, TAS2R40, TAS2R43, TAS2R31,TAS2R45, TAS2R46, TAS2R30,TAS2R19,TAS2R20, TAS2R50, TAS2R60, TAS2R3, TAS2R4, TAS2R16, TAS2R1, TAS2R9,TAS2R8, TAS2R7, TAS2R13, TAS2R10, TAS2R14,TAS2R5,TAS2R38
GOMF_G_PROTEIN_COUPLED_NEUROTRANSMITTER_RECEPTOR_ACTIVITY URL	HRH3, CHRM1, CHRM2, CHRM3, CHRM4, CHRM5, OR10H4, OR10J5, ADRB1, GPR156, DRD4, OR5T2,OR6T1,GABBR1, OR10H3,OR10H2, AC114267.1, OR10H5, GRM1, HRH1,HRH2, HTR1A, HTR1B, HTR1D, HTR1E,HTR1F, HTR2A,HTR2B, HTR2C, HTR4, HTR5A, HTR6, HTR7, OR5T3,OR11H7, OR11H4, OR10J6P, OPRM1, ZNF219, HRH4, GABBR2
GOMF_KAINATE_SELECTIVE_GLUTAMATE_RECEPTOR_ACTIVITY URL	GRIK1, GRIK2, GRIK3, GRIK4, GRIK5
GOMF_ALKANE_1_MONOOXYGENASE_ACTIVITY URL	CYP4F8, CYP4A11, CYP4A22, CYP4F12, CYP4F2
GOMF_DNA_TOPOISOMERASE_ACTIVITY URL	TOP1MT, SPO11, TOP1, TOP2A, TOP2B, TOP3A, TOP3B
GOMF_COPPER_CHAPERONE_ACTIVITY URL	COX17, PARK7, ATOX1, ATP7A, CCS
GOMF_CLASS_I_DNA_APURINIC_OR_APYRIMIDINIC_SITE_ENDONUCLEASE_ACTIVITY URL	APLF, NEIL2, APEX2, APEX1, NTHL1, OGG1, NEIL3, RPS3, NEIL1, ALKBH1
GOMF_DOLICHYL_DIPHOSPHOOLIGOSACCHARIDE_PROTEIN_GLYCOTRANSFERASE_ACTIVITY URL	STT3B, STT3A, OSTC, RPN1, RPN2

Table 4.9: Top 10 gene sets identified as differentiating pre-cancerous and cancerous colon cancer lesions using random forest, MRMR and GSEA.

(a) Gene set ranking using random forest	
Gene Sets	**Gene Names**
GOMF_S_ACYLTRANSFERASE_ACTIVITY URL	YKT6, ZDHHC19, ZDHHC15, DLAT, DLST, FASN, ZDHHC17, ZDHHC20, ZDHHC24, ZDHHC23, ZDHHC5, MCAT, GLUL, ZDHHC22, ZDHHC1, ZDHHC8, ZDHHC21, GOLGA7B ,ZDHHC9, GOLGA7
GOMF_COBALAMIN_BINDING URL	MMACHC, CBLIF, MMAB, MTR, MMUT, CD320, LMBRD1, TCN1, TCN2, CUBN
GOMF_PROTON_CHANNEL_ACTIVITY URL	OTOP1, OTOP3, ASIC5, NOX5, SLC4A11, HVCN1, OTOP2
GOMF_REGULATORY_RNA_BINDING URL	TRIM71, FAM172BP, RC3H1, DHX9, AGO3, AGO4, ELAVL1, C3H7B, TUT4, FMR1, PUM2, DICER1, ZNF346, AGO1, AGO2, ZC3H7A,HNRNPA1, HNRNPA2B1, MECP2, TLR7,SPOUT1
GOMF_ORNITHINE_DECARBOXYLASE_REGULATOR_ACTIVITY URL	AZIN2, OAZ1, OAZ2, AZIN1, OAZ3, PRLR
GOMF_PRE_MRNA_BINDING URL	RNU4ATAC, RNU6ATAC, RNVU1-8, RNVU1-4, RNVU1-14, RNVU1-15, RNVU1-17, RNVU1-3, RNVU1-1, RNVU1-6, RNVU1-19, U2AF1L5, LINC01715, SLU7, PRPF8, CELF1, CELF2, CELF3, U2AF2, OVAAL, DDX5, U2AF1L4, EP300, RBM24, RNU12
GOMF_FATTY_ACID_DERIVATIVE_BINDING URL	ECI2, ACOT7, ACOT12, DBI, OXER1, ALOX5AP, ACOT11, GCDH, ACAD9, HADHA, HMGCL, ACADL, ACADVL, ACBD7, ALDH6A1, PPARG, S100A8, S100A9, SCP2, ACBD3, SOAT1, STX3, ACBD4
GOMF_S_METHYLTRANSFERASE_ACTIVITY URL	INMT, BHMT2, MGMT, ASMT, MTR, BHMT, TPMT
GOMF_BILE_ACID_TRANSMEMBRANE_TRANSPORTER _ACTIVITY URL	SLCO1B1, AKR1C4, SLCO2B1, SLC51B, SLC51A, SLC10A4, SLCO1B3, SLCO1B7, SLC10A6, SLCO1C1, CEACAM1, SLC10A1, SLC10A2, SLCO1A2,ABCB11, ABCC3
GOMF_COMPLEMENT_COMPONENT_C1Q_BINDING URL	CRP, CD93, APCS, PTX3, C1QBP, C4A, CALR, MEGF10
(b) Gene set ranking using MRMR	
Gene Sets	**Gene Names**
GOMF_SPHINGOSINE_N_ACYLTRANSFERASE_ACTIVITY URL	CERS1, CERS3, CERS6, CERS2, CERS4, FAM57B, CERS5
GOMF_MRNA_BINDING_INVOLVED_IN_POSTTRANSCRIPTIONAL _GENE_SILENCING URL	MIR675, MIR298, MIR509-2, MIR892B, MIR876, MIR877, MIR665, MIR873, MIR301B, MIR543, MIR208B, MIR509-3, MIR939, MIR365A, MIR365B, MIR1224, MIR1207, MIR548P
GOMF_ALPHA_AMYLASE_ACTIVITY_RELEASING _MALTOHEXAOSE_ URL	AMY1A, AMY1B, AMY1C, AMY2A, AMY2B
GOMF_AMYLASE_ACTIVITY URL	AMY1A, AMY1B, AMY1C, AMY2A, AMY2B, MGAM
GOMF_RNA_ADENYLYLTRANSFERASE_ACTIVITY URL	TENT5B, TENT5D, TRNT1, TENT5C, TENT5A
GOMF_RRNA_GUANINE_METHYLTRANSFERASE_ACTIVITY URL	BUD23, FTSJ3, TRMT112, MRM3, MRM1
GOMF_PRE_MRNA_5_SPLICE_SITE_BINDING URL	RNU4ATAC, RNU6ATAC, RNVU1-8, RNVU1-4, RNVU1-14, RNVU1-15, RNVU1-17, RNVU1-3, RNVU1-1, RNVU1-6, RNVU1-19, LINC01715, RNU11, RNVU1-18, RNVU1-7, RNU1-3
GOMF_FUCOSE_BINDING URL	FUOM, CLEC17A, ACR, SELP, COLEC11
GOMF_UDP_XYLOSYLTRANSFERASE_ACTIVITY URL	RXYLT1, LARGE2, POGLUT3, XXYLT1, GXYLT1, POGLUT1, XYLT1, XYLT2, GXYLT2, POGLUT2,LARGE1
GOMF_PROTEIN_XYLOSYLTRANSFERASE_ACTIVITY URL	POGLUT3, POGLUT1, XYLT1, XYLT2, POGLUT2
(c) Gene Set ranking using GSEA	
Gene Sets	**Gene Names**
GOMF_3_CHLOROALLYL_ALDEHYDE_DEHYDROGENASE _ACTIVITY URL	ALDH3A1, ALDH3B1, ALDH3B2, ALDH3A2, ALDH1A2
GOMF_UBIQUITIN_LIGASE_INHIBITOR_ACTIVITY URL	FBXO5, RPL5, RPL11, RPS7, RPL23
GOMF_ACROSIN_BINDING URL	POMZP3, SERPINA5, ZP4, ZP2, ZP3
GOMF_GLUTATHIONE_DISULFIDE_OXIDOREDUCTASE _ACTIVITY URL	GLRX3, GSTO2, GLRX, GSR, GLRX2, GLRX5, GSTO1
GOMF_EPOXIDE_HYDROLASE_ACTIVITY URL	EPHX1, EPHX2, LTA4H, RNPEP, EPHX3, AKR7A2
GOMF_MHC_CLASS_IB_RECEPTOR_ACTIVITY URL	KLRC4-KLRK1, LILRB1, CD160, KLRK1, KIR2DL4
GOMF_VASCULAR_ENDOTHELIAL_GROWTH_FACTOR_BINDING URL	FLT1, FLT4, KDR, PDGFRA, PDGFRB, PTN, NRP1
GOMF_TYPE_5_METABOTROPIC_GLUTAMATE_RECEPTOR_BINDING URL	ADORA2A, FYN, DNM3, NECAB2, PRNP
GOMF_FIBRINOGEN_BINDING URL	CDH5, FBLN1, ITGA2B, ITGB3, THBS1
GOMF_COLLAGEN_BINDING_INVOLVED_IN_CELL_MATRIX_ADHESION URL	ITGA11, ITGA1, ITGA2, ITGB1, ITGA10

Table 4.10: Top 10 gene sets identified as differentiating cancerous and non-cancerous breast cancer patients using random forest, MRMR and GSEA.

(a) Gene set ranking for random forest

Gene Set	Gene Names
GOMF_TRACE_AMINE_RECEPTOR_ACTIVITY URL	TAAR9, TAAR1, TAAR6, TAAR8, TAAR5, TAAR2, TAAR3P
GOMF_N_ACETYLGLUCOSAMINE_6_O_SULFOTRANSFERASE_ACTIVITY URL	CHST4, CHST5, CHST6, CHST7, CHST1, CHST2, CHST3
GOMF_FUCOSE_BINDING URL	FUOM, CLEC17A, ACR, SELP, COLEC11
GOMF_LOW_DENSITY_LIPOPROTEIN_PARTICLE_RECEPTOR_BINDING URL	LANCL1, APOA5, AP2M1, CLU, CLTC, CRP, AP2A1, DKK1, MESD, PCSK9, LDLRAP1, DNAJA1, HSPG2, APOB, APOE, LRPAP1, ANKRA2, SYT1, HSP90B1,PICALM, SNX17
GOMF_ACYL_CARNITINE_TRANSMEMBRANE_TRANSPORTER_ACTIVITY URL	SLC25A29, SLC25A48, SLC25A45, SLC25A47, SLC25A20
GOMF_HISTONE_SERINE_KINASE_ACTIVITY URL	PRKAA1, PRKAA2, AURKA, AURKC, VRK1, AURKB
GOMF_OXIDOREDUCTASE_ACTIVITY_ACTING_ON_SINGLE_DONORS_WITH_INCORPORATION_OF_MOLECULAR_OXYGEN URL	ACOT7, ACOT12, DBI, OXER1, ALOX5AP, ACOT11, GCDH, ACAD9, HADHA, HMGCL, ACADL, ACADVL, ACBD7, ALDH6A1,PPARG, S100A8,S100A9,SCP2,ACBD3, SOAT1, STX3,ACBD4,PNPLA3, ACBD6,SOAT2,ACBD5
GOMF_CAMP_DEPENDENT_PROTEIN_KINASE_REGULATOR_ACTIVITY URL	CDO1, IDO2, ETHE1, HAAO, ALOX12, ALOX5, ALOX12B, ALOX15, ALOX15B, P4HA3,HGD, HPD, IDO1, P4HA1, BCO1, TMLHE, ADI1,MIOX, PTGS2, ALOXE3,RPE65,TDO2, BCO2, BBOX1, HPDL,ADO, PIR, P4HA2
GOMF_PHOSPHOLIPASE_D_ACTIVITY URL	PLD4, PLD3, GPLD1, HMOX1, PLD1, PLD2
GOMF_GLUCOSE_SODIUM_SYMPORTER_ACTIVITY URL	SLC5A11, SLC5A10, SLC5A9, SLC5A1, SLC5A2, SLC5A3, SLC5A4

(b) Gene set ranking for MRMR

Gene Set	Gene Names
GOMF_ALPHA_AMYLASE_ACTIVITY_RELEASING_MALTOHEXAOSE_ URL	AMY1A, AMY1B, AMY1C, AMY2A, AMY2B
GOMF_PROTON_CHANNEL_ACTIVITY URL	OTOP1, OTOP3, ASIC5, NOX5, SLC4A11, HVCN1, OTOP2
GOMF_ACYL_CARNITINE_TRANSMEMBRANE_TRANSPORTER_ACTIVITY URL	SLC25A29, SLC25A48, SLC25A45, SLC25A47, SLC25A20
GOMF_PHEROMONE_RECEPTOR_ACTIVITY URL	VN1R2, VN1R3, VN1R4, VN1R5, VN1R17P, VN1R1
GOMF_HISTONE_METHYLTRANSFERASE_ACTIVITY_H4_K20_SPECIFIC_ URL	KMT5A, KMT5B, NSD1, NSD2, KMT5C
GOMF_FUCOSE_BINDING GOMF_HISTONE_METHYLTRANSFERASE_ACTIVITY_H4_K20 _SPECIFIC_ URL	FUOM, CLEC17A, ACR, SELP, COLEC11
GOMF_NOREPINEPHRINE_BINDING URL	ADRA2A, ADRB1, ADRB2, ADRB3, DRD4
GOMF_N_ACETYLGALACTOSAMINE_4_O_SULFOTRANSFERASE_ACTIVITY URL	CHST14, CHST13, CHST11, CHST8, CHST9
GOMF_TRACE_AMINE_RECEPTOR_ACTIVITY URL	TAAR9, TAAR1, TAAR6, TAAR8, TAAR5, TAAR2, TAAR3P
GOMF_ATPASE_INHIBITOR_ACTIVITY URL	PLN, FNIP2, TSC1, ATP5IF1, FNIP1

(c) Gene set ranking for GSEA

Gene Set	Gene Names
GOMF_SPHINGOSINE_N_ACYLTRANSFERASE_ACTIVITY URL	CERS1, CERS3, CERS6, CERS2, CERS4, FAM57B, CERS5
GOMF_SULFATIDE_BINDING URL	TPP1, CLN3, CLN6, PPT1, MANF
GOMF_AMINOACYLASE_ACTIVITY URL	DARS, ASPA, CAT, ACY3, ACY1
GOMF_EPOXIDE_HYDROLASE_ACTIVITY URL	EPHX1, EPHX2, LTA4H, RNPEP, EPHX3, AKR7A2
GOMF_AMIDINE_LYASE_ACTIVITY URL	ADSL, ASL, CHAC2, PAM, GGCT, CHAC1, GGACT
GOMF_MHC_CLASS_IB_RECEPTOR_ACTIVITY URL	KLRC4-KLRK1, LILRB1, CD160, KLRK1, KIR2DL4
GOMF_G_PROTEIN_COUPLED_ADENOSINE_RECEPTOR_ACTIVITY URL	ADORA1, ADORA2A, ADORA2B, ADORA3, P2RY1, P2RY11, P2RY12
GOMF_TYPE_5_METABOTROPIC_GLUTAMATE_RECEPTOR_BINDING URL	ADORA2A, FYN, DNM3, NECAB2, PRNP
GOMF_CCR2_CHEMOKINE_RECEPTOR_BINDING URL	DEFB106A, DEFB106B, CCL2, CCL7, CCR2
GOMF_DNA_TRANSLOCASE_ACTIVITY URL	RAD54B, ATRX, ERCC6L, PBRM1, FBH1

Table 4.11: Top 10 Gene sets for adrenal cancer using random forest, MRMR and GSEA.

(a) Gene set ranking for random forest

Gene Sets	Gene Names
GOMF_STEROID_HORMONE_RECEPTOR_BINDING URL	CDK7, CEBPB, KAT5, ARID5A, PPARGC1A, PADI2, PARK7, PHB2, PPARGC1B, PARP1, SRARP, CTNNB1, DNAAF4, DAXX, DDX5, ZNF366, NR0B1, EP300, ESR1
GOMF_PHOSPHATIDYLINOSITOL_3_4_5_TRISPHOSPHATE_BINDING URL	IQGAP2, FERMT2, ADAP1, ARAP2, ARAP1, COMMD1, AKT1, PHLDA3, WASHC2C, DAPP1, RACGAP1, MYO1B, MYO10, NRGN,ASAP1, ZFYVE1, PLEKHB2, KIF16B
GOMF_RIBOSOME_BINDING URL	PRMT3, RACK1, GCN1, CPEB2, LETM2, EIF5AL1, EEF2, EIF2S1, EIF5A, ETF1, MTIF3, CPEB3, MRPS27, FMR1, MTOR, RICTOR, YTHDF3, LETMD1, GEMIN5, EIF3K, OLA1, SEC61A1
GOMF_PHOSPHATIDYLINOSITOL_PHOSPHATE_BINDING URL	COL4A3BP, IQGAP2, FERMT2, RAB35, ADAP1, WDR45, SNX18, TWF2, TIRAP, OSBPL5, OSBPL8, FCHO2, ARAP2, ARAP1, SNX20, BBS5, CLVS2, AMER1, SYT9, COMMD1
GOMF_TRANSLATION_FACTOR_ACTIVITY_RNA_BINDING URL	TSFM, EIF1, EIF1B, EIF3M, HBS1L, GCN1, COPS5, CPEB2, EIF5AL1, EEF1A1P5, EEF1A1, EEF1A2, EEF1B2, EEF1D, EEF1G, EEF2, EIF2D, EIF1AX, EIF2S1, EIF2B1, EIF2S3
GOMF_NUCLEOBASE_CONTAINING_COMPOUND_TRANSMEMBRANE_TRANSPORTER_ACTIVITY URL	SLC35B1, SLC25A17, SLC35A1, SLC35D2, SLC35A4, SLC25A25, SLC29A1, ABCD1, HNRNPA3, SLC29A4, SLC35D1, SLC35A3, SLC25A41, SLC25A42, SLC25A4, SLC25A5, SLC25A6 ,SLC25A24, SLC29A2, SLC35E4
GOMF_EXTRACELLULAR_MATRIX_STRUCTURAL_CONSTITUENT_CONFERRING_TENSILE_STRENGTH URL	COL1A1, COL1A2, COL2A1, COL3A1, COL4A1, COL4A2, COL4A3, COL4A4, COL4A5, COL4A6, COL5A1, COL5A2, COL6A1, COL6A2, COL6A3, COL7A1, COL8A1, COL8A2
GOMF_L_AMINO_ACID_TRANSMEMBRANE_TRANSPORTER_ACTIVITY URL	SLC25A13, SLC25A15, SLC38A3, SLC7A9, SLC36A4, SLC15A4, SLC25A29, SLC43A2, SLC32A1, CTNS, SLC38A9, SLC36A2, SLC7A13, SLC36A1, SLC7A8, SLC7A11, SLC17A8
GOMF_EXTRACELLULAR_MATRIX_STRUCTURAL_CONSTITUENT URL	EDIL3, ENAM, PRG4, PRG3, SPON1, FBLN5, POSTN, FGL2, EMILIN1, CHI3L1, CTHRC1, PODN, COL1A1, COL1A2, COL2A1, COL3A1, COL4A1, COL4A2, COL4A3, COL4A4, COL4A5, COL4A6, COL5A1, COL5A2, COL6A1, COL6A2
GOMF_PHOSPHOLIPID_BINDING URL	PLA2G4B, PIGK, RASA4B, COL4A3BP, RASGRP1, RASA4, PEMT, CEACAM5, UNC13B, CETP, IQGAP2, FERMT2, RAB35, ADAP1, WDR45, NISCH, ANXA10, PACSIN2

(b) Gene set ranking using MRMR

Gene Sets	Gene Names
GOMF_SPHINGOSINE_N_ACYLTRANSFERASE_ACTIVITY URL	CERS1, CERS3, CERS6, CERS2, CERS4, FAM57B, CERS5
GOMF_MRNA_BINDING_INVOLVED_IN_POSTTRANSCRIPTIONAL_GENE_SILENCING URL	MIR675, MIR298, MIR509-2, MIR892B, MIR876, MIR877,MIR665, MIR939, MIR365A, MIR365B, MIR1224, MIR1207, MIR548P, MIR2355
GOMF_FUCOSE_BINDING URL	FUOM, CLEC17A, ACR, SELP, COLEC11
GOMF_PRE_MRNA_5_SPLICE_SITE_BINDING URL	RNU4ATAC, RNU6ATAC, RNVU1-8, RNVU1-4, RNVU1-14, RNVU1-15, RNVU1-19, LINC01715, RNU11, RNVU1-18, RNVU1-7, RNU1-3, RNU1-2
GOMF_ALPHA_AMYLASE_ACTIVITY_RELEASING_MALTOHEXAOSE_ URL	AMY1A, AMY1B, AMY1C, AMY2A, AMY2BG
GOMF_RRNA_GUANINE_METHYLTRANSFERASE_ACTIVITY URL	BUD23, FTSJ3, TRMT112, MRM3, MRM1
GOMF_AMYLASE_ACTIVITY URL	AMY1A, AMY1B, AMY1C, AMY2A, AMY2B, MGAM
GOMF_UDP_XYLOSYLTRANSFERASE_ACTIVITY URL	RXYLT1, LARGE2, POGLUT3, XXYLT1, GXYLT1, POGLUT1, XYLT1, XYLT2, GXYLT2, POGLUT2, LARGE1
GOMF_PROTEIN_XYLOSYLTRANSFERASE_ACTIVITY URL	POGLUT3, POGLUT1, XYLT1, XYLT2, POGLUT2
GOMF_BENZODIAZEPINE_RECEPTOR_BINDING URL	RIMBP3C, DBI, RIMBP3B, RIMBP3, TSPOAP1

(c) Gene set ranking using GSEA

Gene Sets	Gene Names
GOMF_PROSTAGLANDIN_E_RECEPTOR_ACTIVITY URL	HPGD, PTGER1, PTGER2, PTGER3, PTGER4
GOMF_ATPASE_INHIBITOR_ACTIVITY URL	PLN, FNIP2, TSC1, ATP5IF1, FNIP1
GOMF_TRANSFORMING_GROWTH_FACTOR_BETA_RECEPTOR_ACTIVITY_TYPE_I URL	ACVR1C, BMPR1A, BMPR1B, TGFBR1, ACVR1, ACVR1B, ACVRL1
GOMF_LEUKOTRIENE_C4_SYNTHASE_ACTIVITY URL	ALOX5AP, GSTM4, LTC4S, MGST2, MGST3
GOMF_G_PROTEIN_COUPLED_ADENOSINE_RECEPTOR_ACTIVITY URL	ADORA1, ADORA2A, ADORA2B, ADORA3, P2RY1, P2RY11, P2RY12, RNU4ATAC, RNU6ATAC, RNVU1-8, RNVU1-4, RNVU1-14, RNVU1-15
GOMF_PRE_MRNA_5_SPLICE_SITE_BINDING URL	RNVU1-3, RNVU1-1, RNVU1-6, RNVU1-19, LINC01715, RNU11, RNVU1-7, RNU1-3, RNU1-2, RNU1-1, WEE2-AS1, PRPF39, RNU1-4
GOMF_BETA_N_ACETYLHEXOSAMINIDASE_ACTIVITY URL	OGA, GM2A, HEXD, HEXA, HEXB
GOMF_5_DEOXYRIBOSE_5_PHOSPHATE_LYASE_ACTIVITY URL	POLQ, XRCC6, POLL, XRCC5, HMGA2
GOMF_RNA_DEPENDENT_ATPASE_ACTIVITY URL	DDX17, DDX11, IGHMBP2, YTHDC2, DDX39B
GOMF_STRUCTURAL_CONSTITUENT_OF_POSTSYNAPTIC_ACTIN_CYTOSKELETON URL	ACTBL2, POTEKP, ACTB, ACTG1, INA

4.9.2 Heatmaps for leukaemia, colon, adrenal and breast cancer data sets

The heatmap shows comparative outcomes of the GO terms between the proposed anomaly scores and the state-of-the-art approaches. The color coding of the heatmap reflects the range of different values for the proposed anomaly scores and the enrichment scores for the state-of-the-art approaches.

Experiment setup:

1. Technology: Heatmap.

2. Python packages: pandas, NumPy, matplotlib, pyplot, seaborn, and matplotlib.patches.

3. Input data: Anomaly scores for all four cancer data sets and scores for all state-of-the-art.

Figure 4.29 shows a heatmap for GO terms for the top 10 gene sets for leukaemia datasets using the proposed method and state-of-the-art approaches. The names of the procedures appear in the columns of the heatmap, while the GO terms appear in the rows. From the figure it can be seen that, the proposed anomaly score identified new GO terms, in contrast to the state-of-the-art approaches.

Figure 4.30 shows a heatmap for GO terms for the top 10 gene sets for colon cancer (colorectal cancer (CRC) vs inflammatory bowel disease (IBD)) datasets using the proposed method and state-of-the-art approaches. In contrast to state-of-the-art approaches, the proposed anomaly score identified new GO terms (Figure 4.30).

Figure 4.31 shows a heatmap for GO terms for the top 10 gene sets for breast cancer (healthy vs cancerous patients) datasets using the proposed method and state-of-the-art approaches. From the figure it can be seen that the proposed anomaly scores identified new GO terms from the breast cancer data set in contrast to the state-of-the-art approaches.

Figure 4.32 shows a heatmap for GO terms for the top 10 gene sets for adrenal cancer (adrenal adenoma vs adrenal carcinoma) datasets using the proposed method and state-of-the-art approaches. In contrast to the state-of-the-art approaches, the figure demonstrates that the proposed anomaly score identified new GO terms.

Moreover, the proposed anomaly score and the CSAX, FRaC, and Eigfusion for leukaemia datasets all had the same transmembrane signalling GO term. Amid binding, the same GO term was shared by anomaly score, FRaC, and outlier data for colon cancer. When it comes to breast cancer, the anomaly score, GFS, FRaC, CSAX, PFSNet, and outlier all have the same ceramide binding, protein binding, and lipase inhibitor activity. Choline transport and protein binding are similar between the FRaC and the anomaly score.

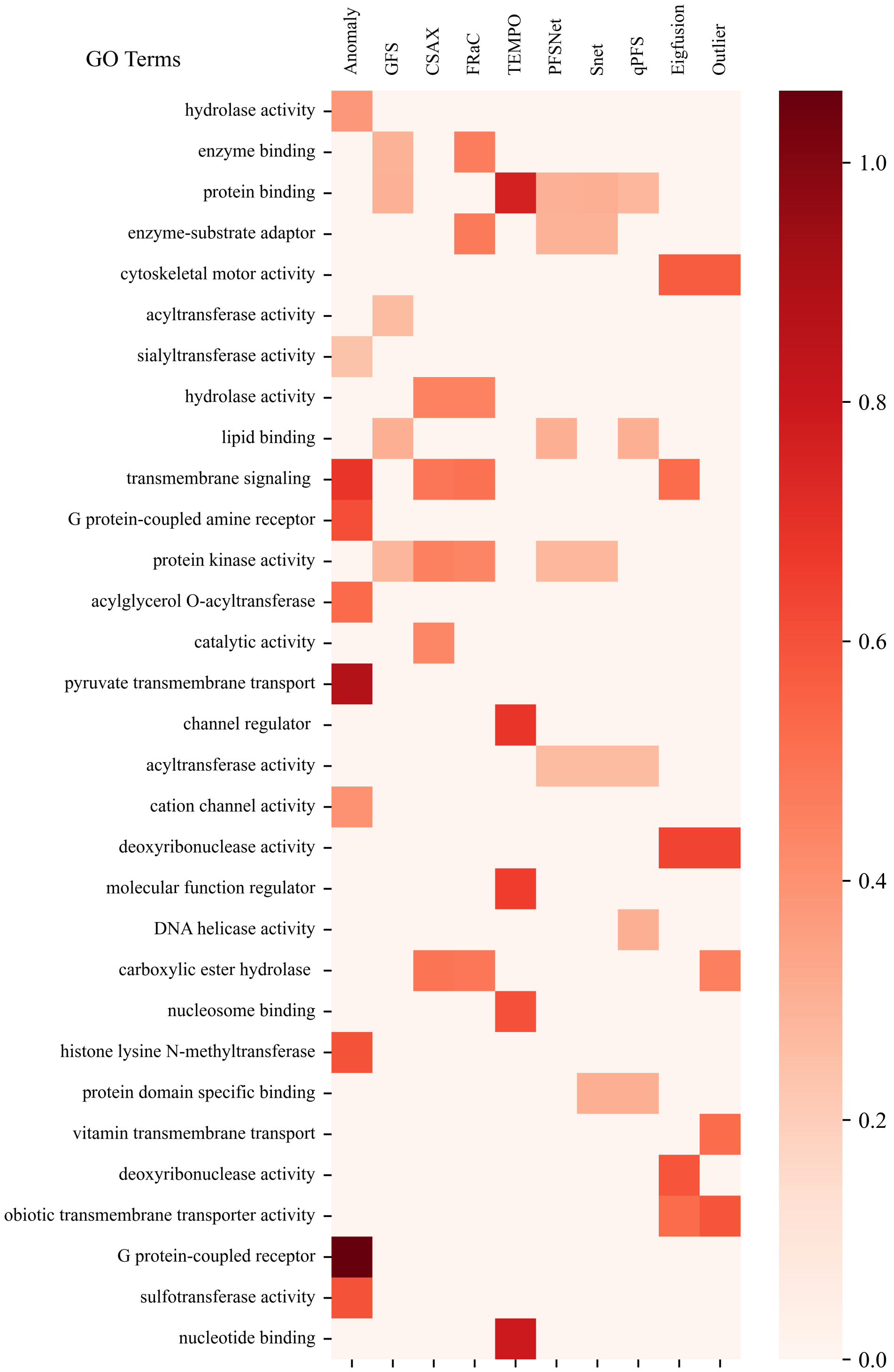


Figure 4.29: Gene ontology terms for the top 10 gene sets for the leukaemia data sets visualized as a heatmap using the proposed method and state-of-the-art approaches.

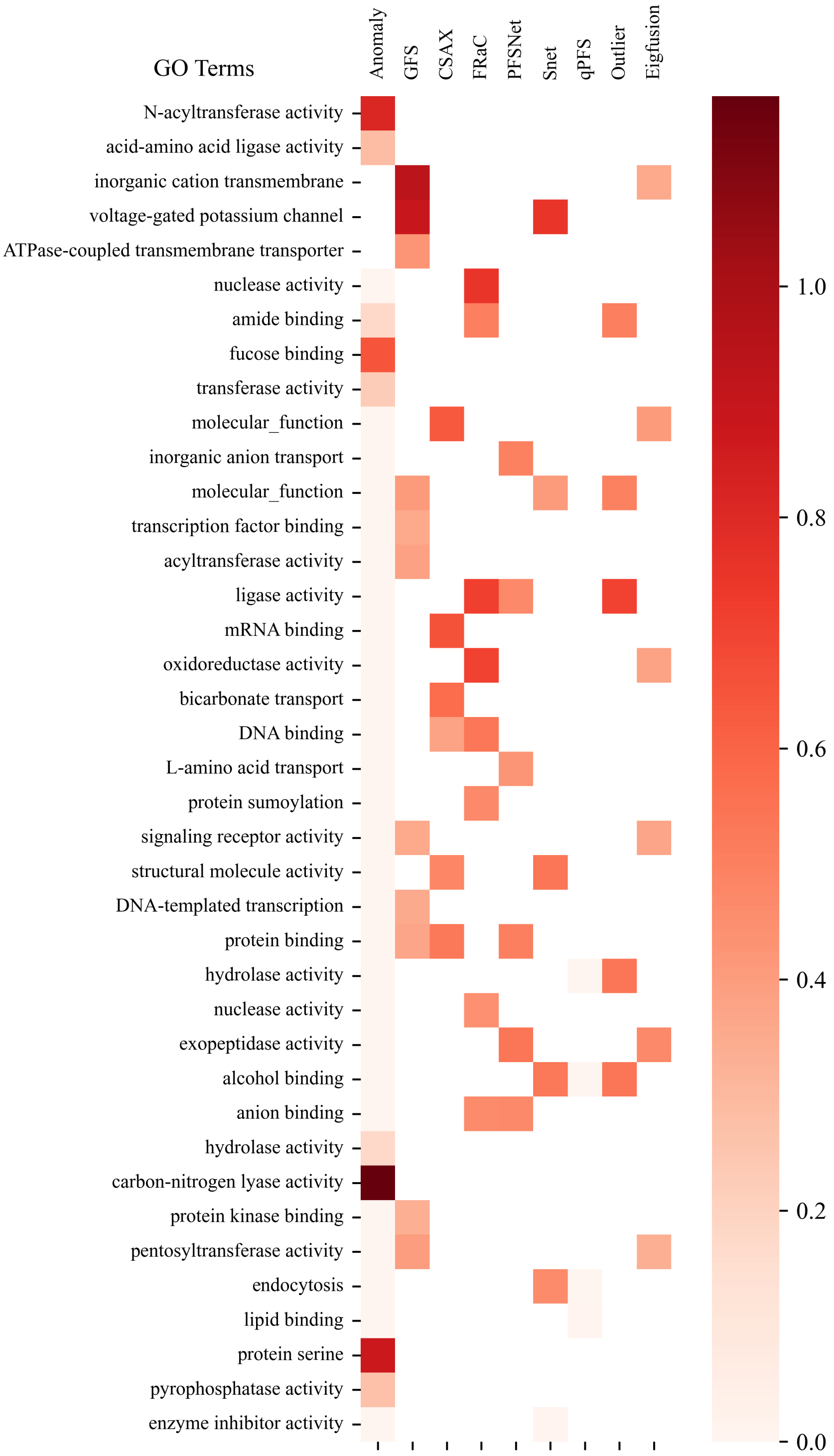


Figure 4.30: Gene ontology terms for the top 10 gene sets for colon cancer (IBD VS CRC) datasets visualized as a heatmap using the proposed method and state-of-the-art approaches.

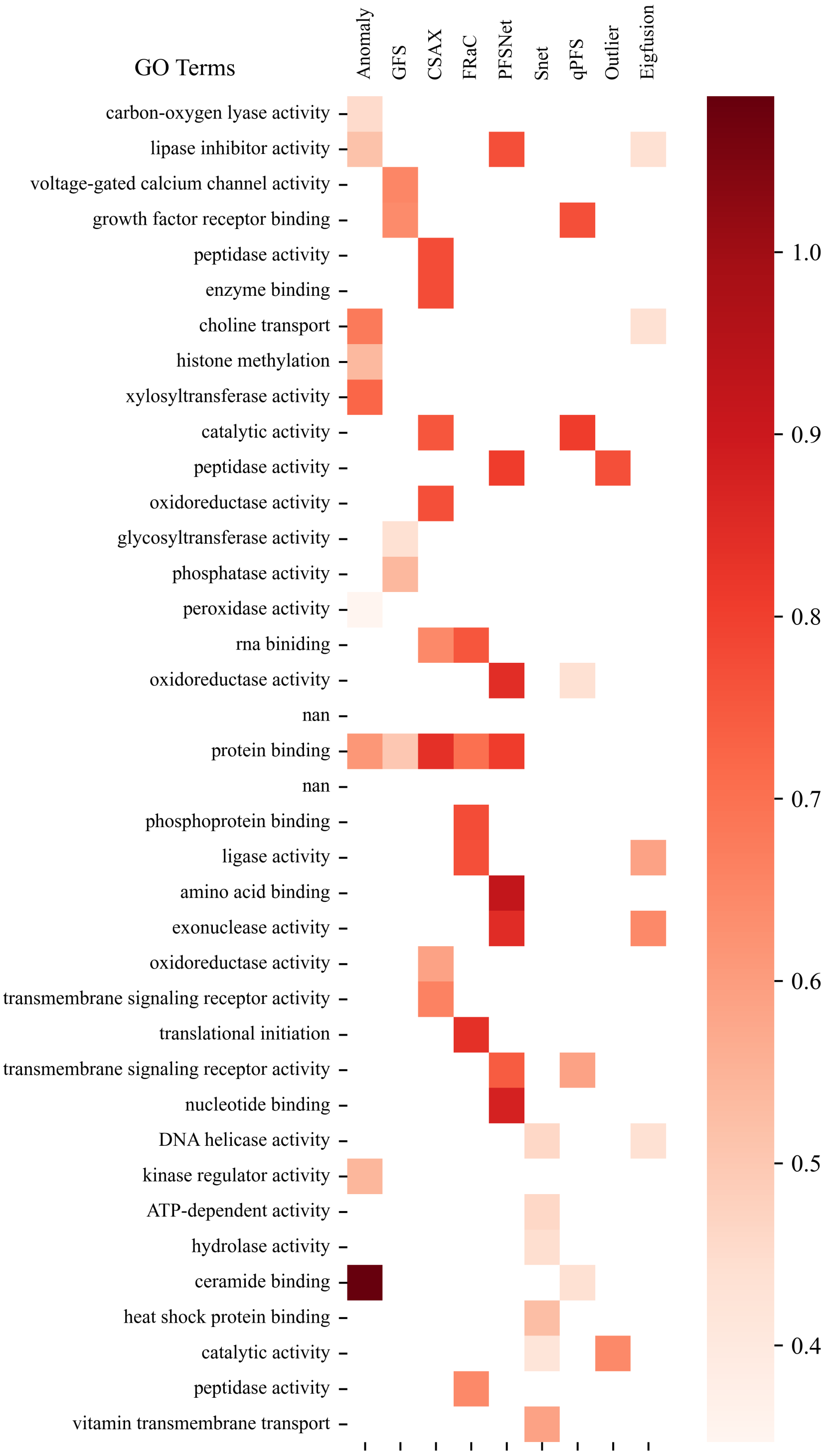


Figure 4.31: Gene ontology terms for the top 10 gene sets for breast cancer (healthy vs cancerous) datasets visualized as a heatmap using the proposed method and state-of-the-art approaches.

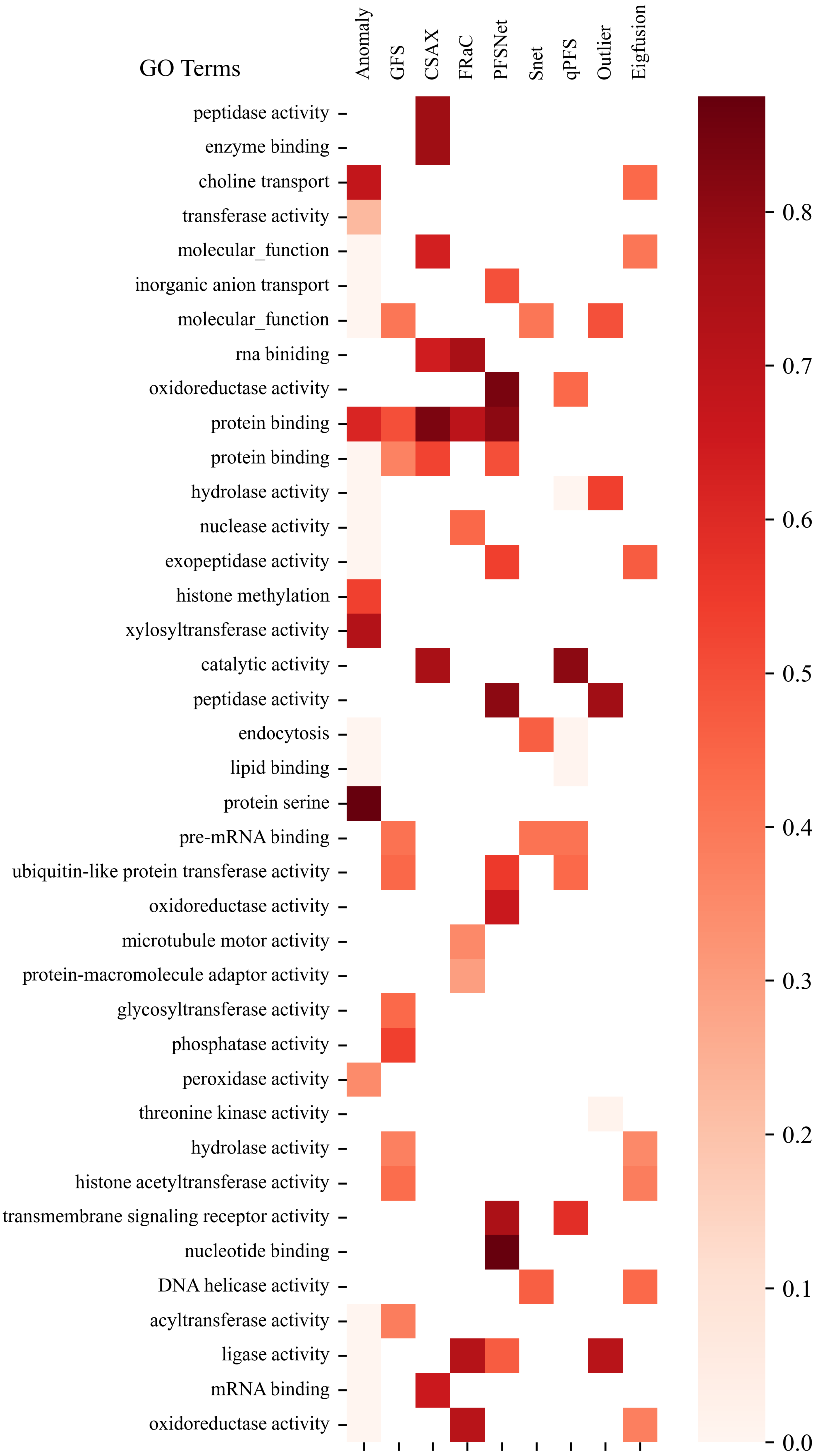


Figure 4.32: Gene ontology terms for the top 10 gene sets for adrenal cancer (adrenal adenoma vs adrenal carcinoma) datasets visualized as a heatmap using the proposed method and state-of-the-art approaches.

4.10 Discussion

From the heatmap, it can be seen that the proposed anomaly scores provide different GO terms than other state-of-the-art approaches. The hypothesis of this thesis, careful integration of all genes of a patient into one gene set will allow measuring the variation in a gene set responsible for the disease. The proposed anomaly score includes all genes to find the variation of each gene set (gene ontology), whereas GSEA uses a small number of genes to find enrichment scores for each gene set.

Most of the gene sets identified by the proposed anomaly scores differ from those identified by state-of-the-art approaches. A few gene sets were consistent between the proposed anomaly scores and the state-of-the-art approaches. Some gene sets differed from the state-of-the-art approaches associated with the prediction of cancer recurrence in leukaemia, breast cancer, colon cancer, and adrenal cancer.

4.11 Conclusion

The ranking of gene sets from two different feature selection approaches showed that the proposed anomaly scores identified new gene sets (biology) than the gene sets from GSEA for leukaemia, breast, colon, and adrenal cancer. In this thesis, the functions of these gene sets were explored in the available literature since there was no scope to validate them in wet labs. From the literature, it appeared that the functions of these gene sets provided an understanding of a disease such as cancer.

In addition, the comparison between heatmap and state-of-the-art approaches showed that the proposed anomaly scores identified new gene sets for leukaemia, breast, colon, and adrenal cancers than the gene sets of the state-of-the-art approaches. Thus, in this thesis, the gene set ranking and the heatmap enabled an individual to understand newly identified gene sets for four different cancers, including leukaemia, breast, colon, and adrenal cancers.

The results of the instance-based LINDA-BN revealed the gene sets associated with the prediction of a patient's medical condition. The functions of these gene sets were examined using the existing literature and their effects on diseases such as cancer. Explainable AI (XAI) provides alternative ways to analyse data that are more understandable and technically equivalent to complex black-box AI approaches. In most cases, implementations of XAI approaches could clarify, in a step-by-step fashion, how the features were interconnected to arrive at a conclusion of predictions and analysis.

When analysing cancer data, the XAI approach achieved an improvement in understanding and interpreting the underlying mechanism of an approach. Using anomaly scores, instance-based LINDA-BN, which is an XAI approach, proved to be a trustworthy mechanism for identifying the gene sets which are responsible for predicting a medical condition of an individual patient.

Consequently, the proposed anomaly scores were able to identify different GO terms compared to the state-of-the-art approaches. The heatmap, a simple and easily interpreted representation that reflects anomaly scores, showed new GO terms in contrast to the state-of-the-art approaches.

In light of the results, it can be concluded that an instance-based LINDA-BN is a useful technique for finding the biological components associated with specific medical conditions for a given patient.

However, a drug could potentially have an impact on the patient's existing medical condition. Looking only on the patient's biology could sometimes leave out important details, such as the drug's effect on the patient's condition. Instance-based LINDA-BN overlooked the association with a drug when predicting a patient's health condition.

Chapter 5

Conclusions and future work

"Imagination is more important than knowledge. Knowledge is limited. Imagination encircles the world". – Albert Einstein.

5.1 Conclusion

This thesis hypothesized that the careful aggregation of gene expression values into gene set anomaly scores would provide opportunities to gain insights from the data analysis of gene expression profiles. A family of techniques was described for this purpose in this thesis, showing that even simple methods like using the z-scores of gene expression values to measure variation and taking the arithmetic mean of variations for each gene set were sufficient to provide a benefit over processing gene expression values directly.

The proposed approach was utilized in the analysis of gene expression data from cancer patients across three datasets. In particular, anomaly scores followed by either PCA or MRMR made clusters of cancer patients visible in scatter plots, clusters associated with treatment outcomes. Also, MRMR identified candidate gene sets with biologically relevant implications. In contrast, when raw gene expression values were analyzed, biologically relevant patterns were not visible. When comparing the distribution of anomaly scores for relapsed and non-relapse patients, a clear distinction was noted. Distributions for non-relapsed patients exhibited a prominent second mode. This hints at an exciting difference that may be exploited to pursue better treatment.

In this thesis, the research hypothesis has been carefully tested based on the results that demonstrate extensive research on this topic. The first contribution of this thesis is the development of a method that generates anomaly scores for each patient with respect to all genes of patient that match genes in gene sets. This is a novel contribution over previous research that considered only a small number of genes from gene sets. This method helps to investigates the relationships between patients and their biology with respect to the variations in gene expression values. Thus, this method makes it possible to understand a patient's genetic makeup and how it relates to their condition, which is a valuable step in understanding the patient's biology.

This thesis concludes that gene set anomaly scores improve the extraction of insights from gene expression data. Using gene sets brings a knowledge-driven aspect that can then be combined with data-driven analysis. This thesis proposed a new analysis tool and new directions for understanding the genetic causes of diseases. In addition, the proposed anomaly scores were able to differentiate between different GO terms when compared to existing methods. In contrast to state-of-the-art approaches, the proposed anomaly scores identified new GO terms, as shown by the heatmap outcomes.

Second finding focuses on patient clustering, with similar patients being grouped together based on their anomaly gene expression values or scores. This strategy of stratifying patients into cohorts based on genetic variations improves understanding of disease patterns and similarities between patients.

In the third contribution, the results show distinct anomaly distributions in patients with relapse, non-relapse, and a medium cancer. The observed differences in histogram distributions for various gene sets suggest that some gene sets may play a distinct role in patients with relapse versus patients without relapse. It indicates that some gene sets anomaly scores do not change sufficiently in patients with relapse, whereas in patients without relapse, some gene sets anomaly scores reflect significant changes in anomaly scores.

Explainable AI (XAI) is an alternative to complicated black-box AI systems that are both more intelligible and technically equal. In the majority of instances, XAI implementations are capable of explaining, step-by-step, how features are interconnected to arrive at a conclusion of predictions and analyses.

The XAI method improved the understanding and interpretation of the underlying mechanism of an approach when analyzing cancer data. Using anomaly scores, instance-based LINDA-BN proved to be a reliable method for identifying the gene sets responsible for predicting the condition of an individual patient.

Finally, instance-based LINDA-BN provides an interpretable predictive outcomes, making it easier to understand which gene sets are associated with relapse and which are not. LINDA-BN assesses the conditional dependencies and independencies to identify the gene sets that are most likely associated with a particular patient's medical condition. This approach enhances interpretability, provides valuable insight into the decision-making process, and potentially improving disease prediction and feature selection. Based on the outcomes, it can be concluded that an instance-based LINDA-BN is a useful approach for determining the patient biologies associated with particular medical conditions for a given patient.

5.1.1 Impact of interpretable AI

LINDA-BN applies both Probabilistic graphical models (PGMs) and Markov blanket strategy to identify assocaited features. PGMs map how different things relate to each other. Just as a map shows us how different cities are connected by roads, PGMs show how different features, in this case gene sets, might be connected or related. The strength of these connections shows how much one gene set can influence another. PGMs show us how changes in one gene set can lead to changes in another.

The Markov Blanket for a gene set includes all other gene sets that directly affect it. By looking at the Markov Blanket for a particular gene set, it can be easily seen which other gene sets might cause changes in that gene set.

5.2 Limitations

As for practical applications, the integration of biological information in the form of gene sets could greatly assist practitioners in identifying cancer patients for appropriate treatment interventions. However, it's essential to remember that these findings should be further investigated in a wet lab to gain a more precise understanding of the biology responsible for cancer or cancer relapse. This would provide practitioners with a more concrete foundation for applying these findings to their treatment strategies.

5.3 Future work

Anomaly scores are used to investigate associations between patients and their biological characteristics through patient embedding. The embedding is shown in several clusters for different patient groups such as relapse, non-relapse, cancer, healthy, adenoma, and carcinoma. This thesis used four different cancer gene expression datasets, namely leukemia, breast cancer, colorectal cancer, and adrenal cancer.

Although anomaly scores are used to investigate the relationship between patients and patient biology, patient embedding did not take into account the influence of treatment planning. As a result, some of the clusters overlapped with patient groups and were not separated into different clusters. In addition, the embedding results showed that the breast cancer datasets were not clustered in a better way than the other three datasets.

Patient embedding with treatment planning and anomaly scores can ensure a strong relationship between patients and patient biology. In addition, treatment planning may improve the clustering among multiple patient groups. Moreover, machine learning techniques, such as neural networks, could be helpful in separating groups of patients with respect to the anomaly scores.

This thesis applied the XAI approach to explain patient biology associated with disease prognosis. Although the XAI shows the patient biology associated with disease prognosis, a major drawback is that aspects related to the patient's medication were not considered. In the absence of medication factors, merely examining gene sets associated with disease prognosis is insufficient to understand a patient's condition.

This thesis did not test whether understanding a model or being able to explain a model leads to better decision-making by physicians or medical experts. Further study is needed to uncover the details of this phenomenon. Since several issues remain unaddressed, a future extension is suggested to include drug factors with gene sets to explain patients' disease prognosis.

It is expected that further improvements will determine whether a better understanding of a model or an explainable model can improve decision-making by practitioners or medical experts through conducting surveys and collecting data.

References

[1] E. D. Green, M. S. Guyer, Charting a course for genomic medicine from base pairs to bedside, Nature 470 (7333) (2011) 204–213.
URL https://www.nature.com/articles/nature09764

[2] F. H. Crick, On protein synthesis, in: Symp Soc Exp Biol, Vol. 12, 1958, p. 8.

[3] Z. Wu, Y. Xie, N. Bucher, S. R. Farmer, Conditional ectopic expression of c/ebp beta in nih-3t3 cells induces ppar gamma and stimulates adipogenesis., Genes & Development 9 (19) (1995) 2350–2363.
URL http://genesdev.cshlp.org/content/9/19/2350.full.pdf

[4] E.-J. Yeoh, M. E. Ross, S. A. Shurtleff, W. K. Williams, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Relling, A. Patel, et al., Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling, Cancer Cell 1 (2) (2002) 133–143.
URL https://www.sciencedirect.com/science/article/pii/S1535610802000326

[5] S. A. Andres, G. N. Brock, J. L. Wittliff, Interrogating differences in expression of targeted gene sets to predict breast cancer outcome, BMC Cancer 13 (1) (2013) 1–18.
URL https://bmccancer.biomedcentral.com/articles/10.1186/1471-2407-13-326

[6] M. Kanehisa, S. Goto, Kegg: kyoto encyclopedia of genes and genomes, Nucleic Acids Research 28 (1) (2000) 27–30.
URL https://academic.oup.com/nar/article/28/1/27/2384332

[7] J. J. Green, J. H. Elisseeff, Mimicking biological functionality with polymers for

biomedical applications, Nature 540 (7633) (2016) 386–394.
URL https://www.nature.com/articles/nature21005

[8] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proceedings of the National Academy of Sciences 102 (43) (2005) 15545–15550.

[9] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, J. P. Mesirov, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles, Proceedings of the National Academy of Sciences 102 (43) (2005) 15545–15550. doi:10.1073/pnas.0506580102.
URL http://www.pnas.org/cgi/doi/10.1073/pnas.0506580102

[10] M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for rna-seq data with deseq2, Genome Biology 15 (12) (2014) 550.

[11] P. Pavlidis, J. Qin, V. Arango, J. J. Mann, E. Sibille, Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex, Neurochemical Research 29 (6) (2004) 1213–1222.

[12] L. Tian, S. A. Greenberg, S. W. Kong, J. Altschuler, I. S. Kohane, P. J. Park, Discovering statistically significant pathways in expression profiling studies, Proceedings of the National Academy of Sciences 102 (38) (2005) 13544–13549.

[13] H.-M. Hsueh, C.-A. Tsai, Gene set analysis using sufficient dimension reduction, BMC Bioinformatics 17 (1) (2016) 74.

[14] F. Al-Shahrour, R. Díaz-Uriarte, J. Dopazo, Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information, Bioinformatics 21 (13) (2005) 2988–2993.

[15] L. Statello, C.-J. Guo, L.-L. Chen, M. Huarte, Gene regulation by long non-coding rnas and its biological functions, Nature Reviews Molecular Cell Biology 22 (2) (2021) 96–118.

[16] J. J. Goeman, S. A. Van De Geer, F. De Kort, H. C. Van Houwelingen, A global test for groups of genes: testing association with a clinical outcome, Bioinformatics 20 (1) (2004) 93–99.

[17] Z. Jiang, R. Gentleman, Extensions to gene set enrichment, Bioinformatics 23 (3) (2007) 306–313.

[18] X. Cui, G. A. Churchill, Statistical tests for differential expression in cdna microarray experiments, Genome Biology 4 (4) (2003) 210.

[19] G. V. Glazko, F. Emmert-Streib, Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets, Bioinformatics 25 (18) (2009) 2348–2354.

[20] B. Efron, R. Tibshirani, et al., On testing the significance of sets of genes, The Annals of Applied Statistics 1 (1) (2007) 107–129.

[21] R. Bellazzi, B. Zupan, Towards knowledge-based gene expression data mining, Journal of Biomedical Informatics 40 (6) (2007) 787–802.

[22] S.-Y. Kim, D. J. Volsky, Page: parametric analysis of gene set enrichment, BMC Bioinformatics 6 (1) (2005) 144.

[23] L. Tian, S. A. Greenberg, S. W. Kong, J. Altschuler, I. S. Kohane, P. J. Park, Discovering statistically significant pathways in expression profiling studies, Proceedings of the National Academy of Sciences 102 (38) (2005) 13544–13549. doi:10.1073/pnas.0506577102.
URL http://www.pnas.org/cgi/doi/10.1073/pnas.0506577102

[24] S.-Y. Kim, D. J. Volsky, PAGE: Parametric analysis of gene set enrichment., BMC Bioinformatics 6 (1) (2005) 144. doi:10.1186/1471-2105-6-144.
URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-6-144

[25] F. Maleki, K. L. Ovens, D. J. Hogan, E. Rezaei, A. M. Rosenberg, A. J. Kusalik, Measuring consistency among gene set analysis methods: A systematic study, Journal of Bioinformatics and Computational Biology 17 (05) (2019) 1940010.

doi:10.1142/S0219720019400109.
URL https://www.worldscientific.com/doi/abs/10.1142/S0219720019400109

[26] R. A. Irizarry, Chi Wang, Yun Zhou, T. P. Speed, Gene set enrichment analysis made simple, Statistical Methods in Medical Research 18 (6) (2009) 565–575. doi:10.1177/0962280209351908.
URL http://journals.sagepub.com/doi/10.1177/0962280209351908

[27] P. Tamayo, G. Steinhardt, A. Liberzon, J. P. Mesirov, The limitations of simple gene set enrichment analysis assuming gene independence, Statistical Methods in Medical Research 25 (1) (2016) 472–487. doi:10.1177/0962280212460441.
URL http://journals.sagepub.com/doi/10.1177/0962280212460441

[28] W. T. Barry, A. B. Nobel, F. A. Wright, Significance analysis of functional categories in gene expression studies: a structured permutation approach, Bioinformatics 21 (9) (2005) 1943–1949.

[29] M. A. Newton, F. A. Quintana, J. A. Den Boon, S. Sengupta, P. Ahlquist, et al., Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis, The Annals of Applied Statistics 1 (1) (2007) 85–106.

[30] I. Dinu, J. D. Potter, T. Mueller, Q. Liu, A. J. Adewale, G. S. Jhangri, G. Einecke, K. S. Famulski, P. Halloran, Y. Yasui, Improving gene set analysis of microarray data by sam-gs, BMC Bioinformatics 8 (1) (2007) 242.

[31] G. K. Smyth, Limma: linear models for microarray data, in: Bioinformatics and Computational Biology Solutions using R and Bioconductor, Springer, 2005, pp. 397–420.

[32] C. Henegar, R. Cancello, S. Rome, H. Vidal, K. Clément, J.-D. Zucker, Clustering biological annotations and gene expression data to identify putatively co-regulated biological processes, Journal of Bioinformatics and Computational Biology 4 (04) (2006) 833–852.

[33] S. Das, A. Rai, D. C. Mishra, S. N. Rai, Statistical approach for gene set analysis with trait specific quantitative trait loci, Scientific Reports 8 (1) (2018) 1–12.
URL `https://www.nature.com/articles/s41598-018-19736-w`

[34] W. Luo, M. S. Friedman, K. Shedden, K. D. Hankenson, P. J. Woolf, Gage: generally applicable gene set enrichment for pathway analysis, BMC Bioinformatics 10 (1) (2009) 161.

[35] H. R. Frost, Z. Li, J. H. Moore, Spectral gene set enrichment (sgse), BMC Bioinformatics 16 (1) (2015) 70.

[36] Y. Rahmatallah, F. Emmert-Streib, G. Glazko, Gene sets net correlations analysis (gsnca): a multivariate differential coexpression test for gene sets, Bioinformatics 30 (3) (2014) 360–368.

[37] M. Reich, T. Liefeld, J. Gould, J. Lerner, P. Tamayo, J. P. Mesirov, Genepattern 2.0, Nature Genetics 38 (5) (2006) 500–501.

[38] Y. Rahmatallah, B. Zybailov, F. Emmert-Streib, G. Glazko, Gsar: Bioconductor package for gene set analysis in r, BMC Bioinformatics 18 (1) (2017) 1–12.

[39] X. Yi, Z. Du, Z. Su, Plantgsea: a gene set enrichment analysis toolkit for plant community, Nucleic Acids Research 41 (W1) (2013) W98–W103.
URL `https://academic.oup.com/nar/article/41/W1/W98/1090873?login=true`

[40] A. L. Tarca, G. Bhatti, R. Romero, A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity, PloS One 8 (11) (2013) e79217.

[41] S. Hänzelmann, R. Castelo, J. Guinney, Gsva: gene set variation analysis for microarray and rna-seq data, BMC Bioinformatics 14 (1) (2013) 1–15.

[42] E. Lee, H.-Y. Chuang, J.-W. Kim, T. Ideker, D. Lee, Inferring pathway activity toward precise disease classification, PLoS Computational Biology 4 (11) (2008) e1000217.

[43] Y. Rahmatallah, F. Emmert-Streib, G. Glazko, Gene set analysis approaches for rna-seq data: performance evaluation and application guideline, Briefings in Bioinformatics 17 (3) (2016) 393–407.

[44] J. H. Littell, J. Corcoran, V. Pillai, Systematic reviews and meta-analysis, Oxford University Press, 2008.

[45] Systematic style literature reviews for education and social sciences, howpublished = https://libraryguides.griffith.edu.au/systematic-literature-reviews-for-education, note = Accessed: 2022-09-21.

[46] A. Fink, Conducting research literature reviews: From the internet to paper, Sage publications, 2019.

[47] S. E. Celniker, L. A. Dillon, M. B. Gerstein, K. C. Gunsalus, S. Henikoff, G. H. Karpen, M. Kellis, E. C. Lai, J. D. Lieb, D. M. MacAlpine, et al., Unlocking the secrets of the genome, Nature 459 (7249) (2009) 927–930.
URL https://www.nature.com/articles/459927a

[48] P. C. Sabeti, P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, C. Cotsapas, X. Xie, E. H. Byrne, S. A. McCarroll, R. Gaudet, et al., Genome-wide detection and characterization of positive selection in human populations, Nature 449 (7164) (2007) 913–918.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2687721/

[49] A. A. Ferrando, D. S. Neuberg, J. Staunton, M. L. Loh, C. Huard, S. C. Raimondi, F. G. Behm, C.-H. Pui, J. R. Downing, D. G. Gilliland, et al., Gene expression signatures define novel oncogenic pathways in t cell acute lymphoblastic leukemia, Cancer Cell 1 (1) (2002) 75–87.

[50] S. Moro, P. Cortez, P. Rita, A data-driven approach to predict the success of bank telemarketing, Decision Support Systems 62 (2014) 22–31.
URL https://www.sciencedirect.com/science/article/abs/pii/S016792361400061X

[51] N. Bolshakova, F. Azuaje, P. Cunningham, A knowledge-driven approach to cluster validity assessment, Bioinformatics 21 (10) (2005) 2546–2547.
URL https://pubmed.ncbi.nlm.nih.gov/15713738/

[52] N. Pallast, M. Diedenhofen, S. Blaschke, F. Wieters, D. Wiedermann, M. Hoehn, G. R. Fink, M. Aswendt, Processing pipeline for atlas-based imaging data analysis of

structural and functional mouse brain mri (aidamri), Frontiers in Neuroinformatics 13 (2019) 42.
URL https://www.frontiersin.org/articles/10.3389/fninf.2019.00042/full

[53] J. Wang, L. Chen, Y. Wang, J. Zhang, Y. Liang, D. Xu, A Computational Systems Biology Study for Understanding Salt Tolerance Mechanism in Rice, PLoS One 8 (6) (2013) e64929.
URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0064929

[54] X. Cui, G. A. Churchill, Statistical tests for differential expression in cDNA microarray experiments., Genome Biology 4 (4) (2003) 210. doi:10.1186/gb-2003-4-4-210.
URL https://genomebiology.biomedcentral.com/articles/10.1186/gb-2003-4-4-210

[55] S. Das, P. K. Meher, A. Rai, L. M. Bhar, B. N. Mandal, Statistical Approaches for Gene Selection, Hub Gene Identification and Module Interaction in Gene Co-Expression Network Analysis: An Application to Aluminum Stress in Soybean (Glycine max L.), PLOS One 12 (1) (2017) e0169605.
URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0169605

[56] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proceedings of the National Academy of Sciences 102 (43) (2005) 15545–15550.
URL https://www.pnas.org/content/102/43/15545.short

[57] G. F. Berriz, O. D. King, B. Bryant, C. Sander, F. P. Roth, Characterizing gene sets with funcassociate, Bioinformatics 19 (18) (2003) 2502–2504.
URL https://academic.oup.com/bioinformatics/article/19/18/2502/194688?login=true

[58] T. D. Dubash, C. M. Hoffmann, F. Oppel, K. M. Giessler, S. Weber, S. M. Dieter,

J. Hüllein, T. Zenz, F. Herbst, C. Scholl, W. Weichert, W. Werft, A. Benner, M. Schmidt, M. Schneider, H. Glimm, C. R. Ball, Phenotypic differentiation does not affect tumorigenicity of primary human colon cancer initiating cells, Cancer Letters 371 (2) (2016) 326–333.
URL https://www.sciencedirect.com/science/article/abs/pii/S0304383515007272

[59] A. Ferrari, A. Vincent-Salomon, X. Pivot, A.-S. Sertier, E. Thomas, L. Tonon, S. Boyault, E. Mulugeta, I. Treilleux, G. MacGrogan, L. Arnould, J. Kielbassa, V. Le Texier, H. Blanché, J.-F. Deleuze, J. Jacquemier, M.-C. Mathieu, F. Penault-Llorca, F. Bibeau, O. Mariani, C. Mannina, J.-Y. Pierga, O. Trédan, T. Bachelot, H. Bonnefoi, G. Romieu, P. Fumoleau, S. Delaloge, M. Rios, J.-M. Ferrero, C. Tarpin, C. Bouteille, F. Calvo, I. G. Gut, M. Gut, S. Martin, S. Nik-Zainal, M. R. Stratton, I. Pauporté, P. Saintigny, D. Birnbaum, A. Viari, G. Thomas, A whole-genome sequence and transcriptome perspective on HER2-positive breast cancers, Nature Communications 7 (1) (2016) 12222.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4947184/9

[60] S. Sdelci, C.-H. Lardeau, C. Tallant, F. Klepsch, B. Klaiber, J. Bennett, P. Rathert, M. Schuster, T. Penz, O. Fedorov, G. Superti-Furga, C. Bock, J. Zuber, K. V. M. Huber, S. Knapp, S. Müller, S. Kubicek, Mapping the chemical chromatin reactivation landscape identifies BRD4-TAF1 cross-talk, Nature Chemical Biology 12 (7) (2016) 504–510.
URL https://pubmed.ncbi.nlm.nih.gov/27159579/

[61] C. A. de Leeuw, B. M. Neale, T. Heskes, D. Posthuma, The statistical properties of gene-set analysis, Nature Reviews Genetics 17 (6) (2016) 353–364.
URL https://pubmed.ncbi.nlm.nih.gov/27070863/

[62] S. Das, A. Rai, D. C. Mishra, S. N. Rai, Statistical Approach for Gene Set Analysis with Trait Specific Quantitative Trait Loci, Scientific Reports 8 (1) (2018) 2391. doi:10.1038/s41598-018-19736-w.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5799309/

[63] R. K. Barman, A. Mukhopadhyay, U. Maulik, S. Das, Identification of infectious disease-associated host genes using machine learning techniques, BMC Bioinformatics 20 (1)

(2019) 736.
URL https://pubmed.ncbi.nlm.nih.gov/31881961/

[64] M. A. Mooney, B. Wilmot, Gene set analysis: A step-by-step guide, American Journal of Medical Genetics Part B: Neuropsychiatric Genetics 168 (7) (2015) 517–527. doi:10.1002/ajmg.b.32328.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ajmg.b.323289

[65] P. F. Sullivan, D. Posthuma, Biological pathways and networks implicated in psychiatric disorders, Current Opinion in Behavioral Sciences 2 (2015) 58–68.
URL https://www.sciencedirect.com/science/article/pii/S2352154614000187

[66] J. I. Nurnberger, D. L. Koller, J. Jung, H. J. Edenberg, T. Foroud, I. Guella, M. P. Vawter, J. R. Kelsoe, Identification of Pathways for Bipolar Disorder: A Meta-analysis, JAMA Psychiatry 71 (6) (2014) 657.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4523227/

[67] S. Roy, R. Kumar, V. Mittal, D. Gupta, Classification models for Invasive Ductal Carcinoma Progression, based on gene expression data-trained supervised machine learning, Scientific Reports 10 (1) (2020) 4113.
URL https://www.nature.com/articles/s41598-020-60740-w

[68] Y. Ma, S. Sun, X. Shang, E. T. Keller, M. Chen, X. Zhou, Integrative differential expression and gene set enrichment analysis using summary statistics for scRNA-seq studies, Nature Communications 11 (1) (2020) 1585.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7101316/

[69] F. Garcia-Garcia, J. Panadero, J. Dopazo, D. Montaner, Integrated gene set analysis for microRNA studies, Bioinformatics 32 (18) (2016) 2809–2816.
URL https://pubmed.ncbi.nlm.nih.gov/27324197/

[70] E. Stovold, D. Beecher, R. Foxlee, A. Noel-Storr, Study flow diagrams in cochrane systematic review updates: an adapted prisma flow diagram, Systematic Reviews 3 (1) (2014) 1–5.

URL https://link.springer.com/article/10.1186/2046-4053-3-54

[71] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information Fusion 58 (2020) 82–115.
URL https://www.sciencedirect.com/science/article/pii/S1566253519308103

[72] S. Seifert, S. Gundlach, O. Junge, S. Szymczak, Integrating biological knowledge and gene expression data using pathway-guided random forests: a benchmarking study, Bioinformatics (Oxford, England) 36 (15) (2020) 4301––4308. doi:10.1093/bioinformatics/btaa483.
URL https://europepmc.org/articles/PMC7520048

[73] J. Parraga-Alava, M. Dorn, M. Inostroza-Ponta, A multi-objective gene clustering algorithm guided by apriori biological knowledge with intensification and diversification strategies, BioData Mining 11 (1) (2018) 16.
URL https://link.springer.com/article/10.1186/s13040-018-0178-4

[74] H. Fyad, F. Barigou, K. Bouamrane, B. Atmani, Obkml-go: Optimized clustering combination with biological knowledge for dna microarray expression data, International Journal of Computing and Digital Systems 10 (2020) 1–12.

[75] Y. Liang, F. Zhang, J. Wang, T. Joshi, Y. Wang, D. Xu, Prediction of Drought-Resistant Genes in Arabidopsis thaliana Using SVM-RFE, PLoS One 6 (7) (2011) e21750.
URL PredictionofDrought-Resistant89GenesinArabidopsisthalianaUsingSVM

[76] L. Shamseer, D. Moher, M. Clarke, D. Ghersi, A. Liberati, M. Petticrew, P. Shekelle, L. A. Stewart, Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015: elaboration and explanation, Bmj 349 (2015).

[77] B. H. W. Chang, W. Tian, GSA-Lightning: ultra-fast permutation-based gene set analysis, Bioinformatics 32 (19) (2016) 3029–3031.
URL https://pubmed.ncbi.nlm.nih.gov/27296982/

[78] L. Geistlinger, G. Csaba, M. Santarelli, M. Ramos, L. Schiffer, N. Turaga, C. Law, S. Davis, V. Carey, M. Morgan, R. Zimmer, L. Waldron, Toward a gold standard for benchmarking gene set enrichment analysis, Briefings in Bioinformatics (2020) bbz158.
URL https://pubmed.ncbi.nlm.nih.gov/32026945/

[79] F. Maleki, K. Ovens, D. J. Hogan, A. J. Kusalik, Gene Set Analysis: Challenges, Opportunities, and Future Research, Frontiers in Genetics 11 (2020) 654.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7339292/

[80] J. Hu, J.-Y. Tzeng, Integrative gene set analysis of multi-platform data with sample heterogeneity, Bioinformatics 30 (11) (2014) 1501–1507.

[81] J. H. Joly, W. E. Lowry, N. A. Graham, Differential Gene Set Enrichment Analysis: A statistical approach to quantify the relative enrichment of two gene sets, Bioinformatics (2020) btaa658.
URL https://pubmed.ncbi.nlm.nih.gov/32692836/

[82] Y. Rahmatallah, F. Emmert-Streib, G. Glazko, Gene Sets Net Correlations Analysis (GSNCA): a multivariate differential coexpression test for gene sets, Bioinformatics 30 (3) (2014) 360–368.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4023302/

[83] K.-L. Tiong, C.-H. Yeang, MGSEA – a multivariate Gene set enrichment analysis, BMC Bioinformatics 20 (1) (2019) 145.
URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2716-6

[84] J. Roder, B. Linstid, C. Oliveira, Improving the power of gene set enrichment analyses, BMC Bioinformatics 20 (1) (2019) 257.

[85] H. Meng, G. Yaari, C. R. Bolen, S. Avey, S. H. Kleinstein, Gene set meta-analysis with Quantitative Set Analysis for Gene Expression (QuSAGE), PLOS Computational Biology 15 (4) (2019) e1006899.
URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006899

[86] M. Neupane, J. N. Kiser, the Bovine Respiratory Disease Complex Coordinated Agricultural Project Research Team, H. L. Neibergs, Gene set enrichment analysis of SNP data in dairy and beef cattle with bovine respiratory disease, Animal Genetics 49 (6) (2018) 527–538.
URL https://pubmed.ncbi.nlm.nih.gov/30229962/

[87] Y. Kong, T. Yu, A Deep Neural Network Model using Random Forest to Extract Feature Representation for Gene Expression Data Classification, Scientific Reports 8 (1) (2018) 16477.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6220289/

[88] J. Arloth, G. Eraslan, T. F. M. Andlauer, J. Martins, S. Iurato, B. Kühnel, M. Waldenberger, J. Frank, R. Gold, B. Hemmer, F. Luessi, S. Nischwitz, F. Paul, H. Wiendl, C. Gieger, S. Heilmann-Heimbach, T. Kacprowski, M. Laudes, T. Meitinger, A. Peters, R. Rawal, K. Strauch, S. Lucae, B. Müller-Myhsok, M. Rietschel, F. J. Theis, E. B. Binder, N. S. Mueller, DeepWAS: Multivariate genotype-phenotype associations by directly integrating regulatory information using deep learning, PLOS Computational Biology 16 (2) (2020) e1007616.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7043350/

[89] A. Allahyar, J. Ubels, J. de Ridder, A data-driven interactome of synergistic genes improves network-based cancer outcome prediction, PLOS Computational Biology 15 (2) (2019) e1006657.
URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006657

[90] X.-H. Zhou, X.-Y. Chu, G. Xue, J.-H. Xiong, H.-Y. Zhang, Identifying cancer prognostic modules by module network analysis, BMC Bioinformatics 20 (1) (2019) 85.
URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2674-z

[91] D. Tong, Y. Tian, T. Zhou, Q. Ye, J. Li, K. Ding, J. Li, Improving prediction performance of colon cancer prognosis based on the integration of clinical and multi-omics data, BMC Medical Informatics and Decision Making 20 (1) (2020) 22.

URL https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1043-1

[92] A. R. Soltis, C. L. Dalgard, H. B. Pollard, M. D. Wilkerson, MutEnricher: a flexible toolset for somatic mutation enrichment analysis of tumor whole genomes, BMC Bioinformatics 21 (1) (2020) 338.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7393734/

[93] Z. Xin, Y. Cai, L. T. Dang, H. M. Burke, J. Revote, H. T. Nim, Y.-F. Li, M. Ramialison, MonaGO: a novel Gene Ontology enrichment analysis visualisation system, preprint, Bioinformatics (Sep. 2020). doi:10.1101/2020.09.27.316067.
URL https://www.biorxiv.org/content/10.1101/2020.09.27.316067v1.abstract

[94] W. Walter, F. Sánchez-Cabo, M. Ricote, GOplot: an R package for visually combining expression data with functional analysis: Fig. 1., Bioinformatics 31 (17) (2015) 2912–2914.
URL https://academic.oup.com/bioinformatics/article-abstract/31/17/2912/184136

[95] S. X. Ge, D. Jung, R. Yao, ShinyGO: a graphical gene-set enrichment tool for animals and plants, Bioinformatics 36 (8) (2020) 2628–2629.
URL https://academic.oup.com/bioinformatics/article-abstract/36/8/2628/5688742

[96] D. Yusuf, J. Lim, W. Wasserman, The Gene Set Builder: collation, curation, and distribution of sets of genes, BMC Bioinformatics 6 (1) (2005) 305.
URL https://link.springer.com/article/10.1186/1471-2105-6-305

[97] K. Rho, B. Kim, Y. Jang, S. Lee, T. Bae, J. Seo, C. Seo, J. Lee, H. Kang, U. Yu, S. Kim, S. Lee, W. K. Kim, GARNET – gene set analysis with exploration of annotation relations, BMC Bioinformatics 12 (Suppl 1) (2011) S25.
URL https://link.springer.com/article/10.1186/1471-2105-12-S1-S25

[98] H. Kang, I. Choi, S. Cho, D. Ryu, S. Lee, W. Kim, gsGator: an integrated web platform for cross-species gene set analysis, BMC Bioinformatics 15 (1) (2014) 13.

[99] E. Ewing, N. Planell-Picola, M. Jagodic, D. Gomez-Cabrero, GeneSetCluster: a tool for summarizing and integrating gene-set analysis results, BMC Bioinformatics 21 (1) (2020) 443.
URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03784-z

[100] G. Wang, D.-H. Oh, M. Dassanayake, GOMCL: a toolkit to cluster, evaluate, and extract non-redundant associations of Gene Ontology-based functions, BMC Bioinformatics 21 (1) (2020) 139.
URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-3447-4

[101] E. Ong, P. Sun, K. Berke, J. Zheng, G. Wu, Y. He, VIO: ontology classification and study of vaccine responses given various experimental and analytical conditions, BMC Bioinformatics 20 (S21) (2019) 704.
URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3194-6

[102] P. Perampalam, F. A. Dick, BEAVR: a browser-based tool for the exploration and visualization of RNA-seq data, BMC Bioinformatics 21 (1) (2020) 221.
URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03549-8

[103] A. L. P. Reyes, T. C. Silva, S. G. Coetzee, J. T. Plummer, B. D. Davis, S. Chen, D. J. Hazelett, K. Lawrenson, B. P. Berman, S. A. Gayther, M. R. Jones, GENAVi: a shiny web application for gene expression normalization, analysis and visualization, BMC Genomics 20 (1) (2019) 745.
URL https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-6073-7

[104] A. Yousif, N. Drou, J. Rowe, M. Khalfan, K. C. Gunsalus, NASQAR: a web-based platform for high-throughput sequencing data analysis and visualization, BMC Bioinformatics 21 (1) (2020) 267.

URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03577-4

[105] J. Zhu, Q. Zhao, E. Katsevich, C. Sabatti, Exploratory Gene Ontology Analysis with Interactive Visualization, Scientific Reports 9 (1) (2019) 7793.
URL https://www.nature.com/articles/s41598-019-42178-x

[106] D. Netanely, N. Stern, I. Laufer, R. Shamir, PROMO: an interactive tool for analyzing clinically-labeled multi-omic cancer datasets, BMC Bioinformatics 20 (1) (2019) 732.
URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3142-5

[107] K. Charmpi, B. Ycart, Weighted kolmogorov smirnov testing: an alternative for gene set enrichment analysis, Statistical Applications in Genetics and Molecular Biology 14 (3) (2015) 279–293.
URL https://www.degruyter.com/view/journals/sagmb/14/3/article-p279.xml

[108] Y. Benjamini, Discovering the false discovery rate, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72 (4) (2010) 405–416.
URL https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1467-9868.2010.00746.x

[109] J. Maksimovic, A. Oshlack, B. Phipson, Gene set enrichment analysis for genome-wide DNA methylation data, preprint, Bioinformatics (Aug. 2020).
URL https://www.biorxiv.org/content/10.1101/2020.08.24.265702v1.abstract

[110] L. Geistlinger, G. Csaba, M. Santarelli, M. Ramos, L. Schiffer, N. Turaga, C. Law, S. Davis, V. Carey, M. Morgan, R. Zimmer, L. Waldron, Toward a gold standard for benchmarking gene set enrichment analysis, Briefings in Bioinformatics 22 (1) (2021) 545–556.
URL https://academic.oup.com/bib/article-abstract/22/1/545/5722384

[111] E. Khodayari Moez, M. Hajihosseini, J. L. Andrews, I. Dinu, Longitudinal linear combination test for gene set analysis, BMC Bioinformatics 20 (1) (2019) 650.

URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3221-7

[112] J. Reimand, R. Isserlin, V. Voisin, M. Kucera, C. Tannus-Lopes, A. Rostamianfar, L. Wadi, M. Meyer, J. Wong, C. Xu, D. Merico, G. D. Bader, Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap, Nature Protocols 14 (2) (2019) 482–517.
URL https://pubmed.ncbi.nlm.nih.gov/30664679/

[113] J. Peng, G. Lu, H. Xue, T. Wang, X. Shang, TS-GOEA: a web tool for tissue-specific gene set enrichment analysis based on gene ontology, BMC Bioinformatics 20 (S18) (2019) 572.
URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3125-6

[114] L. Zhang, S. Gu, Y. Liu, B. Wang, F. Azuaje, Gene set analysis in the cloud, Bioinformatics 28 (2) (2012) 294–295.
URL https://academic.oup.com/bioinformatics/article/28/2/294/197370

[115] M. A. Mooney, B. Wilmot, Gene set analysis: A step-by-step guide, American Journal of Medical Genetics Part B: Neuropsychiatric Genetics 168 (7) (2015) 517–527.
URL https://pubmed.ncbi.nlm.nih.gov/26059482/

[116] P. F. Sullivan, D. Posthuma, Biological pathways and networks implicated in psychiatric disorders, Current Opinion in Behavioral Sciences 2 (2015) 58–68.
URL https://www.sciencedirect.com/science/article/pii/S2352154614000187

[117] J. I. Nurnberger, D. L. Koller, J. Jung, H. J. Edenberg, T. Foroud, I. Guella, M. P. Vawter, J. R. Kelsoe, Identification of pathways for bipolar disorder: a meta-analysis, JAMA Psychiatry 71 (6) (2014) 657–664.
URL https://jamanetwork.com/journals/jamapsychiatry/article-abstract/1859133

[118] K. Wang, H. Zhang, S. Kugathasan, V. Annese, J. P. Bradfield, R. K. Russell, P. M. Sleiman, M. Imielinski, J. Glessner, C. Hou, et al., Diverse genome-wide association

studies associate the il12/il23 pathway with crohn disease, The American Journal of Human Genetics 84 (3) (2009) 399–405.
URL https://www.sciencedirect.com/science/article/pii/S0002929709000652

[119] H. Eleftherohorinou, C. J. Hoggart, V. J. Wright, M. Levin, L. J. Coin, Pathway-driven gene stability selection of two rheumatoid arthritis gwas identifies and validates new susceptibility genes in receptor mediated signalling pathways, Human Molecular Genetics 20 (17) (2011) 3494–3506.
URL https://academic.oup.com/hmg/article/20/17/3494/2527070?login=true

[120] I. Menashe, D. Maeder, M. Garcia-Closas, J. D. Figueroa, S. Bhattacharjee, M. Rotunno, P. Kraft, D. J. Hunter, S. J. Chanock, P. S. Rosenberg, et al., Pathway analysis of breast cancer genome-wide association study highlights three pathways and one canonical signaling cascade, Cancer Research 70 (11) (2010) 4453–4459.
URL https://cancerres.aacrjournals.org/content/70/11/4453.short

[121] A. E. Locke, B. Kahali, S. I. Berndt, A. E. Justice, T. H. Pers, F. R. Day, C. Powell, S. Vedantam, M. L. Buchkovich, J. Yang, et al., Genetic studies of body mass index yield new insights for obesity biology, Nature 518 (7538) (2015) 197–206.
URL https://www.nature.com/articles/nature14177?report=reader

[122] P. Khatri, M. Sirota, A. J. Butte, Ten years of pathway analysis: current approaches and outstanding challenges, PLoS Comput Biol 8 (2) (2012) e1002375.
URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002375

[123] D. W. Huang, B. T. Sherman, R. A. Lempicki, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists, Nucleic Acids Research 37 (1) (2009) 1–13.
URL https://academic.oup.com/nar/article/37/1/1/1026684

[124] D. Kang, J. Lee, S. Choi, K. Kim, An ontology-based enterprise architecture, Expert Systems with Applications 37 (2) (2010) 1456–1464.

[125] M. Matthen, Greek ontology and the'is' of truth, Phronesis (1983) 113–135.

[126] Y. Zhao, J. Wang, J. Chen, X. Zhang, M. Guo, G. Yu, A Literature Review of Gene Function Prediction by Modeling Gene Ontology, Frontiers in Genetics 11 (2020) 400.

[127] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, Gene Ontology: tool for the unification of biology, Nature Genetics 25 (1) (2000) 25–29.

[128] O. G. Troyanskaya, K. Dolinski, A. B. Owen, R. B. Altman, D. Botstein, A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae), Proceedings of the National Academy of Sciences of the United States of America 100 (14) (2003) 8348–8353.

[129] U. Karaoz, T. M. Murali, S. Letovsky, Y. Zheng, C. Ding, C. R. Cantor, S. Kasif, Whole-genome annotation by using evidence integration in functional-linkage networks, Proceedings of the National Academy of Sciences 101 (9) (2004) 2888–2893.

[130] F. Seyednasrollah, A. Laiho, L. L. Elo, Comparison of software packages for detecting differential expression in RNA-seq studies, Briefings in Bioinformatics 16 (1) (2015) 59–70.

[131] A. M. Leslie, O. Friedman, T. P. German, Core mechanisms in ‘theory of mind’, Trends in Cognitive Sciences 8 (12) (2004) 528–533.

[132] Y. Lin, K. Golovnina, Z.-X. Chen, H. N. Lee, Y. L. S. Negron, H. Sultana, B. Oliver, S. T. Harbison, Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual Drosophila melanogaster, BMC Genomics 17 (2016) 28.

[133] O. Folger, L. Jerby, C. Frezza, E. Gottlieb, E. Ruppin, T. Shlomi, Predicting selective drug targets in cancer through metabolic networks, Molecular Systems Biology 7 (1) (2011) 501.

[134] L. M. Heiser, A. Sadanandam, W.-L. Kuo, S. C. Benz, T. C. Goldstein, S. Ng, W. J. Gibb, N. J. Wang, S. Ziyad, F. Tong, N. Bayani, Z. Hu, J. I. Billig, A. Dueregger, S. Lewis, L. Jakkula, J. E. Korkola, S. Durinck, F. Pepin, Y. Guan, E. Purdom, P. Neuvial, H. Bengtsson, K. W. Wood, P. G. Smith, L. T. Vassilev, B. T. Hennessy, J. Greshock, K. E. Bachman, M. A. Hardwicke, J. W. Park, L. J. Marton, D. M. Wolf, E. A. Collisson, R. M. Neve, G. B. Mills, T. P. Speed, H. S. Feiler, R. F. Wooster, D. Haussler, J. M. Stuart, J. W. Gray, P. T. Spellman, Subtype and pathway specific responses to anticancer compounds in breast cancer, Proceedings of the National Academy of Sciences 109 (8) (2012) 2724–2729.

[135] D. W. Huang, B. T. Sherman, Q. Tan, J. Kir, D. Liu, D. Bryant, Y. Guo, R. Stephens, M. W. Baseler, H. C. Lane, R. A. Lempicki, DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists, Nucleic Acids Research 35 (suppl_2) (2007) W169–W175.
URL https://academic.oup.com/nar/article/35/suppl_2/W169/2924156?login=true

[136] K. D. Dahlquist, N. Salomonis, K. Vranizan, S. C. Lawlor, B. R. Conklin, GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways, Nature Genetics 31 (1) (2002) 19–20.
URL https://www.nature.com/articles/ng0502-19

[137] B. R. Zeeberg, W. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine, S. Narasimhan, D. W. Kane, W. C. Reinhold, S. Lababidi, K. J. Bussey, J. Riss, J. Barrett, J. N. Weinstein, GoMiner: a resource for biological interpretation of genomic and proteomic data, Genome Biology 4 (4) (2003) R28.
URL https://genomebiology.biomedcentral.com/articles/10.1186/gb-2003-4-4-r28

[138] F. Al-Shahrour, R. Diaz-Uriarte, J. Dopazo, FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes, Bioinformatics 20 (4) (2004) 578–580.
URL https://academic.oup.com/bioinformatics/article/20/4/578/192603?login=true

[139] D. Martin, C. Brun, E. Remy, P. Mouren, D. Thieffry, B. Jacq, GOToolBox: functional analysis of gene datasets based on Gene Ontology, Genome Biology 5 (12) (2004) R101. URL https://genomebiology.biomedcentral.com/articles/10.1186/gb-2004-5-12-r101

[140] C. I. Castillo-Davis, D. L. Hartl, GeneMerge–post-genomic analysis, data mining, and hypothesis testing, Bioinformatics 19 (7) (2003) 891–892. URL https://academic.oup.com/bioinformatics/article/19/7/891/197817?login=true

[141] Q. Zheng, X.-J. Wang, GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis, Nucleic Acids Research 36 (suppl_2) (2008) W358–W363. URL https://academic.oup.com/nar/article/36/suppl_2/W358/2506335?login=true

[142] G. Bindea, B. Mlecnik, H. Hackl, P. Charoentong, M. Tosolini, A. Kirilovsky, W.-H. Fridman, F. Pagès, Z. Trajanoski, J. Galon, ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks, Bioinformatics 25 (8) (2009) 1091–1093. URL https://academic.oup.com/bioinformatics/article/25/8/1091/324247?login=true

[143] M. D. Robinson, J. Grigull, N. Mohammad, T. R. Hughes, FunSpec: a web-based cluster interpreter for yeast, BMC Bioinformatics 3 (1) (2002) 35. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-3-35

[144] L. A. Martinez-Cruz, A. Rubio, M. L. Martinez-Chantar, A. Labarga, I. Barrio, A. Podhorski, V. Segura, J. L. Sevilla Campo, M. A. Avila, J. M. Mato, GARBAN: genomic analysis and rapid biological annotation of cDNA microarray and proteomic data, Bioinformatics 19 (16) (2003) 2158–2160. URL https://academic.oup.com/bioinformatics/article/19/16/2158/242574?login=true

[145] J. Wang, D. Duncan, Z. Shi, B. Zhang, WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013, Nucleic Acids Research 41 (W1) (2013) W77–W83.

URL https://academic.oup.com/nar/article/41/W1/W77/1105552?login=true

[146] H. Sun, H. Fang, T. Chen, R. Perkins, W. Tong, GOFFA: Gene Ontology For Functional Analysis – A FDA Gene Ontology Tool for Analysis of Genomic and Proteomic Data, BMC Bioinformatics 7 (Suppl 2) (2006) S23.
URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-S2-S23

[147] J. Ye, L. Fang, H. Zheng, Y. Zhang, J. Chen, Z. Zhang, J. Wang, S. Li, R. Li, L. Bolund, J. Wang, WEGO: a web tool for plotting GO annotations, Nucleic Acids Research 34 (Web Server) (2006) W293–W297.
URL https://academic.oup.com/nar/article/34/suppl_2/W293/2505468?login=true

[148] B. Zhang, D. Schmoyer, S. Kirov, J. Snoddy, GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies, BMC Bioinformatics 5 (1) (2004) 16.
URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-5-16

[149] X. Wu, M. A. Hasan, J. Y. Chen, Pathway and network analysis in proteomics, Journal of Theoretical Biology 362 (2014) 44–52.
URL https://www.sciencedirect.com/science/article/abs/pii/S002251931400304X

[150] M. Yi, J. Horton, J. Cohen, H. Hobbs, R. Stephens, WholePathwayScope: a comprehensive pathway-based analysis tool for high-throughput data, BMC Bioinformatics 7 (1) (2006) 30.
URL https://link.springer.com/article/10.1186/1471-2105-7-30

[151] M. A. Newton, F. A. Quintana, J. A. den Boon, S. Sengupta, P. Ahlquist, Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis, The Annals of Applied Statistics 1 (1) (2007) 85–106.
URL https://projecteuclid.org/euclid.aoas/1183143730

[152] W. Cao, Y. Li, D. Liu, C. Chen, Y. Xu, Statistical and Biological Evaluation of Different Gene Set Analysis Methods, Procedia Environmental Sciences 8 (2011) 693–699.
URL https://www.sciencedirect.com/science/article/pii/S1878029611007420

[153] I. Dinu, J. D. Potter, T. Mueller, Q. Liu, A. J. Adewale, G. S. Jhangri, G. Einecke, K. S. Famulski, P. Halloran, Y. Yasui, Improving gene set analysis of microarray data by SAM-GS, BMC Bioinformatics 8 (1) (2007) 242.
URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-242

[154] G. K. Smyth, limma: Linear Models for Microarray Data, in: R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, S. Dudoit (Eds.), Bioinformatics and Computational Biology Solutions Using R and Bioconductor, Springer-Verlag, New York, 2005, pp. 397–420.
URL https://link.springer.com/chapter/10.1007/0-387-29362-0_23

[155] T. Breslin, P. Edén, M. Krogh, Comparing functional annotation analyses with Catmap, BMC Bioinformatics 5 (1) (2004) 193.
URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-5-193

[156] A. Boorsma, B. C. Foat, D. Vis, F. Klis, H. J. Bussemaker, T-profiler: scoring the activity of predefined groups of genes using gene expression data, Nucleic Acids Research 33 (Web Server) (2005) W592–W595.
URL https://academic.oup.com/nar/article/33/suppl_2/W592/2505704?login=true

[157] C. Henegar, R. Cancello, S. Rome, H. Vidal, K. Clément, J.-D. Zucker, Clustering biological annotations and gene expression data to identify putatively co-regulated biological processes, Journal of Bioinformatics and Computational Biology 04 (04) (2006) 833–852.
URL https://www.worldscientific.com/doi/abs/10.1142/s0219720006002181

[158] C. Backes, A. Keller, J. Kuentzer, B. Kneissl, N. Comtesse, Y. A. Elnakady, R. Muller,

E. Meese, H.-P. Lenhof, GeneTrail–advanced gene set enrichment analysis, Nucleic Acids Research 35 (Web Server) (2007) W186–W192. doi:10.1093/nar/gkm323.
URL https://academic.oup.com/nar/article/35/suppl_2/W186/2923179?login=true

[159] S.-B. Kim, S. Yang, S.-K. Kim, S. C. Kim, H. G. Woo, D. J. Volsky, S.-Y. Kim, I.-S. Chu, GAzer: gene set analyzer, Bioinformatics 23 (13) (2007) 1697–1699.
URL https://academic.oup.com/bioinformatics/article/23/13/1697/221402?login=true

[160] D. Wu, G. K. Smyth, Camera: a competitive gene set test accounting for inter-gene correlation, Nucleic Acids Research 40 (17) (2012) e133–e133.
URL https://academic.oup.com/nar/article/40/17/e133/2411151?login=true

[161] W. Luo, M. S. Friedman, K. Shedden, K. D. Hankenson, P. J. Woolf, GAGE: generally applicable gene set enrichment for pathway analysis, BMC Bioinformatics 10 (1) (2009) 161.
URL https://link.springer.com/article/10.1186/1471-2105-10-161

[162] H. R. Frost, Z. Li, J. H. Moore, Spectral gene set enrichment (SGSE), BMC Bioinformatics 16 (1) (2015) 70.
URL https://link.springer.com/article/10.1186/s12859-015-0490-7

[163] H.-M. Hsueh, C.-A. Tsai, Gene set analysis using sufficient dimension reduction, BMC Bioinformatics 17 (1) (2016) 74.
URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-0928-6

[164] M. Reich, T. Liefeld, J. Gould, J. Lerner, P. Tamayo, J. P. Mesirov, GenePattern 2.0, Nature Genetics 38 (5) (2006) 500–501.
URL https://www.nature.com/articles/ng0506-500

[165] J. Rahnenführer, F. S. Domingues, J. Maydt, T. Lengauer, Calculating the Statistical Significance of Changes in Pathway Activity From Gene Expression Data, Statistical

Applications in Genetics and Molecular Biology 3 (1) (2004) 1–29.
URL https://www.degruyter.com/document/doi/10.2202/1544-6115.1055/html

[166] A. L. Tarca, S. Draghici, P. Khatri, S. S. Hassan, P. Mittal, J.-s. Kim, C. J. Kim, J. P. Kusanovic, R. Romero, A novel signaling pathway impact analysis, Bioinformatics 25 (1) (2009) 75–82.

[167] A. Alexeyenko, W. Lee, M. Pernemalm, J. Guegan, P. Dessen, V. Lazar, J. Lehtiö, Y. Pawitan, Network enrichment analysis: extension of gene-set enrichment analysis to gene networks, BMC Bioinformatics 13 (1) (2012) 226.
URL https://link.springer.com/article/10.1186/1471-2105-13-226

[168] E. Glaab, A. Baudot, N. Krasnogor, A. Valencia, TopoGSA: network topological gene set analysis, Bioinformatics 26 (9) (2010) 1271–1272.
URL https://academic.oup.com/bioinformatics/article/26/9/1271/201632?view=extract

[169] P. Martini, G. Sales, M. S. Massa, M. Chiogna, C. Romualdi, Along signal paths: an empirical gene set approach exploiting pathway topology, Nucleic Acids Research 41 (1) (2013) e19–e19.

[170] M. Schena, D. Shalon, R. W. Davis, P. O. Brown, Quantitative monitoring of gene expression patterns with a complementary dna microarray, Science 270 (5235) (1995) 467–470.

[171] D. J. Lockhart, E. A. Winzeler, Genomics, gene expression and dna arrays, Nature 405 (6788) (2000) 827–836.

[172] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, G. K. Smyth, limma powers differential expression analyses for rna-sequencing and microarray studies, Nucleic Acids Research 43 (7) (2015) e47–e47.

[173] S. Ghosh, C.-K. Chan, Analysis of rna-seq data using tophat and cufflinks, in: Methods in Molecular Biology, Vol. 1374, Springer, 2016, pp. 339–361.

[174] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by rna-seq, Nature Methods 5 (7) (2008) 621–628.

[175] Z. Wang, M. Gerstein, M. Snyder, Rna-seq: a revolutionary tool for transcriptomics, Nature Reviews Genetics 10 (1) (2009) 57–63.

[176] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, Y. Gilad, Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays, Genome Research 18 (9) (2008) 1509–1517.

[177] A. Kanitz, F. Gypas, A. J. Gruber, A. R. Gruber, L. Martin, M. Zavolan, Comparative assessment of methods for the computational inference of transcript isoform abundance from rna-seq data, Genome Biology 16 (1) (2015) 1–28.

[178] F. Ozsolak, P. M. Milos, Rna sequencing: advances, challenges and opportunities, Nature Reviews Genetics 12 (2) (2011) 87–98.

[179] D. J. Lockhart, E. A. Winzeler, Expression monitoring by hybridization to high-density oligonucleotide arrays, Nature Biotechnology 18 (12) (2000) 1315–1317.

[180] Z. Wang, M. Gerstein, M. Snyder, Rna-seq: a revolutionary tool for transcriptomics, Nature Reviews Genetics 10 (1) (2009) 57–63.

[181] M. Schena, D. Shalon, R. W. Davis, P. O. Brown, Quantitative monitoring of gene expression patterns with a complementary dna microarray, Science 270 (5235) (1995) 467–470.

[182] F. Ozsolak, P. M. Milos, Rna sequencing: advances, challenges and opportunities, Nature reviews genetics 12 (2) (2011) 87–98.

[183] M. Barnes, J. Freudenberg, S. Thompson, B. Aronow, P. Pavlidis, Experimental comparison and cross-validation of the affymetrix and illumina gene expression analysis platforms, Nucleic acids research 33 (18) (2005) 5914–5923.

[184] Z. Wang, M. Gerstein, M. Snyder, Rna-seq: a revolutionary tool for transcriptomics, Nature reviews genetics 10 (1) (2009) 57–63.

[185] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, Y. Gilad, Rna-seq: An assessment of technical reproducibility and comparison with gene expression arrays, Genome Research 18 (9) (2008) 1509–1517.

[186] J. Pirrello, M. Bourdon, C. Cheniclet, M. Bourge, S. Brown, J.-P. Renaudin, N. Frangne, C. Chevalier, Transcriptome profiling of sorted endoreduplicated nuclei from tomato fruits: how the global shift in expression ascribed to dna ploidy influences rna-seq data normalization and interpretation, Plant Journal 77 (3) (2014) 362–374.

[187] S.-I. Consortium, A comprehensive assessment of rna-seq accuracy, reproducibility and information content by the sequencing quality control consortium, Nature Biotechnology 32 (9) (2014) 903–914.

[188] The microarray quality control (maqc) project shows inter-and intraplatform reproducibility of gene expression measurements, Nature biotechnology 24 (9) (2006) 1151–1161.

[189] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, A. Mortazavi, A survey of best practices for rna-seq data analysis, Genome Biology 17 (1) (2016) 13.

[190] J. Quackenbush, Microarray data normalization and transformation, Nature genetics 32 (2002) 496–501.

[191] D. Van Dijk, R. Sharma, J. Nainys, K. Yim, P. Kathail, A. J. Carr, C. Burdziak, K. R. Moon, C. L. Chaffer, D. Pattabiraman, B. Bierie, L. Mazutis, G. Wolf, S. Krishnaswamy, D. Pe'er, The promises and challenges of spatial transcriptomics, Nature Biotechnology 36 (4) (2018) 310–321.

[192] T. Speed, Statistical analysis of gene expression microarray data, CRC press (2002).

[193] M. T. Zimmermann, The importance of biologic knowledge and gene expression context for genomic data interpretation, Frontiers in Genetics 9 (2018) 670.
URL https://www.frontiersin.org/articles/10.3389/fgene.2018.00670/full

[194] L. Tian, S. A. Greenberg, S. W. Kong, J. Altschuler, I. S. Kohane, P. J. Park, Discovering statistically significant pathways in expression profiling studies, Proceedings of the

National Academy of Sciences 102 (38) (2005) 13544–13549.
URL https://www.pnas.org/content/102/38/13544

[195] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.
URL https://link.springer.com/article/10.1023/A:1010933404324

[196] X. Chen, L. Wang, Integrating biological knowledge with gene expression profiles for survival prediction of cancer, Journal of Computational Biology 16 (2) (2009) 265–278.
URL https://www.liebertpub.com/doi/abs/10.1089/cmb.2008.12TT

[197] M. Reboiro-Jato, R. Laza, H. López-Fernández, D. Glez-Peña, F. Díaz, F. Fdez-Riverola, genensemble: A new model for the combination of classifiers and integration of biological knowledge applied to genomic data, Expert Systems with Applications 40 (1) (2013) 52–63.
URL https://www.sciencedirect.com/science/article/abs/pii/S0957417412008585

[198] N. Bandyopadhyay, T. Kahveci, S. Goodison, Y. Sun, S. Ranka, Pathway-based feature selection algorithm for cancer microarray data, Advances in Bioinformatics 2009 (2009).
URL https://www.hindawi.com/journals/abi/2009/532989/

[199] S. Kim, M. Kon, C. DeLisi, Pathway-based classification of cancer subtypes, Biology Direct 7 (1) (2012) 21.
URL https://biologydirect.biomedcentral.com/articles/10.1186/1745-6150-7-21

[200] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: Nsga-ii, IEEE transactions on Evolutionary Computation 6 (2) (2002) 182–197.
URL https://ieeexplore.ieee.org/abstract/document/996017

[201] F. Glover, Tabu search and adaptive memory programming—advances, applications and challenges, in: Interfaces in Computer Science and Operations Research, Springer, 1997, pp. 1–75.

URL https://link.springer.com/chapter/10.1007/978-1-4615-4102-8_1

[202] J. Dubois-Lacoste, M. López-Ibáñez, T. Stützle, Anytime pareto local search, European Journal of Operational Research 243 (2) (2015) 369–385.
URL https://www.sciencedirect.com/science/article/abs/pii/S0377221714009011

[203] H. Cui, C. Zhou, X. Dai, Y. Liang, R. Paffenroth, D. Korkin, Boosting gene expression clustering with system-wide biological information: a robust autoencoder approach, International Journal of Computational Biology and Drug Design 13 (1) (2020) 98–123.
URL https://www.inderscience.com/info/inarticle.php?artid=105113

[204] C. Zhou, R. C. Paffenroth, Anomaly detection with robust deep autoencoders, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 665–674.
URL https://dl.acm.org/doi/10.1145/3097983.3098052

[205] S. Bauer, J. Gagneur, P. N. Robinson, Going bayesian: model-based gene set analysis of genome-scale data, Nucleic Acids Research 38 (11) (2010) 3523–3532.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2887944/

[206] A. P. Oron, Z. Jiang, R. Gentleman, Gene set enrichment analysis using linear models and diagnostics, Bioinformatics 24 (22) (2008) 2586–2591.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2579710/

[207] M. A. Sartor, G. D. Leikauf, M. Medvedovic, Lrpath: a logistic regression approach for identifying enriched biological groups in gene expression data, Bioinformatics 25 (2) (2009) 211–217.
URL https://academic.oup.com/bioinformatics/article/25/2/211/218259

[208] A. Fagan, A. C. Culhane, D. G. Higgins, A multivariate analysis approach to the integration of proteomic and gene expression data, Proteomics 7 (13) (2007) 2162–2171.
URL https://pubmed.ncbi.nlm.nih.gov/17549791/

[209] C. H. Busold, S. Winter, N. Hauser, A. Bauer, J. Dippon, J. D. Hoheisel, K. Fellenberg, Integration of go annotations in correspondence analysis: facilitating the interpretation of microarray data, Bioinformatics 21 (10) (2005) 2424–2429.
URL https://academic.oup.com/bioinformatics/article/21/10/2424/208235

[210] M. Verbanck, S. Lê, J. Pagès, A new unsupervised gene clustering algorithm based on the integration of biological knowledge into expression data, BMC Bioinformatics 14 (1) (2013) 42.
URL https://link.springer.com/article/10.1186/1471-2105-14-42

[211] M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, D. Haussler, Knowledge-based analysis of microarray gene expression data by using support vector machines, Proceedings of the National Academy of Sciences 97 (1) (2000) 262–267.
URL https://www.pnas.org/content/97/1/262.short

[212] J. A. Nepomuceno, A. Troncoso, I. A. Nepomuceno-Chamorro, J. S. Aguilar-Ruiz, Integrating biological knowledge based on functional annotations for biclustering of gene expression data, Computer Methods and Programs in Biomedicine 119 (3) (2015) 163–180.
URL https://www.sciencedirect.com/science/article/abs/pii/S0169260715000450

[213] J. A. Nepomuceno, A. Troncoso, J. S. Aguilar-Ruiz, Biclustering of gene expression data by correlation-based scatter search, BioData Mining 4 (1) (2011) 3.
URL https://link.springer.com/article/10.1186/1756-0381-4-3

[214] S. Imoto, T. Higuchi, T. Goto, K. Tashiro, S. Kuhara, S. Miyano, Combining microarrays and biological knowledge for estimating gene networks via bayesian networks, Journal of Bioinformatics and Computational Biology 2 (01) (2004) 77–98.
URL https://www.worldscientific.com/doi/abs/10.1142/S021972000400048X

[215] X. Gan, A. W.-C. Liew, H. Yan, Microarray missing data imputation based on a set theoretic framework and biological knowledge, Nucleic Acids Research 34 (5) (2006) 1608–1619.

[216] T. H. Bø, B. Dysvik, I. Jonassen, Lsimpute: accurate estimation of missing values in microarray data with least squares methods, Nucleic Acids Research 32 (3) (2004) e34–e34.
URL https://academic.oup.com/nar/article/32/3/e34/2904603

[217] S. W. Kong, W. T. Pu, P. J. Park, A multivariate approach for integrating genome-wide expression data and biological knowledge, Bioinformatics 22 (19) (2006) 2373–2380.
URL https://pubmed.ncbi.nlm.nih.gov/16877751/

[218] Y. Lu, P.-Y. Liu, P. Xiao, H.-W. Deng, Hotelling's t 2 multivariate profiling for detecting differential expression in microarrays, Bioinformatics 21 (14) (2005) 3105–3113.
URL https://pubmed.ncbi.nlm.nih.gov/15905280/

[219] B. S. Kim, I. Kim, S. Lee, S. Kim, S. Y. Rha, H. C. Chung, Statistical methods of translating microarray data into clinically relevant diagnostic information in colorectal cancer, Bioinformatics 21 (4) (2005) 517–528.
URL https://academic.oup.com/bioinformatics/article/21/4/517/203305

[220] F. G. Kuruvilla, P. J. Park, S. L. Schreiber, Vector algebra in the analysis of genome-wide expression data, Genome Biology 3 (3) (2002) 1–11.
URL https://genomebiology.biomedcentral.com/articles/10.1186/gb-2002-3-3-research0011

[221] A. Szabo, K. Boucher, D. Jones, A. D. Tsodikov, L. B. Klebanov, A. Y. Yakovlev, Multivariate exploratory tools for microarray data analysis, Biostatistics 4 (4) (2003) 555–567.
URL https://pubmed.ncbi.nlm.nih.gov/14557111/

[222] A. C. Culhane, G. Perriere, E. C. Considine, T. G. Cotter, D. G. Higgins, Between-group analysis of microarray data, Bioinformatics 18 (12) (2002) 1600–1608.
URL https://pubmed.ncbi.nlm.nih.gov/12490444/

[223] T. Schlitt, A. Brazma, Modelling gene networks at different organisational levels, FEBS Letters 579 (8) (2005) 1859–1866.
URL https://www.sciencedirect.com/science/article/pii/S0014579305001869

[224] H. De Jong, Modeling and simulation of genetic regulatory systems: a literature review, Journal of Computational Biology 9 (1) (2002) 67–103.
URL https://www.liebertpub.com/doi/abs/10.1089/10665270252833208

[225] Z. Bar-Joseph, Analyzing time series gene expression data, Bioinformatics 20 (16) (2004) 2493–2503.
URL https://academic.oup.com/bioinformatics/article/20/16/2493/236434

[226] S. Liang, S. Fuhrman, R. Somogyi, et al., Reveal, a general reverse engineering algorithm for inference of genetic network architectures, in: Pacific Symposium on Biocomputing, Vol. 3, 1998, pp. 18–29.
URL https://pubmed.ncbi.nlm.nih.gov/9697168/

[227] P. D'haeseleer, S. Liang, R. Somogyi, Genetic network inference: from co-expression clustering to reverse engineering, Bioinformatics 16 (8) (2000) 707–726.
URL https://academic.oup.com/bioinformatics/article/16/8/707/190286

[228] N. Friedman, Inferring cellular networks using probabilistic graphical models, Science 303 (5659) (2004) 799–805.
URL https://science.sciencemag.org/content/303/5659/799.abstract

[229] P. Sebastiani, M. Abad, M. F. Ramoni, Bayesian networks for genomic analysis, Genomic Signal Processing and Statistics 2 (2005) 281–320.
URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.104.3373&rep=rep1&type=pdf

[230] B. Xing, M. J. Van Der Laan, A causal inference approach for constructing transcriptional regulatory networks, Bioinformatics 21 (21) (2005) 4007–4013.

URL https://academic.oup.com/bioinformatics/article/21/21/4007/226730

[231] Y. Wang, T. Joshi, X.-S. Zhang, D. Xu, L. Chen, Inferring gene regulatory networks from multiple microarray datasets, Bioinformatics 22 (19) (2006) 2413–2420.
URL https://academic.oup.com/bioinformatics/article/22/19/2413/240982

[232] S. Li, L. Wu, Z. Zhang, Constructing biological networks through combined literature mining and microarray analysis: a lmma approach, Bioinformatics 22 (17) (2006) 2143–2150.
URL https://academic.oup.com/bioinformatics/article/22/17/2143/274919

[233] G. F. Cooper, E. Herskovits, A bayesian method for the induction of probabilistic networks from data, Machine Learning 9 (4) (1992) 309–347.
URL https://link.springer.com/article/10.1007/BF00994110

[234] P. P. Le, A. Bahl, L. H. Ungar, Using prior knowledge to improve genetic network reconstruction from microarray data, In Silico Biology 4 (3) (2004) 335–353.
URL https://content.iospress.com/articles/in-silico-biology/isb00137

[235] N. Nariai, Y. Tamada, S. Imoto, S. Miyano, Estimating gene regulatory networks and protein–protein interactions of saccharomyces cerevisiae from multiple genome-wide data, Bioinformatics 21 (suppl_2) (2005) ii206–ii212.
URL https://academic.oup.com/bioinformatics/article/21/suppl_2/ii206/227641

[236] A. Bernard, A. J. Hartemink, Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data, in: Biocomputing 2005, World Scientific, 2005, pp. 459–470.
URL https://www.worldscientific.com/doi/abs/10.1142/9789812702456_0044

[237] R. Cao, Y. Dong, K. C. Kural, Integrating literature-based knowledge database and expression data to explore molecular pathways connecting pparg and myocardial

infarction, PPAR Research 2020 (2020).
URL https://www.hindawi.com/journals/ppar/2020/1892375/

[238] K. Lo, A. E. Raftery, K. M. Dombek, J. Zhu, E. E. Schadt, R. E. Bumgarner, K. Y. Yeung, Integrating external biological knowledge in the construction of regulatory networks from time-series expression data, BMC Systems Biology 6 (1) (2012) 101.
URL https://link.springer.com/article/10.1186/1752-0509-6-101

[239] X. Chen, L. Wang, Integrating Biological Knowledge with Gene Expression Profiles for Survival Prediction of Cancer, Journal of Computational Biology 16 (2) (2009) 265–278.
URL https://www.liebertpub.com/doi/abs/10.1089/cmb.2008.12TT

[240] N. Bandyopadhyay, T. Kahveci, S. Goodison, Y. Sun, S. Ranka, Pathway-Based Feature Selection Algorithm for Cancer Microarray Data, Advances in Bioinformatics 2009 (2009) 1–16.
URL https://downloads.hindawi.com/archive/2009/532989.pdf

[241] M. Reboiro-Jato, R. Laza, H. López-Fernández, D. Glez-Peña, F. Díaz, F. Fdez-Riverola, genEnsemble: A new model for the combination of classifiers and integration of biological knowledge applied to genomic data, Expert Systems with Applications 40 (1) (2013) 52–63.
URL https://www.sciencedirect.com/science/article/abs/pii/S0957417412008585

[242] S. Seifert, S. Gundlach, O. Junge, S. Szymczak, Integrating biological knowledge and gene expression data using pathway-guided random forests: a benchmarking study, Bioinformatics 36 (15) (2020) 4301–4308.
URL https://academic.oup.com/bioinformatics/article-abstract/36/15/4301/5836498

[243] M. E. Blazadonakis, M. E. Zervakis, D. Kafetzopoulos, Integration of gene signatures using biological knowledge, Artificial Intelligence in Medicine 53 (1) (2011) 57–71.

URL https://www.sciencedirect.com/science/article/pii/S093336571100073X

[244] P. Minguez, J. Dopazo, Assessing the Biological Significance of Gene Expression Signatures and Co-Expression Modules by Studying Their Network Properties, PLoS One 6 (3) (2011) e17474.
URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0017474

[245] R. Cao, Y. Dong, K. C. Kural, Integrating Literature-Based Knowledge Database and Expression Data to Explore Molecular Pathways Connecting PPARG and Myocardial Infarction, PPAR Research 2020 (2020) 1–6.
URL https://www.hindawi.com/journals/ppar/2020/1892375/

[246] J. A. Nepomuceno, A. Troncoso, I. A. Nepomuceno-Chamorro, J. S. Aguilar-Ruiz, Integrating biological knowledge based on functional annotations for biclustering of gene expression data, Computer Methods and Programs in Biomedicine 119 (3) (2015) 163–180.
URL https://www.sciencedirect.com/science/article/abs/pii/S0169260715000450

[247] K. Lo, A. E. Raftery, K. M. Dombek, J. Zhu, E. E. Schadt, R. E. Bumgarner, K. Yeung, Integrating external biological knowledge in the construction of regulatory networks from time-series expression data, BMC Systems Biology 6 (1) (2012) 101.
URL https://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-6-101

[248] S. W. Kong, W. T. Pu, P. J. Park, A multivariate approach for integrating genome-wide expression data and biological knowledge, Bioinformatics 22 (19) (2006) 2373–2380.
URL https://academic.oup.com/bioinformatics/article-abstract/22/19/2373/241211

[249] J. Parraga-Alava, M. Dorn, M. Inostroza-Ponta, A multi-objective gene clustering algorithm guided by apriori biological knowledge with intensification and diversification strategies, BioData Mining 11 (1) (2018) 16.

[250] H. Cui, C. Zhou, X. Dai, Y. Liang, R. Paffenroth, D. Korkin, Boosting Gene Expression Clustering with System-Wide Biological Information: A Robust Autoencoder Approach, preprint, Bioinformatics (Nov. 2017).

[251] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, D. Haussler, Knowledge-based analysis of microarray gene expression data by using support vector machines, Proceedings of the National Academy of Sciences 97 (1) (2000) 262–267. doi:10.1073/pnas.97.1.262.

[252] K. Santosh, Ai-driven tools for coronavirus outbreak: need of active learning and cross-population train/test models on multitudinal/multimodal data, Journal of Medical Systems 44 (5) (2020) 1–5.
URL https://link.springer.com/article/10.1007/s10916-020-01562-1

[253] P. Ma, C. I. Castillo-Davis, W. Zhong, J. S. Liu, A data-driven clustering method for time course gene expression data, Nucleic Acids Research 34 (4) (2006) 1261–1269.
URL https://academic.oup.com/nar/article/34/4/1261/1337688

[254] Y. Li, K. Kang, J. M. Krahn, N. Croutwater, K. Lee, D. M. Umbach, L. Li, A comprehensive genomic pan-cancer classification using the cancer genome atlas gene expression data, BMC Genomics 18 (1) (2017) 1–13.
URL https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-3906-0

[255] M. Daoud, M. Mayo, A survey of neural network-based cancer prediction models from microarray data, Artificial Intelligence in Medicine 97 (2019) 204–214.
URL https://www.sciencedirect.com/science/article/pii/S0933365717305067

[256] S. Zuo, X. Zhang, L. Wang, A rna sequencing-based six-gene signature for survival prediction in patients with glioblastoma, Scientific Reports 9 (1) (2019) 1–10.
URL https://www.nature.com/articles/s41598-019-39273-4

[257] A. Ciaramella, D. Nardone, A. Staiano, Data integration by fuzzy similarity-based hierarchical clustering, BMC Bioinformatics 21 (S10) (2020) 350.

URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03567-6

[258] X. Dong, Y. Hao, X. Wang, W. Tian, LEGO: a novel method for gene set over-representation analysis by incorporating network-based gene weights, Scientific Reports 6 (1) (2016) 18871.
URL https://www.nature.com/articles/srep18871

[259] S. Yoon, J. Kim, S.-K. Kim, B. Baik, S.-M. Chi, S.-Y. Kim, D. Nam, GScluster: network-weighted gene-set clustering analysis, BMC Genomics 20 (1) (2019) 352.
URL https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-5738-6

[260] S. Aibar, C. Fontanillo, C. Droste, J. De Las Rivas, Functional Gene Networks: R/Bioc package to generate and analyse gene networks derived from functional enrichment and clustering, Bioinformatics 31 (10) (2015) 1686–1688.
URL https://academic.oup.com/bioinformatics/article-abstract/31/10/1686/176902

[261] J. Packer, C. Trapnell, Single-Cell Multi-omics: An Engine for New Quantitative Models of Gene Regulation, Trends in Genetics 34 (9) (2018) 653–665.
URL https://www.biorxiv.org/content/10.1101/864389v2.abstract

[262] F. Buettner, N. Pratanwanich, D. J. McCarthy, J. C. Marioni, O. Stegle, f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq, Genome Biology 18 (1) (2017) 212.
URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1334-8

[263] J. Fan, N. Salathia, R. Liu, G. E. Kaeser, Y. C. Yung, J. L. Herman, F. Kaper, J.-B. Fan, K. Zhang, J. Chun, P. V. Kharchenko, Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis, Nature Methods 13 (3) (2016) 241–244.
URL https://www.nature.com/articles/nmeth.3734

[264] E. Gerrits, Y. Heng, E. W. G. M. Boddeke, B. J. L. Eggen, Transcriptional profiling of microglia; current state of the art and future perspectives, Glia 68 (4) (2020) 740–755.
URL https://onlinelibrary.wiley.com/doi/full/10.1002/glia.23767

[265] T. Tokar, C. Pastrello, I. Jurisica, GSOAP: a tool for visualization of gene set overrepresentation analysis, Bioinformatics 36 (9) (2020) 2923–2925.
URL https://academic.oup.com/bioinformatics/article-abstract/36/9/2923/5715574

[266] S. Rahmati, M. Abovsky, C. Pastrello, I. Jurisica, pathDIP: an annotated resource for known and predicted human gene-pathway associations and pathway enrichment analysis, Nucleic Acids Research 45 (D1) (2017) D419–D426.
URL https://academic.oup.com/nar/article/45/D1/D419/2605696?login=true

[267] G. Yu, L.-G. Wang, Y. Han, Q.-Y. He, clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters, OMICS: A Journal of Integrative Biology 16 (5) (2012) 284–287.
URL https://www.liebertpub.com/doi/abs/10.1089/omi.2011.0118

[268] J. R. Adrian Alexa, topGO (2017).
URL http://ftp.linux.duke.edu/bioconductor.org/dest/packages/3.8/bioc/vignettes/topGO/inst/doc/topGO.pdf

[269] E. Y. Chen, C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. Meirelles, N. R. Clark, A. Ma'ayan, Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool, BMC Bioinformatics 14 (1) (2013) 128.
URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-128

[270] R. A. Fisher, Statistical methods for research workers, in: Breakthroughs in Statistics, Springer, 1992, pp. 66–70.
URL https://link.springer.com/chapter/10.1007/978-1-4612-4380-9_6

[271] A. Agresti, Categorical Data Analysis, 3rd Edition, Wiley, 2018.

[272] D. Nishimura, Biocarta, Biotech Software & Internet Report: The Computer Software Journal for Scient 2 (3) (2001) 117–120.
URL https://www.liebertpub.com/doi/abs/10.1089/152791601750294344

[273] D. Zhang, Q. Hu, X. Liu, K. Zou, E. K. Sarkodie, X. Liu, F. Gao, AllEnricher: a comprehensive gene set function enrichment tool for both model and non-model species, BMC Bioinformatics 21 (1) (2020) 106.
URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-3408-y

[274] L. Sun, S. Dong, Y. Ge, J. P. Fonseca, Z. T. Robinson, K. S. Mysore, P. Mehta, DiVenn: An Interactive and Integrated Web-Based Visualization Tool for Comparing Gene Lists, Frontiers in Genetics 10 (2019) 421.
URL https://www.frontiersin.org/articles/10.3389/fgene.2019.00421/full

[275] M. L. Metzker, Sequencing technologies - the next generation, Nature reviews genetics 11 (1) (2010) 31–46.

[276] M. R. Stratton, P. J. Campbell, P. A. Futreal, The cancer genome, Nature 458 (7239) (2009) 719–724.

[277] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Research 13 (11) (2003) 2498–2504.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC403769/

[278] P. Zeng, X. Zhou, Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models, Nature Communications 8 (1) (2017) 456.
URL https://www.nature.com/articles/s41467-017-00470-2

[279] A. R. Martin, C. R. Gignoux, R. K. Walters, G. L. Wojcik, B. M. Neale, S. Gravel, M. J. Daly, C. D. Bustamante, E. E. Kenny, Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations, The American Journal of Human Genetics

100 (4) (2017) 635–649.
URL https://www.sciencedirect.com/science/article/pii/S0002929717301076

[280] L. S. Mogil, A. Andaleon, A. Badalamenti, S. P. Dickinson, X. Guo, J. I. Rotter, W. C. Johnson, H. K. Im, Y. Liu, H. E. Wheeler, Genetic architecture of gene expression traits across diverse populations, PLOS Genetics 14 (8) (2018) e1007586.
URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1007586&rev=2

[281] A. V. Mikhaylova, T. A. Thornton, Accuracy of Gene Expression Prediction From Genotype Data With PrediXcan Varies Across and Within Continental Populations, Frontiers in Genetics 10 (2019) 261.
URL https://www.frontiersin.org/articles/10.3389/fgene.2019.00261/full

[282] K. L. Keys, A. C. Y. Mak, M. J. White, W. L. Eckalbar, A. W. Dahl, J. Mefford, A. V. Mikhaylova, M. G. Contreras, J. R. Elhawary, C. Eng, D. Hu, S. Huntsman, S. S. Oh, S. Salazar, M. A. Lenoir, J. C. Ye, T. A. Thornton, N. Zaitlen, E. G. Burchard, C. R. Gignoux, On the cross-population generalizability of gene expression prediction models, PLOS Genetics 16 (8) (2020) e1008927.
URL https://journals.plos.org/plosgenetics/article?rev=2&id=10.1371/journal.pgen.1008927

[283] J. J. Fryett, A. P. Morris, H. J. Cordell, Investigation of prediction accuracy and the impact of sample size, ancestry, and tissue in transcriptome-wide association studies, Genetic Epidemiology 44 (5) (2020) 425–441.
URL https://onlinelibrary.wiley.com/doi/full/10.1002/gepi.22290

[284] The 1000 Genomes Project Consortium, A global reference for human genetic variation, Nature 526 (7571) (2015) 68–74.
URL https://www.nature.com/articles/nature15393

[285] S. Das, L. Forer, S. Schönherr, C. Sidore, A. E. Locke, A. Kwong, S. I. Vrieze, E. Y. Chew, S. Levy, M. McGue, D. Schlessinger, D. Stambolian, P.-R. Loh, W. G.

Iacono, A. Swaroop, L. J. Scott, F. Cucca, F. Kronenberg, M. Boehnke, G. R. Abecasis, C. Fuchsberger, Next-generation genotype imputation service and methods, Nature Genetics 48 (10) (2016) 1284–1287.
URL https://www.nature.com/articles/ng.3656

[286] P.-R. Loh, P. Danecek, P. F. Palamara, C. Fuchsberger, Y. A Reshef, H. K Finucane, S. Schoenherr, L. Forer, S. McCarthy, G. R. Abecasis, R. Durbin, A. L Price, Reference-based phasing using the Haplotype Reference Consortium panel, Nature Genetics 48 (11) (2016) 1443–1448.
URL https://www.nature.com/articles/ng.3679

[287] D. Gola, J. Erdmann, B. Müller-Myhsok, H. Schunkert, I. R. König, Polygenic risk scores outperform machine learning methods in predicting coronary artery disease status, Genetic Epidemiology 44 (2) (2020) 125–138.
URL https://onlinelibrary.wiley.com/doi/full/10.1002/gepi.22279

[288] A. Andaleon, L. S. Mogil, H. E. Wheeler, Genetically regulated gene expression underlies lipid traits in Hispanic cohorts, PLOS One 14 (8) (2019) e0220827.
URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0220827

[289] A. R. Tall, D. J. Rader, Trials and Tribulations of CETP Inhibitors, Circulation Research 122 (1) (2018) 106–112.
URL https://www.ahajournals.org/doi/full/10.1161/CIRCRESAHA.117.311978

[290] GTEx Consortium, A. N. Barbeira, S. P. Dickinson, R. Bonazzola, J. Zheng, H. E. Wheeler, J. M. Torres, E. S. Torstenson, K. P. Shah, T. Garcia, T. L. Edwards, E. A. Stahl, L. M. Huckins, D. L. Nicolae, N. J. Cox, H. K. Im, Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics, Nature Communications 9 (1) (2018) 1825.
URL https://www.nature.com/articles/s41467-018-03621-1

[291] S. M. Urbut, G. Wang, P. Carbonetto, M. Stephens, Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions, Nature

Genetics 51 (1) (2019) 187–195.
URL https://www.nature.com/articles/s41588-018-0268-8

[292] GTEx GWAS Working Group, GTEx Consortium, A. N. Barbeira, R. Bonazzola, E. R. Gamazon, Y. Liang, Y. Park, S. Kim-Hellmuth, G. Wang, Z. Jiang, D. Zhou, F. Hormozdiari, B. Liu, A. Rao, A. R. Hamel, M. D. Pividori, F. Aguet, L. Bastarache, D. M. Jordan, M. Verbanck, R. Do, M. Stephens, K. Ardlie, M. McCarthy, S. B. Montgomery, A. V. Segrè, C. D. Brown, T. Lappalainen, X. Wen, H. K. Im, Exploiting the GTEx resources to decipher the mechanisms at GWAS loci, Genome Biology 22 (1) (2021) 49.
URL https://www.biorxiv.org/content/10.1101/814350v2.abstract

[293] A. Belorkar, L. Wong, GFS: fuzzy preprocessing for effective gene expression analysis 17 540. doi:10.1186/s12859-016-1327-8.

[294] K. Noto, C. Brodley, D. Slonim, Anomaly detection using an ensemble of feature models, in: 2010 IEEE International Conference on Data Mining, IEEE, pp. 953–958. doi:10.1109/ICDM.2010.140.

[295] K. Noto, S. Majidi, A. G. Edlow, H. C. Wick, D. W. Bianchi, D. K. Slonim, CSAX: Characterizing systematic anomalies in eXpression data 22 (5) 402–413. doi:10.1089/cmb.2014.0155.

[296] C. M. Pietras, F. Ocitti, D. K. Slonim, TEMPO: Detecting pathway-specific temporal dysregulation of gene expression in disease. doi:10.1101/651018.
URL http://biorxiv.org/lookup/doi/10.1101/651018

[297] Herman Wold., Partial least squares.

[298] C. M. Pietras, L. Power, D. K. Slonim, aTEMPO: Pathway-specific temporal anomalies for precision therapeutics 25 683–694.

[299] A. Conesa, M. J. Nueda, A. Ferrer, M. Talon, maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments 22 (9) 1096–1102. doi:10.1093/bioinformatics/btl056.

[300] Anindya Bhattacharya, Rajat K. De, A methodology for handling a new kind of outliers present in gene expression patterns, in: PReMI'11: Proceedings of the 4th international conference on Pattern recognition and machine intelligence, Lecture notes in computer science, pp. 394–399, meeting Name: International Conference on Pattern Recognition and Machine Intelligence.

[301] S. B. Lyerly, The average spearman rank correlation coefficient, Psychometrika 17 (4) (1952) 421–428.

[302] J. H. Zar, Spearman rank correlation: overview, Wiley StatsRef: Statistics Reference Online (2014).

[303] D. Soh, D. Dong, Y. Guo, L. Wong, Finding consistent disease subnetworks across microarray datasets 12 S15. doi:10.1186/1471-2105-12-S13-S15.

[304] K. Lim, L. Wong, Finding consistent disease subnetworks using PFSNet 30 (2) 189–196. doi:10.1093/bioinformatics/btt625.

[305] W. W. B. Goh, T. Guo, R. Aebersold, L. Wong, Quantitative proteomics signature profiling based on network contextualization 10 (1) 71. doi:10.1186/s13062-015-0098-x.

[306] M. Alshalalfa, T. A. Bismar, R. Alhajj, Detecting cancer outlier genes with potential rearrangement using gene expression data and biological networks 2012 1–13. doi:10.1155/2012/373506.

[307] A. I. Lawal, S. Kwon, O. S. Hammed, M. A. Idris, Blast-induced ground vibration prediction in granite quarries: an application of gene expression programming, anfis, and sine cosine algorithm optimized ann, International Journal of Mining Science and Technology 31 (2) (2021) 265–277.

[308] N. Beebe-Wang, S. Celik, E. Weinberger, P. Sturmfels, P. L. De Jager, S. Mostafavi, S.-I. Lee, Unified ai framework to uncover deep interrelationships between gene expression and alzheimer's disease neuropathologies, Nature Communications 12 (1) (2021) 1–17.

[309] C. Klein, Q. Zeng, F. Arbaretaz, E. Devêvre, J. Calderaro, N. Lomenie, M. C. Maiuri, Artificial intelligence for solid tumour diagnosis in digital pathology, British Journal of Pharmacology 178 (21) (2021) 4291–4315.

[310] M. Yap, R. L. Johnston, H. Foley, S. MacDonald, O. Kondrashova, K. A. Tran, K. Nones, L. T. Koufariotis, C. Bean, J. V. Pearson, et al., Verifying explainability of a deep learning tissue classifier trained on rna-seq data, Scientific Reports 11 (1) (2021) 1–12.

[311] A. Talukder, C. Barham, X. Li, H. Hu, Interpretation of deep learning in genomics and epigenomics, Briefings in Bioinformatics 22 (3) (2021) bbaa177.

[312] M. R. Hassan, M. F. Islam, M. Z. Uddin, G. Ghoshal, M. M. Hassan, S. Huda, G. Fortino, Prostate cancer classification from ultrasound and mri images using deep learning based explainable artificial intelligence, Future Generation Computer Systems (2021).

[313] A. Binder, M. Bockmayr, M. Hägele, S. Wienert, D. Heim, K. Hellweg, M. Ishii, A. Stenzinger, A. Hocke, C. Denkert, et al., Morphological and molecular breast cancer profiling through explainable machine learning, Nature Machine Intelligence 3 (4) (2021) 355–366.

[314] S. Wachter, B. Mittelstadt, L. Floridi, Why a right to explanation of automated decision-making does not exist in the general data protection regulation, International Data Privacy Law 7 (2) (2017) 76–99.

[315] https://www.industry.gov.au/data-and-publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles, accessed: 28-09-2021.

[316] A. Moncada-Torres, M. C. van Maaren, M. P. Hendriks, S. Siesling, G. Geleijnse, Explainable machine learning can outperform cox regression predictions and provide insights in breast cancer survival, Scientific Reports 11 (1) (2021) 1–13.

[317] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI 58 82–115. `doi:10.1016/j.inffus.2019.12.012`.
URL `https://linkinghub.elsevier.com/retrieve/pii/S1566253519308103`

[318] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models 51 (5) 1–42. `doi:10.1145/3236009`.
URL `https://dl.acm.org/doi/10.1145/3236009`

[319] G. Montavon, W. Samek, K.-R. Müller, Methods for interpreting and understanding deep neural networks 73 1–15. doi:10.1016/j.dsp.2017.10.011.
URL https://linkinghub.elsevier.com/retrieve/pii/S1051200417302385

[320] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, B. Baesens, An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models 51 (1) 141–154. doi:10.1016/j.dss.2010.12.003.
URL https://linkinghub.elsevier.com/retrieve/pii/S0167923610002368

[321] A. Fernandez, F. Herrera, O. Cordon, M. Jose del Jesus, F. Marcelloni, Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to? 14 (1) 69–81. doi:10.1109/MCI.2018.2881645.
URL https://ieeexplore.ieee.org/document/8610271/

[322] M. Gleicher, A framework for considering comprehensibility in modeling 4 (2) 75–88. doi:10.1089/big.2016.0007.
URL http://www.liebertpub.com/doi/10.1089/big.2016.0007

[323] R. S. Michalski, A theory and methodology of inductive learning, in: R. S. Michalski, J. G. Carbonell, T. M. Mitchell (Eds.), Machine Learning, Springer Berlin Heidelberg, pp. 83–134. doi:10.1007/978-3-662-12405-5_4.
URL http://link.springer.com/10.1007/978-3-662-12405-5_4

[324] Z. C. Lipton, The mythos of model interpretability.

[325] M. T. Ribeiro, S. Singh, C. Guestrin, ”Why Should I Trust You?”: Explaining the Predictions of Any Classifier, arXiv:1602.04938 [cs, stat]ArXiv: 1602.04938 (Aug. 2016).

[326] Changchun Liu, P. Rani, N. Sarkar, An empirical study of machine learning techniques for affect recognition in human-robot interaction, in: 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, pp. 2662–2667. doi:10.1109/IROS.2005.1545344.
URL http://ieeexplore.ieee.org/document/1545344/

[327] H.-X. Wang, L. Fratiglioni, G. B. Frisoni, M. Viitanen, B. Winblad, Smoking and the occurence of alzheimer's disease: Cross-sectional and longitudinal data in a population-based study 149 (7) 640–644. doi:10.1093/oxfordjournals.aje.a009864.
URL https://academic.oup.com/aje/article-lookup/doi/10.1093/oxfordjournals.aje.a009864

[328] An introduction to statistical learning: with applications in r, OCLC: ocn828488009.

[329] S. Basu, K. Kumbier, J. B. Brown, B. Yu, Iterative random forests to discover predictive and stable high-order interactions 115 (8) 1943–1948. doi:10.1073/pnas.1711236115.
URL http://www.pnas.org/lookup/doi/10.1073/pnas.1711236115

[330] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, W. A. Stahel, Robust Statistics The Approach Based on Influence Functions, OCLC: 904818602.
URL https://nbn-resolving.org/urn:nbn:de:101:1-201502079115

[331] B. Goodman, S. Flaxman, European union regulations on algorithmic decision-making and a "right to explanation" 38 (3) 50–57. doi:10.1609/aimag.v38i3.2741.
URL https://ojs.aaai.org/index.php/aimagazine/article/view/2741

[332] A. Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism prediction instrumentsarXiv:1610.07524.
URL http://arxiv.org/abs/1610.07524

[333] A. Bennetot, J.-L. Laurent, R. Chatila, N. Díaz-Rodríguez, Towards explainable neural-symbolic visual reasoningarXiv:1909.09065.
URL http://arxiv.org/abs/1909.09065

[334] K. Burns, L. A. Hendricks, K. Saenko, T. Darrell, A. Rohrbach, Women also snowboard: Overcoming bias in captioning modelsarXiv:1803.09797.
URL http://arxiv.org/abs/1803.09797

[335] T. Miller, P. Howe, L. Sonenberg, Explainable AI: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciencesarXiv:

1712.00547.
URL http://arxiv.org/abs/1712.00547

[336] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (5) (2019) 206–215.

[337] R. C. Fong, A. Vedaldi, Interpretable explanations of black boxes by meaningful perturbation, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, pp. 3449–3457. doi:10.1109/ICCV.2017.371.
URL http://ieeexplore.ieee.org/document/8237633/

[338] M. T. Ribeiro, S. Singh, C. Guestrin, Model-agnostic interpretability of machine learningarXiv:1606.05386.
URL http://arxiv.org/abs/1606.05386

[339] S. Tan, M. Soloviev, G. Hooker, M. T. Wells, Tree space prototypes: Another look at making tree ensembles interpretablearXiv:1611.07115.
URL http://arxiv.org/abs/1611.07115

[340] Y. Dong, H. Su, J. Zhu, B. Zhang, Improving interpretability of deep neural networks with semantic informationarXiv:1703.04096.
URL http://arxiv.org/abs/1703.04096

[341] J. Burrell, How the machine 'thinks': Understanding opacity in machine learning algorithms 3 (1) 205395171562251. doi:10.1177/2053951715622512.
URL http://journals.sagepub.com/doi/10.1177/2053951715622512

[342] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differencesarXiv:1704.02685.

[343] R. Goebel, A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, P. Kieseberg, A. Holzinger, Explainable AI: The new 42?, in: A. Holzinger, P. Kieseberg, A. M. Tjoa, E. Weippl (Eds.), Machine Learning and Knowledge Extraction, Vol. 11015, Springer International Publishing, pp. 295–303, series Title: Lecture Notes in Computer Science. doi:10.1007/978-3-319-99740-7_21.
URL https://link.springer.com/10.1007/978-3-319-99740-7_21

[344] Learning deep features for discriminative localizationarXiv:1512.04150.

[345] D. Martens, B. Baesens, T. Van Gestel, J. Vanthienen, Comprehensible credit scoring models using rule extraction from support vector machines 183 (3) 1466–1476. doi:10.1016/j.ejor.2006.04.051.
URL https://linkinghub.elsevier.com/retrieve/pii/S0377221706011878

[346] V. Berisha, C. Krantsevich, P. R. Hahn, S. Hahn, G. Dasarathy, P. Turaga, J. Liss, Digital medicine and the curse of dimensionality, NPJ Digital Medicine 4 (1) (2021) 1–8.

[347] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization 128 (2) 336–359. doi:10.1007/s11263-019-01228-7.
URL http://link.springer.com/10.1007/s11263-019-01228-7

[348] P. Hall, On the art and science of machine learning explanationsarXiv:1810.02909.
URL http://arxiv.org/abs/1810.02909

[349] S. Pasricha, XAI eXplainable artificial intelligence [visual data exploration seminar]Publisher: Unpublished. doi:10.13140/RG.2.2.13399.91047.
URL http://rgdoi.net/10.13140/RG.2.2.13399.91047

[350] O. Oni, S. Qiao, Model-agnostic interpretation of cancer classification with multi-platform genomic data, in: Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, ACM, pp. 34–41.

[351] C. Hans, Bayesian lasso regression, Biometrika 96 (4) (2009) 835–845.

[352] K. Fiok, W. Karwowski, E. Gutierrez, M. Wilamowski, Analysis of sentiment in tweets addressed to a single domain-specific twitter account: comparison of model performance and explainability of predictions, Expert Systems with Applications (2021) 115771.

[353] S. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions, arXiv:1705.07874 [cs, stat]ArXiv: 1705.07874 (Nov. 2017).

[354] A. Gramegna, P. Giudici, Shap and lime: An evaluation of discriminative power in credit risk, Frontiers in Artificial Intelligence (2021) 140.

[355] L. Shapley, A value fo n-person games, Ann. Math. Study28, Contributions to the Theory of Games, ed. by HW Kuhn, and AW Tucker (1953) 307–317.

[356] J. Pearl, Probabilistic reasoning in intelligent systems: networks of plausible inference, 2014.

[357] D. Koller, N. Friedman, Probabilistic graphical models: principles and techniques, Adaptive Computation and Machine Learning, MIT Press, Cambridge, MA, 2009.

[358] M. Scutari, C. Vitolo, A. Tucker, Learning Bayesian networks from big data with greedy search: computational complexity and efficient implementation, Statistics and Computing 29 (5) (2019) 1095–1108.

[359] D. Heckerman, D. Geiger, D. M. Chickering, Learning Bayesian Networks: The Combination of Knowledge and Statistical Data, arXiv:1302.6815 [cs] (May 2015).

[360] D. Heckerman, A Tutorial on Learning with Bayesian Networks, in: D. E. Holmes, L. C. Jain (Eds.), Innovations in Bayesian Networks, Vol. 156, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[361] D. Gabbay, J. Woods, Advice on Abductive Logic, Logic Journal of the IGPL 14 (2) (2006) 189–219.

[362] Y. Zeng, J. Luo, S. Lin, Classification using Markov blanket for feature selection, in: 2009 IEEE International Conference on Granular Computing, IEEE, Nanchang, China, 2009, pp. 743–747.

[363] A. Ultsch, J. Hoffmann, M. Röhnert, M. Von Bonin, U. Oelschlägel, C. Brendel, M. C. Thrun, An explainable AI system for the diagnosis of high dimensional biomedical data.

[364] J. H. Friedman, Greedy function approximation: a gradient boosting machine, Annals of statistics (2001) 1189–1232.

[365] C. Molnar, G. König, B. Bischl, G. Casalicchio, Model-agnostic feature importance and effects with dependent features – a conditional subgroup approacharXiv:2006.04628.

[366] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: D. Precup, Y. W. Teh (Eds.), Proceedings of the

34th International Conference on Machine Learning, Vol. 70 of Proceedings of Machine Learning Research, PMLR, 2017, pp. 3145–3153.

[367] P. Choudhary, Datascience.com, Skater: Model agnostic interpretation of machine learning models, https://github.com/oracle/Skater (2018).

[368] W. Samek, G. Montavon, A. Binder, S. Lapuschkin, K.-R. Müller, Interpreting the predictions of complex ML models by layer-wise relevance propagationarXiv:1611.08191.

[369] C. K. Cassel, Dementia in the elderly: An analysis of medical responsibility 94 (6) 802.

[370] P. Croskerry, K. S. Cosby, M. L. Graber, H. Singh, Diagnosis: interpreting the shadows, CRC Press/Taylor & Francis Group.

[371] D. Chakraborty, C. Ivan, P. Amero, M. Khan, C. Rodriguez-Aguayo, H. Başağaoğlu, G. Lopez-Berestein, Explainable artificial intelligence reveals novel insight into tumor microenvironment conditions linked with better prognosis in patients with breast cancer.

[372] A. Moncada-Torres, M. C. van Maaren, M. P. Hendriks, S. Siesling, G. Geleijnse, Explainable machine learning can outperform cox regression predictions and provide insights in breast cancer survival 11 (1) 6968.

[373] D. R. Cox, Regression models and life-tables 34 (2) 187–202.

[374] T. Jansen, G. Geleijnse, M. Van Maaren, M. P. Hendriks, A. Ten Teije, A. Moncada-Torres, Machine learning explainability in breast cancer survival 270 307–311.

[375] I. Bichindaritz, C. Bartlett, G. Liu, Predicting with confidence: A case-based reasoning framework for predicting survival in breast cancer 34 (1).

[376] N. Amoroso, D. Pomarico, A. Fanizzi, V. Didonna, F. Giotta, D. La Forgia, A. Latorre, A. Monaco, E. Pantaleo, N. Petruzzellis, P. Tamborra, A. Zito, V. Lorusso, R. Bellotti, R. Massafra, A roadmap towards breast cancer therapies supported by explainable artificial intelligence 11 (11) 4881.

[377] M. Pellegrini, Accurate prediction of breast cancer survival through coherent voting networks with gene expression profiling 11 (1) 14645.

[378] R. Li, A. Shinde, A. Liu, S. Glaser, Y. Lyou, B. Yuh, J. Wong, A. Amini, Machine learning–based interpretation and visualization of nonlinear interactions in prostate cancer survival (4) 637–646.

[379] M. J. Rho, J. Park, H. W. Moon, J. Kim, C. Lee, C.-S. Kim, S. S. Jeon, M. Kang, J. Y. Lee, Dr. answer AI software for prostate cancer: Explainable variable importance of predicting t stage, in: 2020 International Conference on Computational Science and Computational Intelligence (CSCI), IEEE, pp. 725–730.

[380] A. Consiglio, G. Casalino, G. Castellano, G. Grillo, E. Perlino, G. Vessio, F. Licciulli, Explaining ovarian cancer gene expression profiles with fuzzy rules and genetic algorithms 10 (4) 375.

[381] J. M. Clementino, B. S. Faical, C. C. Bones, C. Traina, M. A. Gutierrez, A. J. M. Traina, Multilevel clustering explainer: An explainable approach to electronic health records, in: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), IEEE, pp. 253–258.

[382] M.-A. Schulz, M. Chapman-Rounds, M. Verma, D. Bzdok, K. Georgatzis, Inferring disease subtypes from clusters in explanation space 10 (1) 12900.

[383] L.-J. Gardiner, A. P. Carrieri, K. Bingham, G. Macluskie, D. Bunton, M. McNeil, E. O. Pyzer-Knapp, Combining explainable machine learning, demographic and multi-omic data to identify precision medicine strategies for inflammatory bowel disease.

[384] Z. Al-Taie, D. Liu, J. B. Mitchem, C. Papageorgiou, J. T. Kaifi, W. C. Warren, C.-R. Shyu, Explainable artificial intelligence in high-throughput drug repositioning for subgroup stratifications with interventionable potential 118 103792.

[385] J. Peng, K. Zou, M. Zhou, Y. Teng, X. Zhu, F. Zhang, J. Xu, An explainable artificial intelligence framework for the deterioration risk prediction of hepatitis patients 45 (5) 61.

[386] M. S. Mellem, M. Kollada, J. Tiller, T. Lauritzen, Explainable AI enables clinical trial patient selection to retrospectively improve treatment effects in schizophrenia 21 (1) 162.

[387] A. M. Antoniadi, M. Galvin, M. Heverin, O. Hardiman, C. Mooney, Development of an explainable clinical decision support system for the prediction of patient quality of life in

amyotrophic lateral sclerosis, in: Proceedings of the 36th Annual ACM Symposium on Applied Computing, ACM, pp. 594–602.

[388] S. M. Lauritsen, M. Kristensen, M. V. Olsen, M. S. Larsen, K. M. Lauritsen, M. J. Jørgensen, J. Lange, B. Thiesson, Explainable artificial intelligence model to predict acute critical illness from electronic health records 11 (1) 3852.

[389] M. Czajkowski, K. Jurczuk, M. Kretowski, Accelerated evolutionary induction of heterogeneous decision trees for gene expression-based classification, in: Proceedings of the Genetic and Evolutionary Computation Conference, ACM, pp. 946–954.

[390] M. R. Karim, M. Cochez, O. Beyan, S. Decker, C. Lange, OncoNetExplainer: Explainable predictions of cancer types based on gene expression data, in: 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), IEEE, pp. 415–422.

[391] M. Kirienko, M. Sollini, M. Corbetta, E. Voulaz, N. Gozzi, M. Interlenghi, F. Gallivanone, I. Castiglioni, R. Asselta, S. Duga, G. Soldà, A. Chiti, Radiomics and gene expression profile to characterise the disease and predict outcome in patients with lung cancer.

[392] A. Anguita-Ruiz, A. Segura-Delgado, R. Alcalá, C. M. Aguilera, J. Alcalá-Fdez, eXplainable artificial intelligence (XAI) for the identification of biologically relevant gene expression patterns in longitudinal human studies, insights from obesity research 16 (4) e1007792.

[393] M. W. Farouq, W. Boulila, Z. Hussain, A. Rashid, M. Shah, S. Hussain, N. Ng, D. Ng, H. Hanif, M. G. Shaikh, A. Sheikh, A. Hussain, A novel coupled reaction-diffusion system for explainable gene expression profiling 21 (6) 2190.

[394] C. Droin, J. El Kholtei, K. B. Halpern, C. Hurni, M. Rozenberg, S. Muvkadi, S. Itzkovitz, F. Naef, Space-time logic of liver gene expression at sub-lobular scale, Nature Metabolism 3 (1) (2021) 43–58.

[395] K. Tan, W. Huang, X. Liu, J. Hu, S. Dong, A hierarchical graph convolution network for representation learning of gene expression data 25 (8) 3219–3229.

[396] V. Bourgeais, F. Zehraoui, M. B. Hamdoune, B. Hanczar, Deep GONet: Self-explainable deep neural network based on gene ontology for phenotype prediction from gene expression data, in: 19th Asia Pacific Bioinformatics Conference (APBC 2021),, pp. 1–19.

[397] M. Yap, R. L. Johnston, H. Foley, S. MacDonald, O. Kondrashova, K. A. Tran, K. Nones, L. T. Koufariotis, C. Bean, J. V. Pearson, M. Trzaskowski, N. Waddell, Verifying explainability of a deep learning tissue classifier trained on RNA-seq data 11 (1) 2641.

[398] H. Park, K. Maruhashi, R. Yamaguchi, S. Imoto, S. Miyano, Global gene network exploration based on explainable artificial intelligence approach 15 (11) e0241508.

[399] X. Huang, Y. Izza, A. Ignatiev, J. Marques-Silva, On efficiently explaining graph-based classifiersarXiv:2106.01350.

[400] H. Chereda, A. Bleckmann, K. Menck, J. Perera-Bel, P. Stegmaier, F. Auer, F. Kramer, A. Leha, T. Beißbarth, Explaining decisions of graph convolutional neural networks: patient-specific molecular subnetworks responsible for metastasis prediction in breast cancer 13 (1) 42. doi:10.1186/s13073-021-00845-7.

[401] C. Agarwal, M. Zitnik, H. Lakkaraju, Towards a rigorous theoretical analysis and evaluation of GNN explanationsarXiv:2106.09078.

[402] A. Holzinger, B. Malle, A. Saranti, B. Pfeifer, Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI 71 28–37. doi:10.1016/j.inffus.2021.01.008.

[403] L. Li, F. Wu, G. Yang, L. Xu, T. Wong, R. Mohiaddin, D. Firmin, J. Keegan, X. Zhuang, Atrial scar quantification via multi-scale CNN in the graph-cuts framework 60 101595. doi:10.1016/j.media.2019.101595.
URL https://linkinghub.elsevier.com/retrieve/pii/S1361841519301355

[404] S. Momin, Y. Fu, Y. Lei, J. Roper, J. D. Bradley, W. J. Curran, T. Liu, X. Yang, Knowledge-based radiation treatment planning: A data-driven method survey, Journal of Applied Clinical Medical Physics 22 (8) (2021) 16–44.

[405] G. O. Consortium, The gene ontology project in 2008, Nucleic Acids Research 36 (suppl_1) (2008) D440–D444.

[406] G. O. Consortium, The gene ontology resource: 20 years and still going strong, Nucleic Acids Research 47 (D1) (2019) D330–D338.

[407] A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, P. Tamayo, The molecular signatures database hallmark gene set collection, Cell Systems 1 (6) (2015) 417–425.

[408] K. Akabe, T. Takeuchi, T. Aoki, K. Nishimura, Information retrieval on oncology knowledge base using recursive paraphrase lattice, Journal of Biomedical Informatics 116 (2021) 103705.

[409] J. Somekh, Model-based pathway enrichment analysis applied to the tgf-beta regulation of autophagy in autism, Journal of Biomedical Informatics 118 (2021) 103781.

[410] A. Kosvyra, E. Ntzioni, I. Chouvarda, Network analysis with biological data of cancer patients: A scoping review, Journal of Biomedical Informatics (2021) 103873.

[411] M. Wong, P. Previde, J. Cole, B. Thomas, N. Laxmeshwar, E. Mallory, J. Lever, D. Petkovic, R. B. Altman, A. Kulkarni, Search and visualization of gene-drug-disease interactions for pharmacogenomics and precision medicine research using genedive, Journal of Biomedical Informatics 117 (2021) 103732.

[412] S. Hänzelmann, R. Castelo, J. Guinney, Gsva: gene set variation analysis for microarray and rna-seq data, BMC Bioinformatics 14 (1) (2013) 7.

[413] S. Carbon, A. Ireland, C. J. Mungall, S. Shu, B. Marshall, S. Lewis, A. Hub, W. P. W. Group, Amigo: online access to ontology and annotation data, Bioinformatics 25 (2) (2009) 288–289.

[414] J. Hastings, Primer on ontologies, in: The Gene Ontology Handbook, Humana Press, New York, NY, 2017, pp. 3–13.

[415] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al., Gene ontology: tool for the unification of biology, Nature Genetics 25 (1) (2000) 25–29.

[416] S. Y. Rhee, V. Wood, K. Dolinski, S. Draghici, Use and misuse of the gene ontology annotations, Nature Reviews Genetics 9 (7) (2008) 509–515.

[417] A. AlSaieedi, A. Salhi, F. Tifratene, A. B. Raies, A. Hungler, M. Uludag, C. Van Neste, V. B. Bajic, T. Gojobori, M. Essack, Des-tcell is a knowledgebase for exploring immunology-related literature, Scientific Reports 11 (1) (2021) 1–11.

[418] B. D. Fulcher, A. Arnatkeviciute, A. Fornito, Overcoming false-positive gene-category enrichment in the analysis of spatially resolved transcriptomic brain atlas data, Nature Communications 12 (1) (2021) 1–13.

[419] Z. Ahmed, E. G. Renart, S. Zeeshan, X. Dong, Advancing clinical genomics and precision medicine with gvviz: Fair bioinformatics platform for variable gene-disease annotation, visualization, and expression analysis, Human Genomics 15 (1) (2021) 1–9.

[420] S. Y. Rhee, V. Wood, K. Dolinski, S. Draghici, Use and misuse of the gene ontology annotations, Nature Reviews Genetics 9 (7) (2008) 509–515.

[421] Y. Chen, F. J. Verbeek, K. Wolstencroft, Establishing a consensus for the hallmarks of cancer based on gene ontology and pathway annotations, BMC Bioinformatics 22 (1) (2021) 1–20.

[422] A. Anaissi, P. J. Kennedy, M. Goyal, D. R. Catchpoole, A balanced iterative random forest for gene selection from microarray data, BMC Bioinformatics 14 (1) (2013) 1–10.

[423] Oncogenomics db national cancer institute, https://omics-oncogenomics.ccr.cancer.gov/cgi-bin/JK, accessed: 22-08-2021.

[424] K. D. Pruitt, T. Tatusova, G. R. Brown, D. R. Maglott, Ncbi reference sequences (refseq): current status, new features and genome annotation policy, Nucleic Acids Research 40 (D1) (2012) D130–D135.

[425] C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, C. Desmedt, D. Larsimont, F. Cardoso, H. Peterse, D. Nuyten, M. Buyse, M. J. Van de Vijver, J. Bergh, M. Piccart, M. Delorenzi, Gene expression profiling in breast cancer:understanding the molecular basis of histologic grade to improve prognosis, JNCI: Journal of the National Cancer Institute 98 (4) (2006) 262–272.

[426] O. Galamb, B. Wichmann, F. Sipos, S. Spisák, T. Krenács, K. Tóth, K. Leiszter, A. Kalmár, Z. Tulassay, B. Molnár, Dysplasia-Carcinoma Transition Specific Transcripts in Colonic Biopsy Samples, PLoS One 7 (11) (2012) e48547.

[427] J. H. Heaton, M. A. Wood, A. C. Kim, L. O. Lima, F. M. Barlaskar, M. Q. Almeida, M. C. B. V. Fragoso, R. Kuick, A. M. Lerario, D. P. Simon, I. C. Soares, E. Starnes, D. G. Thomas, A. C. Latronico, T. J. Giordano, G. D. Hammer, Progression to adrenocortical tumorigenesis in mice and humans through insulin-like growth factor 2 and -catenin, The American Journal of Pathology 181 (3) (2012) 1017–1033.

[428] T. T. Giang, T.-P. Nguyen, Q. T. Pham, D. H. Tran, A combination model of robust principal component analysis and multiple kernel learning for cancer patient stratification, in: Soft Computing: Biomedical and Related Applications, Springer, 2021, pp. 21–33.

[429] E. O. Omuya, G. O. Okeyo, M. W. Kimwele, Feature selection for classification using principal component analysis and information gain, Expert Systems with Applications 174 (2021) 114765.

[430] L. Sun, T. Yin, W. Ding, Y. Qian, J. Xu, Feature selection with missing labels using multilabel fuzzy neighborhood rough sets and maximum relevance minimum redundancy, IEEE Transactions on Fuzzy Systems (2021).

[431] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, B. P. Feuston, Random forest: a classification and regression tool for compound classification and qsar modeling, Journal of Chemical Information and Computer Sciences 43 (6) (2003) 1947–1958.

[432] F. H. M. Oliveira, A. R. P. Machado, A. O. Andrade, On the use of t-distributed stochastic neighbor embedding for data visualization and classification of individuals with parkinson's disease 2018 1–17. doi:10.1155/2018/8019232.

[433] K. Hinata, A. M. Gervin, Y. Jennifer Zhang, P. A. Khavari, Divergent gene regulation and growth effects by nf-κb in epithelial and mesenchymal cells of human skin, Oncogene 22 (13) (2003) 1955–1964.

[434] M. N. Chamorro, D. R. Schwartz, A. Vonica, A. H. Brivanlou, K. R. Cho, H. E. Varmus,

Fgf-20 and dkk1 are transcriptional targets of β-catenin and fgf-20 is implicated in cancer and development, The EMBO journal 24 (1) (2005) 73–84.

[435] J. D. Cahoy, B. Emery, A. Kaushal, L. C. Foo, J. L. Zamanian, K. S. Christopherson, Y. Xing, J. L. Lubischer, P. A. Krieg, S. A. Krupenko, et al., A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function, Journal of Neuroscience 28 (1) (2008) 264–278.

[436] D. A. Barbie, P. Tamayo, J. S. Boehm, S. Y. Kim, S. E. Moody, I. F. Dunn, A. C. Schinzel, P. Sandy, E. Meylan, C. Scholl, et al., Systematic rna interference reveals that oncogenic kras-driven cancers require tbk1, Nature 462 (7269) (2009) 108–112.

[437] S. Na, L. Xumin, G. Yong, Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm (Apr. 2010).

[438] C. Papadimitriou, Algorithms, complexity, and the sciences 111 (45) 15881–15887. doi:10.1073/pnas.1416954111.

[439] S. Devi, K. Selvam, S. Rajagopalan, An abstract to calculate big o factors of time and space complexity of machine code, in: International Conference on Sustainable Energy and Intelligent Systems (SEISCON 2011), IET, pp. 844–847. doi:10.1049/cp.2011.0483.

[440] T. H. Cormen, C. E. Leiserson, R. L. Rivest, C. Stein, Introduction to Algorithms, 3rd Edition, MIT Press, 2009.

[441] D. E. Knuth, The Art of Computer Programming, Volume 1: Fundamental Algorithms, Addison-Wesley, 1997.

[442] Algorithms and complexity: 10th international conference, CIAC 2017, athens, greece, may 24-26, 2017, proceedings. doi:10.1007/978-3-319-57586-5.

[443] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[444] R. de Souza Pinto, A. C. Botazzo Delbem, F. J. Monaco, Characterization of runtime resource usage from analysis of binary executable programs 71 1133–1152. doi:10.1016/j.asoc.2017.12.040.

[445] D. B. Phi, K. N. Trong, V. H. Nguyen, A runtime approach for estimating resource usage, in: Proceedings of the Fourth Symposium on Information and Communication Technology - SoICT '13, ACM Press, pp. 261–266. doi:10.1145/2542050.2542091.

[446] I. E. Commission, IEC 60027-2: Letter symbols to be used in electrical technology - Part 2: Telecommunications and electronics, IEC, 1998, standard defining binary prefixes Ki, Mi, Gi.

[447] J. VanderPlas, Python data science handbook: Essential tools for working with data, " O'Reilly Media, Inc.", 2016.

[448] J. Moore, C. Allan, S. Besson, J.-M. Burel, E. Diel, D. Gault, K. Kozlowski, D. Lindner, M. Linkert, T. Manz, et al., Ome-ngff: a next-generation file format for expanding bioimaging data-access strategies, Nature methods 18 (12) (2021) 1496–1498.

[449] S. Warnat-Herresthal, et al., Scalable prediction of acute myeloid leukemia using high-dimensional machine learning and blood transcriptomics, iScience 23 (1) (2020) 100780. doi:10.1016/j.isci.2019.100780.

[450] Y. Wang, et al., Distinct clinical and biological characteristics of acute myeloid leukemia with higher expression of long noncoding rna kiaa0125, Annals of Hematology 100 (2) (2021) 487–498. doi:10.1007/s00277-020-04358-y.

[451] A. J. Gentles, S. K. Plevritis, R. Majeti, A. A. Alizadeh, Association of a leukemic stem cell gene expression signature with clinical outcomes in acute myeloid leukemia, JAMA 304 (24) (2010) 2706–2715. doi:10.1001/jama.2010.1862.

[452] D. W. Aha, D. Kibler, M. K. Albert, Instance-based learning algorithms, Machine learning 6 (1) (1991) 37–66.

[453] C. Moreira, Y.-L. Chou, M. Velmurugan, C. Ouyang, R. Sindhgatta, P. Bruza, Linda-bn: An interpretable probabilistic approach for demystifying black-box predictive models, Decision Support Systems 150 (2021) 113561.

[454] K. M. Ahmed, C. Y. Tsai, W.-H. Lee, Derepression of hmga2 via removal of zbrk1/brca1/ctip complex enhances mammary tumorigenesis, Journal of Biological Chemistry 285 (7) (2010) 4464–4471.

[455] H. Yu, D. L. Simons, I. Segall, V. Carcamo-Cavazos, E. J. Schwartz, N. Yan, N. S. Zuckerman, F. M. Dirbas, D. L. Johnson, S. P. Holmes, et al., Prc2/eed-ezh2 complex is up-regulated in breast cancer lymph node metastasis compared to primary tumor and correlates with tumor proliferation in situ, PloS one 7 (12) (2012) e51239.

[456] J. H. Cook, G. E. Melloni, D. C. Gulhan, P. J. Park, K. M. Haigis, The origins and genetic interactions of kras mutations are allele-and tissue-specific, Nature communications 12 (1) (2021) 1–14.

[457] C. Zhu, L. Li, B. Zhao, The regulation and function of yap transcription co-activator, Acta biochimica et biophysica Sinica 47 (1) (2015) 16–28.

[458] https://www.gsea-msigdb.org/gsea/msigdb/cards/KRAS.50_UP.V1_UP.

[459] https://www.gsea-msigdb.org/gsea/msigdb/cards/GLI1_UP.V1_UP.

[460] https://www.gsea-msigdb.org/gsea/msigdb/cards/BCAT.100_UP.V1_UP.

[461] https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?geneSetName=SINGH_KRAS_DEPENDENCY_SIGNATURE.

[462] GeneCards®: The Human Gene Database, howpublished = https://www.genecards.org/cgi-bin/carddisp.pl?gene=znf75d, note = Accessed: 2022-10-05.

[463] K. Qin, D. Jian, Y. Xue, Y. Cheng, P. Zhang, Y. Wei, J. Zhang, H. Xiong, Y. Zhang, X. Yuan, Ddx41 regulates the expression and alternative splicing of genes involved in tumorigenesis and immune response, Oncology reports 45 (3) (2021) 1213–1225.

[464] GeneCards®: The Human Gene Database, howpublished = https://www.genecards.org/cgi-bin/carddisp.pl?gene=znf75d, note = Accessed: 2022-10-05.

[465] GeneCards®: The Human Gene Database, howpublished = https://www.genecards.org/cgi-bin/carddisp.pl?gene=wbp2&keywords=wbp2, note = Accessed: 2022-10-05.

[466] S. Chen, H. Wang, Z. Li, J. You, Q.-W. Wu, C. Zhao, C.-M. Tzeng, Z.-M. Zhang, Interaction of wbp2 with erα increases doxorubicin resistance of breast cancer cells by modulating mdr1 transcription, British journal of cancer 119 (2) (2018) 182–192.

[467] GeneCards®: The Human Gene Database, howpublished = https://www.genecards.org/cgi-bin/carddisp.pl?gene=spag1&keywords=spag1, note = Accessed: 2022-10-05.

[468] M. R. Knowles, L. E. Ostrowski, N. T. Loges, T. Hurd, M. W. Leigh, L. Huang, W. E. Wolf, J. L. Carson, M. J. Hazucha, W. Yin, et al., Mutations in spag1 cause primary ciliary dyskinesia associated with defective outer and inner dynein arms, The American Journal of Human Genetics 93 (4) (2013) 711–720.

[469] GeneCards®: The Human Gene Database, howpublished = https://www.genecards.org/cgi-bin/carddisp.pl?gene=rhoq&keywords=rhoq, note = Accessed: 2022-10-05.

[470] S. J. Heasman, A. J. Ridley, Mammalian rho gtpases: new insights into their functions from in vivo studies, Nature reviews Molecular cell biology 9 (9) (2008) 690–701.

[471] K. Hashimoto, H. Ochi, S. Sunamura, N. Kosaka, Y. Mabuchi, T. Fukuda, K. Yao, H. Kanda, K. Ae, A. Okawa, et al., Cancer-secreted hsa-mir-940 induces an osteoblastic phenotype in the bone metastatic microenvironment via targeting arhgap1 and fam134a, Proceedings of the National Academy of Sciences 115 (9) (2018) 2204–2209.

[472] GeneCards®: The Human Gene Database, howpublished = https://www.genecards.org/cgi-bin/carddisp.pl?gene=arhgap1&keywords=arhgap1, note = Accessed: 2022-10-05.

[473] GeneCards®: The Human Gene Database, howpublished = https://www.genecards.org/cgi-bin/carddisp.pl?gene=arhgap1&keywords=arhgap1, note = Accessed: 2022-10-05.

[474] C. Lee, E. Bongcam-Rudloff, C. Sollner, W. Jahnen-Dechent, L. Claesson-Welsh, Type 3 cystatins; fetuins, kininogen and histidine-rich glycoprotein, Front Biosci 14 (2009) 2911–2922.

[475] GeneCards®: The Human Gene Database, howpublished = https://www.genecards.org/cgi-bin/carddisp.pl?gene=arhgap1&keywords=arhgap1, note = Accessed: 2022-10-05.

[476] M. Stoffel, S. A. Duncan, The maturity-onset diabetes of the young (mody1) transcription factor hnf4α regulates expression of genes required for glucose transport and metabolism, Proceedings of the National Academy of Sciences 94 (24) (1997) 13209–13214.

[477] J. Elia, X. Gai, H. Xie, J. Perin, E. Geiger, J. Glessner, M. D'arcy, R. Deberardinis, E. Frackelton, C. Kim, et al., Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes, Molecular psychiatry 15 (6) (2010) 637–646.

[478] R. Devon, S. Anderson, P. Teague, W. Muir, V. Murray, A. Pelosi, D. Blackwood, D. Porteous, The genomic organisation of the metabotropic glutamate receptor subtype 5 gene, and its association with schizophrenia, Molecular psychiatry 6 (3) (2001) 311–314.

[479] N. Ramoz, C. Boni, A. M. Downing, S. L. Close, S. L. Peters, A. M. Prokop, A. J. Allen, M. Hamon, D. Purper-Ouakil, P. Gorwood, A haplotype of the norepinephrine transporter (net) gene slc6a2 is associated with clinical response to atomoxetine in attention-deficit hyperactivity disorder (adhd), Neuropsychopharmacology 34 (9) (2009) 2135–2142.

[480] R. E. Urwin, B. H. Bennetts, B. Wilcken, B. Lampropoulos, P. J. Beumont, J. D. Russell, S. L. Tanner, K. P. Nunn, Gene-gene interaction between the monoamine oxidase a gene and solute carrier family 6 (neurotransmitter transporter, noradrenalin) member 2 gene in anorexia nervosa (restrictive subtype), European Journal of Human Genetics 11 (12) (2003) 945–950.

[481] J. Fishman-Lobell, J. E. Haber, Removal of nonhomologous dna ends in double-strand break recombination: the role of the yeast ultraviolet repair gene rad1, Science 258 (5081) (1992) 480–484.

[482] A. Carr, H. Schmidt, S. Kirchhoff, W. Muriel, K. Sheldrick, D. Griffiths, C. Basmacioglu, S. Subramani, M. Clegg, A. Nasim, The rad16 gene of schizosaccharomyces pombe: a homolog of the rad1 gene of saccharomyces cerevisiae, Molecular and Cellular Biology 14 (3) (1994) 2029–2040.

[483] R. Peoples, L. Perez-Jurado, Y.-K. Wang, P. Kaplan, U. Francke, The gene for replication factor c subunit 2 (rfc2) is within the 7q11. 23 williams syndrome deletion., American journal of human genetics 58 (6) (1996) 1370.

[484] V. N. Noskov, H. Araki, A. Sugino, The rfc2 gene, encoding the third-largest subunit of the replication factor c complex, is required for an s-phase checkpoint in saccharomyces cerevisiae, Molecular and Cellular Biology 18 (8) (1998) 4914–4923.

[485] G. Kirov, I. Zaharieva, L. Georgieva, V. Moskvina, I. Nikolov, S. Cichon, A. Hillmer, D. Toncheva, M. J. Owen, M. C. O'Donovan, A genome-wide association study in 574 schizophrenia trios using dna pooling, Molecular psychiatry 14 (8) (2009) 796–803.

[486] M. V. Yusenko, A. Nagy, G. Kovacs, Molecular analysis of germline t (3; 6) and t (3; 12) associated with conventional renal cell carcinomas indicates their rate-limiting role and supports the three-hit model of carcinogenesis, Cancer genetics and cytogenetics 201 (1) (2010) 15–23.

[487] S. Salunkhe, N. Chandran, P. Chandrani, A. Dutt, S. Dutt, CytoPred: 7-gene pair metric for AML cytogenetic risk predictiondoi:10.1093/bib/bby100.
URL https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bby100/5144166

[488] https://ftp.ncbi.nlm.nih.gov/geo/series/gse14nnn/gse14468/matrix/.

[489] https://ftp.ncbi.nlm.nih.gov/geo/series/gse12nnn/gse12417/matrix/.

[490] M. F. Lara, R. García-Escudero, S. Ruiz, M. Santos, M. Moral, A. B. Martínez-Cruz, C. Segrelles, C. Lorz, J. M. Paramio, Gene profiling approaches help to define the specific functions of retinoblastoma family in epidermis, Molecular Carcinogenesis: Published in cooperation with the University of Texas MD Anderson Cancer Center 47 (3) (2008) 209–221.

[491] J. D. Cahoy, B. Emery, A. Kaushal, L. C. Foo, J. L. Zamanian, K. S. Christopherson, Y. Xing, J. L. Lubischer, P. A. Krieg, S. A. Krupenko, et al., A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function, Journal of Neuroscience 28 (1) (2008) 264–278.

[492] M. Borkin, K. Gajos, A. Peters, D. Mitsouras, S. Melchionna, F. Rybicki, C. Feldman, H. Pfister, Evaluation of artery visualizations for heart disease diagnosis, IEEE transactions on visualization and computer graphics 17 (12) (2011) 2479–2488.