A Robust Unified Graph Model Based on Molecular Data Binning for Subtype Discovery in High-dimensional Spaces

by

Muhammad Sadiq Hassan Zada

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy

July 2, 2023

College of Science and Engineering
University of Derby

A Robust Unified Graph Model Based on Molecular Data Binning for Subtype Discovery in High-dimensional Spaces

by

Muhammad Sadiq Hassan Zada

Director of Studies: Stephan Reiff-Marganiec
1st Supervisor: Wajahat Ali Khan

School of Computing and Engineering

July 2, 2023

Faculty of Science and Engineering
University of Derby

Acknowledgments

First and foremost, I want to express my gratitude to Prof. Stephan Reiff-Marganiec and Dr. Wajahat Ali Khan for their great guidance, consistent support, and patience throughout my PhD studies. Their vast expertise and wealth of experience have inspired me throughout my academic research and daily life. I would also want to thank Prof. Ashiq Anjum and Dr. Bo Yuan from the University of Leicester for their advice and guidance with my research.

Also, Dr. Syed Usama Khalid Bukhari, Dr. Maqbool Hussain, and all of my colleagues at the College of Science and Engineering deserve my heartfelt gratitude. Their friendly assistance and support have made my studies and life in the UK a delightful experience. And a particular thanks to Prof. Rahman Ali from the University of Peshawar for his encouragement through tough times, without which I would have given up on my studies long ago.

I would like to express my gratitude to my parents, my fiancée, and my whole family. It would have been difficult for me to finish my studies without their wonderful understanding and encouragement throughout my PhD studies.

Finally, I would like to take this opportunity to thank everyone who has contributed to my Ph.D. journey over the years!

Abstract

Machine learning (ML) is a subfield of artificial intelligence (AI) that has already revolutionised the world around us. It is a widely employed process for discovering patterns and groups within datasets. It has a wide range of applications including disease subtyping, which aims to discover intrinsic subtypes of disease in large-scale unlabelled data. Whilst the groups discovered in multi-view high-dimensional data by ML algorithms are promising, their capacity to identify pertinent and meaningful groups is limited by the presence of data variability and outliers. Since outlier values represent potential but unlikely outcomes, they are statistically and philosophically fascinating.

Therefore, the primary aim of this thesis was to propose a robust approach that discovers meaningful groups while considering the presence of data variability and outliers in the data. To achieve this aim, a novel robust approach (ROMDEX) was developed that utilised the proposed intermediate graph models (IMGs) for robust computation of proximity between observations in the data. Finally, a robust multi-view graph-based clustering approach was developed based on ROMDEX that improved the discovery of meaningful groups that were hidden behind the noise in the data.

The proposed approach was validated on real-world, and synthetic data for disease subtyping. Additionally, the stability of the approach was assessed by evaluating its performance across different levels of noise in clustering data. The results were evaluated through Kaplan-Meier survival time analysis for disease subtyping. Also, the concordance index (CI) and normalised mutual information (NMI) are used to evaluate the predictive ability of the proposed clustering model. Additionally, the accuracy, Kappa statistic and rand index are computed to evaluate the clustering stability against various levels of Gaussian noise. The proposed approach outperformed the existing state-of-the-art approaches MRGC, PINS, SNF, Consensus Clustering, and Icluster+ on these datasets. The findings for all datasets were outstanding, demonstrating the predictive ability of the proposed unsupervised graph-based clustering approach.

Declaration

I declare that the thesis here submitted is original, except for the sources explicitly acknowledged. This thesis as a whole or any part of it has not been previously submitted for the same degree or for a different degree other than the Doctor of Philosophy (Ph.D.) at the University of Derby. Parts of this thesis have appeared in papers, I authored or co-authored.

Muhammad Sadiq Hassan Zada

University of Derby, 2023

Table of Contents

List of Tables and Figures

List of Publications

Journal Papers

1. Zada, Muhammad Sadiq Hassan, Bo Yuan, Wajahat Ali Khan, Ashiq Anjum, Stephan Reiff-Marganiec, and Rabia Saleem. "A unified graph model based on molecular data binning for disease subtyping." Journal of Biomedical Informatics 134 (2022): 104187.

Conference Papers

1. Zada, Muhammad Sadiq Hassan, Bo Yuan, Ashiq Anjum, Muhammad Ajmal Azad, Wajahat Ali Khan, and Stephan Reiff-Marganiec. "Large-scale Data Integration Using Graph Probabilistic Dependencies (GPDs)." In 2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT), pp. 27-36. IEEE, 2020.

Chapter 1

Introduction

The complexity of genomics-related big data has led to the adoption of state-of-art technologies like artificial intelligence (AI) and machine learning (ML) for performing out class in healthcare services [1]. These technologies have shown promise in improving healthcare services by analysing complex molecular data and guiding tailored treatments for individual patients, resulting in better health outcomes [2] [3]. Although AI technology has made significant advancements, the treatment of cancer remains challenging due to the substantial genetic variability observed among affected individuals. To estimate gene variability, unsupervised approaches are often employed due to the difficulty of generating labels for such a vast amount of data [4]. Recently, disease subtyping has emerged as a promising approach to better understand the molecular basis of cancer by incorporating individual gene variability and health data into advanced unsupervised clustering algorithms. This provides crucial insights for better understanding the disease progression among affected individuals. Furthermore, the field of cancer treatment requires a comprehensive analysis of multi-modal data including clinical, and other data types acquired through advanced techniques such as genomics, transcriptomics, proteomics, and metabolomics. These diverse modalities pose an additional challenge for subtyping algorithms, which must develop effective integrative approaches to enable their analysis.

While machine learning advancements have accelerated the discovery of data-driven insights for cancer patients, most existing approaches are limited to single modalities, leading to underdeveloped methods for integrating multi-modal data [5]. Leveraging the integration of multiple modalities opens up opportunities to advance precision oncology [5]. The integration of multi-modal data presents inherent challenges; however, it offers a promising avenue for enhancing cancer risk stratification. Notably, when genomics and clinical imaging data are effectively combined, the predictive capacity of machine learning models can be significantly improved [6]. Likewise, the integration of genomic, medical imaging, and histopathologic features holds tremendous promise for improving the predictive capacity of immunotherapy response [7].

The rest of this chapter is organised as below: This chapter will introduce the research conducted in this thesis by first discussing the background and context of the research, followed by the problem statement, the research aims, objectives, and the significance of the research, followed by the research validation methods, and finally, the layout of the thesis.

1.1 Background

High-dimensional genomics data is typically generated in large quantities using omics technologies. This high-dimensional data captures the essence of biological organisms and their interactions. Unsupervised learning is widely adopted for genomics data analysis. It carries crucial information about the individual gene variability which helps in understanding the biological process which is vital for disease subtyping. For example, integrating genomes and transcriptomics data together using a full graph model allows clustering algorithms to identify clinically relevant disease subtypes based on proximity and similar gene connection patterns [8]. This is an important and basic task in molecular biology and precision medicine. Graph-based techniques have demonstrated promising results in accurate molecular subtyping using various types of gene interactions [9, 10]. Furthermore, the integrated view of various high-dimensional omics data can provide important insights for better understanding disease progression [11, 12].

Spectral clustering is a graph-based unsupervised learning algorithm, it has roots in graph theory and is commonly used to address various aspects of subtyping [13, 14, 15]. One of the aspects is to use similarity matrices to organise observations (patients) into coherent groupings [16]. The similarity matrix depicts an affinity graph of the patients produced from omics data. These similarity matrices are frequently generated with similarity kernels (SK) and represented by an affinity network [17, 18]. A similarity kernel takes a pair of observations in $\mathbb{R}^n$ and generates a number value that quantifies the pair's similarity. That is, K (X, Y) $\rightarrow \mathbb{R}$, where K is a kernel and X, Y $\in \mathbb{R}^n$. The distance function, which computes the distances between a pair of observations in a given metric space, is a critical component of the similarity kernel.

By grouping patients based on their genetic or molecular characteristics, physicians can develop tailored treatments that are specific to each patient's needs, maximising the chances of success and minimising the risk of side effects. Additionally, the patient grouping can help to identify common biomarkers that are associated with particular diseases or conditions, providing valuable insights into the underlying mechanisms and potential targets for treatment.

1.2 Motivation

Genetic changes within the genome give rise to cancer. When specific genes responsible for regulating essential functions like cell growth undergo alterations, they become activated and expressed at unusually high levels. This abnormal gene expression leads to uncontrolled cell growth, resulting in the formation of tumours. When gene expression data is visualised through density plots, it becomes possible to identify abnormal gene expression values which are located at the extreme tails of the distribution, significantly distant from the mean. The clustering of this data provides the capability to detect genetic alterations for each patient and further classify tumours into more precise subtypes, facilitating the development of tailored therapies and target drug design. However, clustering data that includes extreme gene expression values can result in the formation of spurious clusters. Specifically, these extreme values can have a significant impact on the calculation of distances between observations from the data, complicating the disease subtyping process. To estimate the distance between observations in data, distance functions often utilise classic statistical functions such as mean and standard deviation. Whilst, these statistical functions perform well on compact and isolated clusters, these are sensitive to outliers [19, 20, 21]. Outliers influence the distance between observations in the interquartile range (IQR) [22, 23]. Even a single highly influential outlier has a massive impact on these classic statistical functions [24]. The lack of robustness to outliers restricts these techniques' ability to provide a reliable measure for estimation. A possible solution is to eliminate outliers before computing pairwise distances. Outliers, however, can sometimes provide useful information about unusual behaviour [24]. As a result, deleting outliers that contain such valuable information might have a negative impact on the assessment of data variability and distances. Several distance methods are available to solve the outlier observation problem. One of them is the average distance function, which deals with outliers in ”n” dimensions. Others use weights to provide relative priority to each attribute. However, weights are computed differently depending on the kind of dataset, and activity [19]. Feature scaling is another way to combat the robustness of distance-matching functions, but their disadvantage is that the magnitude of the distance remains constant. Furthermore, while certain ranking-based algorithms are resistant to outliers, they lack scale, and variability, and neglect the degree of proximity between values. These approaches, in particular, are unable to determine how much better or worse one value is than another. As a result, a large quantity of critical information included in the data is lost [21].

1.3 Problem Statement

Outliers and extreme values are types of data points in scientific research that can have a significant impact on the results of statistical analysis. They differ, however, in terms of their fundamental causes. Outliers are data points in a sample that differ significantly from the rest of the data. Outliers may occur as a result of human error, measurement or data entry error, these can indicate unusual or unexpected behaviour. Extreme values, on the other hand, are data points located at the extreme ends of the data range. Extreme values, unlike outliers, are not always erroneous; these might simply indicate the upper or lower bounds of what is feasible for a certain variable. Extreme values and outliers can significantly affect statistical analysis by distorting the distribution of data and exerting influence on measures of central tendency and variability. In the context of genomics data, which inherently involves complexities such as high-dimensionality versus limited observations, data variability, and the presence of extreme values and outliers, the performance of machine learning (ML) algorithms is constrained in achieving the desired level of robustness for uncovering clinically relevant subtypes. [25] [26] [27] [28]. These properties of the genomics data present issues for both supervised ML models (which frequently leads to model overfitting) and clustering techniques that employ the similarity graph to put observations into coherent groups. The complexity of genomics data in these circumstances leads to a sparse patient similarity graph, making it difficult for clustering algorithms to group some of the patients [29] [30]. The above difficulties impede the identification of disease subgroups defined by clinical characteristics, such as survival [31]. Legacy disease subtyping approaches achieve robustness without considering extreme values and data variability in clustering omics data. The extreme values need proper attention, otherwise, these will pose a negative impact on the results. Addressing these characteristics can add extra dimensions to robustness and provide useful information in molecular subtype discovery. Moreover, disease subtyping comes with high-dimensional multi-view (gene expression, DNA methylation, and miRNA) data. It requires integrative analysis to discover subtypes not only at a single view but as a whole that can take into account facts from other views [29]. This stresses the need for robust approaches to minimise the influence of noise (extreme values, data variability) in discovering distinct groups.

1.4 Aim and Scope of the Research

In this thesis, robust statistical measures are investigated for unsupervised clustering algorithms to synthesize the structurally relevant observations in omics data. These robust measures will be introduced to spectral clustering algorithms for the effective discovery of disease subtypes.

Robust statistical measures are designed to be resilient against the influence of extreme values and outliers. Robust measures offer reliable and stable estimates of central tendency and variability, ensuring the accuracy of the analysis even when extreme values are present in the data.

1.4.1 Aim

The research aims to develop a robust approach to enhance the discovery of distinct groups in complex high-dimensional data for disease subtyping.

1.4.2 Research Objectives

1. To develop a novel distance function that exhibits robustness. This function is designed to address the challenges posed by data variability and extreme values when computing proximity between observations within the intermediate graph models. It ensures reliable and accurate measurements in high-dimensional spaces, promoting more effective analysis and subtyping.

2. Develop Intermediate Graph Models (IMGs), to enhance the identification of similarities between pair of samples. These IMGs represent the topological graph structure of the data, which aids in identifying patterns and relationships within the dataset. By incorporating IMGs into our approach, we aimed to improve the performance of distance metrics in accurately measuring the similarities between samples.

3. To develop a disease subtyping approach based on this robust distance function to accurately discover disease subtypes defined by clinical differences, such as survival outcomes. This approach leverages the robustness of the proposed distance function to enable the precise discovery of distinct disease subtypes, ultimately leading to aid in personalised treatment strategies.

4. The final objective aim to demonstrate the robustness, accuracy, and effectiveness of the proposed disease subtyping approach across diverse datasets. Therefore, the proposed disease subtyping approach will be validated using a range of datasets, including Genomics, Synthetic, and Generic machine learning datasets. Extensive evaluation will be conducted using various metrics, such as Cox-proportional hazards (Cox P-value) for survival analysis, Concordance statistics (CI) to assess the predictive ability of the approach, and Normalized Mutual Information (NMI) and Clustering purity to evaluate the clustering performance of the approach. Finally, the stability of the proposed approach against the noise will be evaluated by introducing various levels of noise in the data.

1.5 Significance, Contribution and Benefits of this Research

Clustering analysis plays a vital role in various domains, including pattern recognition, consumer segmentation, and disease subtyping. Particularly, discovering disease subtypes based on similar molecular and clinical characteristics from multi-view high-dimensional data is crucial for targeted drug design, clinical diagnosis, and treatment selection. Therefore, this research contributes to the existing knowledge in clustering analysis by introducing robust statistical approaches to effectively discover disease subtypes in complex high-dimensional omics data.

The research makes several significant contributions. Firstly, it develops intermediate graph models (IMG) from omics data, enabling the integration of structurally meaningful measures into coherent groups for improved subtype separability. Secondly, a novel robust function (ROMDEX) is proposed, which utilises IMGs to address the challenges posed by data variability and extreme values when computing proximity between observations in high-dimensional spaces. This robust function enhances the accuracy of clustering analysis. Thirdly, the proposed robust function is incorporated into a novel disease subtyping framework, enabling the accurate discovery of disease subtypes characterised by clinical differences, such as survival outcomes. This framework provides valuable insights for actionable target drug design and personalised treatment strategies. Lastly, the proposed approach is validated using datasets from genomics, synthetic, and generic machine learning domains, ensuring its applicability and effectiveness across different data types and domains.

Beyond the biomedical field, a robust clustering strategy has broad applications across business organisations. It can optimise operations by offering better consumer segmentation and revealing hidden insights in noisy datasets. The proposed robust clustering approach is generic and can be applied to various domains. For example, in clinical trials, it can aid in evaluating the efficacy of new medicines.

1.6 Validation of the Research

The evaluation of results on omics data for disease subtyping is based on Kaplan-Meier survival time analysis, which is validated using statistical tests e.g., Cox-proportional hazard (Cox p-value). We also included concordance statistics for evaluating the fitted survival model on five TCGA datasets. The concordance index (CI) is used to evaluate the predictive ability of the survival model. The CI values of the fitted survival model for all the datasets are impressive which demonstrates the predictive ability of the proposed unsupervised graph-based disease subtyping.

1.7 Layout of the Thesis

Chapter 2 – Background

In this chapter, the background related to the proposed research is provided.

Chapter 3 – Literature Review

In this chapter, existing machine learning algorithms for disease subtyping are investigated and their analysis is performed. Also, the limitations of each approach are discussed.

Chapter 4 – Methodology

In this chapter, the overall multi-view graph-based clustering pipeline is provided and explained. This is followed by the proposed approach for robust similarity graph construction. Also, the evaluation metrics are described in detail.

Chapter 5 – ROMDEX

In this chapter, the mathematical background for the proposed approach is provided. Also, the effect of extreme values and data variability in clustering multi-view high-dimensional data are reviewed and a novel robust statistical solution is provided. Also, a novel end-to-end algorithm is proposed for robust graph-based clustering and disease subtyping.

Chapter 6 – Results

The proposed methodology is validated on real-world and synthetic cancer datasets in this chapter, and the outcomes were compared to multiple baseline disease subtyping approaches.

Chapter 7 – Discussions

In this chapter, the research findings are analysed and interpreted, and the significance of the results is explained.

Chapter 8 – Conclusion & Future works

Finally, in this chapter, the conclusion, and future works are discussed.

Chapter 2

Background

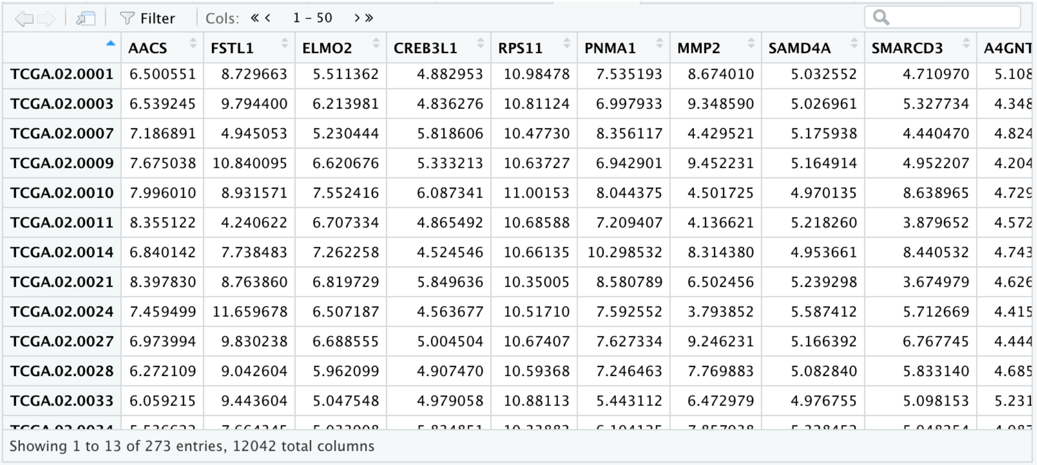
2.1 Introduction

This chapter provides background information on the proposed research. The chapter begins by defining disease subtyping—it is the process of identifying different types of diseases based on the symptoms, causes, and treatments. It is a crucial step in the diagnosis and treatment of diseases, as it can help doctors choose the most effective course of treatment for each patient. To understand the disease subtyping process, existing approaches and techniques are discussed. Clustering analysis is commonly used to discover subtypes of diseases. Therefore, how clustering analysis has been used so far in the study of disease subtypes is also discussed. Clustering provides insights into the similarities and differences between different types of diseases. Additionally, clustering analysis can be relatively easy to use and interpret, making it a valuable approach for researchers and clinicians alike. The backbone of clustering is similarity measures; therefore, various similarity measures (distance metrics), their limitations, and similarity kernels are also highlighted at the end of this chapter.

2.2 Dataset Overview

The dataset used for disease subtyping typically encompasses multiple modalities, such as genomics, transcriptomics, and medical imaging. For our analysis, we will focus on genomics data, specifically gene expression, DNA methylation, and MicroRNA. This chapter will concentrate on a single view or modality, namely gene expression. In subsequent chapters, we will propose a robust approach that analyses each individual view separately before integrating all the views to perform disease subtyping.

Furthermore, we will incorporate clinical data, including survival information, to evaluate the proposed approach. These clinical variables will provide valuable insights during the analysis. To illustrate this, the figure (2.1) below displays the gene expression data extracted from glioblastoma multiforme cancer (GBM).



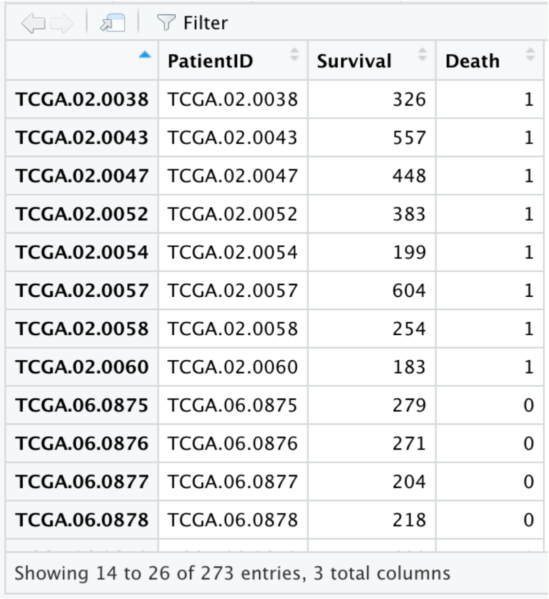
	AACS	FSTL1	ELMO2	CREB3L1	RPS11	PNMA1	MMP2	SAMD4A	SMARCD3	A4GNT
TCGA.02.0001	6.500551	8.729663	5.511362	4.882953	10.98478	7.535193	8.674010	5.032552	4.710970	5.108
TCGA.02.0003	6.539245	9.794400	6.213981	4.836276	10.81124	6.997933	9.348590	5.026961	5.327734	4.348
TCGA.02.0007	7.186891	4.945053	5.230444	5.818606	10.47730	8.356117	4.429521	5.175938	4.440470	4.824
TCGA.02.0009	7.675038	10.840095	6.620676	5.333213	10.63727	6.942901	9.452231	5.164914	4.952207	4.204
TCGA.02.0010	7.996010	8.931571	7.552416	6.087341	11.00153	8.044375	4.501725	4.970135	8.638965	4.729
TCGA.02.0011	8.355122	4.240622	6.707334	4.865492	10.68588	7.209407	4.136621	5.218260	3.879652	4.572
TCGA.02.0014	6.840142	7.738483	7.262258	4.524546	10.66135	10.298532	8.314380	4.953661	8.440532	4.743
TCGA.02.0021	8.397830	8.763860	6.819729	5.849636	10.35005	8.580789	6.502456	5.239298	3.674979	4.626
TCGA.02.0024	7.459499	11.659678	6.507187	4.563677	10.51710	7.592552	3.793852	5.587412	5.712669	4.415
TCGA.02.0027	6.973994	9.830238	6.688555	5.004504	10.67407	7.627334	9.246231	5.166392	6.767745	4.444
TCGA.02.0028	6.272109	9.042604	5.962099	4.907470	10.59368	7.246463	7.769883	5.082840	5.833140	4.685
TCGA.02.0033	6.059215	9.443604	5.047548	4.979058	10.88113	5.443112	6.472979	4.976755	5.098153	5.231

Showing 1 to 13 of 273 entries, 12042 total columns

Figure 2.1: The gene expression data obtained from GBM cancer. In this figure, rows represent samples and columns represent genes.

The data shown in figure (2.1) has 12042 gene expression values measured on 273 samples. The gene expression data extracted from GBM cancer is unlabelled, indicating that the true number of clusters or disease subtypes is unknown. To assess the quality of clusters (disease subtypes) generated from this data, we will utilise clinical data that includes survival information. The following figure (2.2) depicts this data.

In the figure (2.2) *Survival* column denotes the survival time in days while the *Death* column denotes the outcome variable of interest. It is worth noting that the samples in the clinical data figure (2.2) are the same as the samples in the gene expression data figure (2.1).



	PatientID	Survival	Death
TCGA.02.0038	TCGA.02.0038	326	1
TCGA.02.0043	TCGA.02.0043	557	1
TCGA.02.0047	TCGA.02.0047	448	1
TCGA.02.0052	TCGA.02.0052	383	1
TCGA.02.0054	TCGA.02.0054	199	1
TCGA.02.0057	TCGA.02.0057	604	1
TCGA.02.0058	TCGA.02.0058	254	1
TCGA.02.0060	TCGA.02.0060	183	1
TCGA.06.0875	TCGA.06.0875	279	0
TCGA.06.0876	TCGA.06.0876	271	0
TCGA.06.0877	TCGA.06.0877	204	0
TCGA.06.0878	TCGA.06.0878	218	0

Showing 14 to 26 of 273 entries, 3 total columns

Figure 2.2: The clinical data obtained from GBM cancer samples. In this figure, rows represent samples and columns represent clinical data.

2.3 Disease Subtyping

Disease subtyping is the process of identifying different sub-types of a disease within a population [32] [29] [25]. It is a way of further understanding how a disease progresses and identifying which groups of people are most at risk. Disease subtypes can be identified using either an expert-driven or data-driven approach [33]. In the expert-driven approach, domain experts, including clinicians and researchers, utilise their expertise and knowledge to define subtypes based on one or more criteria. This approach relies on the profound insights and understanding gained through years of clinical experience and scientific research. To classify patients into subtypes, experts thoroughly examine a vast amount of data including clinical records, medical imaging, genetic information, laboratory tests, and other relevant data sources. However, this manual examination of extensive data can place a significant burden on experts and may introduce the possibility of human error. Technological advancements have made it possible to use large omics data to identify subtypes more accurately [32] [34] [35]. Therefore, data-driven approaches for disease subtyping are gaining significant research attention. Data-driven approaches leverage the power of machine learning (ML) to efficiently integrate and analyse vast amounts of genetic and clinical data for the discovery of disease subtypes. These approaches have the potential to revolutionise our understanding of diseases. Given a large data set of people with a certain disease, machine learning algorithms can be used to identify patterns and group people together based on their similarities [36] [37]. This process can reveal previously unknown disease subtypes and aid in our understanding of how the disease progresses.

Data-driven disease subtyping approaches are still in their early stages, but they have the potential to transform our understanding of diseases and improve our ability to treat them. These approaches can broadly be categorised based on the number of datatypes they take into account during subtyping. The approaches which only consider a single datatype a.k.a, single-view are known as single-view analysis approaches. On the other hand, the approaches which consider multiple datatypes simultaneously during the subtyping process are known as integrative analysis approaches [36]. These approaches are briefly described in the following section.

2.3.1 Single-view Analysis of Omics for Subtype Discovery

Disease sub-typing datasets often come in multi-view formats such as gene expression, DNA methylation, and MicroRNA, which are high-dimensional. The approaches focusing solely on a single view, such as gene expression, for disease subtyping are referred to as single-view analysis. However, considering multiple views is important as they offer complementary information, and neglecting their integration could result in the loss of valuable insights [38] [39]. To address this challenge, integrative analysis is gaining popularity as these methods consider all the views during subtyping.

2.3.2 Integrative Analysis of Omics for Subtype Discovery

An integrative analysis is a type of data analysis that combines information from multiple sources or views to gain a more complete understanding of a phenomenon [25]. Integrative approaches can be used to study anything from disease subtypes to social networks. Integrative approaches have widely been adopted in bioinformatics for subtyping diseases [40] [34] [35]. The integrative analysis is important as it allows us to see relationships between different data views that we might not be able to see if we were only looking at one set of data [41] [42]. By combining data from multiple views, we can get a more complete picture and make better predictions about future events.

2.4 Clustering Analysis a key driver for Disease Subtype Discovery

In recent years, doctors have been able to subtype diseases into more and more specific categories. This has led to better treatments and outcomes for patients and a greater understanding of the disease itself. In recent years, the introduction of new high-throughput sequencing technologies has enabled the rapid and low-cost characterisation of genomes.

Clustering analysis is a machine-learning approach used to group a set of data points with similar characteristics so that they can be more easily analysed [43]. It is also, widely used in data mining applications to discover patterns in data that would otherwise be difficult to find [43]. In machine learning, it plays a significant role in understanding the relationship between different data points.

Clustering is applied in various domains, but in this thesis, we will focus on clustering from a bioinformatics perspective. One application of clustering in bioinformatics is disease subtyping [44]. Disease subtyping is the task of dividing a disease into different subtypes based on common characteristics [32] [29].

Clustering and other statistical methods are used for subtyping to understand the complex structure of diseases [45]. This helps in finding new disease subtypes and assists experts to understand how diseases progress. It is also used to predict patient outcomes and identify possible treatments [46] [47]. Particularly, for cancer diseases that are caused by multiple factors, and we might be interested in knowing the most common factors. The clustering of patients based on these factors can pinpoint the causing factors for specific diseases. Clustering algorithms that have widely been used for disease subtyping are categorised and included in the following section. In this thesis, the clustering approaches are broadly categorised as Partitional-based, Hierarchical, and Consensus-based clustering. These are briefly reviewed in the following section.

2.4.1 Partitional Clustering

Partitional clustering is a clustering approach that uses similarity to divide observations within a data set into multiple groups. These approaches require the number of clusters (k) to divide the observations into k sets of groups. Following are the partitional-based clustering approaches.

K-Means Clustering

The most common type of clustering algorithm is the k-means algorithm [48]. The k-means algorithm selects a number of clusters (k) and then assigns each data point to one of the clusters at random. The centre of each cluster is computed once all data points have been allocated to a cluster. This is done by taking the mean of all the data points in the cluster. Each data point is then reassigned to the cluster whose centre is closest to it. This process is repeated until there are no more changes in the assignment of data points to clusters. However, extreme values and outliers can influence the clustering results generated by k-means algorithm.

Spectral Clustering

The most popular methods for subtyping disease are methods that use some form of clustering. There are a variety of different clustering methods, but one of the most popular for disease subtyping is spectral clustering [16]. Spectral clustering is a method that uses the eigenvectors of a similarity matrix to group data points together. The similarity matrix is created using a kernel function, which measures the similarity between data points. The eigenvectors of the similarity matrix are used to create clusters, and each data point is assigned to the cluster that is most similar to it. Spectral clustering has been shown to be effective for subtyping cancer diseases [49] [13] [50]. The results showed that spectral clustering could accurately group patients with similar clinical outcomes together and that the resulting subtypes were clinically meaningful.

Clustering Gene Expression Data using Partitional Clustering

In order to gain a basic understanding of how partitional clustering operates on the gene expression data (figure 2.1), we applied k-means clustering using the default settings shown in figure (2.3). Furthermore, to enhance our comprehension and visualise the clustering outcomes, we focused on two specific genes that are known to be associated with GBM cancer.

In Figure (2.3), the k-means clustering partitioned the samples into three clusters based on two selected genes. However, it is evident that this data is challenging to cluster as some samples are widely separated from the rest. Additionally, determining the optimal number of clusters, denoted as K, is difficult for this dataset. Although the number of clusters, K, was provided as an input beforehand, there are machine learning techniques available that can estimate the potential number of clusters in the data. These techniques can then be utilised to cluster the data using any of the clustering algorithms discussed in this chapter. Several of these techniques will be described in this chapter. Subsequently, we will employ these techniques to estimate the number of clusters and re-cluster the data accordingly.

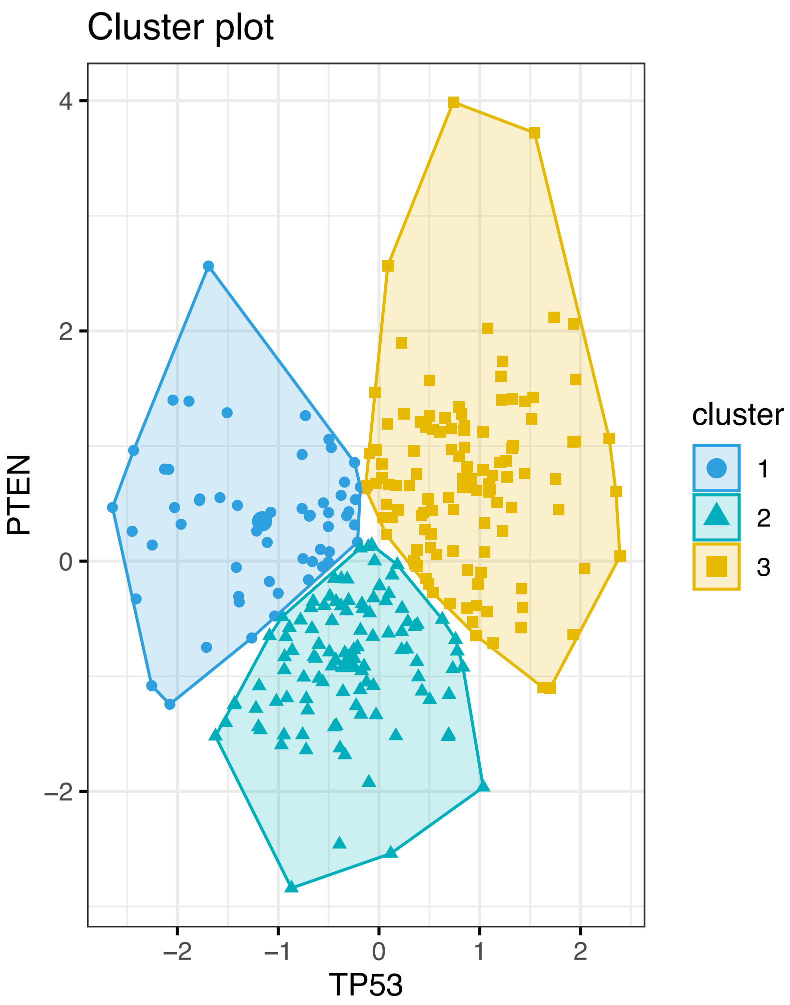


Figure 2.3: The K-means clustering results on the gene expression data (figure 2.1) on two genes namely PTEN, and TP53.

2.4.2 Hierarchical Clustering

Hierarchical clustering can be used to discover subtypes of disease in molecular data. This method may be used to group data points according to their symptoms, demographics, or even genetic information. Hierarchical clustering is a cluster analysis approach that aims to create a hierarchy of data points in the cluster. [51]. This approach starts with all samples in a single cluster and then successively splits or merges clusters until a desired level of granularity is achieved. An advantage to hierarchical clustering is that it visualises the structure of the data. This can be helpful in understanding the relationships between different disease subtypes.

Agglomerative

Agglomerative clustering is a type of hierarchical cluster analysis that produces a hierarchy of clusters. It is a bottom-up approach, where each data point is treated as a single cluster, and clusters are then merged together until all points are in one cluster. This process can be represented by a dendrogram, which shows the order of merging and the resulting cluster structure. First, a similarity matrix is computed, which contains the pairwise similarities between all data points. Then, the algorithm proceeds to merge the most similar pairs of points into clusters, until all points are in one cluster. The order in which the elements are merged can have a significant impact on the results of the algorithm, so care must be taken when choosing an agglomerative clustering algorithm. One advantage of agglomerative clustering is that it can be used to construct dendrograms, which can be helpful for visualising the structure of complex data sets. Additionally, this approach does not require that the number of clusters is specified upfront, as is necessary with k-means clustering.

Divisive

Divisive clustering is a type of hierarchical cluster analysis that involves the recursive division of a dataset into smaller and smaller clusters. It is a top-down approach. The divisive clustering algorithm starts by assigning all of the data points to a single cluster. It then iteratively splits the cluster into two smaller clusters until each cluster contains only one data point.

Clustering Gene Expression Data using Hierarchical Clustering

To gain a deeper understanding of how hierarchical clustering functions, we applied it to the gene expression data (Figure 2.1) using the same two gene measurements. The initial step of hierarchical clustering involves generating a dendrogram as shown in Figure (2.4), which is a tree structure representing the relationships between data points. Figure (2.4) illustrates this dendrogram with varying levels. Subsequently, the dendrogram can be utilised to cluster the data points, or samples, into distinct groups.

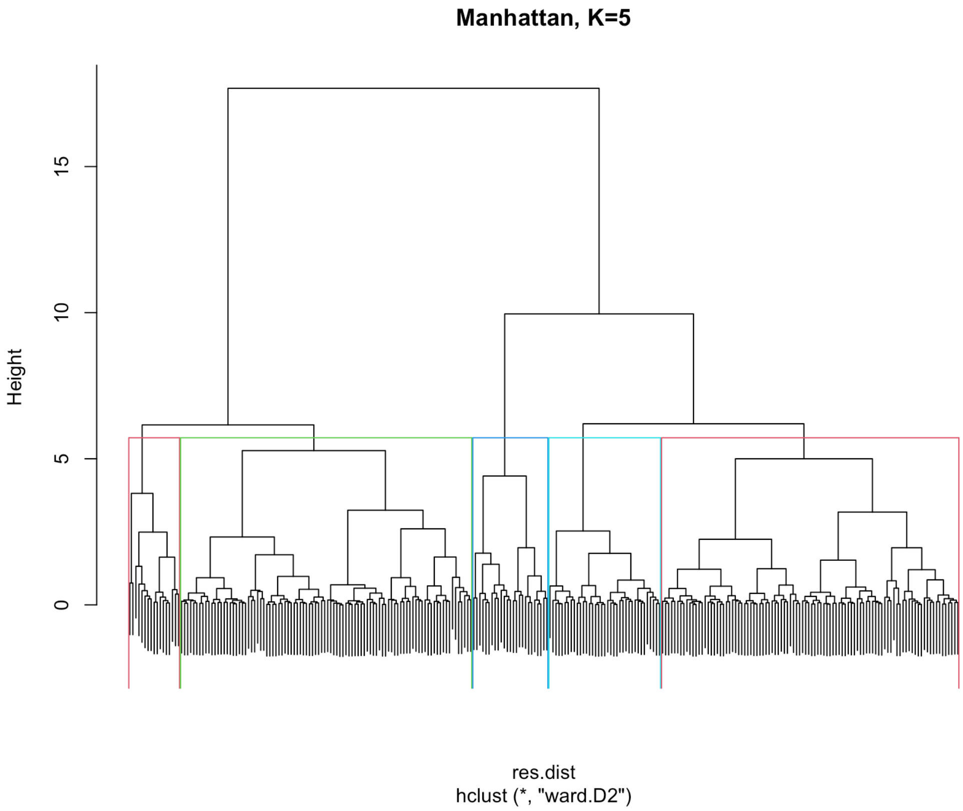


Figure 2.4: The dendrogram obtained from the gene expression data (Figure 2.1) for hierarchical clustering.

To cluster the samples represented in the dendrogram a technique known as *cutree* is employed. This technique generates a vector with group memberships when k or h is a scalar. In this case, each column of the vector corresponds to the elements of k or h. The parameter k represents the number of clusters, while h indicates the heights at which the tree should be cut. Figure (2.5) shows the hierarchical clustering results which are obtained after applying *cutree* technique on this dendrogram with $k = 2, 3, 4, 5$, and Manhattan distance.

To determine the optimal number of clusters, k, various techniques are commonly utilised, including the Elbow method [52], average Silhouette [53], and Gap statistic [54]. These techniques provide valuable insights for selecting the most appropriate number of clusters in a given dataset. We applied these approaches to determine the optimal number of clusters on the gene expression data. The

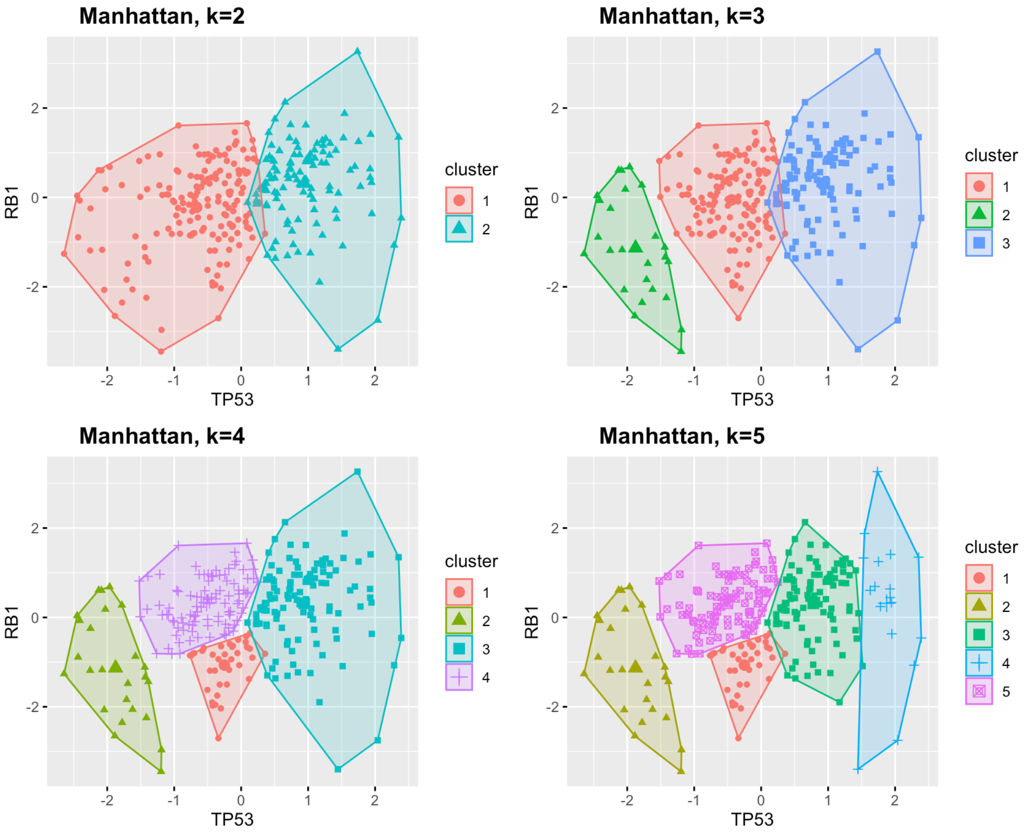


Figure 2.5: The hierarchical clustering results obtained after using *cutree* technique on the dendrogram (Figure 2.4). We used $k = 2, 3, 4, 5$ and the Manhattan distance to obtain these clusters.

results produced by these approaches are shown in the following Figure (2.6). In Figure (2.6) the top-left plot shows the Elbow method which computes the total within sum of square between the data points and the top-right plot shows the average Silhouette method while the bottom-left plot shows the Gap statistics.

The Elbow method demonstrates that as the number of clusters increases, the total within sum of squares (WSS) value decreases. The WSS value reaches its highest point when K equals 1. By examining the figure, we can see a sharp decline in the graph as the number of clusters, K, increases, resulting in an elbow-like shape. The optimal number of clusters, K, is typically determined at the point where the graph begins to flatten out and move parallel to the X-axis. Although it can be somewhat challenging to precisely identify this point from this plot alone, in this case, a possible value for K is six. Similarly, Average Silhouette suggests K=2, and the Gap statistic method suggests K=4.

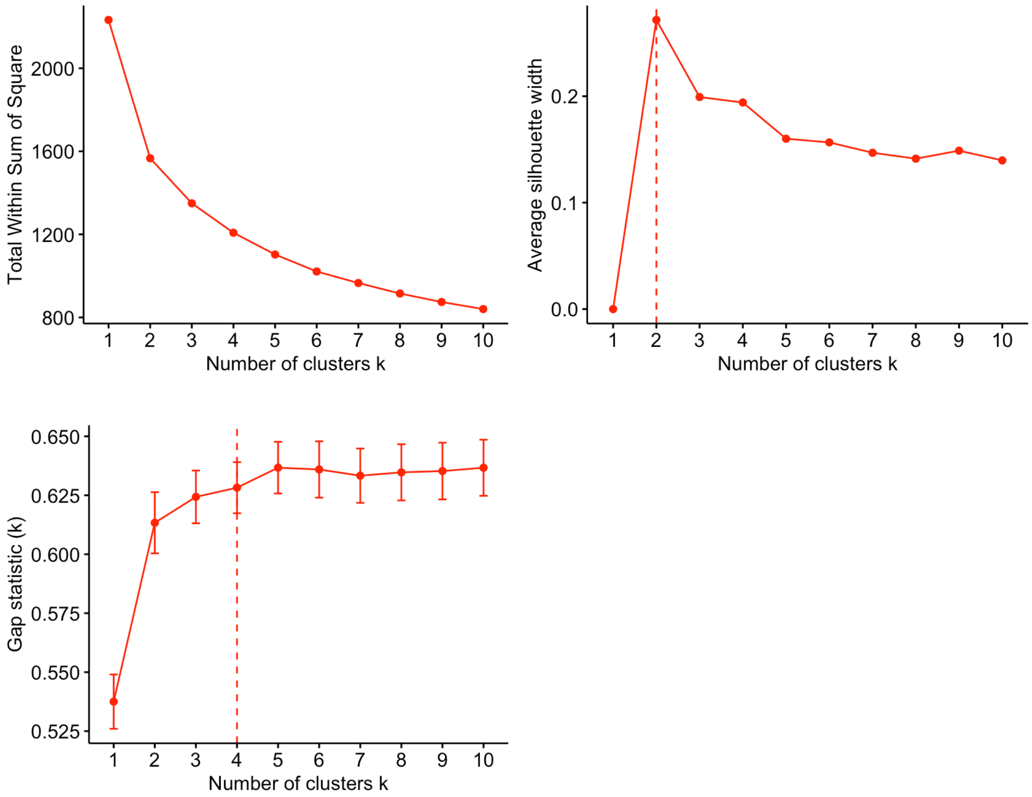


Figure 2.6: The optimum K value determined by Elbow method, Average Silhouette and Gap statistic on gene expression data (Figure 2.1).

2.4.3 Consensus Clustering

Consensus Clustering is a method of combining multiple individual clustering into a single, robust clustering [55] [56] [57]. It is a form of unsupervised classification, which aims to find natural classes or clusters by using data from repeated clustering runs. It computes how often pairs of samples are grouped together and then uses the resulting pairwise ”consensus rates” to visualise clusters, compare the clustering stability, and determine the number of clusters to be created. Consensus clustering is least affected by outliers and it is more robust to changes in the data than other methods. It has been widely and successfully used for disease subtyping [58] [59] [60]. To perform consensus clustering, first, it needs to generate a set of individual clusters. Once a set of individual clusters is generated, those are then combined using a consensus function. The most popular method to combine clustering is the plurality rule. The plurality rule simply takes the most common label from all of the individual clusters and assigns it as the label for the consensus clustering. This method is easy to implement and understand, but it may not always give the best results. Another method that is often used is called the average linkage rule. The average linkage rule works by finding the pair of points that are closest together and assigning them to the same group.

In general, for clustering algorithms, it is important to tune their parameters so that these can work well with the data. For instance, to determine the number of clusters k. In many cases, the desired number of clusters is known ahead of time. However, in other cases, it must be determined experimentally. Common methods for choosing the number of clusters are the Elbow method [52], average Silhouette [53], and Gap statistic [54] as shown in the Figure (2.6).

2.4.4 Evaluation and Assessment of Clustering

There are a variety of ways to evaluate and assess the performance of a clustering algorithm [61]. In general, we want to know how well the algorithm has clustered the data points, and whether or not the clusters make sense from the domain perspective. One common approach is to use an external criterion, such as expert knowledge, to compare the clusters generated by the algorithm with known groupings [61][62]. Another approach is to use a measure of similarity between data points within a cluster, and between data points in different clusters, to assess the quality of the clusters. In either case, it is important to have a clear understanding of the problem that is being tried to solve with clustering, in order to choose an appropriate evaluation method.

A variety of evaluation metrics have been proposed in the literature, where each metric evaluates the generated clusters from a different aspect. The widely used cluster evaluation metrics in biomedical informatics particularly for disease subtyping are described in detail at the end of the methodology chapter.

Characteristics of Good Clustering Analysis: a good clustering analysis should be able to identify different groups of data points with similar characteristics. It should be able to handle different types of data, including both numerical and categorical data. The clusters that are generated should be well-defined and clearly separated from each other. Additionally, the clustering algorithm should be able to run quickly and efficiently on large-scale and high-dimensional datasets.

2.5 Similarity Measures

Similarity measures play a significant role in clustering. As the Clustering algorithms group data points together based on their similarity. Therefore, similar data points are more likely to be assigned to the same group or cluster. Firstly, a similarity metric is selected to measure how similar two data points are. Common similarity metrics include Euclidean distance, Manhattan distance, and Cosine similarity [63] [64] [65]. For a similarity measure to be a valid metric, it should satisfy all the axioms of the metric space. Below in this section, we provide a formal definition of metric space and it is axioms.

2.5.1 Distance Metrics and the Metric Space

In mathematics, a metric space is a set for which distances between all members of the set are defined. More specifically, given any two members of the set, there is a real number associated with their distance that satisfies certain properties [66]. Distances in a metric space can be measured using any one of a variety of different distance measures, such as Euclidean distance or Manhattan distance. Formally, a metric space is an ordered pair (X, d) where X is a set and d is a metric or distance function defined between every pair of elements of X e.g., $d : X \times X \to \mathbb{R}^+$. Suppose we have two observations x, y in n-dimensional space e.g., x,y $\in$ X where $x = \{x_i, i \in 1, ..., n\}$, and $y = \{y_i, i \in 1, ..., n\}$. The distance function d must satisfy the following properties, called axioms, to be considered a valid metric.

1. $d(x, y) \geqslant 0 \ \forall \ x, y \in X$ (Positive Semi-definite)

2. $d(x, y) = 0 \iff x = y$ (Identity of Indiscernible)

3. $d(x, y) = d(y, x) \; \forall \; x, y \in X$ (Symmetry)

4. $d(x, z) \leqslant d(x, y) + d(y, z) \; \forall \; x, y, z \in X$ (Triangle Inequality)

The first axiom, non-negativity or Positive semi-definite, states that the distance between any two points must be greater than or equal to zero. This makes sense intuitively; the distance between two points can never be negative. The second axiom, Identity of Indiscernible, states that the distance from a point to itself must always be zero. The third axiom, symmetry, says that the distance between two points is the same in either direction. So if the distance from x to y is m, then the distance from y to x is also m. Again, this makes intuitive sense. The fourth and final axiom is the triangle inequality. This states that the sum of the distances of any two sides of a triangle must be greater than or equal to the length of the third side. So in other words, if you have three points x, y, and z, then the distance from x to z must be less than or equal to the sum of the distances from x to y, and from y to z. Intuitively, this just means that it's never shorter to go around a corner than it is to go straight ahead. These four axioms define what it means for a function to be a metric on a set of points. Together they ensure that the function satisfies all of these axioms. There are many different ways to measure distances between points in a metric space. The most common distance measures are Euclidean distance and Manhattan distance. Euclidean distance is the straight-line distance between two points, while Manhattan distance is the sum of the absolute values of the differences in the coordinates of two points. Other common distance measures include Chebyshev distance and Minkowski distance.

Which distance measure is best to use depends on the specific application. In some cases, one measure may be more appropriate than another. For example, Euclidean distance is often used when working with data that lie on a regular grid, such as an image. Manhattan distance may be more appropriate when working with data that is not regularly spaced, or have high dimensions.

Example 1. Let X $= \mathbb{R}^n$ and $x, y \in$ X where $x = \{x_i, i \in 1, ..., n\}$, and $y = \{y_i, i \in 1, ..., n\}$, Euclidean distance between these points is defined as follows:

$$d_2(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2.1}$$

In the equation above x, y are two points in Euclidean space where x_i, y_i are Euclidean vectors denoting the initial point or the origin of Euclidean space and finally, n denotes the dimensions of the space. Note that the subscript in d_2 denotes that the Euclidean distance is L_2-norm. It is the most common type of distance metric and is also known as the L_2-norm or simply the Euclidean norm. There are many ways to measure the distance between data points in space. Euclidean distance is the common metric which is the straight-line distance between two points [63]. However, this method can be limiting when you're working with data that are not evenly distributed. In this thesis, we'll take a look at another distance measure called the Manhattan distance [64]. This method is often used in data clustering, as it can help to accurately group together data points in high-dimensional spaces [67]. We'll also explore how the Manhattan distance can be used to construct robust similarity graphs from high-dimensional data. Other measures that are sometimes used include cosine similarity and Jaccard similarity. Cosine similarity is a measure of how similar two vectors are and Jaccard similarity is a measure that tells how many items two sets have in common.

Example 2. Let X $= \mathbb{R}^n$ and $x, y \in$ X where $x = \{x_i, i \in 1, ..., n\}$, and $y = \{y_i, i \in 1, ..., n\}$, Manhattan distance between these points is defined as follows:

$$d(x, y) = \|x - y\|_{L1} = \sum_{i=1}^{n} |x_i - y_i| \tag{2.2}$$

It computes the distance between two points $x, y \in X$ as the sum of the absolute differences of their Cartesian coordinates, and it satisfies all of the four properties of a metric defined above. The Manhattan distance function (L_1 norm) is preferred for high dimensional data compared to the Euclidean distance function (L_2 norm) [67]. In general, it is important to choose a distance measure that reflects the underlying structure of the data. This will ensure that clustering results are meaningful and interpretable.

Limitations of Distance Measures

There are a few limitations to consider when using distance measures for clustering. Firstly, the choice of distance measure can be critical and any given measure may not be appropriate for all data sets. Secondly, even when an appropriate distance measure is used, the results can be sensitive to outliers. And finally, clusters found by a distance-based method may not have a clear interpretation. In addition, the notion of similarity, and distance which is crucial for clustering, becomes qualitatively less meaningful. In a detailed behavioural examination of the distance functions (L_k norm) it has been shown that the problem of meaningfulness is sensitive to the value k [67].

2.5.2 Similarity Kernels

Differentiating between different subtypes of diseases is an important task for both clinicians and researchers. However, it can be difficult to accurately identify disease subtypes, especially when there are many different types of diseases. One way to tackle this problem is to use similarity kernels. Similarity kernels are mathematical tools that can be used to compare two objects and measure their similarity. In the context of disease subtyping, similarity kernels can be used to compare different diseases and measure their similarities. There are many different types of similarity kernels, but we will present a few below in this chapter. A kernel is a similarity measure between two data points. It quantifies the similarity between two data points in terms of a distance metric, such as Euclidean distance or Manhattan distance. Kernels are often used in machine learning algorithms, such as support vector machines, to find patterns in data. Kernels can be classified into two types: linear kernels and nonlinear kernels. Linear kernels are defined by a dot product between two vectors, while nonlinear kernels are defined by a transformation of the input vectors. Common examples of nonlinear kernels include the polynomial kernel and the Radial Basis Function (RBF) kernel. The use of similarity kernels has been found to be very successful in many machine learning applications, such as image recognition and classification, text categorisation, and document clustering. In fact, most modern machine learning algorithms make use of some form of similarity kernel in order to learn from data. In the following section, both linear and non-linear kernels are reviewed.

Kernel Trick

The kernel trick is a method used in machine learning to implicitly map data into a high-dimensional space which enables linear models to solve nonlinear problems. The trick is to compute the inner products of the data points in the new space, which are then used as features in the standard linear model. This mapping is done implicitly, meaning that it does not require any explicit knowledge of the high-dimensional structure of the data. The kernel trick has been shown to be very effective in practice and has been used to develop some of the most successful machine-learning algorithms. In order to apply the kernel trick, we need a kernel function.

Definition 2.1. A kernel function is defined as follows:

$$K(x, y) = \langle \phi(x), \phi(y) \rangle \tag{2.3}$$

Here, K is a kernel function, and $x, y \in \chi$ are data points in χ-dimensional space. The mapping function ϕ maps the data points x, y from input space χ to another higher-dimension space υ e.g., $\phi : \chi \rightarrow \upsilon$. The angle brackets $\langle ., . \rangle$ define a proper inner product of the data points which makes this computation efficient.

The most popular kernel functions are the linear, polynomial, Radial Basis Function (RBF), and KNN kernels. Once the kernel function is selected, then it can be applied to transform the data points into higher-dimensional space. After this transformation has been performed, the linear model can be used on the transformed data points to get improved performance. In the following section we briefly explain the above-mentioned similarity kernels:

Linear Kernel

The linear kernel is one of the most commonly used kernels for support vector machines SVMs. It works well for data that is linearly separable, which means that the data points can be separated by a line (or hyperplane in higher dimensions). The linear kernel is also one of the simplest kernels, which makes it easier to interpret the results of the classification.

A linear kernel is a similarity function that measures the similarity between two vectors by their dot product. In other words, it measures how much two vectors are linearly correlated.

The linear kernel is often used in machine learning algorithms that require a similarity measure particularly, it is commonly used in text classification.

Definition 2.2. The linear kernel is defined as:

$$K(x, y) = x^T y + c \tag{2.4}$$

The linear kernel is one of the simplest forms of kernels. Here $x^T y$ denotes an inner product of the points x, y with an optional constant parameter c. It is usually preferred for high-dimensional datasets. Note that the linear kernel is not mapping the input data points into higher-dimensional spaces and is therefore used for problems that are linearly separable.

Polynomial kernel

A polynomial kernel is a type of nonlinear kernel that can be used in machine learning algorithms. It is a function that takes in two inputs (x and y) and outputs a value that represents the similarity between the two inputs.

Definition 2.3. The polynomial kernel is defined as:

$$K(x, y) = (ax^T y + c)^d \tag{2.5}$$

The degree of the polynomial kernel e.g., d determines how complex the function is. A higher-degree polynomial kernel will be more complex and will be able to capture more subtle relationships between the inputs, but it may also overfit the data. The polynomial kernel can be used with any classification algorithm, but it is commonly used with SVMs. The polynomial kernel has several advantages over other kernels, such as the linear kernel. It can model non-linear decision boundaries, which is important for many applications.

Radial Basis Function (RBF) kernel

The Radial Basis Function (RBF) is also a well-known kernel. The RBF kernel works by mapping data points onto a higher dimensional space, where they can be more easily separated into different classes. This makes it ideal for use in disease subtyping, as it can help to clearly differentiate between the subtypes.

Definition 2.4.

$$K(x, y) = e^{(-\frac{\|x-y\|^2}{2\sigma^2})} \tag{2.6}$$

Here, $\|x - y\|$ is the Euclidean distance function, and σ is the bandwidth of the kernel to associate local kNN graph structure. In contrast to other types of kernels, RBF kernel is widely used for disease subtyping.

Laplacian Kernel

A Laplacian kernel is a similarity kernel that can be used to measure the similarity between two data points. The Laplacian kernel has its own advantages over other similarity kernels. First, it is scale-invariant, meaning that it does not change if the data points are scaled up or down. Second, it is translation-invariant, meaning that it does not change if the data points are shifted. Third, it is able to capture both local and global similarities between data points. Fourth, it has a closed-form expression, making it easy to compute. Finally, it is differentiable with respect to both data points, making it possible to use gradient-based optimisation methods to learn model parameters.

Definition 2.5. The Laplacian kernel is defined as:

$$K(x, y) = e(-\gamma \|x - y\|_1) \tag{2.7}$$

Laplacian kernel is a variation of the radial basis function (RBF). Here, $\|x - y\|_1$ is a Manhattan distance between the input points x, y. These similarity kernels have the ability to generate fully connected similarity graphs out of the omics data. These similarity graphs are then used by the various clustering algorithms to group similar observations to help identify subtypes of diseases.

KNN Kernel

One of the most basic machine learning algorithms is KNN or k-nearest neighbours. It is a non-parametric classification and regression approach. The approach computes the distance between a new data point and all previous training data points. The label of the majority of the data point's neighbours is then applied to it. Kernels are used in machine learning to transform data so that it can be more easily separable. This transformation allows for non-linear decision boundaries, which can improve performance on certain datasets. When using a kernel with KNN, it is important to choose an appropriate kernel function and set the hyper-parameters correctly.

2.5.3 Similarity Graph

A similarity graph is a mathematical representation of how similar two objects are to each other. The similarity between two objects is calculated by using a similarity measure such as a similarity kernel, and the resulting similarity score is represented as a point on the graph. The closer the two points are to each other, the more similar the objects are. The similarity graph can be used to compare any two objects, but it is most commonly used in data mining and machine learning applications. For example, a machine learning algorithm may use a similarity graph to calculate the similarity between diseases in a population for subtyping.

Definition 2.6 (Graph)**.** A graph is defined as follows:

$$G = (V, E) \tag{2.8}$$

where G is a graph, V is the set of all vertices a.k.a, nodes, and $E \subset V \times V$ is the set of edges a.k.a, links between the vertices.

A key element of a similarity graph is a similarity measure e.g., $sim(.,.) \rightarrow \mathbb{R}$, which defines the similarity between the vertices in a graph. This can be achieved using similarity kernels as these have the ability to compute a similarity value for every pair of data points (vertices) in a set e.g., $s_{ij} = sim(v_i, v_j)$, where v_i, $v_j \in V$ and s_{ij} is the similarity score between the vertex v_i, and v_j. The similarity score between the pair of vertices in a graph can be represented through a weight on the edge. Thus, we can represent the pairwise similarity between the vertices using a weighted graph. Where the edge carrying the weight denotes the similarity between its source and a target vertex.

Definition 2.7 (Similarity Graph)**.** A similarity graph is formally defined as follows:

$$G = (V, E, W) \tag{2.9}$$

where the additional W is a set of weights on the edges denoting the similarity between the pair of vertices e.g., $w_{ij} \in W$, denotes the similarity score between the vertex v_i, and v_j in the graph.

The similarity graph is represented with a weighted similarity matrix. In the following, we provide a graphical illustration of the matrix representation of data, distance, and similarity. These play a key role in graph-based clustering analysis. We aim to transform the data (omic view) into a distance matrix and this distance matrix is then transformed into a similarity matrix (similarity graph). The final similarity graph is used as an input for disease subtyping.
Let's assume that X is a dataset with m observations and n measurements e.g., $X \rightarrow \mathbb{R}^{m\times n}$.

Example 3 (Data Matrix)**.** Lets, for this particular example, take $m = 5$, and $n = 3$ then the data matrix X, can be represented as follows:

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43} \\ x_{51} & x_{52} & x_{53} \end{bmatrix} \tag{2.10}$$

where each $x_{ij} \in X$, and the subscript i, j denotes the position of the element in X located on i^{th} row and j^{th} column. Where rows denote the samples a.k.a, observations or data points, and the columns denote features a.k.a, measurements.

This data matrix can be transformed into a distance matrix using any of the valid metrics defined above in this chapter e.g., Euclidean distance or Manhattan distance. When a distance metric is applied, then the data matrix $X \rightarrow \mathbb{R}^{m\times n}$ will be transformed into another matrix $D \rightarrow \mathbb{R}^{m\times m}$ e.g., distance matrix. which has $|rows| = |columns| = m$, and m is the number of observations in data matrix X.

Using Equation (2.1) which represents the Euclidean distance can be used to transform this data matrix into the following distance matrix. The distance matrix D is defined as follows:

Example 4 (Distance Matrix)**.**

$$D = \begin{bmatrix} 0 & & & & \\ d_{21} & 0 & & & \\ d_{31} & d_{32} & 0 & & \\ d_{41} & d_{42} & d_{43} & 0 & \\ d_{51} & d_{52} & d_{53} & d_{54} & 0 \end{bmatrix} \tag{2.11}$$

where each $d_{ij} \in D$, and denotes the distance score between the i^{th}, and j^{th} sample in X. Here both rows and columns denote the distance between a pair of observations e.g., data points. Each element denotes a distance score computed based on the features. Note that the diagonal elements are zero because for a valid matric a distance from a point to itself is zero e.g., the 2^{nd} axiom of a valid metric (Identity of Indiscernible). Similarly, the upper diagonal values are empty this is because of the 3^{rd} axiom (Symmetry) of the valid metric.

This distance matrix D can be transformed into a similarity matrix S using any of the similarity measures defined above in this chapter. By using Equation (2.6) which defines the RBF kernel can be used to transform this distance matrix into a similarity graph which is represented through the following similarity matrix S as follows:

Example 5 (Similarity Matrix)**.**

$$S = \begin{bmatrix} 1 & & & & \\ s_{21} & 1 & & & \\ s_{31} & s_{32} & 1 & & \\ s_{41} & s_{42} & s_{43} & 1 & \\ s_{51} & s_{52} & s_{53} & s_{54} & 1 \end{bmatrix} \tag{2.12}$$

where each $s_{ij} \in S$, and denotes the similarity score between the i^{th}, and j^{th} sample in X. Here both rows and columns denote the similarity between a pair of observations e.g., data points. Each element denotes a similarity score. Note that the diagonal elements are 1 because for this particular example we assume that the maximum similarity is 1, which naturally occurs with a point and itself. Similarly, the upper diagonal values are empty this is because of the 3^{rd} axiom (Symmetry) of the valid metric. With respect to the similarity graph, this matrix represents a fully connected similarity graph. This similarity graph can be used as an input to the clustering algorithms to generate clusters or find subtypes of a disease.

2.6 Summary

Disease subtyping helps to better understand how a disease develops and pinpoint the populations that are most vulnerable to it. Recent technological developments have made it possible to use massive omics data more precisely define subtypes using data-driven methodologies. These approaches have the potential to revolutionise our understanding of diseases. Data-driven disease subtyping approaches are still in their early stages, but they have the potential to transform our understanding of diseases and improve our ability to treat them. The data-driven approaches that only consider a single view like gene expression to subtype diseases are known as single view analysis. Whilst these approaches are often simple and do not need any integration, information can be lost by neglecting other views. In contrast, data-driven approaches based on integrative analysis combine information from multiple sources or views to gain a more complete understanding of a phenomenon.

Clustering algorithms are used with data-driven disease subtyping to group patients with similar characteristics together. In the context of diseases, this means identifying a group of patients who seem to share many common traits. This information can be helpful for developing more targeted treatments for each disease subtype. Similarity measures play a significant role in clustering. A kernel is a measure of similarity between two data points. The kernel trick is a technique that allows linear models to solve nonlinear problems by implicitly mapping data into a high-dimensional space. Kernels measure the similarity of two data points using a distance metric such as Euclidean distance or Manhattan distance. Distance measures, on the other hand, can be sensitive to outliers. Furthermore, in high-dimensional spaces, the concept of similarity and distance, which is critical for clustering, loses qualitative significance. Similarity kernels produce similarity graphs which are graphical models that encode the pairwise similarities between objects in a dataset. The nodes in the similarity graph represent the objects, and the edges represent the similarities between them. Similarity graphs can be used to cluster objects into groups.

Chapter 3

Literature Review

3.1 Introduction

In this chapter, the current state of research on disease subtyping is reviewed. The chapter begins by providing an overview of the findings from the scientific literature, organised by topics relevant to the proposed research. By evaluating the current state of research, this chapter demonstrates how the research findings relate to each other and what gaps in the research exist. This chapter is not exhaustive but rather gives an overview of the major approaches, challenges, and limitations to set the stage for the proposed research on the topic.

The success of clustering depends on the quality of the generated clusters. Therefore, evaluating the importance of clustering is essential to ensure that the right groups are being formed and that the process is effective. Therefore, we will explore evaluation methods both based on external and internal criteria in the next chapter for assessing the quality of clustering for disease subtyping.

3.2 Current State of Disease Subtyping

Omics technology creates large amounts of high-dimensional data that encompass essential information about biological entities. The integration of generated omics data into a useful unified model is difficult yet necessary for biological investigations. In molecular subtyping, for example, integration aids in identifying specific gene variations in patients across several views.

Nguyen et al. presented perturbation clustering for data integration and subtyping (PINS) to overcome these issues [39]. PINS is based on the notion that genuine subtypes stay constant even when their dimensions alter slightly. Therefore, it injected Gaussian noise into data repeatedly and partitioned the patients using k-means clustering for variable k values. Clustering stability is assessed by comparing the least impacted partitions to the clusters obtained from the original omics data. Nonetheless, the PINS architecture has been rigorously verified on a large number of cancer samples, with outstanding results. However, further work is needed to reduce its computational time complexity and to develop a technique for distinguishing between different data kinds in a multi-view dataset.

Similarly, a multi-view robust graph-based clustering (MRGC) is proposed to mitigate the influence of noise in omics data [9]. To limit the effect of noise on high-dimensional individual omics-views, MRGC learns robust latent representations for each view. Similarity matrices are learned using these latent representations. Finally, a consensus technique is used to generate a final unified similarity graph for subtyping molecular diseases. On both general machine learning and omics datasets, the suggested technique produced excellent clustering results. A case study is also carried out to highlight the biological importance of the MRGC, especially in hepatocellular cancer. Despite its strong clustering performance, the suggested method might be improved in terms of learning appropriate hyper-parameter values and preserving stability on omics datasets.

A consensus-guided graph auto-encoder (CGGA) is developed to overcome the integration difficulty [8]. CGGA is made up of two steps. First, it uses graph auto-encoders to learn a feature matrix for each specific data type. It is a versatile strategy that incorporates structural as well as particular feature information into the learning process in order to successfully learn clinically important cancer subtypes. Second, the learnt feature matrices are utilised to generate similarity matrices, and ultimately, a unified consensus matrix is generated by repeating the suggested two-step procedure a few times. CGGA showed great improvement in learning cancer subtypes for diverse cancer conditions; nevertheless, performance may be enhanced further by minimising its reliance on configurable hyper-parameters.

Molecular disease subtyping algorithms are widely unsupervised; however, Liu et al. used a supervised approach and suggested a revolutionary survival supervised graph clustering (S2GC) methodology [11]. S2GC learns survival analysis embedding and patient similarity graphs together for molecular subtyping. The similarity graph is continually adjusted by assessing the survival time as labelled training data, resulting in an ideal similarity graph for subtyping. S2GC demonstrated promising findings and outperformed several existing unsupervised subtyping algorithms on multi-omics cancer data. However, the benefit comes at the expense of considerably costly computation and upfront human interaction.

The goal of clustering techniques for disease subtyping in general is to classify patients into coherent groups based on their underlying commonalities revealed by omics and clinical data. As a result, creating an accurate patient similarity graph is critical in molecular disease subtyping. High dimensionality, noise, and data variability are the obstacles to similarity graph building. As a result, the spectral clustering and current similarity graph-building algorithms are studied in the next section.

3.3 Integrative Approach for Disease Subtyping

A data-driven approach for disease subtyping is proposed which effectively integrated motor and non-motor characteristics of Parkinson's Disease (PD) patients [31]. It adopted the standard disease subtyping approach and constructed a similarity matrix for each characteristic. These similarity matrices represented the pairwise patient similarity graph based on these clinical characteristics e.g., motor, and non-motor. Secondly, an optimum single similarity graph is constructed with a few iterations of the fusion mechanism on the individual graphs. This integrated graph is then used for the discovery of PD subtypes. Hierarchical clustering is applied, and subtypes are discovered. The discovered subtypes are evaluated by computing the demographic differences, clinical differences, and genetic differences between the subtypes via statistical testing.

Gillenwater et al. proposed the largest multi-omics subtyping pipeline for chronic obstructive pulmonary disease (COPD) [25]. The approach enlightened the importance of the multi-omics clustering stage. In the proposed approach the multi-omics profiles were first analysed individually (post-clustering) and later the subtypes were discovered, secondly all the profiles were integrated and treated equally (pre-clustering) and the results were analysed. The analysis provided insights into the clustering stages and found that the post-clustering for the COPD disease discovered clean subtypes for each omics profile differentiated via unique clinical signatures. While the pre-clustering approach was unable to find well-defined subtypes which had obvious clinical differences. Also, the research findings identified some comorbidity and suggested their inclusion in multi-omics clustering as these could play a vital role in subtyping.

Yin et al. proposed an unsupervised multi-view clustering approach to discover subtypes of complex diseases [40]. They proposed an approach to discover subtypes of complex diseases which has clinical and biological significance. In the proposed research the genomics data is integrated with clinical phenotypes for clustering analysis. The key novelty in the proposed research was to use genotype-predicted gene expression levels rather than raw SNPs for disease subtyping. The authors claimed that using the gene-based approach as proposed in this research is faster and requires less memory compared to the SNP-based approaches. Further to this, the approach is able to link the functional impact of SNPs to genes which may help in easier interpretability of the discovered subtypes. Another advantage of using the gene-based approach is that it is easy to inject expression levels in various tissues, while it is difficult to achieve the same using the SNPs-based approaches.

Most of the intermediate integration models proposed for disease subtyping assume full datasets containing complete disease information for analysis. However, acquiring full datasets for all patients might not be easy due to various associated costs. Therefore, the authors proposed NEMO a neighbourhood-based clustering method that is able to perform subtyping on partially available data [34]. In addition, NEMO is fast compared to the existing integration systems as it does not require iterative optimisation. Even though NEMO is based on existing similarity graph-based clustering such as SNF but it has added advantages such as it is fast and supports partially available data. NEMO is evaluated on full datasets, partial datasets, and synthetically generated datasets. It is clear from the results that the NEMO has advantages in terms of simplicity, time efficiency, and support for datasets with missing data.

In this review, authors emphasised the challenges of learning from heterogeneous and noisy data [35]. Some of the existing machine learning models proposed for disease subtyping are based on a single data type or view. However, most of the data from subtyping come in multiple views e.g., gene expression, DNA methylation, and MicroRNA, etc... as such learning from only a single view will lead to incomplete understanding or model overfitting. In these scenarios, it is necessary to consider all the views associated with a disease for a better understanding of the process. Also, as these views complement each other, therefore, an integrative analysis will eventually infer missing data in one view from other views consequently it will reduce the noise in data. Therefore, the authors extended the empirical risk minimisation (ERM) model by introducing Multi-view learning and proposed Multi-view empirical risk minimisation (MV-ERM). MV-ERM generalises the modelling and application aspects of the ERM. In MV-ERM authors introduced the concept of Multi-view learning in ERM in a unified mathematical framework for a better understanding of complex processes.

Authors in [41] state that existing multi-omics data integration systems rely on joint statistical modelling and are based on the strong assumption about the data distribution and feature selection, which makes them sensitive to noise and slight changes in measurements. To address these challenges a robust approach is proposed called perturbation clustering for data integration and disease subtyping (PINSPlus) [41]. PINSPlus optimises two algorithms of the previously proposed PINS framework. The extension makes it robust to noise and bias. Specifically, it extends two algorithms of the PINS e.g., a) perturbation clustering and b) subtyping omics data. Perturbation clustering continuously adds Gaussian noise to the data and performs clustering for a variable number of clusters. The clusters returned by each are examined, and the number of clusters that give the most stable connectivity is considered optimal. In this way, the slight changes in measurements that occur through noise have the least effect on the final subtyping results.

To address the integration challenge in multi-omics data for disease subtyping an integrative network fusion (INF) is proposed [42]. INF is a network-based subtyping framework that leverages the intermediate results of both early integration approaches and intermediate integration approaches such as SNF to find the optimal set of predictive variables. Firstly, INF takes top-ranked features from the early integration approach by using a classifier. Secondly, it fetches the highly predictive features from the data integrated by SNF. Finally, the two sets of top-ranked features are intersected, and a random forest (RF) classifier is trained on the intersection of two sets of high-ranked features. Through this, INF achieves an effective way of the multi-omics data integration system. Furthermore, to reduce the computational bottleneck, INF introduces an approximate data analysis (DAP) pipeline with the least effect on the final results.

A novel hierarchical data fusion and integrative clustering approach called HC-fused is proposed [68]. HC-fused applies a two-step process to integrate multi-view datasets. It first creates a network-structured view of the data by clustering each data type with hierarchical clustering. Then a novel hierarchical data fusion technique is proposed to effectively integrate the constructed networks. Compared to other multi-view data integration approaches, HC-fused has the added advantage of taking into account the contribution of individual omics data in the construction of an integrated view. The HC-fused approach generates results that are transparent, easy to understand and interpret.

Rappoport et al. Proposed a multi-omics clustering by non-exhaustive types (MONET) [69]. MONET algorithm is designed to discover patient modules from multi-omics data. It extends the existing MATISSE algorithm which is used to identify gene modules and further generalises its algorithms to be used for multi-omics clustering tasks. MONET repeatedly uses a subset of omics data to discover patient modules, in this way it discovers a common structure among the patient modules. MONET is successfully used to discover patient modules which are clinically and biologically relevant. In addition to that, MONET is found useful in various other biomedical tasks such as discovering gene modules, and cells from single-cell data.

Kamoun et al. adopted an unsupervised approach and proposed a comprehensive classification model for the discovery of localised prostate cancer (PCa) subtypes [70]. The model was applied on three molecular levels e.g., DNA copy number, DNA methylation, and mRNA expression for subtyping. The proposed approach identified three molecular subtypes of localised prostate cancer defined by clinical, genomics, epigenomics, and transcriptomics features. Finally, the subtype-associated risks were measured using survival data extracted from the cohorts and the Cox regression model.

NEMO is a well-known partial multi-omics data integration approach based on network embedding. However, NEMO makes the strong assumption of common omics data for integrating a pair of samples. The assumption results in the removal of a large number of samples from the dataset. In addition, during the integration, NEMO takes the average of the individual similarity networks which results in lower accuracy because of the different scale and edge weights among the different similarity networks. To address these limitations an improved version of the NEMO is proposed based on multiple similarity network embedding called MSNE [71]. Through multiple embedding, MSNE efficiently integrates partial omics data for the discovery of disease subtypes. MSNE follows manifold learning and effectively learns the integrated similarity of samples via a random walk on multiple sample networks with partially available data. The integrated similarity is then transformed into a low-dimensional space representation for improved performance. MSNE is extensively evaluated on synthetic full and partially available data, and for realisation, it is evaluated on real-world omics data. MSNE outperforms many existing omics data integration approaches.

3.4 Graph-based Approaches

There is a lot of excitement in the medical community about using a graph-based approach to disease subtyping. This approach offers many potential benefits, including the ability to more accurately identify different subtypes of diseases and the potential to develop more targeted treatments. One of the key advantages of this approach is that it can help to identify previously unknown disease subtypes. By looking at the interactions between different genes, proteins, and other molecules, researchers can get a better understanding of how diseases develop and progress. This information can then be used to develop more targeted and effective treatments. Another advantage of this approach is that it has the potential to improve our understanding of how different subtypes of diseases respond to treatment. Currently, many treatments are developed without taking into account the heterogeneity of disease subtypes. However, by using a graph-based approach, it may be possible to develop treatments that are more effective against specific subtypes of diseases.

Overall, the use of a graph-based approach to disease subtyping has a lot of promise. It helps to better understand diseases and develop more targeted and effective treatments. In graph-based approaches, a fully connected similarity graph of observations is constructed prior to the clustering algorithms. This similarity graph is then provided as an input to the specified clustering algorithms to discover subtypes of diseases based on their similarity and connectivity patterns. These similarity graphs are usually generated using similarity kernels. The following section highlights some of the similarity kernels that can be used with clustering algorithms.

3.4.1 Pairwise Similarity Kernels

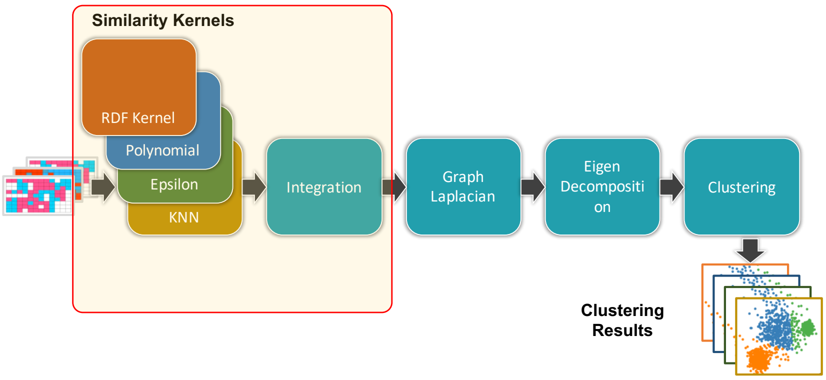


Figure 3.1: Spectral clustering pipeline for disease subtyping.

Similarity Kernels (SK) have been extensively explored in machine learning (ML). Among others the most frequently used kernels for structured data include the Laplacian kernel and Gaussian radial basis function (RBF) kernel [72, 17]. The difference between RBF and Laplacian kernel is that the former uses Euclidean norm while the latter uses 1-norm [73]. These kernels serve as a good similarity measure for noiseless data [73]. For noisy data, these kernels might not produce good similarity results and this is because of the distance metrics such as Euclidean distance employed in these kernels. These distance functions are not robust to noise such as extreme values and data variability.

3.4.2 Similarity graph Construction

In existing literature often kernel-based approaches are adopted to model the similarity between the objects in the dataset through a pairwise similarity network. The distance functions used with kernels work well with categorical and normally distributed, symmetric numerical data. However, on highly skewed continuous features with different scales and variability, the extreme values dominate the smaller values. The distance functions used to generate similarity networks are not robust to extreme values, variability, and scales. On the other hand, ranking-based methods resolve the extreme values but suffer from scale and variability. Moreover, it is not a good similarity measure as it ignores the strength of proximity between values. The majority of these distance metrics are usually based on central tendency measures (mean, median, and mode) however, for the highly skewed features we need more information to explain data variability or dispersion. Because a single extreme value or an outlier has an unbounded influence on these central tendency measures [24]. Consequently, these can adversely affect the estimation of distances.

Below in this section, we present famous approaches for disease subtyping based on graph-based approaches.

Krishnagopal et al. proposed a novel multi-layer graph-based trajectory clustering (TC) [29]. The TC algorithm discovers subtypes via variable clusters which are based on similarities in trajectories. It models a bipartite graph from patient and variable interactions, which is then used to track patient membership from multiple layers of co-expressed cluster variables. Finally, similar trajectories are clustered to discover subtypes. The proposed algorithm is a variable-centric approach that considers disease progression as a function of the outcome variable. However, the approach is restricted to the availability of high-quality variables. Moreover, the approach is not applicable to data with high variability and small quantity.

Ramazzotti et al. proposed cancer integration via multikernel learning (CIMLR) [74]. CIMLR is a novel disease subtyping approach that integrates multi-omics data using multiple kernels for molecular subtype discovery. The proposed approach learns the similarity between a pair of patients (samples) in multi-omics data by integrating results from multiple Gaussian kernels defined over each data type. The final constructed patient similarity matrix contains block structures that are used for dimensionality reduction and clustering. The authors claim that CIMLR is scalable to many other data types, and does not assume equal importance for the datatypes. CIMLR is evaluated on many multi-omics data and outperformed many existing approaches on speed, accuracy, and survival time prediction of the patients.

3.5 Spectral Clustering for Disease subtyping

Real-world datasets are often large in dimensions, noisy, and include data variability, their integration breaks the condition of data integrity [75, 76, 77]. Furthermore, certain high-dimensional multi-view datasets, such as genomics, contain complementary disease information [78]. Finding clinically meaningful subtypes is made more difficult by the integration of complementary information from these diverse perspectives [78, 38]. Disease subtyping is a critical and challenging step in precision medicine that divides patients into well-defined risk categories based on clinical and molecular characteristics [79, 80]. To address these problems, numerous techniques such as hierarchical clustering, model-based, matrix factorization, and spectral clustering are being researched in order to discover the heterogeneity of these diseases [68]. The use of spectral clustering yielded considerable results in the discovery of subgroups and related survival rates. The complementary nature of multiple points of view makes spectral clustering a feasible possibility. It seeks to turn multiple views into separate graphs by creating pairwise similarity matrices using kernels and then integrating these graphs into a single similarity graph to enhance clustering performance [81, 78]. Figure (3.1) depicts the many phases of the spectral clustering workflow. The first two steps play a considerable role in the generation of a holistic view and improvement in the ultimate clustering performance.

Similarity kernels are widely used with spectral clustering for the transformation of individual source data into graphs [78, 38]. These individual graphs are then integrated using similarity fusion to construct a single similarity graph [38]. These kernels create an adjacency matrix by computing pairwise similarity between the samples. Each entry in the matrix represents the strength of similarity between the pairs. For instance, Similarity Network Fusion (SNF) [38] is considered state-of-the-art in cancer disease subtyping. SNF transforms each view into a graph using a scaled exponential similarity kernel. These graphs are then combined with iterative network diffusion before being utilised for subtyping and survival prediction. Affinity Network Fusion (ANF) [78] is proposed to address the computational difficulty of SNF. In contrast to iterative diffusion, ANF employs one and two-step random walks to provide smoothed views, which are then fused using their weighted average.

Similarity kernels, on the other hand, are unstable and susceptible to changes in hyper-parameters and numerical measurements [28]. Any modifications to the type of kernel or its settings will almost certainly result in different clustering results. One probable explanation is that the kernels are based on distance functions, which perform better with categorical and normally distributed, symmetric data. When used to heavily skewed data, it generates noisy and ineffective graphs. Because distance functions are not resilient to noise, as a result, data variability, and extreme values overwhelm the rest when computing pairwise distances between samples. Therefore, more information is required to explain data variability and dispersion for the robust construction of the graph.

Spectral clustering techniques are widely used for disease subtyping because these approaches work well for multi-view complementary datasets. As there are always limitations for even the best approaches similarly, one of the limitations associated with the spectral clustering approach is the tuning of hyper-parameters of the kernels used with these approaches. Therefore, in this article, a fast density-aware spectral clustering (Spectrum) for disease subtyping is proposed which self-tunes the kernel parameters [13]. In the proposed approach an adaptive density-aware kernel is used to compute similarity matrices for each view of the dataset. This kernel is specifically designed for continuous datasets and for this to work on multi-view datasets, the data points need to match in different views.

The advantage of the proposed approach compared to other spectral clustering-based approaches is that first, it improves the local links between the samples in higher-density regions, moreover, it self-tunes the kernel parameters for better results. Finally, the constructed similarity matrix is used to form clusters or identify subtypes.

Rafique et al. proposed an approach that aims to increase separability among survival curves of samples. Therefore, proposed a robust approach that assigns weights to every gene [82]. The weights are assigned using the median absolute deviation (MAD) of each gene. The MAD score is calculated for every gene and weights are assigned in such a way that the genes with greater MAD scores received higher weights. The idea behind the MAD score was that the genes for which the samples have greater variability receive a greater MAD score. MAD holds variability information for every gene which helps the subsequent clustering algorithm to find better-separated patient clusters. The proposed approach computes the MAD score in such a way that the gene for which the patients have greater variability will receive greater weight. The genes having greater weight scores will help in the upcoming dimensionality reduction and clustering analysis tasks and consequently help in increasing the separability in survival curves.

3.6 Neural Networks for Disease Subtyping

Deep learning-based approaches for disease subtyping are gaining rising attention. Mostly, these approaches are based on Neural Networks (NN) following auto-encoders for low-dimensional feature representation. Autoencoding is a way of learning dense representations of data. In the context of disease subtyping, autoencoding can be used to learn a latent representation of disease subtypes. This representation can then be used for downstream tasks such as classification and prediction.

Neural networks (NN) are capable of a wide range of tasks such as pattern recognition, classification, and prediction. They've been used to solve issues in a variety of domains, including medical, finance, and manufacturing. Although the application of neural networks for disease subtyping is still in its early stages, initial results have been encouraging. A neural network was employed in one study to classify breast cancer patients into molecular subgroups. Based on their gene expression patterns, the neural network was able to effectively categorise the patients into the correct subtype.

The use of neural networks for disease subtyping has the potential to improve the accuracy of diagnosis and treatment. as well as provide a more personalised approach to medicine. For disease subtyping, the neural network-based approach makes use of auto-encoders for low-dimensional representation. Autoencoders are neural networks that are trained to encode data into a low-dimensional latent space. The encoder part of the network learns to compress the data into this latent space, while the decoder part learns to reconstruct the data from the latent space back to the original dimensionality.

One advantage of using autoencoders for disease subtyping is that they can be trained on unlabeled data. This is because the autoencoder only needs to learn a representation of the data, not necessarily any specific labels. Another advantage is that autoencoders can learn complex nonlinear relationships in the data. This is because they are not limited by predefined categories or labels. The downside of autoencoders is that they can be difficult to train, and sometimes do not converge on a good solution.

Following are the approaches for disease subtyping based on deep learning.

To tackle the time irregularity challenges in the standard LSTM architecture for healthcare, a novel T-LSTM architecture is proposed [83]. T-LSTM address this challenge by considering the time-elapsed between consecutive elements of a sequence. The proposed approach improved the standard LSTM by considering the time irregularities. To incorporate the T-LSTM in disease subtyping an unsupervised approach is proposed utilising the T-LSTM. Temporal patient data is used to learn a single representation by mapping sequential records of samples to a representation capturing the dependencies. The learned representations are then used to identify patient clusters for disease subtyping.

The authors in [32] argued that existing auto-encoder (AE) based disease subtyping approaches are not generalised toward other diseases and these are usually proposed for a specific disease type. Therefore, they proposed a robust auto-encoder-based approach for disease subtyping which gain information bottleneck by using both strategies of reducing the number of neurons on the hidden layers and penalising the activation functions inside the layers [32]. The aforementioned strategies are utilised to discover the global structure of the multi-omics data by identifying robust latent space representation of the integrated omics data. The proposed approach performed survival-based feature selection first, and these features are then fed to the AE to learn embedded feature representation. Finally, spectral clustering is performed on the learned feature representations to discover disease subtypes. Generalisation is achieved through excessive tuning of the models' hyper-parameters. The proposed approach is evaluated on five available cancer datasets. Finally, a comparative analysis is performed with existing early-integration, intermediate-integration, and late-data-integration approaches.

Deep learning-based approaches for multi-omics data integration and disease subtyping are mostly based on auto-encoder (AE) or variational auto-encoder (VAE). In the disease subtyping literature, AE and VAE are mostly based on either single-input (SI) or multi-input (MI) models. The problem with SI modelling of multi-omics data is that SI ignores the difference between the data distribution and the number of features that are of vital importance for discovering accurate subtypes. To address these limitations, MI-based modelling is adopted. MI-based approaches model each type of omics data individually and then apply statistical modelling on each type to reveal subtypes. However, modelling multi-omics data using MI-based approaches is challenging due to the inherent complexity of these data types.

To resolve these issues, Yang et.al suggested subtype-GAN a generative adversarial network based on the MI model [36]. subtype-GAN can incorporate a variety of omics data, including copy number, DNA methylation, gene expression, and miRNA data. A multi-input multi-output network and a generative adversarial network are combined in the suggested method. Finally, consensus clustering is used to determine the number of subtypes and the class label for each observation. The suggested technique is used for BRCA data to determine its subtypes. The suggested subtype-GAN is assessed and compared with existing techniques with outstanding results.

One of the challenges for integrative disease subtyping approaches is the missing omics data. The multi-view datasets often lack the full set of views which creates hurdles to integrating multiple datasets having missing data with several missing patterns. One option to handle the missing data is to remove it from the dataset however, the removal of missing data can significantly reduce the sample size. Another option is to impute the missing values with mean imputation however, the mean imputation severely distorts the data distribution. To address these challenges for omics data integration Lee et al. proposed a deep variational information bottleneck (IB) approach for learning from incomplete data called deepIMV [84]. The proposed deepIMV is able to learn from inter-view interaction, and intra-view interactions and efficiently integrate multi-omics data with missing patterns. DeepIMV consists of view-specific encoders, view-specific predictors, product-of-experts(PoE) components, and multi-view predictors to effectively learn from incomplete data. The proposed deepIMV models the joint representation as PoE for integration which is utilised by the multi-view predictor component to predict the target labels for samples. DeepIMV is extensively evaluated on multiple omics data and compared against state-of-the-art in subtyping with multi-omics data.

Disease-subtyping datasets are usually high-dimensional with the essence of the information in them. If not all these measurements are considered then the approach results in lower accuracy, on the other hand, considering all these measurements cost computational time, and challenges in learning from such high-dimensional data. To address the challenges associated with learning from high-dimensional data a graph-based neural network approach called multiGATAE is proposed [85]. multiGATAE is a deep-learning-based graph neural network approach that considers all features in the learning process. The proposed approach learns feature embedding by defining a graph-based auto-encoder which includes a graph attention network, and omics-level attention mechanism. The learned embedding is then used for clustering omics data and discovering disease subtypes.

Graph convolutional neural networks (GCN) are gaining research attention in graph-based high-dimensional clustering. Therefore, in this work a novel consensus-guided model for clustering is proposed which is based on graph autoencoder (GAE) called scGAC [86]. scGAC takes full advantage of the graph structure to learn node-level graph embedding. In the proposed work first, top-level features are used to extract genetic information. This resolves many limitations associated with traditional PCA-based feature selection such as the risk of distortion. It then uses GAEs to learn feature embedding which takes into account the graph topology and node features in the learning process. A set of similarity matrices are then learned using the top-level features from the first step. The strength of association between the cells is predicted by retaining the linear and nonlinear manifolds of the data via the linear fusion of two distance functions. The learned similarity matrices are provided back to the GAE to assist the feature learning process. These steps are iterated several times to construct a single similarity matrix from the set of similarity matrices. The proposed approach has the ability to obtain information from the data more comprehensively and address the limitations of the traditional non-deep learning-based data processing methods.

Although deep learning architectures have demonstrated high performance on numerous problems, they frequently encounter challenges when applied to small sample-size data. For instance, disease subtyping commonly involves analysing genetics data, which often consists of a small number of patient samples compared to a large number of features measured [8][9][39]. Such datasets are characterised as high-dimensional with limited samples and noisy, posing challenges for deep learning models [87]. Generally, the performance of deep learning algorithms in recognising patterns is dependent on the dataset size, smaller datasets make these models less powerful and accurate. Despite the prevalence of this issue and efforts to address it, comprehensive studies dedicated to this crucial aspect of machine learning are lacking [87]. This is a common challenge faced by deep learning models when confronted with a small sample size (limited number of patients) which greatly impact their performance. [88]. Most of the deep learning architectures are prone to overfitting on small training samples, leading to sub-optimal performance when tested on new and unseen samples [89].

3.7 Limitations and Challenges in Disease Subtyping

There are several limitations and challenges in disease subtyping that need to be considered when conducting research in this area. First, there is a lack of standardisation in the way that diseases are classified and subtyped. This can make it difficult to compare results across studies. Second, there is a lack of agreement on the best methods for disease subtyping. This can lead to different researchers using different methods, which can make it difficult to compare results. Third, disease subtypes can change over time, making it difficult to track changes in disease prevalence or incidence. Finally, some diseases are very rare, making it difficult to find enough cases for study.

Following in Table 3.1 are the summary of common limitations associated with existing disease subtyping approaches.

Table 3.1: Limitations and Challenges in Disease Subtyping

State-of-the-art	Limitation	Reference
PINS and TC	Data types weighted equally	[39, 29]
CC[1], PINS	Computation Intensive	[55, 35, 39]
IntPD and TC	Lack of complete data	[31, 29]
IntPD and TC	Lack of clinical domain knowledge	[31, 29]
NEMO	Datasets Assumptions	[34]
AD	Regularization Parameter γ	[90]
ES[2]	Feature space constraint	[30]
MPE[3]	Missing subtype data	[45]
YinInt, AE[4]	Sensitivity to outliers	[40, 25]
CC and TC	Data variability	[55, 29]
CoINcIDE	Compatible Datasets	[91]

All data types are weighted equally

Another weakness with most of the disease subtyping approaches is that all data types are equally weighted when creating subtypes, which may not always be appropriate [39][29]. In some cases, it might be more meaningful to differentiate between different types of data when determining subtypes. For example, if we have a dataset with both gene expression and DNA methylation, it will support adding different weights to each type of data. This would allow giving more weight to the data type which is more important to the type of disease being subtyped.

Computation Intensive

Disease subtyping becomes computationally intensive, especially when we focus on accuracy. For instance, PINS achieved spectacular results but it is slower than Consensus Clustering [57] [55], and SNF [38] since it needs more time to perform the following additional analyses on large data sets [39] [28]. Firstly, it needs to do perturbations and repeated clustering to find subtypes of a disease against which small changes in molecular data have little or no effect. Secondly, it runs k-means multiple times to make sure that the results are stable and reproducible. In addition, the Multiview-learning based approaches are both computationally expensive and memory-demanding [35]. For instance, multiview learning applications based on alignment methods requires twice as high data memory as required by the non-multiview learning approaches. this is due to the pairwise processing strategy of the alignment-based methods. On the other hand, multiview learning, based on factorisation methods requires high time and space complexities. This is due to the simultaneous processing of available multiview datasets.

Lack of complete data

Data-driven approaches are limited by the availability and quality of the dataset. The lack of large-size datasets affects the quality and robustness of these approaches [29]. Another limitation of the data-driven approaches is that some diseases have access to limited samples which limits the generalisability of these approaches [31]. It is important to remember that these methods can't necessarily be generalised to other types of diseases. The study did not have a complete set of data and outcome variables for all patients, which may have limited the findings. Additionally, some of the clinical data was missing, which could also impact the results.

Lack of clinical domain knowledge

Various research findings suggested that the data-driven approaches should be used with caution as these studies are usually not based on any knowledge from clinical domain experts. Therefore, this caveat needs to be kept in mind while using data-driven approaches. These approaches should never be used in isolation – they should always be used in conjunction with clinical knowledge and expertise [29] [31].

The assumption about the dataset

Similarly, some disease subtyping approaches make assumptions about the data that at least one omics is common in every pair of samples [34]. However, this is a strong assumption that is often violated in the real-world dataset. The second limitation is the choice of K in nearest neighbours in KNN, which is implicitly assumed that the cluster sizes are equal for all samples.

Challenges that obstruct Autoencoders to perform effective disease subtyping

Autoencoders are a type of neural network that is used to learn efficient representations of data. They are similar to other types of neural networks, but they have a special architecture that allows them to learn efficiently. However, AEs are sensitive to hyperparameters, which are parameters that control the learning process. Therefore, it is important to carefully tune the learning rate in order to achieve good results with AE [32]. Moreover, AE may not perform optimally when applied to datasets with a small number of samples and high dimensionality.

It is difficult to find compatible data types for integrative analysis

One challenge with integrative analysis is that it can be difficult to find data sets that are compatible with each other. Another challenge is that the results of an integrative analysis can be complex and hard to interpret.

Even though multi-omics data integration has the potential to improve clustering solutions, it is still a difficult task to uncover all of the complementary information contained within the data. This is because each type of omics data (genomic, proteomic, metabolomic, etc.) contains its own unique information that needs to be taken into account. Furthermore, it is often difficult to effectively integrate all of these different types of data [32].

Outliers, data variability, and extreme values lead to unstable results

Outliers and data variability can have a significant impact on clustering results. In some cases, they can completely change the cluster structure. These can be very influential when calculating distances between points. The more influential an outlier is, the more it can impact the distance calculations and, as a result, the clustering results. This is a common issue with disease subtyping approaches which leads to unstable results in most cases. This is because of their sensitivity to outliers, which affect the entire subtyping results [25]. In addition, the distance functions used with clustering are usually based on classical statistical procedures such as mean, standard deviation, and variance to estimate the distance between the observations in the data. Whilst these central tendency measures perform well on compact and isolated clusters, they are susceptible to outliers [19, 20, 21]. These outliers have an impact on the distances of a pair of data points in the interquartile range (IQR) [22, 23]. Even a single highly influential outlier has an unbounded impact on these techniques [24]. The lack of resilience to outliers restricts these techniques' ability to provide reliable measures for estimation. A simple solution is to eliminate outliers before computing pairwise distances. Outliers, on the other hand, can often give useful information about unusual behaviour [24]. As a result, deleting outliers that contain such valuable information might have a negative impact on the assessment of data variability and distances.

Therefore, in this thesis, a novel robust graph-based clustering approach is proposed for disease subtyping, that is able to address the data variability and extreme values challenges on high-dimensional data. The proposed approach neutralises the influence of extreme values and data variability in an effective manner.

3.8 Summary

Numerous data-driven approaches have been proposed to address the disease subtyping problem in Multiview omics data. The challenges exist on various levels for instance: 1) High-dimensionality vs fewer samples, 2) lack of complete data, 3) outliers, data variability, and extreme values, 4) significance or weightage of each view, and 5) Data compatibility. Whilst the integration of the produced omics data in a meaningful unified model is a challenge, it is crucial for biological studies.

Most disease subtyping approaches, for example, equally weight all data types (views), which is not always appropriate [39][29]. When determining subtypes, it may be more meaningful in some cases to differentiate between different types of data. Likewise, when the focus is given to accuracy, disease subtyping becomes computationally demanding. PINS, for example, achieved spectacular results but is computationally expensive [39]. Furthermore, the Multiview-learning-based approaches are both computationally and memory-demanding [35]. Moreover, the availability and quality of the dataset limit the quality of data-driven approaches. The lack of a large number of samples has an impact on the quality and robustness of these approaches [29].

Another limitation is that some diseases have access to limited samples, limiting their generalisability [31]. Similarly, some disease subtyping approaches make data assumptions that at least one omics is shared by every pair of samples [34]. This, however, is a strong assumption that is frequently violated in the real-world dataset. Moreover, autoencoders are hyperparameter sensitive. As a result, in order to achieve good results with AE, it is critical to carefully tune the learning rate [32]. Another difficulty with integrative analysis is finding data sets that are compatible with one another. Despite the fact that multi-omics data integration has the potential to improve clustering solutions, integrating and uncovering all of the complementary information contained within the data remains a difficult task [32].

Finally, outliers and data variability can significantly affect clustering results. When calculating distances between points, these can have a significant impact. This is a common problem with disease subtyping approaches, which leads to instability in the majority of cases. This is due to their sensitivity to outliers, which have an impact on the overall subtyping results [25]. The statistical procedures work well on compact and isolated clusters, but they are vulnerable to outliers [19, 20, 21]. These outliers influence the distance between observations in the interquartile range (IQR) [22, 23]. Even a single highly influential outlier can have a massive impact on these classical central tendency measures [24]. Outliers, on the other hand, can sometimes provide useful information about unusual behaviour [24]. As a result, removing outliers that represent such valuable information can have a negative impact on the meaningfulness of the results.

Chapter 4

Methodology

4.1 Overview

In the last few years, high-throughput sequencing projects have dramatically increased the number of high-dimensional genomics data generated in biomedical research. This has created a need for new statistical approaches to help researchers make sense of the data, and to aid clinicians in deciding how best to diagnose or treat patients.

However, the characteristics of genomics data for disease subtyping give rise to several challenges. Firstly, the data is high-dimensional with a limited number of samples. This makes it difficult for machine learning models, such as supervised or deep learning, to effectively learn from the data. Secondly, the data is multi-modal, meaning it contains complementary information across different modalities that needs to be integrated to gain a comprehensive understanding. Additionally, the presence of extreme values and data variability further complicates the discovery of subtypes, as these characteristics often lead to spurious clusters. To address these challenges, we propose a methodology that tackles each of these issues systematically in the disease subtyping pipeline.

Firstly, we address the challenge of extreme values and data variability by employing a novel robust distance function. This function incorporates robust statistical techniques, such as statistical quartiles and the Freedman-Diaconis estimator, to define a topological structure (e.g., graph) for each modality. This approach enhances resilience against extreme values and data variability. Secondly, in order to compute distances or similarities in high-dimensional data, we utilise the

$L_1 - norm$ distance metric. This metric is preferred over $L_2 - norm$ distance metrics as it is more resilient to the challenges posed by high dimensionality. The third challenge, learning from high-dimensionality versus few samples, is addressed through an unsupervised graph-based clustering approach. We employ Spectral clustering, which minimises a ratio cut on a connected graph, enabling efficient learning and subtype discovery even with limited samples. This approach proves advantageous compared to non-graph-based or deep learning models. Lastly, the challenge of multi-modal data is tackled by utilising similarity network fusion (SNF), a well-known and widely used approach in the bio-informatics domain for integrating multi-view or multi-modal data. This method constructs fully connected similarity graphs for each modality and iteratively integrates them using a novel network fusion technique.

The proposed disease subtyping approach takes into account these challenges, making it the most suitable and comprehensive solution for the problem at hand. By addressing the issues of extreme values, data variability, high dimensionality, few samples, and multi-modal data integration, our methodology provides a robust framework for effective disease subtyping. The overall framework is validated on multiple TCGA cancer datasets, synthetic data, and generic machine-learning datasets. The results were compared with multiple baseline clustering approaches. The results are extensively evaluated using various clustering evaluation metrics as described in the last section of this chapter.

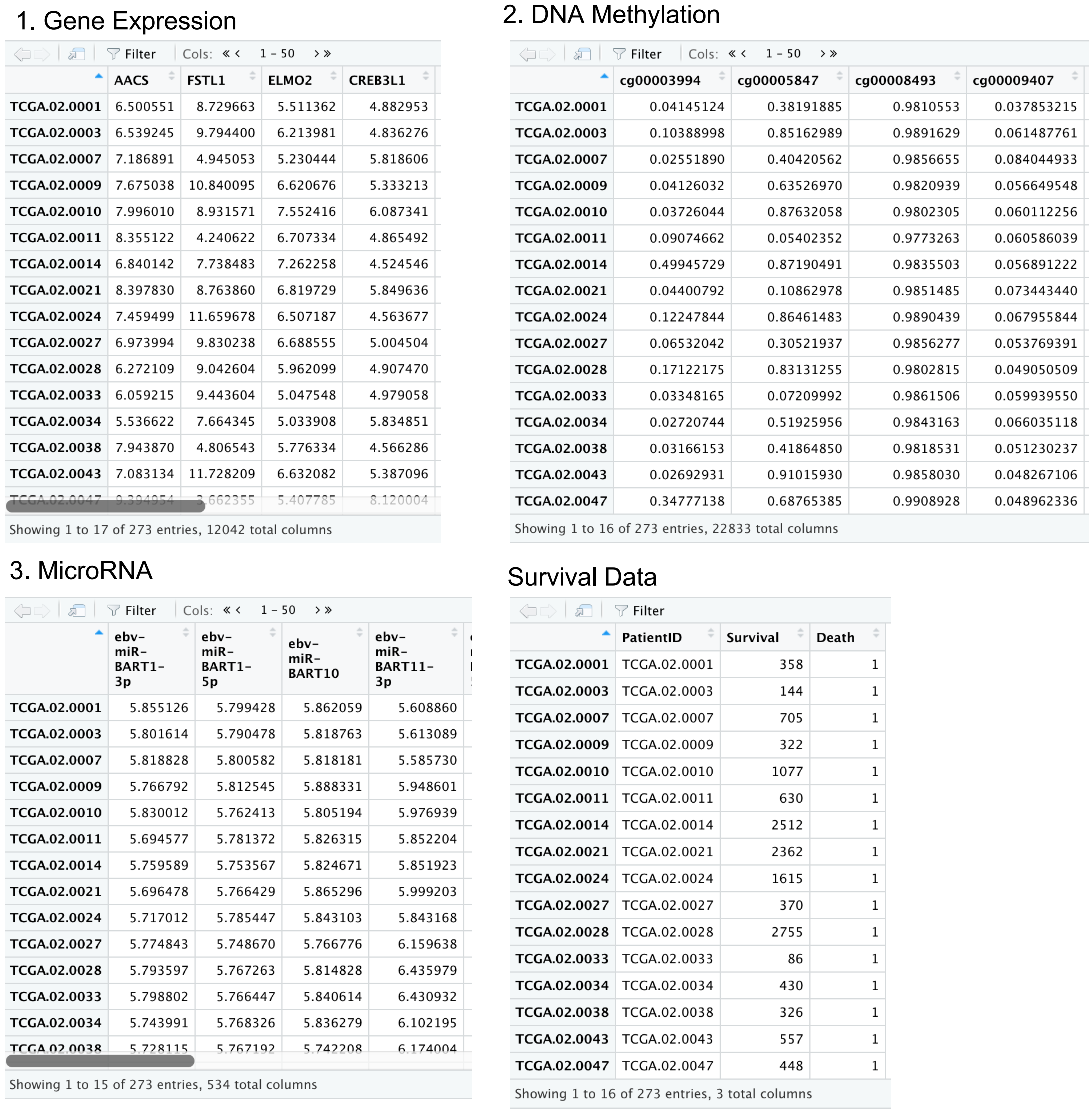
4.1.1 Exploring The Disease Subtyping Data

We selected five cancer disease datasets as used in [28, 39], which were taken from TCGA [1]

The selected five datasets include Kidney Renal Clear Cell Carcinoma (KIRC), Glioblastoma Multiforme (GBM), Lung Squamous Cell Carcinoma (LUSC), Breast Invasive Carcinoma (BRCA), and Colon Adenocarcinoma (COAD). These are

[1] https://www.cancer.gov/tcga

multi-view high-dimensional datasets, that consist of Gene Expression, DNA Methylation, and MicroRNA data. The following Figure (4.1) shows the data for GBM cancer. It includes three types of data (Gene Expression, DNA Methylation, and MicroRNA) which need to be analysed and integrated for disease subtyping.



1. Gene Expression

	AACS	FSTL1	ELMO2	CREB3L1
TCGA.02.0001	6.500551	8.729663	5.511362	4.882953
TCGA.02.0003	6.539245	9.794400	6.213981	4.836276
TCGA.02.0007	7.186891	4.945053	5.230444	5.818606
TCGA.02.0009	7.675038	10.840095	6.620676	5.333213
TCGA.02.0010	7.996010	8.931571	7.552416	6.087341
TCGA.02.0011	8.355122	4.240622	6.707334	4.865492
TCGA.02.0014	6.840142	7.738483	7.262258	4.524546
TCGA.02.0021	8.397830	8.763860	6.819729	5.849636
TCGA.02.0024	7.459499	11.659678	6.507187	4.563677
TCGA.02.0027	6.973994	9.830238	6.688555	5.004504
TCGA.02.0028	6.272109	9.042604	5.962099	4.907470
TCGA.02.0033	6.059215	9.443604	5.047548	4.979058
TCGA.02.0034	5.536622	7.664345	5.033908	5.834851
TCGA.02.0038	7.943870	4.806543	5.776334	4.566286
TCGA.02.0043	7.083134	11.728209	6.632082	5.387096
TCGA.02.0047	9.394954	3.662355	5.407785	8.120004

Showing 1 to 17 of 273 entries, 12042 total columns

2. DNA Methylation

	cg00003994	cg00005847	cg00008493	cg00009407
TCGA.02.0001	0.04145124	0.38191885	0.9810553	0.037853215
TCGA.02.0003	0.10388998	0.85162989	0.9891629	0.061487761
TCGA.02.0007	0.02551890	0.40420562	0.9856655	0.084044933
TCGA.02.0009	0.04126032	0.63526970	0.9820939	0.056649548
TCGA.02.0010	0.03726044	0.87632058	0.9802305	0.060112256
TCGA.02.0011	0.09074662	0.05402352	0.9773263	0.060586039
TCGA.02.0014	0.49945729	0.87190491	0.9835503	0.056891222
TCGA.02.0021	0.04400792	0.10862978	0.9851485	0.073443440
TCGA.02.0024	0.12247844	0.86461483	0.9890439	0.067955844
TCGA.02.0027	0.06532042	0.30521937	0.9856277	0.053769391
TCGA.02.0028	0.17122175	0.83131255	0.9802815	0.049050509
TCGA.02.0033	0.03348165	0.07209992	0.9861506	0.059939550
TCGA.02.0034	0.02720744	0.51925956	0.9843163	0.066035118
TCGA.02.0038	0.03166153	0.41864850	0.9818531	0.051230237
TCGA.02.0043	0.02692931	0.91015930	0.9858030	0.048267106
TCGA.02.0047	0.34777138	0.68765385	0.9908928	0.048962336

Showing 1 to 16 of 273 entries, 22833 total columns

3. MicroRNA

	ebv-miR-BART1-3p	ebv-miR-BART1-5p	ebv-miR-BART10	ebv-miR-BART11-3p
TCGA.02.0001	5.855126	5.799428	5.862059	5.608860
TCGA.02.0003	5.801614	5.790478	5.818763	5.613089
TCGA.02.0007	5.818828	5.800582	5.818181	5.585730
TCGA.02.0009	5.766792	5.812545	5.888331	5.948601
TCGA.02.0010	5.830012	5.762413	5.805194	5.976939
TCGA.02.0011	5.694577	5.781372	5.826315	5.852204
TCGA.02.0014	5.759589	5.753567	5.824671	5.851923
TCGA.02.0021	5.696478	5.766429	5.865296	5.999203
TCGA.02.0024	5.717012	5.785447	5.843103	5.843168
TCGA.02.0027	5.774843	5.748670	5.766776	6.159638
TCGA.02.0028	5.793597	5.767263	5.814828	6.435979
TCGA.02.0033	5.798802	5.766447	5.840614	6.430932
TCGA.02.0034	5.743991	5.768326	5.836279	6.102195
TCGA.02.0038	5.728115	5.767192	5.742208	6.174004

Showing 1 to 15 of 273 entries, 534 total columns

Survival Data

	PatientID	Survival	Death
TCGA.02.0001	TCGA.02.0001	358	1
TCGA.02.0003	TCGA.02.0003	144	1
TCGA.02.0007	TCGA.02.0007	705	1
TCGA.02.0009	TCGA.02.0009	322	1
TCGA.02.0010	TCGA.02.0010	1077	1
TCGA.02.0011	TCGA.02.0011	630	1
TCGA.02.0014	TCGA.02.0014	2512	1
TCGA.02.0021	TCGA.02.0021	2362	1
TCGA.02.0024	TCGA.02.0024	1615	1
TCGA.02.0027	TCGA.02.0027	370	1
TCGA.02.0028	TCGA.02.0028	2755	1
TCGA.02.0033	TCGA.02.0033	86	1
TCGA.02.0034	TCGA.02.0034	430	1
TCGA.02.0038	TCGA.02.0038	326	1
TCGA.02.0043	TCGA.02.0043	557	1
TCGA.02.0047	TCGA.02.0047	448	1

Showing 1 to 16 of 273 entries, 3 total columns

Figure 4.1: The GBM cancer data with three types of data (gene expression, DNA methylation, and microRNA) from 273 samples, along with survival information for clustering quality assessment.

As can be seen from the Figure (4.1) each data type has the same number of samples (rows) but the number of measurements (columns) is different for each view. The gene expression data has 12042, the DNA methylation has 22833, and the microRNA data has 534 measurements.

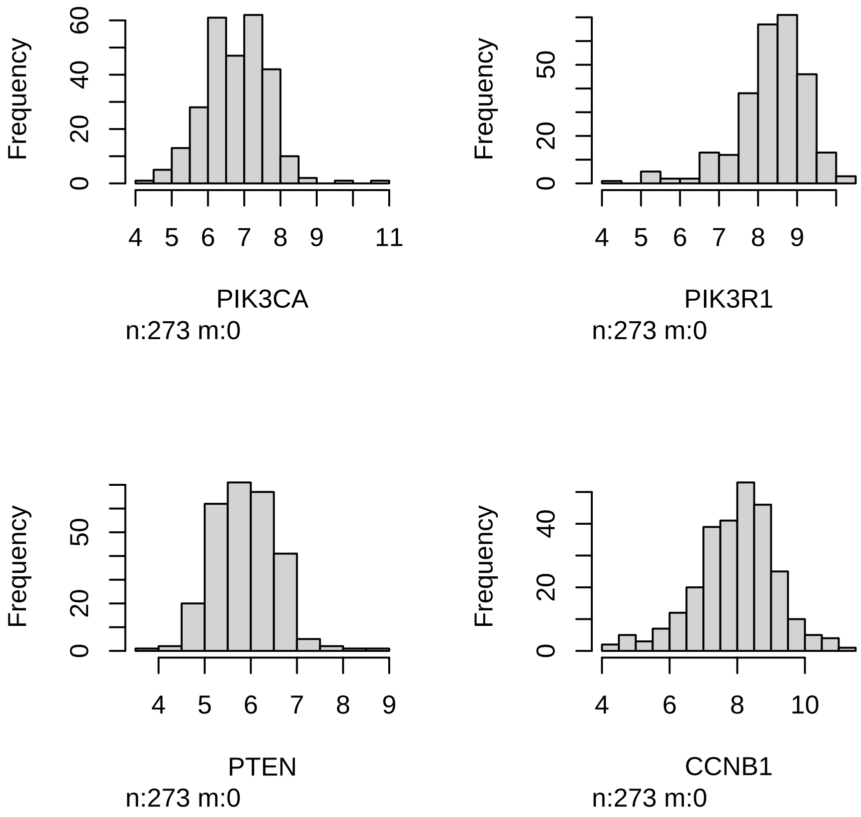


Figure 4.2: The histogram for four measurements (genes) taken form gene expression view of GBM

To get a general understanding of the distribution of gene expression values, we plotted a few measurement values from gene expression data in the Figure (4.2). It is evident from this figure that the measurement values exhibit a high degree of skewness. These skewed values pose a challenge for clustering algorithms as they can lead to suboptimal clusterings. To address this issue, we will propose an approach in this chapter that aims to mitigate the impact of these extreme values. By implementing this approach, we aim to minimise the influence of such extreme values on the clustering algorithms. The following chapter will delve into the details of this proposed method, offering a comprehensive solution for handling extreme values in the context of clustering analysis.

Now to get an understanding of how these measurement or gene expression values are distributed for samples we plotted a scatter plot for pair of samples as

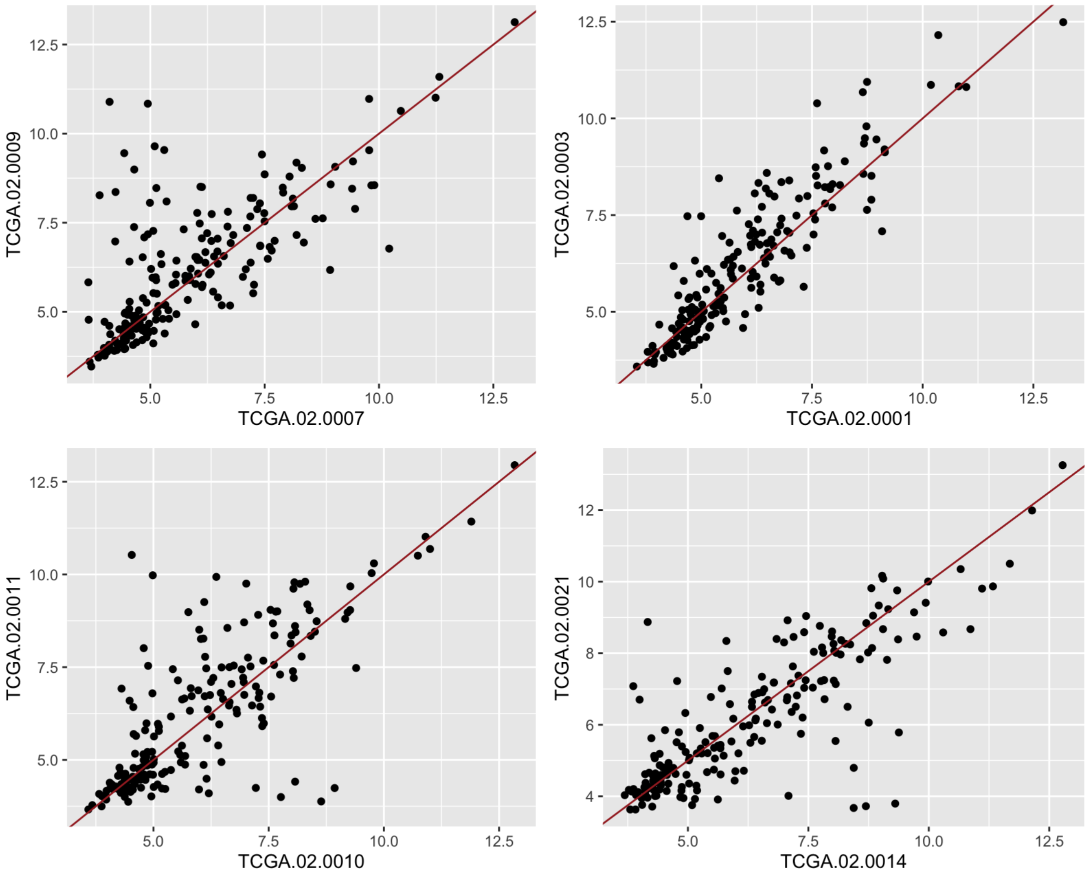


Figure 4.3: Plotting the genetic similarity between a pair of samples based on the gene expression measurements.

shown in Figure (4.3). The measurement values for these pair of samples exhibit a significant degree of scatter or dispersion. This scattering poses a challenge for distance functions used to calculate the similarity between samples. The wide range of values and their distribution across the measurement space can impact the accuracy of similarity computations. Therefore, it becomes crucial to address this issue and develop strategies to mitigate the influence of scattered measurements on distance-based similarity calculations.

In the following Figure (4.4), a heatmap is generated to visualise the gene expression and DNA methylation data of GBM cancer, aiming to identify common genetic patterns. Additionally, a dendrogram is plotted to explore the potential hierarchical clustering of samples, with the intention of grouping them into four distinct clusters. These visualisations provide valuable insights into the underlying structure of the data and facilitate the identification of shared genetic characteristics among samples.

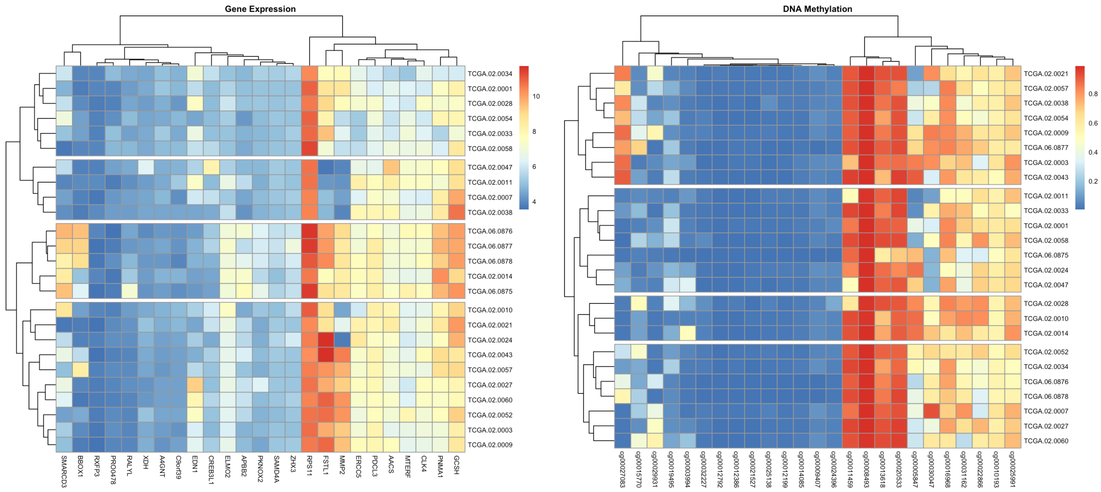


Figure 4.4: The visualisation of heatmaps for GBM data, including gene expression and DNA methylation, for the first 25 samples and first 25 measurements, with a dendrogram using a cutree of four.

The heatmaps in Figure (4.4) exhibit distinct genetic patterns among the samples. The colour coding scheme represents the strength of measurement values, with blue indicating minimum values and red denoting the highest values. Furthermore, a dendrogram is displayed alongside the heatmap, illustrating the clustering of samples into four groups using a cutree value of four. These visualisations employ the R library pheatmap to represent the clusters in Gene Expression and DNA Methylation data. The primary objective of this research is to integrate these modalities into a unified view and cluster samples based on their genetic similarities.

4.2 Frequently used Notations

Table (4.1) provides a concise and organised compilation of frequently used notations. It presents an extensive array of symbols, abbreviations, and acronyms that will be frequently encountered in the following sections. It simplifies the process of understanding and interpreting complex notations. Each entry in this table is accompanied by a succinct yet informative description, enabling readers to quickly grasp the meaning and context of the notation they encounter.

Table 4.1: Frequently used notations

Notation	Meaning
M	$M = \{X_i, i \in 1, 2, .., t\}$ denotes a multi-modality or multi-view dataset. Where each modality or view is denoted with X_i, where i represents the i^{th} view or modality.
W	$W = \{w^k, k \in 1, 2, 3\}$ denotes the set of the estimated width vectors. Each feature vector is divided into three partitions and for each partition, the width is estimated using the XFD estimator.
x^k	A superscript denotes the partition number e.g., $x^k = k^{th}$ partition of a vector. Likewise, $x_i^k = k^{th}$ partition of the i^{th} feature.
w^1, w^2, w^3	w^1, w^2, w^3 denotes the width estimated on the partitions based on Q_1, iqr, and Q_4 respectively. where Q_1, iqr, and Q_4 denotes the 1^{st} quartile, interquartile and 4^{th} quartile respectively.
Q_1, iqr, Q_4	$\forall x_i \in Q_1 \leqslant \forall y_i \in iqr \leqslant \forall z_i \in Q_4$ and $x_i, y_i, z_i \subseteq X$
w^k	$w^k = \{w_i^k, i \in 1, ..., n\}$, n is the size of dimensions and w_i^k denotes the estimated width value of feature i on k^{th} partition.
B, β^k	B denotes the set of vectors containing an estimated number of buckets for each partition e.g. $B = \{\beta^k, k \in 1, 2, 3\}$, and $\beta^k = \{\beta_i^k, i \in 1, ..., n\}$, and β_i^k denotes the estimated number of buckets of feature i on k^{th} partition.
$\phi(.)$	$\phi(.)$ facilitates the grouping of elements into buckets.
$G = (V, E)$	$G = (V, E)$, denotes a general or single type graph
$G = (V_t, E_{tt'})$	denotes a multi-typed graph, with t type of vertices.

4.3 Problem Definition & Formulation

Suppose that given a multi-view dataset M, consisting of t views or modalities. Each view is represented by a matrix e.g., $X_i \in \mathbb{R}^{m \times n}$ e.g, m samples and n features, where $i \in \{1, 2, \ldots, t\}$. We will often use X without subscripts to denote any view in the multi-view dataset M. Similarly, x represents any row vector or sample in X where x_i^j represents the element at i^{th} row and j^{th} column of the view X. The multi-view dataset M can then be represented as a collection or set of t matrices, as follows: $M = \{X_1, X_2, \ldots, X_t\}$. In the multi-view dataset M, each view X_i has the same number of observations (samples) across the views, but the number of measurements (features) might be different in different views. The goal is to identify the cohorts of patients (samples) within each individual view first, and then integrate all of the views within a dataset M, and identify the cohorts of patients on the integrated view.

To achieve this goal, we need to create an $m \times m$ pairwise distance matrix for each individual view X_i using a robust distance function in such a way where m denotes the number of observations (patients) and each entry of this matrix e.g., $(e_{i,j}) \in \mathbb{R}^{m \times m}$ denotes the distance (dissimilarity) between the patient i, and j. This distance matrix needs to be transformed into a similarity matrix for the disease subtyping. To do so, this distance matrix is provided to a similarity kernel e.g., Radial Basis Function (Gaussian Kernel), which transforms it to another $m \times m$ similarity matrix, where each entry of this matrix $(e_{i,j}) \in \mathbb{R}^{m \times m}$ denotes the similarity between the patient i, and j. This similarity matrix is represented via a similarity graph e.g., $S = (V, E)$, where V denotes the vertices (patients), and E denotes the set of edges between the vertices. Each edge carries a real value denoting the strength of similarity between the patients. Finally, **A)** The similarity graph S is provided for the graph-based clustering to identify cohorts of patients in S. **B)** All the similarity graphs for each view within a dataset M are integrated using the similarity network fusion (SNF) [38]. It generated a single similarity graph for each dataset. This integrated similarity graph is provided for graph-based clustering to identify cohorts of patients on the integrated similarity graph.

4.4 Multi-view Graph-based Clustering

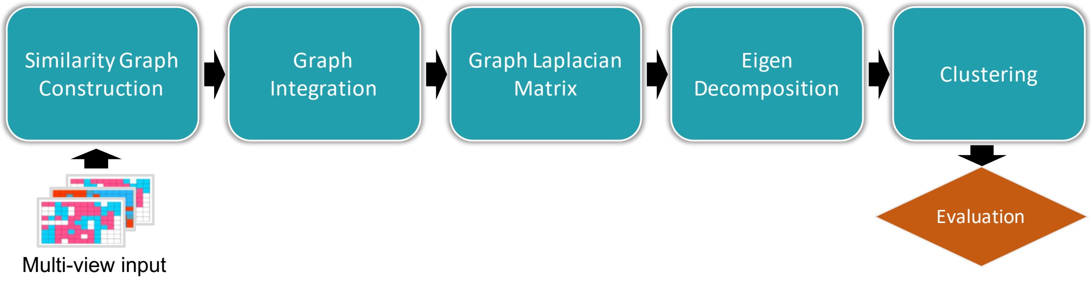


Figure 4.5: Multi-view graph-based clustering pipeline

Multi-view graph-based clustering pipeline is shown in Figure (4.5). Given a multi-view dataset, a graph-based clustering approach such as Spectral Clustering uses the following stages to partition the dataset into k clusters. First, it builds a similarity graph for each view. Second, it integrates all of the produced similarity graphs into a single graph. Third, it constructs the Laplacian matrix from the similarity graph. Fourth, in the eigen-decomposition, it computes the eigenvectors and eigenvalues of the Laplacian matrix: This stage computes the vectors and values that define the clusters. Finally, k-means clustering is applied to the eigenvectors selected in the previous stage, this stage allocates each data point to a cluster. Following the generation of the clusters, numerous clustering evaluation approaches can be employed to analyse and validate the quality of the formed clusters. The evaluation techniques used in this research are explained in the last section of this chapter.

The input to this pipeline is a multi-view dataset. A multi-view (also known as multiple datatypes) dataset contains more than one view for each observation (also called a data point). Each view represents a different aspect of the observation. For instance, in the case of cancer diseases, we need to extract multiple types of data from each patient, such as gene expression, DNA methylation, and MicroRNA, to better understand the disease. While disease subtyping can be accomplished with a single view, such as gene expression alone. However, these views might include complementary information about the disease and thus their integration can improve the subtyping results.

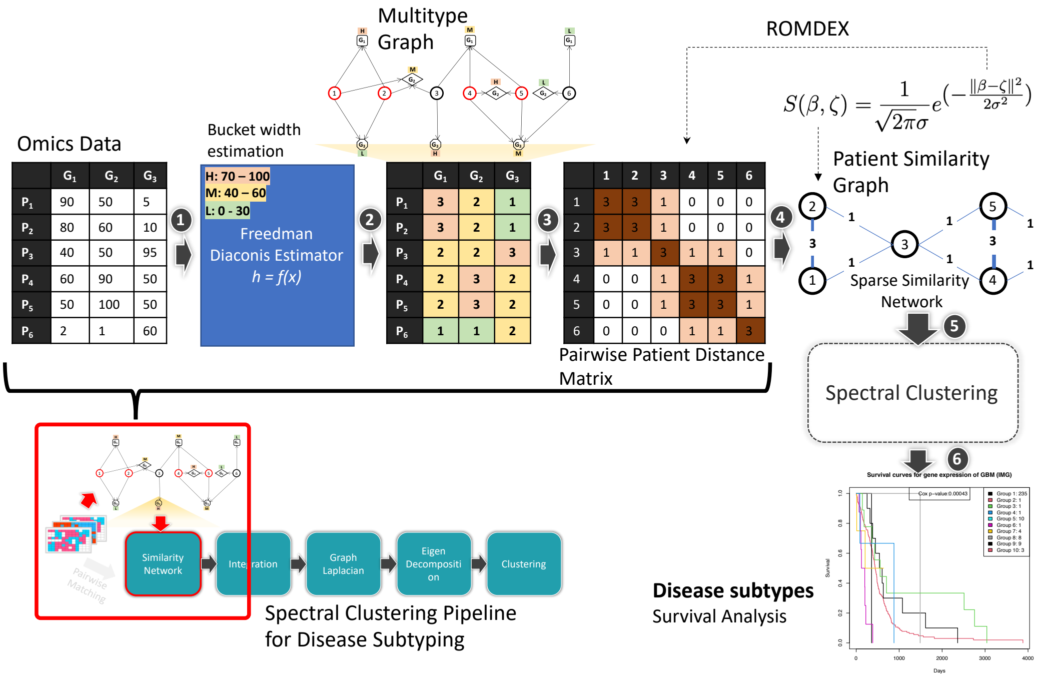
The first two phases are critical to the overall clustering quality. These procedures must be carefully followed in order to build clusters that are accurate representations of the data. For example, if the created similarity graph does not accurately match the actual data, it will be unable to generate correct clusters from it in the following stages. Therefore, accurate graph construction is essential for data mining which helps extract useful information from complex datasets and is a vital technique for a wide range of machine-learning applications. The same can be said for the graph integration in the second stage. High dimensionality, extreme values, and data variability are challenges that obstruct the construction of accurate similarity graphs from the data.

4.4.1 Robust Similarity Graph Construction

In this section, the architecture for the proposed novel similarity graph construction approach is provided with details on each stage. It effectively constructs similarity graphs from each view and then integrates the constructed similarity graphs into a single graph. The proposed approach is embedded into the multi-view graph-based clustering pipeline and proposed a novel clustering framework that can be used for clustering, disease subtyping, object detection, and classification. Figure (4.6), shows the proposed approach for similarity graph construction from the multi-view datasets. The following steps are followed in sequence to generate a robust similarity graph for clustering:

1. Preprocessing multi-view dataset

2. Statistical data binning (buckets) on each view



	G_1	G_2	G_3
P_1	90	50	5
P_2	80	60	10
P_3	40	50	95
P_4	60	90	50
P_5	50	100	50
P_6	2	1	60

	G_1	G_2	G_3
P_1	3	2	1
P_2	3	2	1
P_3	2	2	3
P_4	2	3	2
P_5	2	3	2
P_6	1	1	2

	1	2	3	4	5	6
1	3	3	1	0	0	0
2	3	3	1	0	0	0
3	1	1	3	1	1	0
4	0	0	1	3	3	1
5	0	0	1	3	3	1
6	0	0	0	1	1	3

$$S(\beta, \zeta) = \frac{1}{\sqrt{2\pi\sigma}} e^{(-\frac{\|\beta-\zeta\|^2}{2\sigma^2})}$$

Figure 4.6: Proposed methodology.

3. Construction of intermediate multi-typed graph models (IMG) from buckets

4. Robust construction of similarity graphs from IMG

5. Integration of the constructed similarity graphs

Preprocessing multi-view dataset

The following pre-processing steps have been performed on this high-dimensional data. The first question is about which features have been included in the clustering, depending on their availability across the patients. For instance, in each view, any feature with more than 60% missing values across the observations is excluded from the analysis. And the remaining missing values for the features are replaced with the mean value. Similarly, any observation missing from any of the views is excluded from the analysis. As this approach is integrative, therefore, it assumes that all the observations are available across the views.

Statistical data binning (buckets) on each view

Statistical binning groups data points into bins (also called buckets) and provides insights that would not be apparent from raw data. In addition, it reduces the amount of data that needs to be processed, stored, and analysed. As omics datasets are often complex and contain extreme values with huge data variability between them, therefore binning can improve the accuracy of estimates and predictions by reducing noise and outliers in the data. Thus, the rationale behind the statistical binning was to minimise the impact of data variability and outliers on the distance metrics, and similarity graphs and to improve the overall accuracy of the clustering results.

An important feature of data binning is bin width estimation. Common approaches used for bin width estimation are Sturge's rule [92], Scott's rule [93], Rice's rule [94], and Freedman Diaconis estimator (FD) [95]. Sturge's rule and Rice's rule do not take into account the data variability and extreme values. Similarly, Scott's rule is based on standard deviation, therefore, it is not robust to outliers. These statistical procedures for width estimation work best for small and normally distributed data, however, real-world datasets are often huge and skewed. Moreover, these techniques create equal-width bins which are not useful for complex and noisy data. Equal-width buckets lead to skewed results, it is especially true when data contain variability and outliers.

On the other hand, the FD estimator does consider data variability, as it is based on statistical quartiles which are robust to outliers. whilst, the FD estimator is robust against data variability, it computes an optimal fixed bin width on the inter-quartile range (IQR) only. The fixed bin width in a skewed distribution ignores the outliers or extreme values. This leads to the loss of valuable data and fixed-size buckets failing to handle outliers effectively. Therefore, we proposed an extended version of the FD estimator to address this limitation. The extended version works by dividing each feature into three defined intervals e.g., Q_1, IQR and Q_4. Afterwards, a bin width is estimated on each interval, resulting in multiple variable-sized bin widths. Using this division the highly influential outliers (if any) will be in Q_1, and/or Q_4. The rationale behind this division was to minimise the influence of outliers (e.g., data points on the extreme ends) by grouping them into relevant buckets. The buckets constructed at this stage are then transformed into a multi-typed graph. The following section describes how the process of transforming these buckets into a multi-typed graph.

Construction of intermediate multi-typed graph models (IMG)

Graphs provide a way to represent data in a structured manner that preserves the relationships between different elements. This allows for certain inferences to be made about the data based on its structure. In addition to the content similarities graphs add semantics such as the relationship between different elements, the strength of the relationships and the types of relationships between the various elements. All of this information is important in understanding the data set as a whole. Without it, it would be left with a much less informative representation of the same data. In addition, graphs are used to find patterns, trends, and outliers in data sets, and can be used to understand the structure of complex data.

This information is vital for constructing similarity graphs that best represent the underlying structure of the data. As a result, it contributes to better overall clustering outcomes. Therefore, the rationale behind the graph-based representation of the data at the intermediate level was to consider the structure of the data as well, rather than just the individual elements. As it allows for a more global view of the data and helps in identifying patterns and relationships that may not be apparent from looking at individual elements. Therefore, we proposed an intermediate multi-typed graph model (IMG) to represent the structural information of the high-dimensional data.

IMG is built from the buckets constructed in the previous stage. Note that IMG is defined at an intermediate level which is then used to construct the final similarity graph. In the IMG graph, the vertices (also called nodes) are the buckets which contain one or more measurements (also called feature values). It contains $n + 1$ type of vertices, where n is the dimensionality of the view (data type) and the number of vertices is equivalent to the number of buckets constructed in the previous stage. In IMG, each feature type is denoted with a particular type of vertex and additionally there is a vertex type denoting the data points (also called observation). Any two data points having their measurements co-occurred in the same bucket will be connected through an edge in IMG. This IMG modelling is then used in the following stage to construct similarity graphs for clustering.

Robust construction of similarity graphs from IMG

An important factor in determining the quality of a clustering algorithm is its ability to find groups of similar data points. Therefore, similarity graphs are useful in graph-based clustering, as they provide an effective way to measure vertex similarities. This is used to cluster vertices together based on their similarity in the graph. The IMGs constructed in the previous stage is used to construct this similarity graph. The IMG graph modelling shows the similarity and structural information both at the feature level and at the data points level. This modelling can answer questions like how closely connected a pair of features or data points are. It can also find overlapping features and data points. Therefore, a distance metric is defined over the IMG vertices as a measure of similarity. The similarity information about the data points is extracted from the IMG based on the feature similarities to construct a similarity graph. This is achieved by first defining a similarity metric over the IMG vertices to construct a pairwise distance matrix. Finally, a similarity kernel is defined over this distance metric to construct a fully connected similarity graph. This similarity graph is represented through a weighted graph. In the weighted graph, the nodes represent the data points while the weighted edges represent the pairwise similarities. A weight over the edge connecting a pair of data points represents the strength of connectivity between two data points.

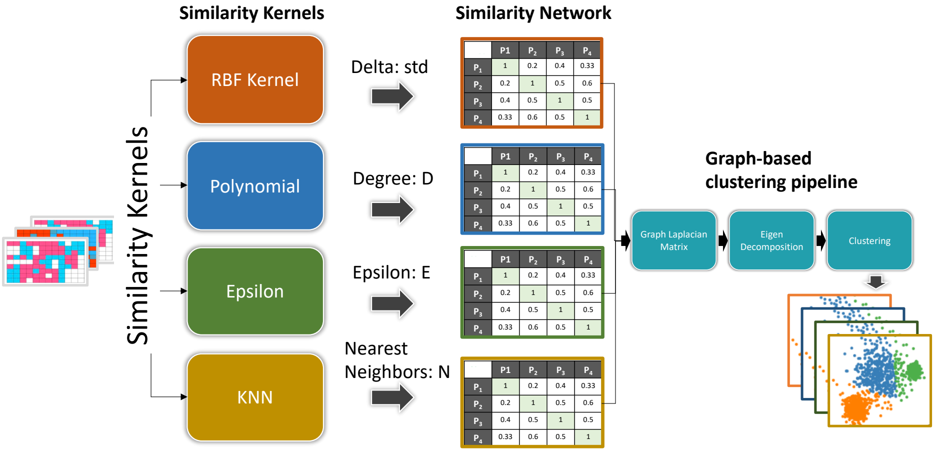


Figure 4.7: It shows the construction of similarity networks from each view X_i. These similarity networks can be constructed by employing any of the kernels shown in this figure. The constructed similarity networks can then be provided to a graph-based clustering algorithm like spectral clustering to group samples into K groups.

Fig (4.7) shows various kernels that are often employed as a similarity measure, such that K(x,x') is large when x and x' are more similar and vice versa [96]. The similarity kernel is defined over the distance metric to construct a similarity graph (also called network) as shown in Fig (4.7). The similarity graph in the figure is represented through a similarity matrix where the rows and columns denote data points and the values inside each cell denote the similarity between a pair. Note that the diagonal values are all 1^s which represents the similarity of a data point with itself (also called self-edge in graphs). The final clustering is based on this similarity graph. In addition, in the case of multi-view datasets, an IMG and hence a similarity graph are constructed for each view. These similarity graphs are then combined to generate a single similarity graph. As a result, clustering is done on both individual similarity graphs and the integrated similarity graph. The following section describes the process of integrating these similarity graphs.

4.4.2 Integration of the Constructed Similarity Graphs

The importance of disease subtyping is the type of data that is collected about the disease. The data-driven subtyping is usually achieved on gene expression, DNA methylation, and miRNA data types. In some cases, the subtypes of the disease are discovered on a single data type however, in most cases, these data types contain complementary information about the disease. Therefore, their integration into a single similarity graph is of vital importance for the discovery of the subtypes of a disease. For this reason, the individual similarity graphs that are being constructed from the individual data types (views) are integrated into a single similarity graph using the similarity network fusion SNF [38]. SNF integrates the similarity graphs in a nonlinear manner that considers the similarities, and complementary pieces of information across the similarity graphs. It iteratively integrates these graphs with a network fusion technique to construct the final integrated similarity graph. In order to integrate the constructed similarity graphs, the following similarity network fusion [38] approach is used:

$$G^{(X)} = N^{(X)} \times \left[\frac{\Sigma_{k \neq X} G^{(k)}}{t-1}\right] \times (N^{(X)})^T \tag{4.1}$$

where, $X = 1, 2, .., t$, denotes the number of views e.g., (data types). G, is a fully connected similarity graph constructed for each view X, while N, is a local affinity retrieved through KNN, which contains the similarity information about k nearest neighbours for each data point in G.

The mathematical details of the proposed approach for similarity graph construction and the overall graph-based clustering pipeline are provided in the next chapter. While the following section describes the evaluation metrics used to evaluate the accuracy performance of the proposed multi-view graph-based clustering approach.

4.5 Evaluation Metrics

The aim of clustering evaluation methods is to provide insight into how well the algorithm is performing. This can be useful for both debugging purposes and for comparing different algorithms against each other. It is important to remember, however, that no single evaluation metric will give a perfect picture - it is often necessary to use a combination of different methods in order to get a complete understanding of how an algorithm is performing. Following are the clustering evaluation metrics that are usually used in combination with disease subtyping.

4.5.1 Survival Analysis

Survival analysis is a type of statistical analysis that is often used in disease subtyping. This method looks at the length of time that patients survive after being diagnosed with a disease. This information can be used to create models that predict how long a patient is likely to live and what factors may influence their survival. Survival analysis can be used to compare different subgroups of patients with a disease. For example, survival analysis could be used to compare the survival rates of patients with different subtypes of cancer. This information can help doctors choose the best treatment options for each patient. To achieve this, time-to-event data is modelled. An event of interest in biological studies is "death". However, this information might not be available to all patients after the end of the follow-up studies, which affects the survivability results. The phenomenon in which the occurrence of the event for some patients is unknown is known as censoring. For instance, some patient may stop follow-up, or a different event other than the event of interest occurs with some patients or the event do not occur after the end of the follow-up study.

In bioinformatics, Survival analysis is used in different ways in this thesis however we look into it from two perspectives. Firstly, from the Kaplan-Meier perspective, which is used to describe the survival time of patients belonging to a group. Secondly, from the Cox proportional hazards perspective, which is used to see the impact of categorical, or quantitative variables on survival.

Kaplan-Meier curves

Kaplan-Meier curves are commonly used in survival analysis, which is a branch of statistics that deals with the study of data relating to the time until some event occurs. The Kaplan-Meier curve is a graphical representation of the estimated probability of surviving over time. The Kaplan-Meier curve can be used to compare the survival functions of two or more groups of subjects. For example, if we wanted to compare the survival probabilities of two groups of patients with different types of cancer, we could use a Kaplan-Meier curve. To create a Kaplan-Meier curve, we first need to calculate the survival function for each individual in our data set. This can be done by using the following formula:

Definition 4.1 (Kaplan-Meier curves)**.**

$$S(t_i) = \frac{a_i - e_i}{a_i} \times S(t_{i-1}) \tag{4.2}$$

Here, $S(t_i)$ is the survival probability at time i, and a_i denotes the number of members that were alive at time t_i, and e_i is the number of members that were dead at time t_i, and finally, $S(t_{i-1})$ is the survival probability at time t_{i-1}.

Note that it is a recurring formula and initially at time t_0 all the members of the follow-up were alive therefore, $S(t_0) = 1$, where $t_0 = 0$, denotes the beginning time.

Cox proportional hazards

Cox proportional hazard is a statistical model used to assess the effect of one or more covariates on the time of an event. It is widely used for survival analysis. This model allows to estimate the hazard function, which is the probability of an event occurring at a given time, and to identify the factors that influence the hazard function. The model also allows to compare the hazard function between different groups of individuals. For example, we can use the Cox proportional hazards model to compare the survival rates of men and women or to compare the survival rates of different ethnic groups. The formula for Cox proportional hazards is given as follows:

Definition 4.2 (Cox proportional hazards).

$$h(t) = h_0(t)exp(a_1x_1 + a_2x_2 + \ ... \ + a_nx_n) \tag{4.3}$$

Here, $x = \{x_i, i \in 1, ..., n\}$ denotes a set of n covariates, and the coefficient a with a covariate denotes its impact. Similarly, t denotes the survival time, $h(t)$ denotes the expected hazard at time t, while $h_0(t)$ denotes the baseline hazard when all the covariates are zero.

Survival analysis is one method that can be used in disease subtyping. Other evaluation methods, such as concordance index, normalised mutual information (NMI), and clustering purity, can also be used to evaluate subtyping results. Ultimately, disease subtyping aims to improve our understanding of diseases and better tailor treatments to each patient's needs.

4.5.2 Concordance Index (CI)

The concordance index (CI) [97], or C-Statistics is often used to measure how well a model predicts time to an event. In disease subtyping, CI assigns a risk score to each patient, a higher risk score means a shorter time for an event. In medical settings, the event is usually disease or death. Therefore, the high-risk score means the patients will soon encounter the event. The CI index is computed by dividing the number of concordant pairs by the total number of evaluation pairs. The total number of evaluation pairs is usually the sum of the total concordant and discordant pairs. The CI index can be formulated in a formula given below which is taken from [98]:

Definition 4.3.

$$CI = \frac{\sum_{i \neq j} 1\{\lambda_i < \lambda_j\} 1\{T_i > T_j\} d_j}{\sum_{i \neq j} 1\{T_i > T_j\} d_j} \tag{4.4}$$

Here, λ denotes the risk score whereas the attached subscript denotes the patient index for which the risk score is calculated. Similarly, T_i, T_j denotes the time to an event for patient i, j, respectively. A pair of patients i, j is **concordant** if $\lambda_i > \lambda_j$ and $T_i < T_j$. While it is **discordant** if $\lambda_i > \lambda_j$ and $T_i > T_j$.

4.5.3 Normalised Mutual Information (NMI)

Normalised mutual information (NMI) is a measure of the similarity between two variables. It is often used to evaluate clustering quality. NMI is typically expressed as a value between 0 and 1, where 0 indicates no similarity and 1 indicates perfect similarity. To get an NMI score, first, compute the mutual information (MI) between the two data sets. MI is a measure of how much information is shared between two variables. Once MI has been computed, NMI can be obtained by the following equation:

Definition 4.4 (Normalised Mutual Information)**.**

$$NMI(Y, Y') = \frac{2 \times I(Y; Y')}{H(Y) + H(Y')} \tag{4.5}$$

Here, Y, and Y' denotes the set of true labels and cluster labels respectively. $I(Y; Y')$, denotes the mutual information between the true and cluster labels. $H(Y)$, and $H(Y')$ denote the entropy of true and cluster labels respectively.

NMI is a useful metric for comparing clusterings because it takes into account both the intra-cluster and inter-cluster similarities. A high NMI value is preferred as it indicates that the generated clusters are of high quality.

4.5.4 Clustering Purity

A cluster is considered to be pure if all of the points within it belong to the same class. If there are points from multiple classes in a cluster, it is impure. Measuring clustering purity can be useful for determining how well a clustering algorithm is performing. It is often used to evaluate the clustering quality of the clustering algorithms. Clustering purity is defined in the following equation:

Definition 4.5 (Clustering Purity).

$$Purity = \frac{1}{N}\sum_{i=1}^{k} max_j |y'_i \cap y_j| \tag{4.6}$$

Here, N denotes the number of observations, k denotes the number of clusters, and y'_i denotes the i^{th} cluster in C and y_j is the class j.

Clustering purity is an important metric, it evaluates the quality of clusters. A high degree of purity e.g., 1, indicates that the clusters are well-defined and that the members of each cluster share similar characteristics. On the other hand, a low degree of purity usually well below 0.5 indicates that the clusters are not well-defined and that their members are more heterogeneous. A clustering purity of around 0.5, is no better than random clustering.

4.5.5 Robustness and Stability Evaluation

In order to assess the robustness and stability of our proposed approach in the presence of noise, we conducted experiments using a well-known Iris clustering dataset [99], which provides ground truth labels. To introduce noise, we systematically added Gaussian noise to the dataset, gradually increasing the noise level. This allowed us to evaluate the performance of our approach on noisy data by employing various evaluation metrics, including Accuracy, Kappa statistic [100], and Rand index [101].

Gaussian noise is generated by introducing random values (noise) with a zero-mean and a standard deviation denoted as sigma. By adjusting the sigma value, we could control the level of noise in the dataset. After adding noise to the data we clustered this noisy data using the proposed approach. The predicted clustering labels are then compared with ground truth and robustness is evaluated using accuracy, kappa statistic and rand index. The Kappa statistic, which measures inter-rater agreement, was utilised to quantify the level of agreement beyond chance between the predicted labels and the ground truth labels. It ranges from

-1 to 1, with higher values indicating stronger agreement and a more reliable approach. Similarly, we employed the Rand index to evaluate the agreement in clustering, which takes values between 0 and 1. A Rand index value of 0 suggests no agreement in the clustering of any pair of elements, while a value of 1 indicates perfect agreement.

Chapter 5

ROMDEX

5.1 Introduction

This chapter focused on statistical methods that are being incorporated into graph theory to help improve the robustness of disease subtyping by grouping patients with similar characteristics from noisy data. This chapter is organised as follows:

In the first section, the definition and explanation of the preliminary knowledge for the proposed approach are provided. It provides definitions of statistical techniques related to the proposed approach, such as statistical binning, quarterlies, and interquartile ranges. The multi-typed graph model created using statistical binning for robust clustering on high-dimensional data is then defined. This is followed by the definition of the proposed Romdex function. First, a brief overview of the problem will be given. Next, the mathematical formulation of the robust distance metric will be presented. The proposed approach is designed to be robust to outliers and noise. The construction of robust similarity graphs from high-dimensional omics data is provided which can effectively handle noise and data variability. Finally, a comprehensive flow-chart and algorithm is outlined that connects all the techniques proposed and discussed in this thesis for robust disease subtyping on high-dimensional data.

5.2 ROMDEX - A Robust Metric for Data Variability & Extreme Values

This section provides details on the novel robust approach for similarity graph construction from multi-view datasets. This begins with an assumption about the data on which a distance metric is defined. This is followed by the creation of variable-sized buckets through statistical data binning from the multi-view dataset. Followed by the construction of intermediate multi-typed graphs (IMG) from these buckets. Finally, the construction of a similarity graph from IMG for graph-based clustering and disease subtyping. Similarity graphs are useful in graph-based clustering, as they provide an effective way to measure vertex similarities. The proposed approach exploits the underlying relationships and therefore finds meaningful clusters in data. This is used to cluster vertices together based on their similarity in the graph.

5.2.1 Mapping the Approach

Assumptions

Let's assume a set of m observations $\mathrm{X} = \{x_i : x_i \in X \subseteq \mathbb{R}^n\}^m$. The pair (X, d) is a metric space where d is called a metric a.k.a distance function. Suppose we have two observations, x, $y \in \mathrm{X}$ where $x = \{x_i, i \in 1, ..., n\}$, and $y = \{y_i, i \in 1, ..., n\}$.

Properties

Metric learning for numeric data aims to learn a distance metric $d(.,.) : \mathrm{X}{\times}\mathrm{X} \rightarrow \mathbb{R}^+$ for all observations in X that satisfies the following properties:

1. $d(x, y) \geqslant 0$, Positive Semi-definite

2. $d(x, y) = 0 \iff x = y$, Identity of Indiscernible

3. $d(x, y) = d(y, x)$, Symmetry

4. $d(x, z) \leqslant d(x, y) + d(y, z)$, Triangle Inequality

Limitation of Manhattan Distance

A typical metric to calculate the distance between observations is the Manhattan distance which can be defined as:

$$d(x, y) = \|x - y\|_{L1} = \sum_{i=1}^{n} |x_i - y_i| \tag{5.1}$$

It computes the distance between two points $x, y \in X$ as the sum of the absolute differences of their Cartesian coordinates, and it satisfies all of the four properties of a metric defined above. The Manhattan distance function (L_1 norm) is preferred for high dimensional data compared to the Euclidean distance function (L_2 norm) [67].

However, it lacks the elements required to be deemed a robust metric. A measure is said to be robust if it is insensitive to extreme values. This necessitates the synthesis of structurally relevant elements in order to mitigate the impact of highly influential outliers in distance metrics. As a result, a solution must be developed to synthesise the relevant measurements into imminent buckets in order to calculate the distance between them. As a result, in this part, all of the notions mentioned above are merged and a novel robust function for clustering is proposed.

This subsection is organized as follows:

1) An extended Freedman Diaconis estimator is proposed to estimate multiple variable-sized buckets widths on data.

2) An intermediate multi-typed graph (IMG) is proposed to represent the constructed buckets through a graph-based representation.

3) Finally, steps 1), and 2) are embedded in a distance function and proposed a novel robust function called ROMDEX for Clustering in general, and Disease subtyping.

5.2.2 Extended Freedman Diaconis Estimator

Statistical binning is a method of grouping data points into bins or buckets. It is used to reduce the effect of outliers, and noise in data. The common approach to data binning is to use equal-width bins, which group data points into buckets of equal width. A common reason for binning data is to improve the accuracy of results. This is especially true when working with statistical methods like clustering. Clustering aims to find groups of similar data points however, if the data is not evenly distributed, it can be difficult to find these groups. Binning the data can help to even out the distribution and make it easier to find clusters. In the following section, we define the widely used bin width estimation approaches.

Statistical data binning

In the literature, various statistical methods have been proposed, aiming to determine optimal bucket size through probability density estimation. The questions of optimal buckets and width are critical for constructing reliable similarity networks in the proposed research. One common approach to probability density estimation is histograms, which display the frequency of each value per independent feature through 'bins'. Bin width is inversely proportional to the number of bins. Good trade-offs are possible between the width and number of bins by employing accurate estimations of how a given feature is distributed. Common approaches used for bin width estimation are Sturge's rule [92], Scott's rule [93], Rice's rule [94], and Freedman Diaconis estimator (FD) [95].

Let's x denotes the vector, and n denotes the length of x, e.g., $n = |x|$, then the bucket width w, can be estimated using Sturge's rule as follows:

$$k = \lceil log_2 \, n \rceil + 1 \tag{5.2}$$

$$w = \frac{MAX(x) - MIN(x)}{\lceil log_2 \, n \rceil + 1} \tag{5.3}$$

Sturge's rule works best for small, normally distributed, and symmetrical data. However, it does not tack into account the variability and extreme values.
Similarly, Scott's rule is defined as follows:

$$h = \frac{3.49\sigma}{\sqrt[3]{n}} \tag{5.4}$$

Scott's rule works well for large datasets. As Scott's rule is based on the standard deviation, therefore, it is not robust to outliers.
Likewise, the Rice rule is defined as follows:

$$k = \lceil 2\sqrt[3]{n}\, \rceil \tag{5.5}$$

Similarly, the rice rule overestimates the number of bins and does not consider the data variability. Generally, the challenge in data binning is the estimation of bucket width, while the limitation is the equal-width bins. The bucket width is usually estimated using classical statistical procedures which are not robust to outliers. Similarly, equal-width buckets lead to skewed results, it is especially true when data contain variability and outliers. In the following section, we define the robust bin width estimation approach.

Freedman Diaconis Estimator

FD estimator [95] uses IQR instead of standard deviation which is robust to outliers. FD estimator is defined as:

$$w^2 = 2 \times \frac{IQR(x)}{\sqrt[3]{n}} \tag{5.6}$$

FD estimator computes the optimal bin width, which is proportional to the interquartile range (IQR) and inversely proportional to the cube root of N. It considers both data variability and size. It uses IQR instead of standard deviation which is robust to outliers. FD estimator, estimates buckets widths on inter-quartile Range which is comparatively robust to outliers than other bucket width estimation methods such as Sturges' [92], Scott's [93], and Rice rule [94]. Important concepts to the FD estimator are the statistical quartiles and interquartiles which are defined below:

Definition 5.1 (Quartile)**.** In statistics, quartile describes the division of data points into four segments of approximately equal size.

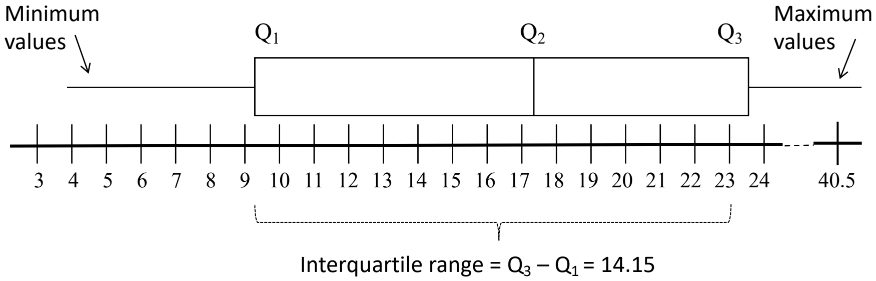


Figure 5.1: The division of a vector into statistical quartiles

According to Def. (5.1) a feature vector v, with n data points is sorted and divided into four parts in such a way that each part consists of approximately a quarter e.g., 25% of the total data points as shown in Fig. 5.1. The partitions Q_1, Q_2, Q_3, and Q_4 denotes the 1^{st}, 2^{nd}, 3^{rd}, and 4^{th} quartile respectively. The interquartile (IQR) is then computed by taking the difference between the 3^{rd} and 1^{st} quartile e.g., $iqr = Q_3$ - Q_1, and consists of the maximum distribution (50%) of the data points. In this thesis, a vector v is divided into three parts e.g., Q_1, iqr, and Q_4, and the data points in each part are grouped into buckets using variable-sized bucket widths. Using this division the highly influential outliers (if any) will be in Q_1, and/or Q_4. The rationale behind this division is to minimise the influence of outliers (e.g., data points on the extreme ends) by grouping them into relevant buckets.

Limitations of FD Estimator

The only caveat in the FD estimator is that it computes an optimal fixed bin width on IQR e.g., (25th to 75th percentile) of the data. The FD Estimator estimates fixed-size bucket width based on IQR only. The feature of fixed size buckets in a highly skewed distribution determines outliers as anything outside the range of $Q_1 - 1.5 \times IQR$ and $Q_3 + 1.5 \times IQR$. This results in empty buckets on either side of the extreme ranges. It is because the variation between the values on extreme ends is higher than the variation in the IQR range. This results in the loss of valuable data and fixed-size buckets not handling the variations in a more effective way.

The proposed approach aims to group elements into multiple optimum numbers of buckets to minimise the influence of the outliers on the data distribution. This can be achieved using an accurate estimation of the true underlying probability distribution of the observations. Therefore, the proposed approach extended Freedman Diaconis (FD) estimator [95] as it considers both data variability and size.

Extending FD Estimator to Multiple Variable Sized Bucket

The proposed methodology aims to achieve a robust metric by incorporating statistical binning and quartiles into the distance function. This necessitates a grasp of the extreme minimums and maximums of a highly skewed distribution. The extreme minimums are in the first quartile (Q_1), while the extreme maximums are in the fourth (Q_4). Therefore, for each dimension $x_i \in x$ we estimate width on Q_1, IQR and Q_4 as w^1, w^2, and w^3 respectively. Based on the estimated widths, the values in these designated quartiles are divided into buckets. Finally, the distances between the buckets are calculated using the proposed method.

FD Estimator generates fixed bucket sizes, which might result in data loss, therefore, it is extended to variable bucket sizes. The variable bucket sizes for the measurements are computed in the IQR, lower quartile, and upper quartile.

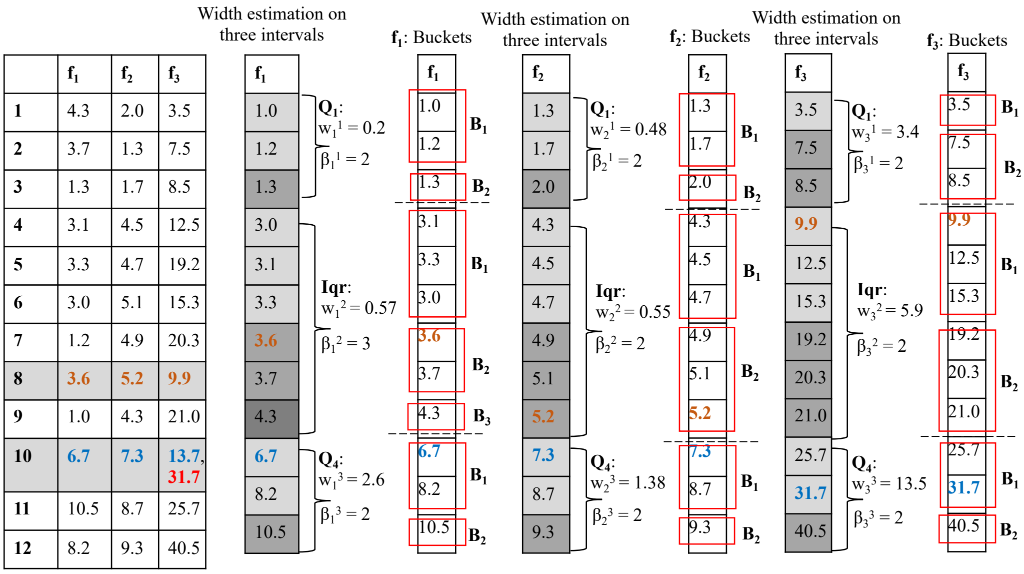
$$w^1 = 2 \times \frac{IQR(Q_1(x))}{\sqrt[3]{m_1}} \tag{5.7}$$

$$w^3 = 2 \times \frac{IQR(Q_4(x))}{\sqrt[3]{m_3}} \tag{5.8}$$

In the above equations, m_1, m_3 denotes the corresponding sizes for the first quartile and fourth quartile respectively. Where, Eq. (5.7) estimates the bucket width on the 1^{st} quartile (Q_1) of the data while Eq. (5.8) estimates the bucket width on the 4^{th} quartile (Q_4) of the data, thus returns bucket widths for the lower and upper extreme ranges respectively. Using this strategy, the number of buckets for IQR, Q_1, and Q_4 are estimated based on equation (5.6), (5.7), and (5.8) respectively. Let's say, w^1, w^2, and w^3 denote the estimated width on 1^{st} quartile, interquartile, and 4^{th} quartile of the feature vector v respectively, then the bucket number for any data point $x_i \in x$ is computed by dividing it on its respective estimated width.

Figure (5.2) provides a visual depiction of the bucket construction process from a single view, denoted as X_i. The figure includes an example view represented as a matrix with 12 samples and 3 features. To construct the buckets, each feature vector (column vector) is initially divided into three predefined intervals: Q_1 (first quartile), Iqr (interquartile range), and Q_4 (fourth quartile). Next, the proposed extended Freedman-Diaconis estimator is employed to estimate the width of each interval. These estimated widths are denoted as w_i^j, where i represents the feature number and j indicates the interval number. Once the bucket width is determined, it is used as a sliding window over the elements within the interval. Elements that occur within the same window are grouped together into the same bucket. In the illustrated Figure (5.2), two buckets are created in the first (Q_1) and third (Q_4) intervals for feature f_1, while three buckets are generated in the second (Iqr) interval.

This bucket construction process is repeated for each feature vector in the dataset, resulting in the creation of distinct buckets corresponding to different intervals and features. In summary, Figure (5.2) provides a visual explanation of the step-by-step bucket construction process from a single view, encompassing partitioning feature vectors into intervals, estimating widths, sliding window operations, and grouping elements into buckets. In addition, the generated buckets are represented by an intermediate multi-typed graph, allowing graph theoretical techniques to leverage the structural information of the data (IMG). The following section describes the way to represent buckets using IMG.



	f1	f2	f3
1	4.3	2.0	3.5
2	3.7	1.3	7.5
3	1.3	1.7	8.5
4	3.1	4.5	12.5
5	3.3	4.7	19.2
6	3.0	5.1	15.3
7	1.2	4.9	20.3
8	3.6	5.2	9.9
9	1.0	4.3	21.0
10	6.7	7.3	13.7, 31.7
11	10.5	8.7	25.7
12	8.2	9.3	40.5

Width estimation on three intervals
f1
1.0
1.2
1.3
3.0
3.1
3.3
3.6
3.7
4.3
6.7
8.2
10.5
Q1: $w_1^1 = 0.2$ $\beta_1^1 = 2$
Iqr: $w_1^2 = 0.57$ $\beta_1^2 = 3$
Q4: $w_1^3 = 2.6$ $\beta_1^3 = 2$
f1: Buckets
f1
1.0 1.2 B1
1.3 B2
3.1 3.3 3.0 B1
3.6 3.7 B2
4.3 B3
6.7 8.2 B1
10.5 B2
Width estimation on three intervals
f2
1.3
1.7
2.0
4.3
4.5
4.7
4.9
5.1
5.2
7.3
8.7
9.3
Q1: $w_2^1 = 0.48$ $\beta_2^1 = 2$
Iqr: $w_2^2 = 0.55$ $\beta_2^2 = 2$
Q4: $w_2^3 = 1.38$ $\beta_2^3 = 2$
f2: Buckets
f2
1.3 1.7 B1
2.0 B2
4.3 4.5 4.7 B1
4.9 5.1 5.2 B2
7.3 8.7 B1
9.3 B2
Width estimation on three intervals
f3
3.5
7.5
8.5
9.9
12.5
15.3
19.2
20.3
21.0
25.7
31.7
40.5
Q1: $w_3^1 = 3.4$ $\beta_3^1 = 2$
Iqr: $w_3^2 = 5.9$ $\beta_3^2 = 2$
Q4: $w_3^3 = 13.5$ $\beta_3^3 = 2$
f3: Buckets
f3
3.5 B1
7.5 8.5 B2
9.9 12.5 15.3 B1
19.2 20.3 21.0 B2
25.7 31.7 B1
40.5 B2

Figure 5.2: It shows an example data with n features and m rows. Each column vector (feature) is divided into three intervals e.g., Q_1, Iqr, and Q_4. A width w is estimated on each interval to group samples into buckets in each interval based on the similarity of feature value.

5.2.3 Intermediate Multi-typed Graph (IMG)

Definition 5.2. Intermediate Multi-typed Graph (IMG):

$G = (V_t, E_{tt'}, l_V, l_E, F_G)$, is a multi-typed graph, where:

- V_t, denotes a set of finite vertices of type t, $(t = 1 \; to \; c)$.
- L_V, denotes a set of vertex labels
- $E_{tt'}$, represents a set of edges between the vertices of type t, t' where $(t \neq t')$.
- Each vertex, $v, v \in V$, is assigned a list of finite pair of attributes from F_G.
- $F_G(v) = (A_1, a_1), (A_2, a_2) \; \ldots \; (A_n, a_n)$.

A multi-typed graph is created from the constructed buckets. Graph-based approaches exploit the structure of the data to find similarities between data points.

By examining the structure of the graph, it is possible to identify overarching patterns and trends. Therefore, we proposed an intermediate graph model (IMG), defined on the buckets computed from the high-dimensional data. IMG will be used by the proposed approach to construct a robust high-quality similarity graph for clustering in general and disease subtyping. Furthermore, the structure-based treatment of the data provides a way to create sparse matrices with well-separated clusters. Besides content similarity, IMG also captures structural information in the final similarity graph, which can be utilised to avoid a pairwise matching situation or generate sparse matrices with better-separated clusters. The pictorial representation of this IMG graph is provided in Fig (5.3).

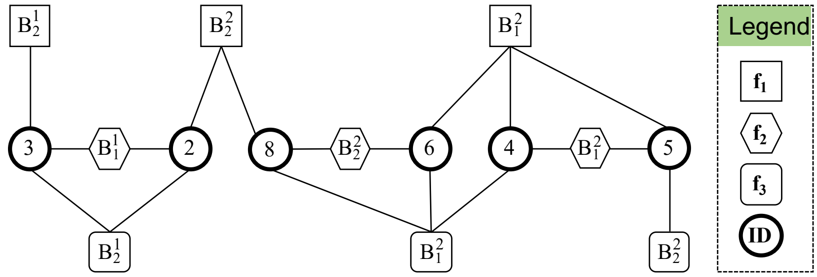


Figure 5.3: A fragment of the intermediate multi-typed graph (IMG)

Moreover, to enable graph theoretical techniques to compute distances in a *topological space*, we propose an intermediate graph representation of the constructed buckets. The constructed graph is called an intermediate multi-typed graph (IMG) as shown in Fig. (5.3). The graph in Fig. (5.3), shows a fragment of the buckets constructed in Fig. (5.2). IMG consists of n types of vertices, where n denotes the dimensions of the dataset. In Fig. (5.3), $n = 3$ since it consists of three different types of vertices, each represented with a distinct symbol. Note that each vertex in the graph is called a supernode as it may contain a set of similar values that co-occurred in a single bucket.

IMG modelling handles data variability, and skewness and supports the construction of a high-quality graph. IMG brings-in useful information from the data and reflects it in the graph for improved performance. In addition to the content similarity, this graph model brings-in structural information in the final similarity graph. Further to this, the structural information can be utilised to avoid pairwise matching, and generate sparse matrices with better-separated clusters.

Example 6. Let's assume that $X \rightarrow \mathbb{R}^{5\times3}$ represents the gene expression measurements of five patients measured on three proteins e.g., (features). Firstly, the extended FD estimator will be applied on each feature vector (e.g., protein) separately to construct a set of buckets for that feature as shown in Fig (5.2). The type of vertices in IMG is analogous to the number of features in X. Secondly, there will be three types of vertices (as there are three proteins in X) in IMG, and the number of vertices (buckets) for each type of vertex will be determined by the extended FD estimator. Finally, the in-memory representation of this graph will require three 5×5 adjacency matrices, each for a single type of vertex.
The mathematical representation of this IMG graph is given below:

$$G = (V_t, E_{tt'}) \tag{5.9}$$

Where, V_t, denotes a set of finite vertices of type t, ($t = 1$ *to* 3 in this case), and $E_{tt'}$, represents a set of edges between the vertices of type t, t' where $(t \neq t')$. Note that for simplicity we do not assume any labels for edges or vertices. For each type of vertex V_t, where $t = 1, 2, 3$, we require a 5×5 adjacency matrix as shown below:

$$V_1 = \begin{bmatrix} 0 & & & & \\ e_{21} & 0 & & & \\ e_{31} & e_{32} & 0 & & \\ e_{41} & e_{42} & e_{43} & 0 & \\ e_{51} & e_{52} & e_{53} & e_{54} & 0 \end{bmatrix} \tag{5.10}$$

where each $e_{ij} \in E_t$, and denotes the presence or absence of an edge (link) between the i^{th}, and j^{th} patient in this adjacency matrix. Note that the diagonal elements are zero this is because of the absence of the *loop* (also called a self-loop or a buckle) between a vertex and itself. Similarly, the upper triangular values are empty this is because of the *undirected* property of the graph. Similarly, an adjacency matrix will be constructed for each feature vector e.g., V_2, V_3.

Here, each adjacency matrix e.g., V_t, denotes the structural connectivity of patients based on feature t. The mathematical representation of this matrix is given by $G_1 = (V, E)$. The structural similarity of patients' overall features can be represented through IMG as shown in Fig(5.3), and Eq (5.9). IMG is the combined (integrated) view of a set of t graphs e.g., G_1, G_2, G_3 in this particular example. Now, graph theoretical approaches can exploit the structural connectivity from IMG to find similarities between the patients on either single or multiple measurements (features or proteins). The proposed approach will rely on IMG in order to construct a robust, high-quality similarity graph.

5.2.4 Robust Similarity Graph Construction

The proposed research computes the distance between the pair of feature vectors e.g., x, y as the sum of the absolute differences of their corresponding buckets. With this definition the Manhattan distance becomes as follows:

$$d(x, y) = \sum_{i=1}^{n} \left| \frac{x_i}{w_i} - \frac{y_i}{w_i} \right| \tag{5.11}$$

where w is a width vector and each w_i consists of the width estimated on i^{th} dimension e.g., $w = \{w_i, i \in 1, ..., n\}$. The width is estimated using the Freedman Diaconis estimator (FD estimator). Therefore, dividing the variable x_i on i^{th} estimated width will generate the bucket number for x_i inclusion. In the case of highly skewed features and data variability, the estimated width might not generate accurate buckets. To solve this problem we need to divide each feature into defined intervals e.g., (Q_1, iqr, and Q_4), and then estimate the width of each interval independently. In this way, the extreme values e.g., (Q_1, Q_4) are separated from the maximum distribution e.g., (iqr) and hence the effect of data variability is minimized. Now we have three set of width vectors $W = [w^1, w^2, w^3]$, where each $w^k \in W$ and $W = \{w^k, k \in 1, 2, 3\}$. With this modification we have the following equation:

$$d(x, y) = \sum_{i=1}^{n} \left| \frac{x_i}{w_i^p} - \frac{y_i}{w_i^q} \right| \tag{5.12}$$

where, w_i^p, w_i^q denotes the estimated width of i^{th} feature of x, y on p^{th}, and q^{th} partition respectively. As each feature vector is divided into three defined intervals e.g., Q_1, iqr, and Q_4, and on each interval, the bucket width is estimated e.g. w^1, w^2, and w^3 therefore, $p, q \in 1, 2, 3$. We define a function $\phi(.)$, which assigns each data point to its respective bucket number. $d(x, y) =$

$$\sum_{i=1}^{n} \left| \phi\left(\frac{x_i^p - min(x^p)}{w_i^p}\right) - \phi\left(\frac{y_i^q - min(y^q)}{w_i^q}\right) \right| \tag{5.13}$$

$\phi(.)$ generates the respective bucket number for x_i, and y_i. The aim behind this approach is to group values in each feature into an estimated number of buckets and then compute the distance between the buckets. To achieve this we need the total number of buckets in each partition. If a width w, is estimated on a vector v, then the total number of buckets in v, is computed with the following equation:

$$\beta_N = \frac{max(v) - min(v)}{w} \tag{5.14}$$

where β_N, is the total number of buckets in v. To facilitate the process we do not need to explicitly group the elements into buckets, but we can compute the bucket number for any element whenever it is needed with the following equations.

$$\beta_i^p = \phi\left(\frac{x_i^p - min(x^p)}{w_i^p}\right) \tag{5.15}$$

$$\zeta_i^q = \phi\left(\frac{y_i^q - min(y^q)}{w_i^q}\right) \tag{5.16}$$

β_i^p, and ζ_i^q denotes the bucket number for the x_i, y_i feature in the p^{th}, q^{th} partition respectively. As the bucket numbers are always integers, therefore, the decimal value generated by the function $\phi(.)$ is rounded up to the nearest integer. The bucket numbers in each partition are in increasing order where, β_N denotes the maximum bucket number, which is equivalent to the last bucket number in a partition. Therefore, there could be nine possible scenarios which we cover with the following axioms:

When $x_i \in$ lower Quartile and $y_i \in$ upper Quartile

1. $d(x,y) = |\beta_i^1 - (\zeta_i^2 + \beta_N^1)| \iff x_i \in Q_1, y_i \in iqr$
2. $d(x,y) = |\beta_i^1 - (\zeta_i^3 + \beta_N^1 + \beta_N^2)| \iff x_i \in Q_1, y_i \in Q_4$
3. $d(x,y) = |\beta_i^2 - (\zeta_i^3 + \beta_N^2)| \iff x_i \in iqr, y_i \in Q_4$

When $x_i \in$ upper Quartile and $y_i \in$ lower Quartile

4. $d(x,y) = |(\beta_i^2 + \beta_N^1) - \zeta_i^1| \iff x_i \in iqr, y_i \in Q_1$
5. $d(x,y) = |(\beta_i^3 + \beta_N^1 + \beta_N^2) - \zeta_i^1| \iff x_i \in Q_4, y_i \in Q_1$
6. $d(x,y) = |(\beta_i^3 + \beta_N^2) - \zeta_i^2| \iff x_i \in Q_4, y_i \in iqr$

When both $x_i, y_i \in$ same Quartile

7. $d(x,y) = |\beta_i^1 - \zeta_i^1| \iff x_i, y_i \in Q_1$
8. $d(x,y) = |\beta_i^2 - \zeta_i^2| \iff x_i, y_i \in iqr$
9. $d(x,y) = |\beta_i^3 - \zeta_i^3| \iff x_i, y_i \in Q_4$

where β_N^1, β_N^2, and β_N^3 are the total number of buckets in the partitions defined by Q_1, iqr, and Q_4 respectively.

Now, by adding the bucket numbers from the lower quartile partitions the final distance metric becomes:

$$d(\beta,\zeta) = \sum_{i=1}^{n} \left| \left(\beta_i^p + \sum_{k=q}^{p-1} \beta_i^k \right) - \left(\zeta_i^q + \sum_{k=p}^{q-1} \zeta_i^k \right) \right| \tag{5.17}$$

The final ROMDEX function proposed in Eq. (5.17) covers all of the nine axioms. These axioms can broadly be categorised into three categories as follows:

1) when $x_i \in$ lower quartile and $y_i \in$ upper quartile. In this case, the sum of the number of buckets on each partition from p to $q-1$ is added to the y_i. Where, p, q denotes the partition number such that $x_i \in p$, and $y_i \in q$. Note that the summation becomes irrelevant if $q \leqslant p$.

2) when $x_i \in$ upper quartile and $y_i \in$ lower quartile. In this case, the sum of the number of buckets on each partition from q to $p-1$ is added to the x_i. Where, p, q denotes the partition number such that $x_i \in p$, and $y_i \in q$. Note that the summation becomes irrelevant if $p \leqslant q$.

3) when $x_i, y_i \in$ same partition. In this case, there is no need to add anything as both $x_i, y_i \in$ same partition. Note that in this case $p = q$, therefore both the summations on a maximum number of buckets become irrelevant.

The ROMDEX function proposed in Eq. (5.17) is robust to outliers. It synthesizes relevant features into buckets in such a way that minimizes the influence of outliers on the final computed distance. A critical part of the proposed approach is the bucket construction as shown in Eq.(5.15), and Eq.(5.16). The internal process of these equations is visually depicted in Fig. 5.2. In the figure, each feature vector is divided into partitions based on three defined intervals e.g., (Q_1, iqr, Q_4). In each partition, the elements are grouped into buckets based on their corresponding estimated bucket width. The process moves the highly influential elements to the extreme buckets. It can be seen that the relevant elements in each partition are nicely grouped into multiple buckets using Eq.(5.15), and Eq.(5.16). In Fig. (5.2), B_i denotes the bucket number which contains the i^{th} feature. The proposed distance function constructs these buckets and computes the distances between the buckets instead of individual elements, which is robust to outliers.

The distance matrix computed in Eq (5.17) is transformed into a similarity graph for clustering. The clustering algorithms take the similarity matrix of the graph and generate clusters based on the similarity of the vertices. A similarity matrix is defined as a symmetric matrix S, such that $S_{i,j} > 0$ represents the strength of similarity between patient i, and j. A standard clustering algorithm such as KNN is used to separate the group of patients on the graph. In order to generate a fully connected similarity graph, the ROMDEX function proposed in Eq. (5.17) is embedded into the Gaussian function as below.

$$S(\beta, \zeta) = \frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{\|\beta-\zeta\|^2}{2\sigma^2}\right)} \tag{5.18}$$

In Eq. (5.18), $\|\beta - \zeta\|$ is the ROMDEX function proposed in Eq. (5.17), and σ is the bandwidth of the kernel to associate local KNN graph structure. For multi-view datasets, which is usually the case with disease subtyping datasets, a separate similarity graph will be created for each type of view a.k.a, data type. For instance, if m different data types are given, then m similarity matrices for data types $v = 1, 2, ...m$, are constructed using Eq (5.18).

The individual similarity graphs are used for single-view analysis, clustering, and disease subtyping. However, for integrative analysis, these similarity graphs can be integrated into a single unified similarity graph using any graph integration approach such as ANF[78], or SNF[38].

5.3 Multi-view Clustering & Disease Subtyping

Disease subtyping typically involves preparing multi-view datasets that include information on the diseases being studied. After that the similarity graphs for each type of data (e.g., view) that shows the similarity between the patients based on their disease status are constructed. At this stage, the graph-based clustering algorithm is applied to the individual similarity graphs to group the diseases into distinct clusters. On the other hand, for integrated clustering, these similarity graphs need to be integrated into a single similarity graph prior to applying any clustering algorithm. Therefore, in this section, the integration of the constructed similarity graphs is explained. Furthermore, the working mechanism of the Spectral clustering algorithm on the integrated similarity graph is shown. Finally, the end-to-end algorithm which connects all the proposed techniques into a single framework for robust multi-view clustering and disease subtyping is proposed.

5.3.1 Spectral Clustering and Disease Subtyping

Disease subtyping is widely achieved with spectral clustering, which aims to identify sub-types of a disease on a patient similarity graph. In this thesis, the spectral clustering for disease subtyping is applied on the similarity graph G with normalised graph Laplacian L as below:

$$L^{norm} = I - D^{-1}AD^{-1} \tag{5.19}$$

where I is the identity matrix, and A is the adjacency matrix of the similarity graph G generated using Eq. (5.18). In the above Eq. (5.19), D denotes the diagonal matrix, where each diagonal element is the row sum of the adjacency matrix A computed using the following equation:

$$D_{ii} = \sum_{j} A_{ij} \tag{5.20}$$

Spectral clustering [102] algorithms aim to minimise RatioCut [103]. In terms of clustering, the goal is to minimise the capacity of a cut in a graph. Capacity defines the strength of edges between the vertices in two different partitions of the similarity graph G. For instance if V denotes a set of all vertices in G, and P, P', denotes two partitions of the vertices in V, where P', is the complement of P e.g., $P' = (V - P)$, and $P, P' \in V$. The total capacity of this partition is defined as the strength of the total number of edges that exist between these two sets. Therefore, an optimal cut is found through an objective function RatioCut, by solving the optimisation problem below:

$$\begin{aligned} &\min_{P \in R^{n \times C}} Trace(P^T L P) \\ &\text{s.t. } P^T P = I \end{aligned} \tag{5.21}$$

where, P, and L, denote the partition matrix, and normalised Laplacian matrix respectively.

In the following section, an end-to-end algorithm for the proposed disease subtyping approach is outlined. It connects all the techniques proposed and discussed in this thesis at a comprehensive level. The resulting framework enables robust single-view, integrated-view clustering, and disease subtyping in high-dimensional data.

5.3.2 End to End Algorithm

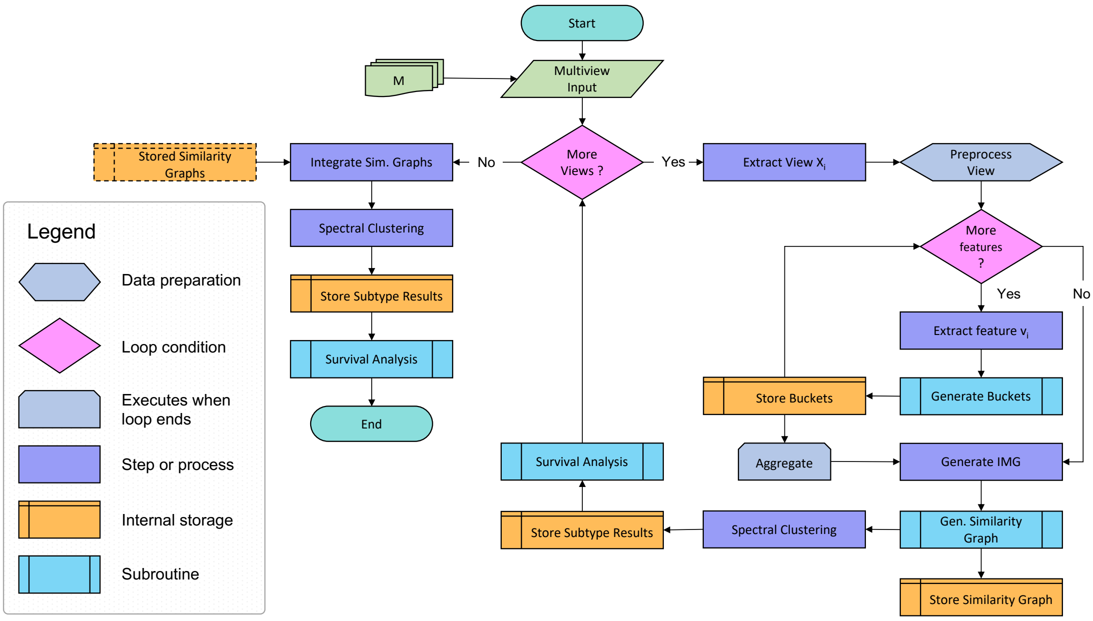


Figure 5.4: Flowchart of the methodology.

In this section first, a comprehensive flowchart is provided which visually depicts the process flow of the proposed methodology. Secondly, an end-to-end algorithm is provided which provides algorithmic and technical details to the proposed disease subtyping process. The flowchart illustrating the proposed methodology is presented in Figure (5.4). The chart is accompanied by a legend that provides an explanation of the different processing stages involved in the methodology, aiding in the comprehension of the workflow.

The technical details of the flowchart are shown in the algorithm (1). It requires a multi-view data set M, consisting of t views or modalities. Each view is represented by a matrix e.g., $X_i \in \mathbb{R}^{m \times n}$ e.g, m samples and n features, where i $\in \{1, 2, \ldots, t\}$. We will often use X without subscripts to denote any view in the multi-view dataset M. Similarly, x represents any row vector or sample in X where x_i^j represents the element at i^{th} row and j^{th} column of the view X. The multi-view dataset M can then be represented as a collection or set of t matrices, as follows: $M = \{X_1, X_2, \ldots, X_t\}$. Each X_i needs to be clustered into K subtypes. First, a robust similarity graph is constructed for each view e.g., X_i of the dataset X, which is clustered individually into K subsets using Spectral Clustering. After that, the similarity graphs for all views are integrated into a single graph, and Spectral Clustering is applied to the integrated similarity graph. The other requirements are the number of clusters e.g., K, and the Gaussian kernel parameter σ. This algorithm returns a set of similarity graphs e.g., L_G, and clustering labels e.g., L_C for each individual view X_i respectively. Additionally, it returns the integrated similarity graph G, and the clustering labels L_C, which are computed on the integrated graph G.

The algorithm (1), begins by initialising the L_G, and L_C to empty sets in line 3. In (line 5) each view is pre-processed according to the pre-processing steps explained in the results chapter. In (line 7) three variable-sized bucket widths are estimated for each dimension of the view x_i. This is achieved using the proposed extended FD estimator (see Eq. 5.7, 5.6, 5.8). In (line 8) a set of buckets are generated for each dimension using the estimated widths e.g., $w^{1,2,3}$. These buckets are then added to the list L_B. At the end of the loop (see line 6 to 10) the list L_B consists of all the buckets for view X_i generated on its individual dimensions.

The generated buckets are transformed and represented through the proposed intermediate multi-typed graph model (IMG). It is achieved in (line 11), where

G_X denotes the IMG graph of the view X_i. There will be n type of vertices in G_X as there are n dimensions to the X_i. Each vertex of the IMG denotes a bucket, while the vertex type represents one of the dimensions of X_i from which it is created (see Sec. 5.2.3, Example 6, Eq. 5.9).

Algorithm 1 Robust Multi-view Clustering for Disease Subtyping

```
Require: $M, K, \sigma$
Ensure: $L_G, L_C, G, C$
 1: $M = \{X_i, i \in 1, 2, .., t\}$, $X_i \in \mathbb{R}^{m \times n}$
 2: $K$, $\sigma$ are the clustering, and kernel parameters respectively.
 3: Initialise:  $L_G = L_C = [\,]$
 4: for view $X_i \in M$ in $M$ do
 5:     $X_i \leftarrow pre\_process(X_i)$
 6:     for vector $v \in X_i$ in $X_i$ do
 7:         $w^{1,2,3} \leftarrow xFD\_estimaor(v)$          ▷ (using Eq. 5.7, 5.6, 5.8)
 8:         $B_v \leftarrow generate\_buckets(w^{1,2,3}, v)$
 9:         $L_B \leftarrow add\_to\_list(B_v)$
10:     end for
11:     $G_X \leftarrow IMG(L_B)$                              ▷ (using Eq. 5.9)
12:     $D \leftarrow ROMDEX(vertices(G_X))$                  ▷ (using Eq. 5.17)
13:     $G_S \leftarrow Gaussian\_kernel(D, \sigma)$          ▷ (using Eq. 5.18)
14:     $L_G \leftarrow add\_to\_list(G_S)$
15:     # Single-view clustering
16:     $C_S \leftarrow SpectralClustering(G_S, K)$          ▷ (Sp. Clustering [102])
17:     $L_C \leftarrow add\_to\_list(C_S)$
18: end for
19: # Integrative-view clustering
20: $G \leftarrow network\_fusion(L_G)$                     ▷ (SNF [38])
21: $C \leftarrow SpectralClustering(G, K)$                  ▷ (Sp. Clustering [102])
```

For graph-based clustering, a similarity graph representing the data-point similarities is clustered into K desired clusters. This similarity graph plays a vital role which is often computed from the pairwise distance matrix of the data points. Therefore, the distance matrix is created using the proposed ROMDEX function (see Eq. 5.17). ROMDEX works on the vertices of the IMG graph constructed above. This is achieved at the line (line 12) of the algorithm (1). This distance matrix is then transformed into a similarity graph G_S using the Gaussian kernel (see Eq. 5.18), which is achieved in line (13). Here, G_S represents the similarity graph of the view X_i. Similarly, a G_S is created for each view X_i in each iteration and added to the list L_G for later integrative clustering. The similarity graph

G_S, for each view is clustered into K subsets using the spectral clustering (see line 16), and the cluster labels are added to the list L_C corresponding to the view X_i. This is used for the evaluation and validation of the clusters generated for the single view.

At the end of this loop (line 4 to 18), for the dataset D, a set of similarity graphs G_S, and a set of clustering labels L_C corresponding to each view X_i are generated. This means, there will be t, similarity graphs, and t set of clustering labels as there are total t views in the dataset D. The second part of this algorithm is the integrated view clustering of the dataset D. Therefore, the individual similarity graphs from L_G are integrated into a single similarity graph G, using similarity network fusion [38]. This is achieved in (line 20) of the algorithm. The integrated similarity graph G, fuses the similarity between the data points from each similarity graph G_S. The integrated similarity graph G, is then clustered into K subsets using the spectral clustering in (line 21) of this algorithm. Finally, the clustering labels computed for the G are stored in C for the evaluation and validation of the clusters generated for the integrated views.

5.4 Summary

High-dimensional omics data limits the ability of current bioinformatics approaches to analyse the data for a variety of reasons. Outliers, extreme values, and data variability are examples of such limitations, which limit the robustness of these approaches. Therefore, a novel robust approach for disease subtyping is proposed, which allows for robust clustering in the presence of outliers in high-dimensional data. On each view of omics data, the proposed approach performed binning. Existing statistical binning methods are best suited for small, normally distributed datasets. Also, these fail to take the extreme values and variability of the data into account. To avoid this, the proposed approach provided a statistical solution that extended the FD estimator but created variable size buckets along three defined intervals e.g., Q1, IQR, and Q4 that account for extreme values on either end. The rationale behind this division was to minimise the influence of outliers (e.g., data points on the extreme ends) by grouping them into relevant buckets. The constructed buckets were then combined to form the Intermediate Multi-typed Graph (IMG). These IMG were then used to generate a robust high-quality similarity graph for clustering in general and disease subtyping in particular. The structure-based data treatment allows for the creation of sparse matrices with well-separated clusters. In addition to content similarity, IMG captures structural information in the final similarity graph. A robust function is then defined on the vertices (buckets) of IMG to compute a pairwise distance matrix. This distance matrix is then embedded in a Gaussian kernel to generate a patient similarity graph for each view. In addition, similarity network fusion (SNF), is used to iteratively integrate all of the constructed similarity graphs. Finally, an end-to-end algorithm is proposed for Robust Multi-view Clustering and Disease Subtyping.

Chapter 6

Results and Evaluation

6.1 Revisiting the Research Objectives

The aim of this research was to investigate a robust statistical approach that could effectively identify disease subtypes in high-dimensional data with data variability and extreme values. In particular, the research focuses on four main objectives: firstly, the development of an Intermediate Graph Models (IMGs) to represent the topological graph structure of the data, that aids in identifying patterns and relationships within the dataset. Secondly, to develop a novel robust function (Romdex) utilising IMG to address the data variability and extreme values challenges in finding proximity between observations in high-dimensional spaces. Thirdly, to develop a robust disease subtyping approach based on Romdex for the accurate discovery of disease subtypes defined by clinical differences, such as survival. Finally, validation of the proposed approach on genomics, synthetic, and generic machine learning datasets.

6.2 Introduction

This chapter provides an overview of the various types of data sets used in this thesis for experiments. The nature of the datasets, their types, and their dimensions are extremely important when performing statistical analysis. Therefore, it also goes over the various dimensions and types of these data sets. The chapter then presents some statistical analysis of the data sets. This includes both descriptive statistics and inferential statistics. When performing descriptive statistics, it is important to understand the measures of central tendency and dispersion. Central tendency measures include the mean, median, and mode. Dispersion measures include the range, variance, and standard deviation. These have been shown through detailed box plots in the descriptive statistics section. These measures provide a good idea of what the data looks like and whether or not there are any outliers. Inferential statistics, on the other hand, allows making predictions about a population based on a sample. This includes correlation analysis between various measurements (features). In this section, correlation is measured from both aspects e.g., between continuous measurements and between continuous and categorical measurements. After that, the experimental procedure is presented. This includes setting up the hypothesis (null and alternate), the significance of the test, the laboratory setup for conducting experiments, and then evaluation metrics. These tests allow us to determine if the results are statistically significant or if they could have occurred by chance alone. Finally, the evaluation of the method on Omics data, synthetic data, and generic machine learning data is examined and compared with state-of-the-art in the field.

6.3 Datasets Information

We used a variety of real-world and synthetic datasets to evaluate and validate the proposed research. Five multi-view high-dimensional omics datasets taken for a specific cancer disease are among the real-world datasets. For further validation of the research, synthetic data was generated based on real-world omics data with and without extreme values. Finally, the robustness and stability of the proposed approach is evaluated by introducing various levels of noise in the clustering data.

6.3.1 Datasets Overview and Shape

The datasets that have been included in this thesis for evaluation and validation of the proposed work are all multi-view and high-dimensional. A multi-view dataset includes more than one view (a.k.a., data type), where each view measures a different aspect of the problem. Moreover, the data points across all views of a dataset are the same but the number and type of measurements across the views are different. Moreover, each view has the shape of $\mathbb{R}^{m \times n}$, with m, rows (a.k.a, observations, data points) which are the same across all the views, and n columns (a.k.a., measurements, features) which are different across the views. These matrices are then transformed into individual distance matrices of the shape $\mathbb{R}^{m \times m}$, denoting the pairwise distance between the data points. The distance matrix for each view is then transformed into another similarity matrix of the same shape e.g., $\mathbb{R}^{m \times m}$, denoting the pairwise similarities between the data points. Prior to applying the spectral clustering or finding subtypes of a disease, for each view, a data matrix of the shape $\mathbb{R}^{m \times n}$, a distance matrix of the shape $\mathbb{R}^{m \times m}$, and a similarity matrix of the shape $\mathbb{R}^{m \times m}$ is constructed. Finally, all the similarity graphs generated for each view of a dataset are integrated into a single similarity graph of a shape $\mathbb{R}^{m \times m}$.

6.3.2 Omics Data

The proposed approach is evaluated through extensive experiments on multiple genomics datasets. These datasets are selected for five cancer diseases from [28, 39], which are taken from TCGA [1].

The selected five datasets include Kidney Renal Clear Cell Carcinoma (KIRC), Glioblastoma Multiforme (GBM), Lung Squamous Cell Carcinoma (LUSC), Breast Invasive Carcinoma (BRCA), and Colon Adenocarcinoma (COAD). These are multi-view high-dimensional datasets, that consist of Gene Expression, DNA Methylation, and MicroRNA data. In addition, clinical data was also available with each dataset, which included the following information: survival data, age, gender, tumour status, tumour stage, and histological type.

Table (6.1) depicts the dimensionality of each dataset on each view. The table's

[1] https://www.cancer.gov/tcga

	LUSC	GBM	BRCA	KIRC	COAD
Expression	$110_r \times 12042_c$	$273_r \times 12042_c$	$172_r \times 20100_c$	$124_r \times 17974_c$	$146_r \times 17062_c$
Methylation	$110_r \times 23348_c$	$273_r \times 22833_c$	$172_r \times 22533_c$	$124_r \times 23165_c$	$146_r \times 24454_c$
MicroRNA	$110_r \times 706_c$	$273_r \times 534_c$	$172_r \times 718_c$	$124_r \times 590_c$	$146_r \times 710_c$

Table 6.1: Shape of each high-dimensional dataset.

row names represent the datasets, whereas the column names represent the views within each dataset. The dimensionality of the view is represented in *rows* × *columns* notations in each cell. As can be seen, the number of measurements (a.k.a, features) far outnumbers the number of observations (a.k.a, data points or samples). The comparative results generated on these datasets are provided in Table (6.2). The details and explanation for the generated results are provided in the corresponding evaluation section.

6.3.3 Synthetic Data

Synthetic data has been generated using the synthpop library. The data is generated based on the miRNA view of the GBM data.

The Synthpop package in R is used to produce the synthetic data [104]. This tool is used to create a fictitious version of real MicroRNA data. In addition to the MicroRNA data, clinical data such as survival, death, gender, and age were submitted for synthesis. The package includes a *survctree* method for synthesis survival time analysis. Because the default options were accepted, the data generation procedure was largely automated. The synthetically created MicroRNA data set included five MicroRNA variables and 110 clinical observations. Figure (6.1) depicts the characteristics of the synthetic data. The skewness coefficient for each feature is generated using a statistical library *moments*, which denotes the skewness direction and strength.

The results generated on the synthetic data are provided in Figure (6.10), and Figure (6.11). The details and explanation for the generated results are provided in the corresponding evaluation section.

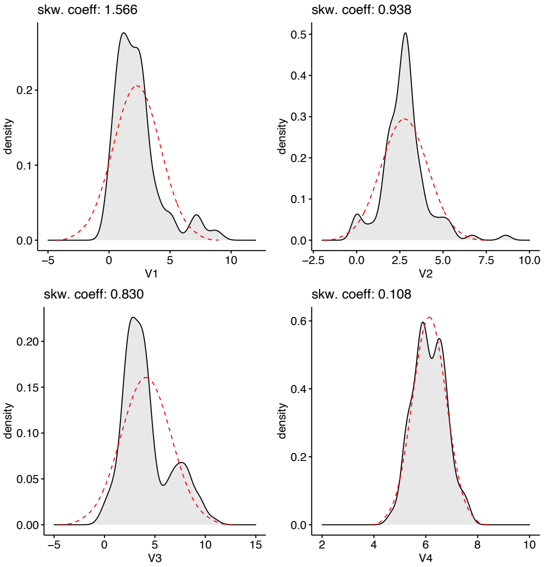


Figure 6.1: Characteristics of the synthetic data

6.4 Statistical Analysis and Visualisation

6.4.1 Descriptive Statistics

Descriptive statistics are used to summarise and describe interesting information about the data. They can be used to describe the distribution of data, and to calculate measures of central tendency and dispersion. Box plots are a type of descriptive statistic that can be used to visualise the distribution of data. It is also known as a box and whisker plot. Figure (6.2), shows a box plot for the selected few gene expression values from the GBM, and COAD cancer data. It consists of five elements: the minimum value, the first quartile, the median, the third quartile, and the maximum value. Each of these values corresponds to a point on the graph. The values on the x-axis denote the expression value, whereas the names on the y-axis are the gene names. The boxes inside the figure for each gene represent the interquartile range (IQR), which is the difference between the first and third quartiles. The horizontal lines inside the boxes denote the median value. Similarly, the lines extending from the boxes are called whiskers. They extend from the edges of the boxes to the minimum and maximum values. The purpose of this plot is to give a visual representation of how the gene expression values are distributed among the patients and across the datasets (e.g., GBM, COAD). It is helpful for comparing distributions of data sets. In addition, these plots visually represent the outliers and extreme values denoted as dots at the far ends of the whisker lines. Interestingly, from Figure (6.2), we can see that almost every gene has many extreme expression values, lying far outside the IQR range. As, in statistics, outliers are defined as observations that are more than $1.5 \times IQR$ below Q1 or more than $1.5 \times IQR$ above Q3. Since extreme values represent potential but unlikely outcomes, they are statistically and philosophically more fascinating. In these situations, any similarity metric based on central tendency measures such as mean, median or variability elements such as standard deviation, and variance is severely affected by these extreme values.

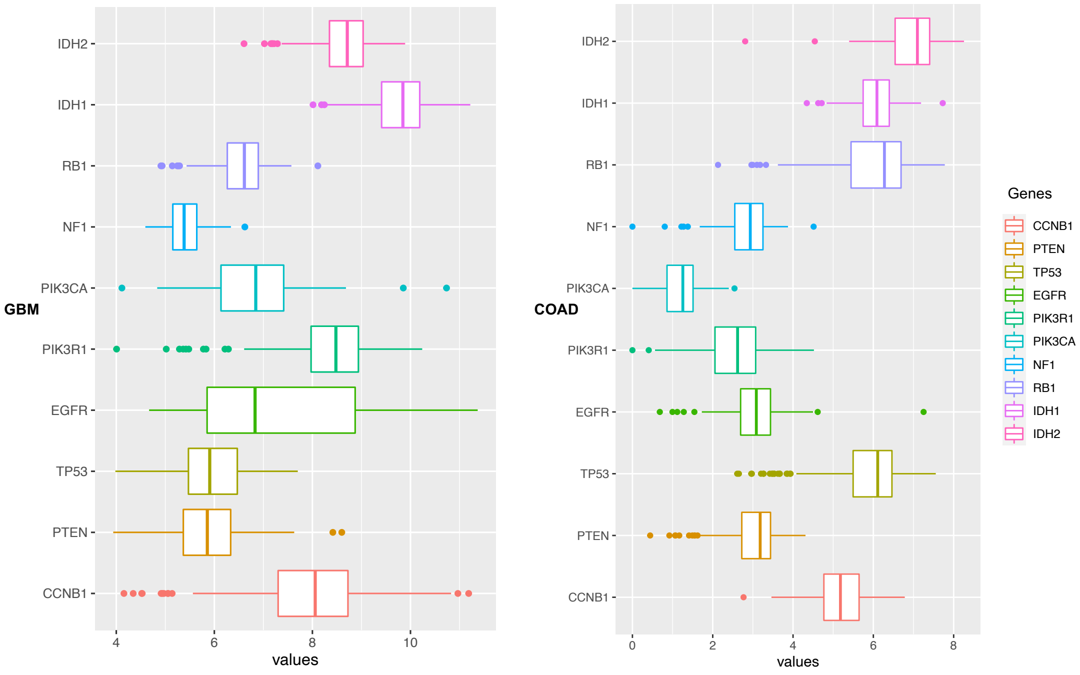


Figure 6.2: Box plot for selected genes.

In addition, while conducting statistical analysis it is important to validate the observed characteristics of the data through multiple approaches. Therefore, to ensure whether the values at the edge of whisker lines in Figure (6.2) are truly outliers? we performed some more statistical analysis as shown in Figure (6.3). This figure is generated using the gene expression values of CCNB1, and PIK3R1 of the COAD dataset. The plot shows an ordered squared Mahalanobis distance of the observations against the empirical distribution function. Figure (6.3) shows the outliers using four sub-plots. In the top-right plot, the chi-square p-value is plotted along with vertical lines analogous to the chi-square quantile (0.975), and adjusted quantile. Similarly, the plot at the top-left is the actual distribution. Likewise, the plot at the bottom-left shows the outliers detected by the chi-square distribution, while the plot at the bottom-right shows the outliers detected by adjusted quantile. The outliers in this figure are shown in red colour, where each value denotes the index of the outlier in the data.

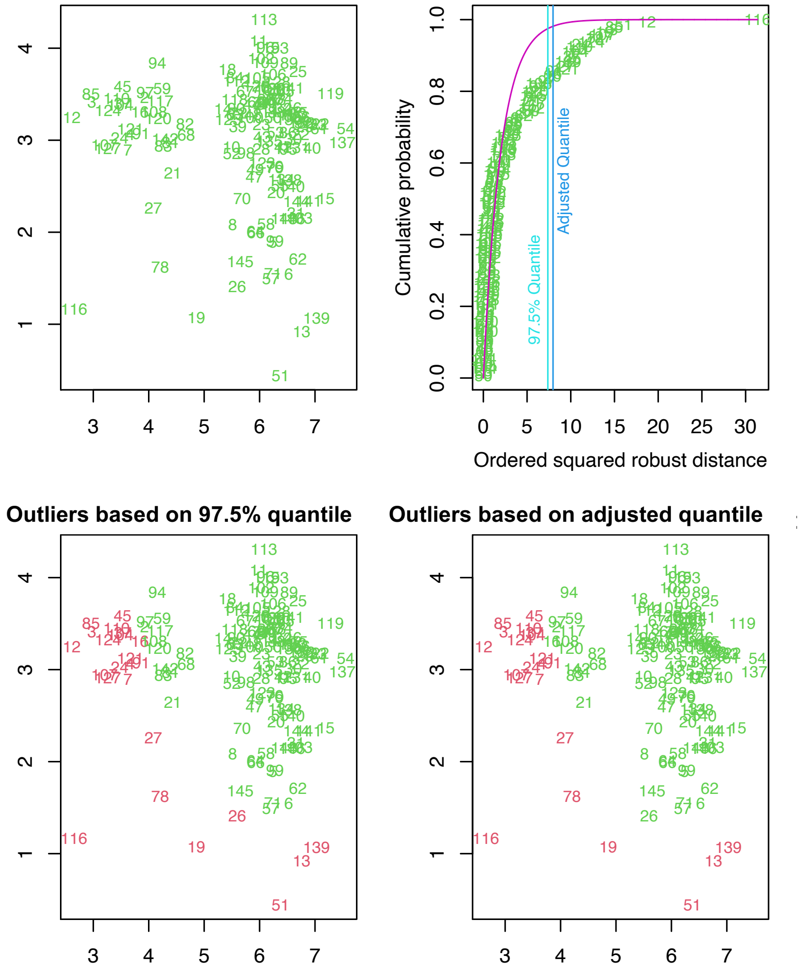


Figure 6.3: Extreme values and outliers

6.4.2 Inferential Statistics

Inferential statistics are often used in data science to test the correlation between the measurements. For instance, to understand the role of genes in disease, scientists often measure the expression of thousands of genes in patients' cells. This gene expression data is then used in inferential statistics to determine the correlation between genes.

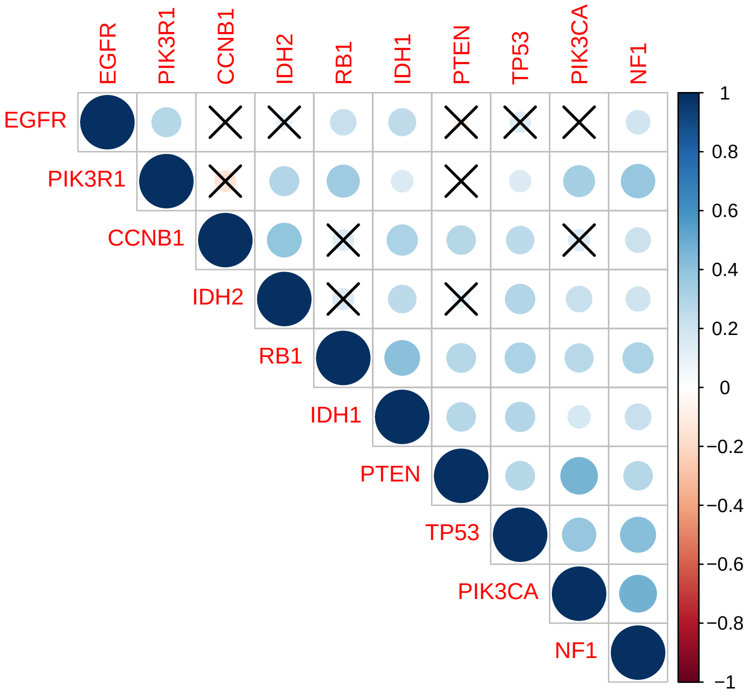


Figure 6.4: Statistically significant correlation between genes in GBM.

Correlation analysis is a statistical technique from inferential statistics used to examine the relationship between two measurements. The relationship between two measurements is typically represented by a linear equation, and the strength of that relationship is known as the correlation coefficient. Correlation coefficients can range from -1 to 1, with -1 indicating a perfect negative correlation and 1 indicating a perfect positive correlation. A correlation of 0 indicates that there is no relationship between the two variables. Correlation analysis can be used to determine whether there is a statistically significant relationship between two measurements. If the correlation coefficient is significantly different from 0, then we can conclude that there is a relationship between the two measurements.

Figure (6.4), shows the correlation between expression values of various genes taken from GBM data. The slider, in the figure, shows the range of correlation where the row and column names denote the genes. The size of the circles in cells denotes the strength of the correlation between the genes, it could either be a positive, or negative correlation. For instance, the pairs (CCNB1, IDH2), (PIK3R1, PIK3CA), (PIK3R1, NF1), (PTEN, PIK3CA), and (PIK3CA, NF1) are highly correlated. In addition, the cells with cross symbols denote that the relationship is insignificant. For instance, the correlation between the pairs (EGFR, CCNB1), and (EGFR, IDH2) is insignificant. The maximum circle sizes can be seen on the diagonal because it represents the self-correlation.

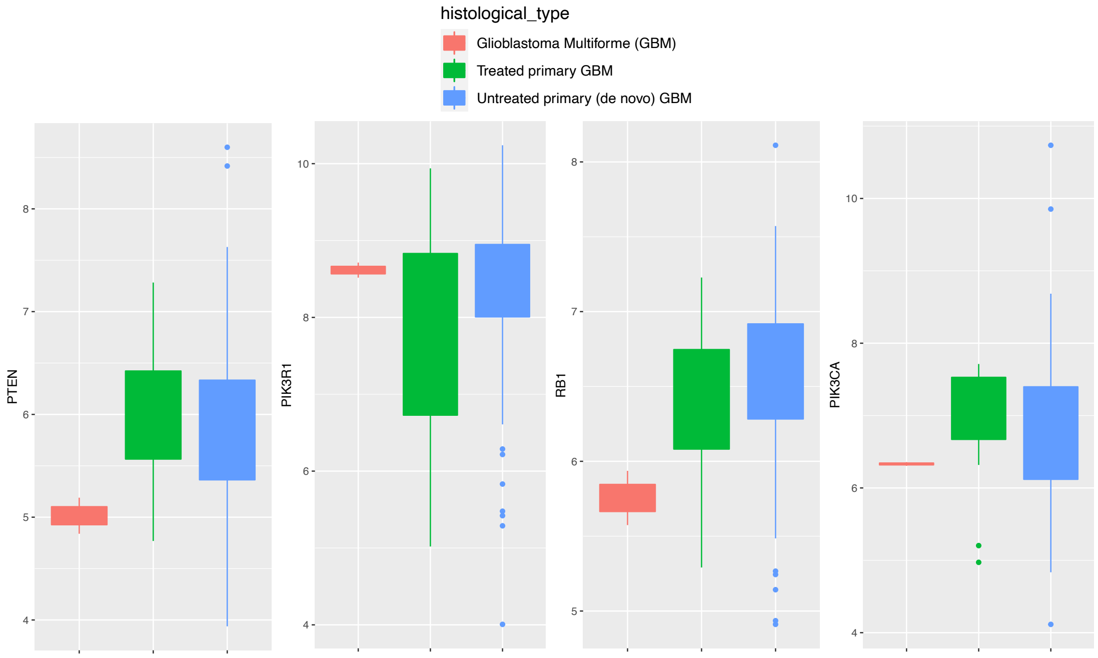


Figure 6.5: Distribution of gene expression values among histological types

In addition, more sophisticated statistical methods need to be used which take into account other factors that could be influencing the gene expression levels. For instance, it is statistically, interesting to see the correlation between a gene expression, and one or more categorical features such as histological type, tumour status, or tumour stage. It can provide a better understanding of influencing factors for expression levels.

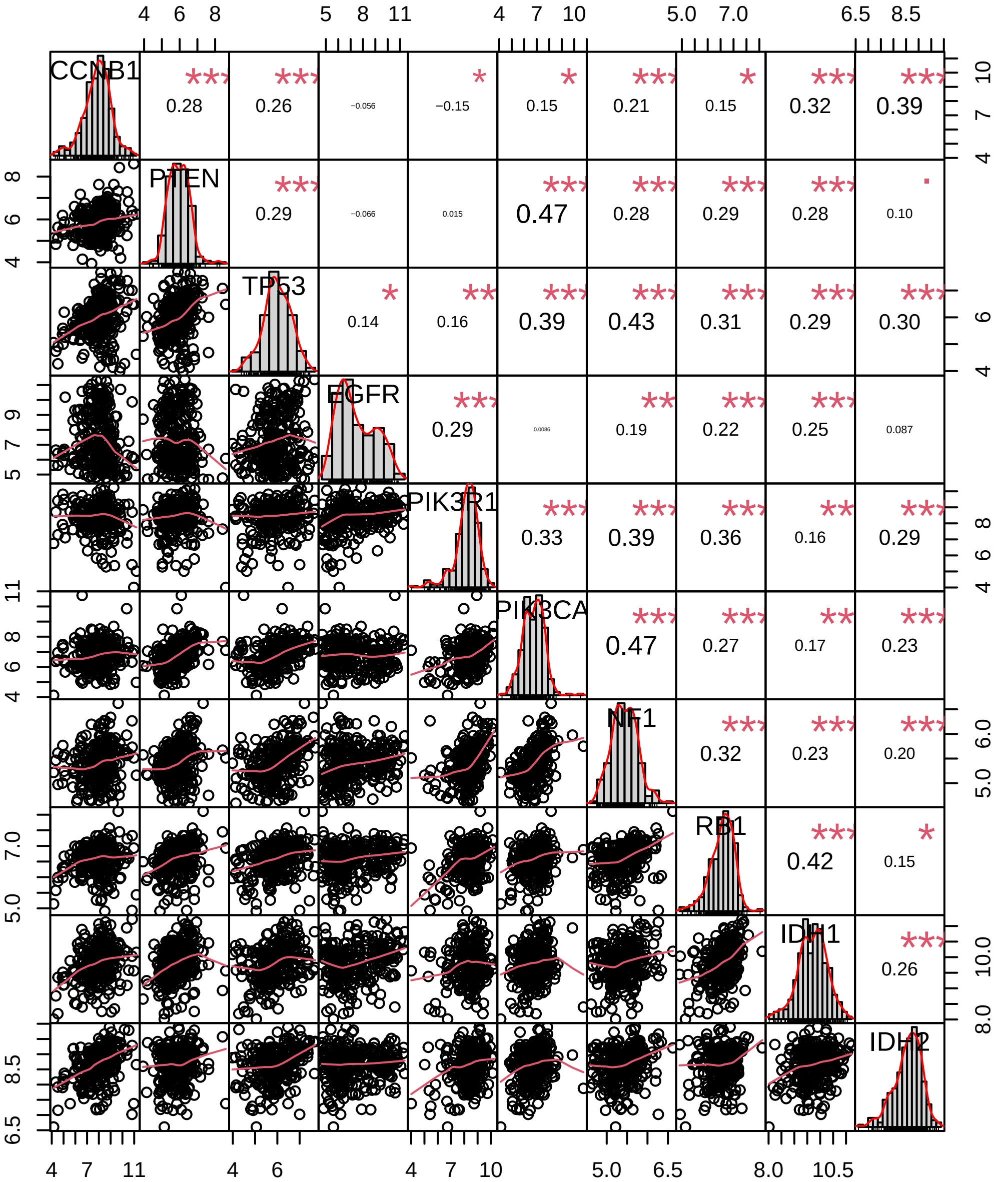


Figure 6.6: Performance analytics - inferential statistics.

In Figure (6.5), the expression values of various genes (known to be associated with GBM cancer) are plotted against the histological types e.g., glioblastoma multiforme, treated primary GBM, and untreated primary (de novo) GBM. The figure shows how the expression values are distributed among these histological types. From the figure, most of the extremely high or low expression values can be seen with the type untreated-primary (de novo) GBM.

For more detailed insights on correlation, a performance analytics plot is provided in Figure (6.6), showing various information about the pairwise correlation between the selected genes. In this figure, for each pair, we plotted a correlation line, histogram, and the significance of the correlation. Note that, the asterisk in each cell denotes the significance of the correlation, where the more the asterisk the stronger the correlation between the pair. The significance of the correlation between each pair is computed using the P-value, denoting the level of marginal significance within a statistical hypothesis test for the correlation. In this figure, the cells without any asterisks denote that the correlation between the pair is insignificant. Additionally, the values inside each cell denote the correlation value, and the sign of the coefficient indicates the direction of the correlation (e.g., negative, or positive).

6.5 Experimental Procedure

6.5.1 Hypothesis

Hypothesis testing is a method used in order to make decisions about whether or not to accept or reject a hypothesis. It is a statistical procedure used to assess whether or not a hypothesis is supported by given data. This is based on statistical evidence, and the decision that is made is either accepting or rejecting the null hypothesis. Therefore, to conduct a hypothesis test for the subtypes identified by the proposed approach from the omics data, first, we specify a null and then an alternative hypothesis as below:

Null Hypothesis (H_0): There is no significant evidence of a difference in survival times between the identified groups.

Alternate Hypothesis (H_a) : There is significant evidence of a difference in survival times between the groups.

The null hypothesis (H_0) is intriguing because it indicates that there is no significant difference in survival times between different groups of patients. If this is the case, then (H_0) will be accepted. This contradicts the assumptions made in alternate hypothesis (H_a) which is in favour of the research conducted in this thesis. (H_a) assumes that there is strong evidence in clinical data about the survival time differences between patients in different subtypes (groups). A significance test (see the section below) will be performed to determine whether there is significant evidence of the differences in survival times between the patients in different groups identified by the proposed approach. If this is the case, then (H_0) will be rejected and (H_a) will be accepted.

Test of Significance

We conducted a significance test, to see if it can provide enough evidence to reject the null hypothesis (H_0). The p-value has been used for testing the null hypothesis. It is a key concept in hypothesis testing. The smaller the p-value, the stronger the evidence against the null hypothesis.

There are two types of errors that can be made when performing hypothesis testing: Type I and Type II. A Type I error occurs when the null hypothesis is rejected when it should have been accepted. A Type II error occurs when the null hypothesis is accepted when it should have been rejected. The p-value helps to control for these errors by giving us a measure of how likely they are to occur. It is important to note that the p-value is not the same as the significance level (alpha). The significance level is the probability of rejecting the null hypothesis when it is actually true. The p-value is used to determine whether or not to reject the null hypothesis. If the p-value is less than the significance level (0.05), then the null hypothesis is rejected and the alternative hypothesis is accepted. If the p-value is greater than the significance level, then the null hypothesis cannot be rejected.

This experiment's evaluation metric was survival time analysis. Cox-proportional hazards (cox p-value) are used to estimate survival time [105]. It is commonly used to assess the outcomes of spectral clustering for disease subtyping. The Cox P-value is a popular statistical model for examining the association between a patient's survival time and one or more predictor variables. This model is used to compare the survival times of two groups of patients at the same time. For p-values less than 0.05, the null hypothesis that the cohorts have the same survival is rejected since there is significant evidence of a difference in survival times. As a result, the lowest values are favoured because they suggest better grouping.
To carry out the test of significance in cancer data, Cox-proportional hazards (Cox P-value) is used for survival time analysis. For more details on determining significance in survival data using p-values, the readers are referred to a recent study [106].

6.5.2 Experimental Setup

Experiment Pipeline

Each individual view of the cancer data is a $\mathbb{R}^{n \times m}$ matrix with m observations and n measurements. The number of measurements varies between views of the same data, but the observations remain consistent across all views. Following preprocessing, an $\mathbb{R}^{m \times m}$ similarity matrix representing the patients' similarity graph is constructed for each individual view using the proposed approach. Survival curves and the cox p-value are generated for the clustering results on each similarity graph. During the integration phase, all of the constructed similarity graphs for each dataset are combined into a single similarity graph using the SNF [38], and survival curves and cox p-value for the integrated view are generated. Table (6.2) in the evaluation section contains the p-value for each experiment. The best p-values obtained by disease subtyping approaches on any dataset are shown in bold font in the Table. Several survival plots were also created and are available in the evaluation section.

Furthermore, two steps of pre-processing are conducted before analysing the gene expression, methylation, and miRNA of five distinct cancers: To begin, any biological feature with more than 35% missing values across patients in any data type was discarded. The values in each data type are then normalised using the normalisation technique adopted in SNF [38].

Experimental Laboratory Setup

In addition, the experiments are conducted in the following laboratory setup:

Machine: iMac (Retina 4K, 21.5-inch)
Operating System: macOS Big Sur
Processor: 3.1 GHz Quad-Core Intel Core i5
Memory: 8 GB 1867 MHz DDR3
Storage: Macintosh HD 1 TB
Graphics: Intel Iris Pro
Programming Language - R
Programming Environment: R-Studio

Evaluation Metrics

The evaluation of results on omics data for disease subtyping is based on Kaplan-Meier survival time analysis, which is validated using statistical tests e.g., Cox-proportional hazard (Cox p-value). Furthermore, we added concordance statistics for the evaluation of the fitted survival model on five TCGA datasets. The concordance index (CI) is used to evaluate the predictive ability of the survival model. The CI values of the fitted survival model for all the datasets are impressive which demonstrates the predictive ability of the proposed unsupervised graph-based disease subtyping. In addition, the proposed method is also validated on generic machine learning datasets (vision datasets) and evaluated using normalised mutual information (NMI), clustering purity, and clustering accuracy.

6.6 Evaluation

In this section, the proposed approach is extensively evaluated using various types of datasets and evaluation metrics with state-of-the-art approaches. To begin, the following section evaluates the proposed approach on Omics (cancer datasets) using Cox-proportional hazards for survival analysis.

6.6.1 Evaluation on Genomics Data

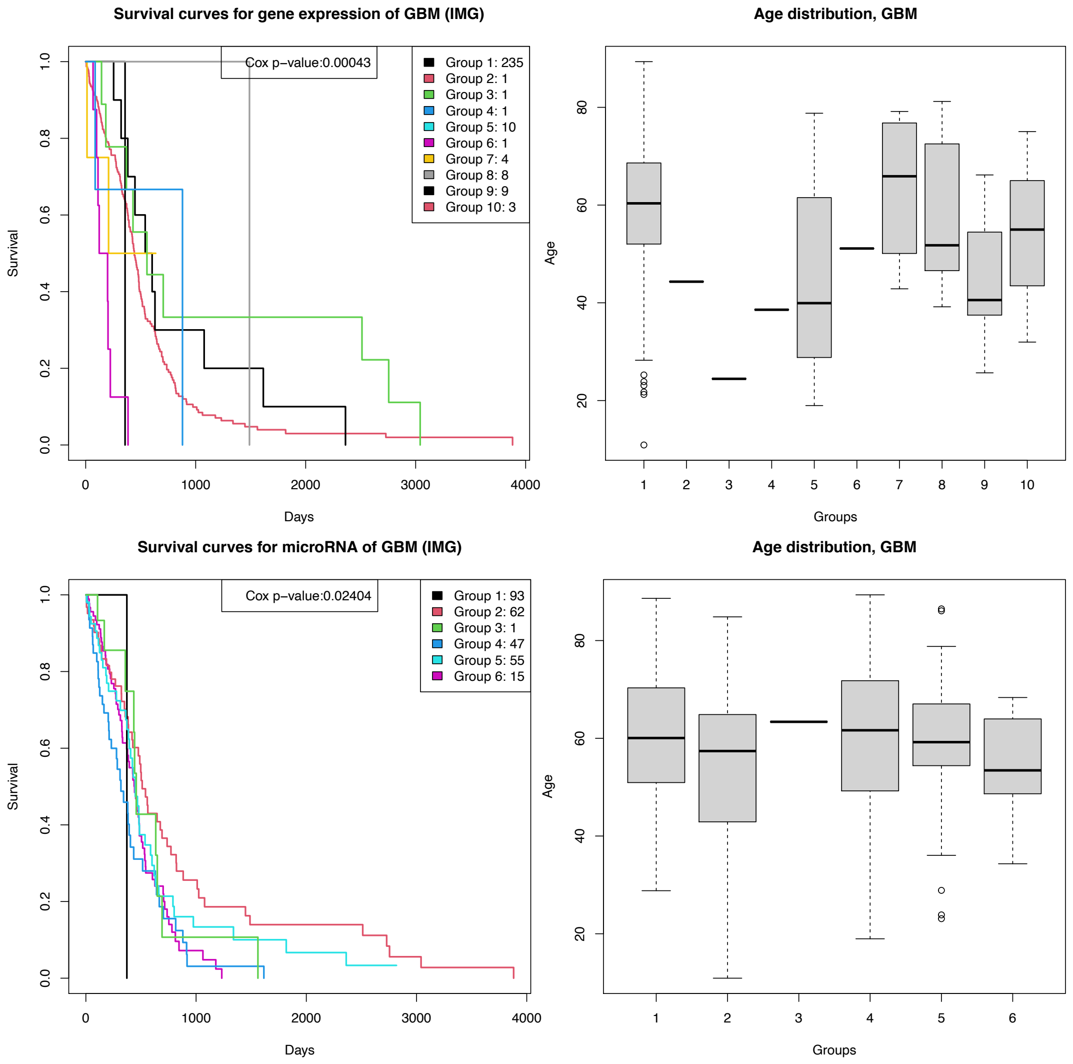


Figure 6.7: Kaplan-Meier survival curves the groups identified by the proposed approach in GBM.

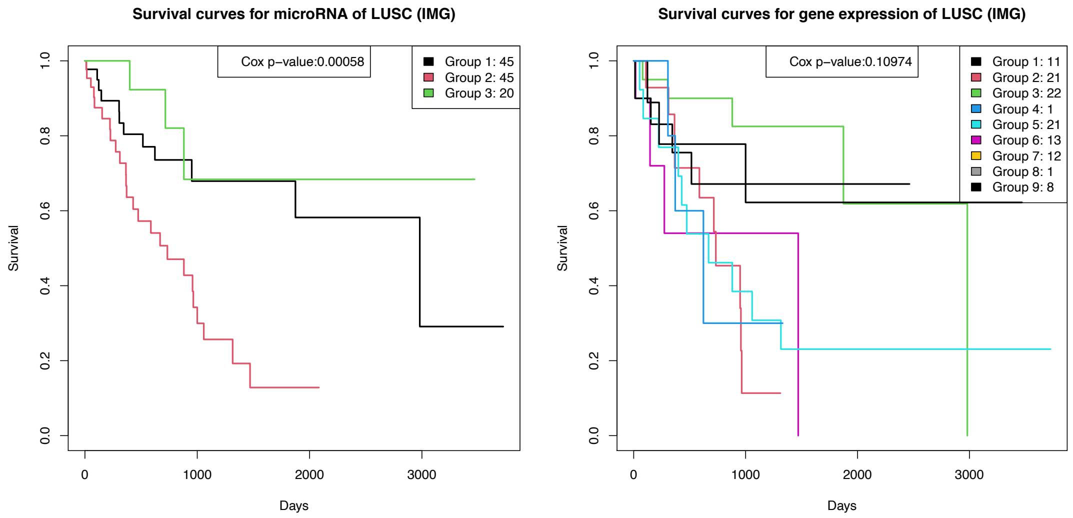


Figure 6.8: Kaplan-Meier survival curves the groups identified by the proposed approach in LUSC.

The proposed approach is tested on these datasets and compared against the state-of-the-art approaches for disease subtyping. We compared the results of the proposed approach with the state-of-the-art approaches in disease subtyping from [39]. These include, Perturbation Clustering (PINS) [39], Similarity Network Fusion (SNF) [38], Consensus Clustering (CC) [55], and iCluster+ [107]. In addition, we compared the proposed approach with MRGC, which is a robust graph-based disease subtyping approach [9].

The Kaplan-Meier survival curves for GBM patients on Gene Expression and MicroRNA views obtained by the proposed technique are displayed in Fig (6.8). There is a noticeable difference between the curves of different groups in this illustration. On the MicroRNA view, the curves for group 2 are dropping at a slower rate than the other groups, indicating a higher survival rate. All of the patients in this category are adults aged 43 to 65. Figure (6.9) shows the survival curves for GBM data using the PINS method. As can be observed, the proposed method for gene expression has a lower p-value than the PINS. Likewise, the survival curves for LUSC patients are depicted in Fig (6.8). The proposed technique found three groups on microRNA view in this image with a Cox p-value of 0.00058. In the figure, the survival curves for groups 1 and 3 are declining at a considerably slower rate than group 2, indicating that these two groups have

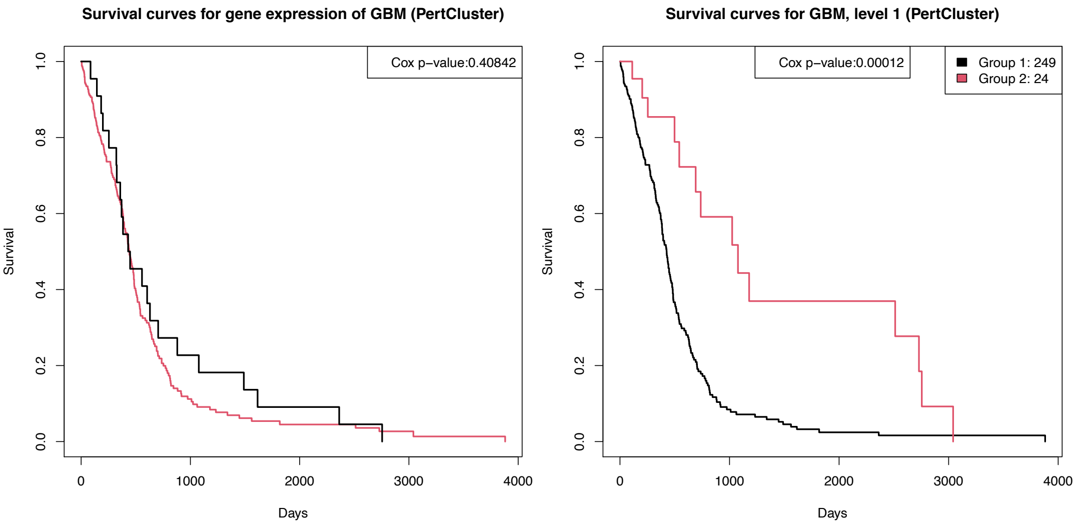


Figure 6.9: PINS - Kaplan-Meier survival curves for GBM patients on individual gene expression view and on integrated view.

Dataset	Views	ROMDEX	MRGC	PINS	CC	SNF	Icluster+
KIRC	Gene Expression	**0.00181**	0.003109	0.176	0.073	0.219	0.072
	DNA Methylation	**0.00171**	0.00323	0.111	0.128	0.577	0.14
	MicroRNA	**0.00758**	0.0084	0.138	0.509	0.138	NA
	Integration	0.00676	**8.5E-05**	0.00013	0.104	0.138	0.077
GBM	Gene Expression	**0.00043**	0.002106	0.408	0.281	0.992	0.056
	DNA Methylation	0.03004	0.00508	**0.0001**	0.001	0.017	0.003
	MicroRNA	**0.02404**	0.0406	0.086	0.526	0.401	0.09
	Integration	0.00307	0.0467	**0.000087**	0.039	0.062	0.076
LUSC	Gene Expression	0.10974	**0.00182**	0.125	0.782	0.095	0.588
	DNA Methylation	**0.0106**	0.0244	0.019	0.129	0.376	0.606
	MicroRNA	**0.00058**	0.0077	0.117	0.938	0.001	NA
	Integration	0.01084	**0.00608**	0.0097	0.794	0.428	0.36
BRCA	Gene Expression	**0.00692**	0.00766	0.902	0.114	0.969	0.101
	DNA Methylation	0.0698	**0.000012**	0.048	0.578	0.878	0.083
	MicroRNA	**0.04905**	0.07623	0.218	0.142	0.105	NA
	Integration	0.03963	**0.01496**	0.034	0.667	0.398	0.416
COAD	Gene Expression	0.05281	**0.00006**	0.113	0.048	0.148	0.29
	DNA Methylation	0.13944	0.0486	0.741	**0.034**	0.389	0.194
	MicroRNA	0.18799	**0.0175**	0.452	0.318	0.131	NA
	Integration	0.08898	**0.0602**	0.201	0.225	0.296	0.445

Table 6.2: The evaluation and comparison with baseline methods using cox p-value.

Concordance Statistics					
Datasets	KIRC	GBM	LUSC	BRCA	COAD
CI	0.7081	0.6599	0.6775	0.7361	0.8405

Table 6.3: The evaluation of the fitted survival model using the Concordance index (CI).

a higher survival rate. In this case, the p-value is 0.00058, which is significantly less than 0.05, and so the null hypothesis that the groups have the same hazard is rejected. The p-value of 0.00058 indicates that there is a significant difference in survival between the groups. Table (6.2), shows a comprehensive comparison of the proposed method to state-of-the-art disease subtyping methodologies on five TCGA multi-view cancer datasets.

As seen in the table (6.2), the proposed method outperformed the baseline approaches on the majority of the individual view data. Except for the GBM data, the MRGC algorithm performed the best on the integrated data. The PINS method, on the other hand, performed well on both the integrated and methylation views of the GBM data. Table (6.2) shows the most significant p-values with bold numerals. The proposed method consistently received good scores on the microRNA views.

On the KIRC dataset, the proposed method consistently produced the best p-value on all views, whereas MRGC earned the best p-value on the integrated view. On the BRCA dataset, the proposed method had the best p-value for gene expression and microRNA, whereas the MRGC took first place for DNA methylation. When compared to other techniques, the Consensus Clustering algorithm (CC) earned the best p-value on the Methylation view of the COAD dataset. The MRGC remained the closest rival for the proposed subtyping approach. Overall, the proposed method performed well in terms of the p-value when compared to the baseline approaches.

Furthermore, an assessment statistic, such as the concordance index (CI), is employed to test the fitted survival model's prediction capacity [108]. A decent predictive model has a CI value larger than 0.7. The CI values for the five TCGA

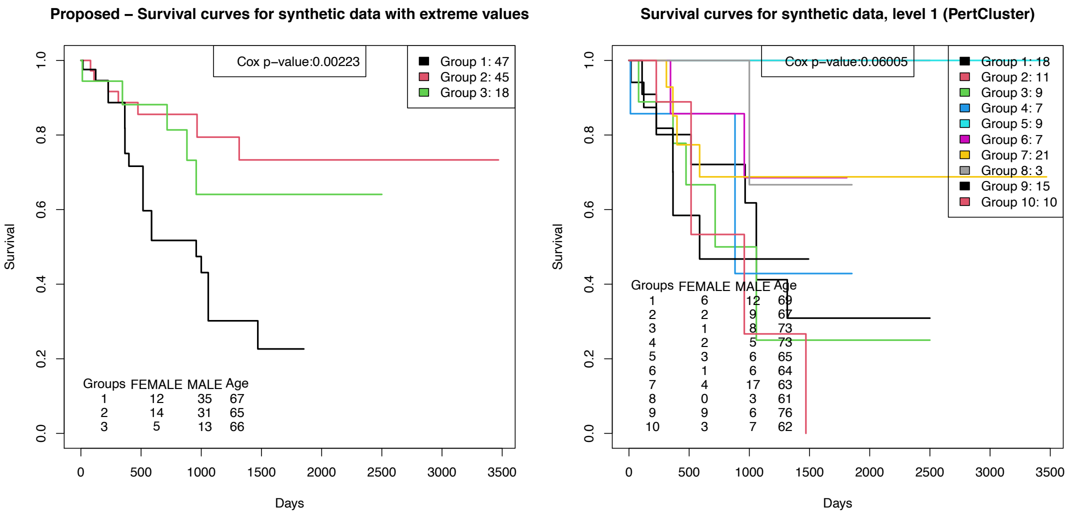


Figure 6.10: Proposed vs PINS - Results on synthetic data with extreme values

datasets are displayed in the table (6.3). Overall, the proposed method produced outstanding results, notably on the KIRC, BRCA, and COAD datasets. The research conducted by Terry et.al [108] is recommended for further information on the concordance index for survival models.

6.6.2 Validation of Clustering Performance on Synthetic Data

Based on the actual MicroRNA data, synthetic data with fever variables is developed to verify the robustness of the proposed approach. First, the experiment is conducted on synthetic data with a non-normal distribution and extreme values, as is common with real omics data. Second, the synthetically created data is transformed to map it to a roughly normal distribution. The outcomes are created in both scenarios and compared using various approaches.

Synthetic Data Generation and Transformation

The proposed approach is compared to state-of-the-art perturbation clustering (PINS) using this synthetic data (see Fig. 6.1) with extreme values. Figure 6.10 depicts the outcomes. Furthermore, we used several statistical transformations

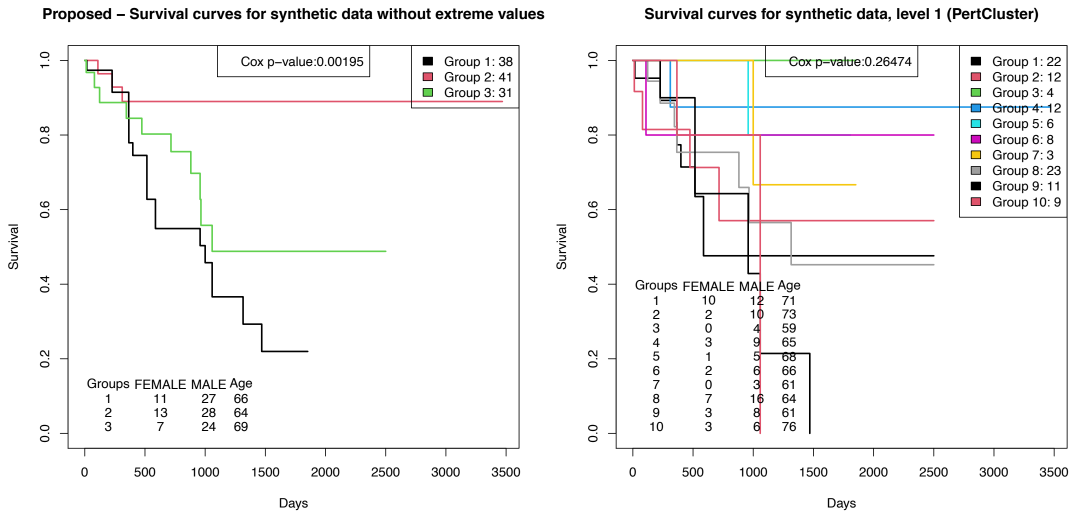


Figure 6.11: Proposed vs PINS - Results on the transformed synthetic data

to the generated data. The goal was to reduce skewness from the data and bring it closer to the normal distribution. As a result, we used square root and log transformations to approximate the normal distribution of the data. We transformed the data using typical heuristics based on the normality violation. Based on the normality violation, we applied the following transformation methods. Applied square-root transformation for variables V1, V3, and V4, since they were more positively skewed. We applied the log transformation for the variable V2. Following the transformation, the skewness coefficients were as follows: V1 (0.145), V2 (0.117), V3 (-0.201), and V4 (-0.0083). As we can see, the transformation reduced the skewness dramatically. Following these modifications, we assessed the proposed approach's performance on the cleaned synthetic dataset and compared it to existing state-of-the-art methodologies. The findings are depicted in Fig (6.11). The next section explains the results using synthetic data.

6.6.3 Results on Synthetic Data

The performance of the proposed approach is examined using both synthetic data with extreme values and synthetic data without extreme values. The findings for data with and without extreme values are displayed in Fig (6.10) and Fig (6.11), respectively. We chose the PINS method for the comparison and displayed the results side by side for the proposed and PINS, as shown in Fig (6.10) and (6.11). The proposed method attained a p-value of 0.0022 on the extreme values in Fig (6.10) and identified three patient groups with unique survival curves; however, the survival curves for the PINS were overlapping. Similarly, using data with no outliers, the proposed approach produced a much better p-value of 0.00195 than the PINS, as shown in Fig (6.11). The findings show that the suggested technique is least influenced by extreme values.

6.6.4 Robustness and Stability of Romdex Against Noise

To assess the stability of the proposed romdexClustering algorithm against data variability and extreme values, we conducted experiments by repeatedly adding Gaussian noise to the input data. Gaussian noise introduces random values with a zero-mean and a standard deviation (sd) determined by the noise level. To control the amount of noise, we sequentially increased the standard deviation from 1% to 25%. In order to evaluate stability, we employed several metrics: Accuracy, Kappa statistics, and the Rand index. Since these metrics require ground truth, we utilised the Iris dataset [99], a commonly used dataset in machine learning for clustering, which provides ground truth labels.

We performed two types of experiments to test the stability of romdexClustering. Firstly, after each iteration of adding noise, we computed the values of the evaluation metrics and recorded their values for each noise level. These values were plotted in Figure (6.12). Since the noise values are randomly generated with the same level of noise, multiple repetitions of the experiment yield different random values for the same noise level. To mitigate biases caused by these random values, we repeated the iterations of adding noise for each level of noise.

Subsequently, we computed the minimum, average, and maximum values of each metric achieved by romdexClustering across all iterations for each noise level. These values were then plotted in the following three Figures (6.13, 6.14, 6.15) corresponding to Accuracy, Kappa statistic and Rand index, which provide a

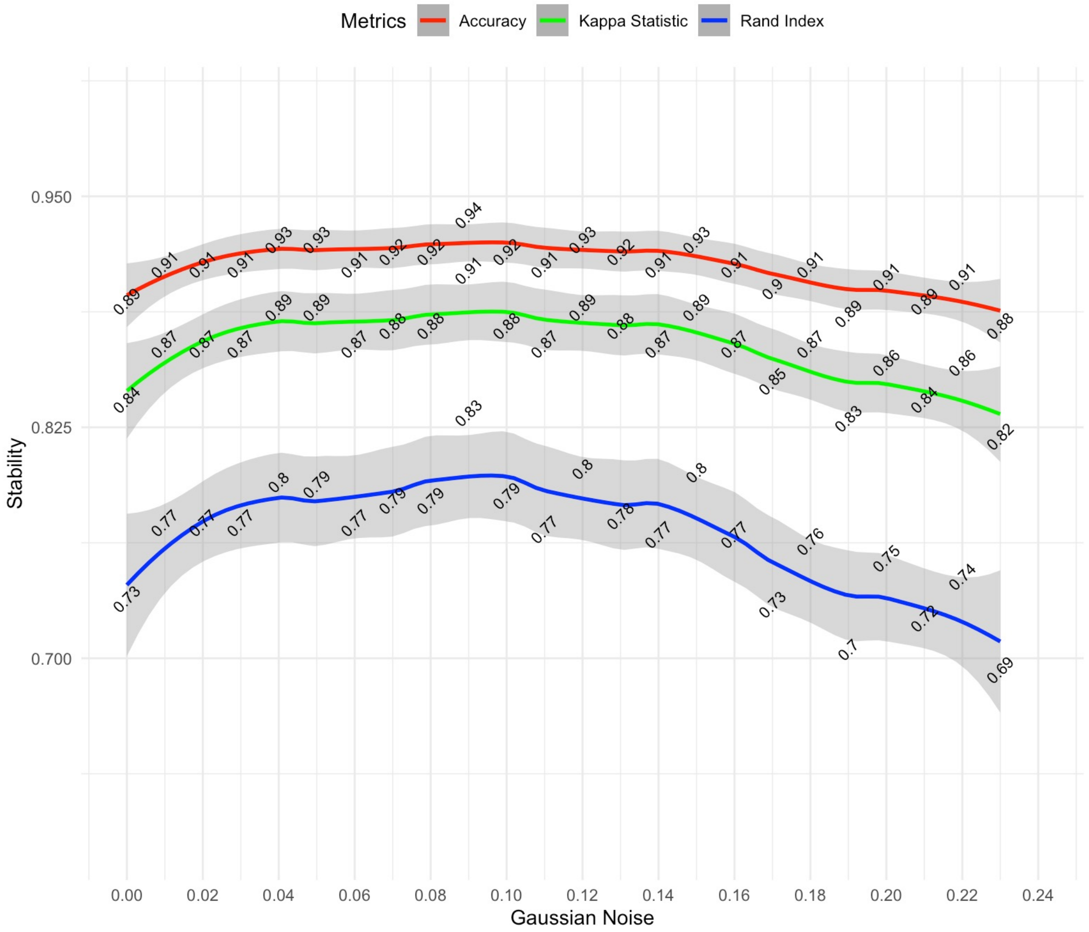


Figure 6.12: The stability of Romdex clustering against noise. Gaussian noise is added in each iteration in increasing order. The evaluation metrics, including Accuracy, Kappa statistic, and Rand index, are then computed. The X-axis represents the noise level, while the y-axis depicts the impact of each noise level on the evaluation metrics. This analysis provides insights into how Romdex performs under different levels of noise, allowing us to assess its stability.

comprehensive visualisation of the stability analysis.

Figure (6.12) illustrates three line plots representing the Accuracy, Kappa statistic, and Rand index. These results were generated using the following settings:

First, the Iris dataset was loaded into R-studio and assigned to a dataframe, excluding the ground truth column. The ground truth labels were stored in a variable called "true label".

Next, a loop was defined to iterate over Gaussian noise levels, ranging from 0% to 25% in increasing order. A noise level of 0% represents the dataset without any noise, and the loop increments the noise level by 1% in each iteration.

In Figure (6.12), the X-axis displays the noise levels as 0.01, 0.02, ..., 0.25, corresponding to 1%, 2%, ..., and 25% respectively. For each iteration, a specific level of Gaussian noise (determined by the iteration) was generated, maintaining the same shape as the dataset. This noise was then added to the dataset. The resulting noisy dataset was then clustered using the proposed romdexClustering algorithm.

The predicted clustering labels were compared to the ground truth labels using the Accuracy, Kappa statistic, and Rand index. The computed values were recorded for each noise level, and the loop moved to the next iteration. This process was repeated twenty-five times. Finally, the recorded values were plotted in Figure (6.12). The figure demonstrates the stability of the proposed clustering algorithm against noise. From the figure, it can be observed that the addition of increasing noise has minimal impact on the results. There is no significant decline in any of the metrics as the noise level increases. Instead, the lines representing the metrics remain relatively stable up to a noise level of sixteen percent, after which they show a slow and slight decline.

The lines in the figure have been smoothed using a smoothening function in R, but the actual values are also displayed on each point for a better understanding of the stability. It is noteworthy that the accuracy and kappa statistic exhibit only a minor decline compared to the Rand index.

Overall, the figure demonstrates the robustness of the romdex clustering algorithm against noise, indicating its stability in the presence of varying noise levels.

To ensure an unbiased evaluation of the proposed method in the presence of randomness, we conducted an extensive evaluation by iterating over the same noise level multiple times. To achieve this, we implemented an inner loop within the outer loop that iterates over the noise levels.

For each noise level defined by the outer loop, the inner loop iterates ten times, generating a noise sample for the same noise level in each iteration. Within each inner loop iteration, the generated noise is added to the dataset, and the resulting

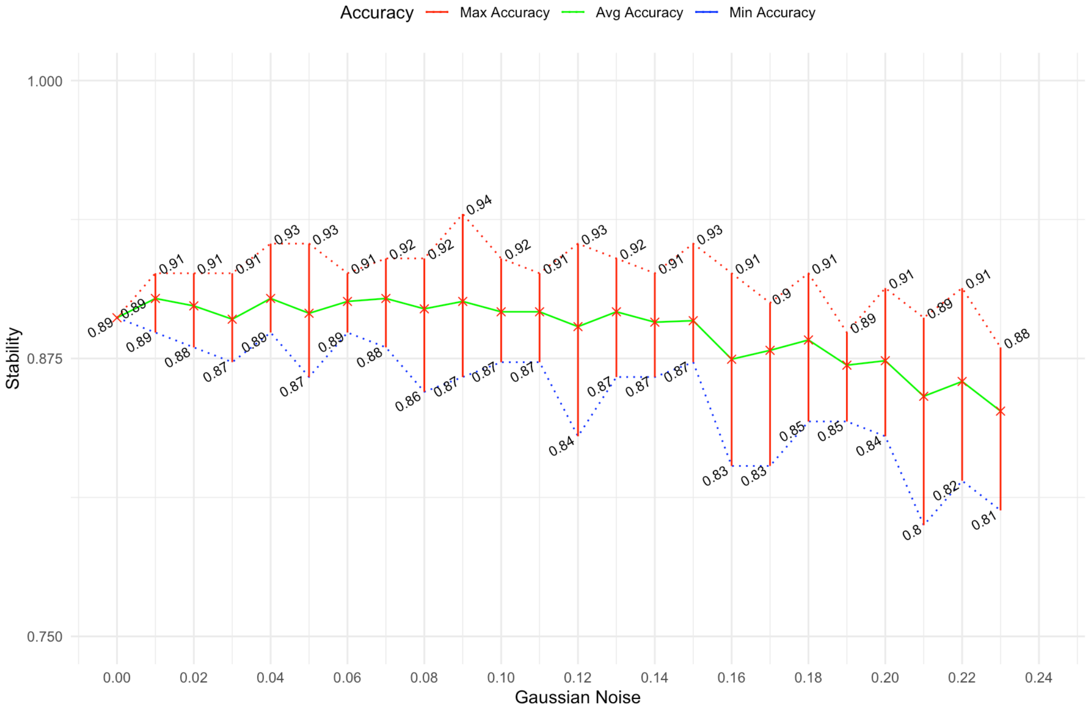


Figure 6.13: To mitigate biases, each noise level is repeatedly added multiple times within each iteration, and the accuracy is calculated for each instance of noise addition. Subsequently, the minimum, average, and maximum values of the accuracy metric are computed for each noise level over all iterations, and these values are displayed in the figure.

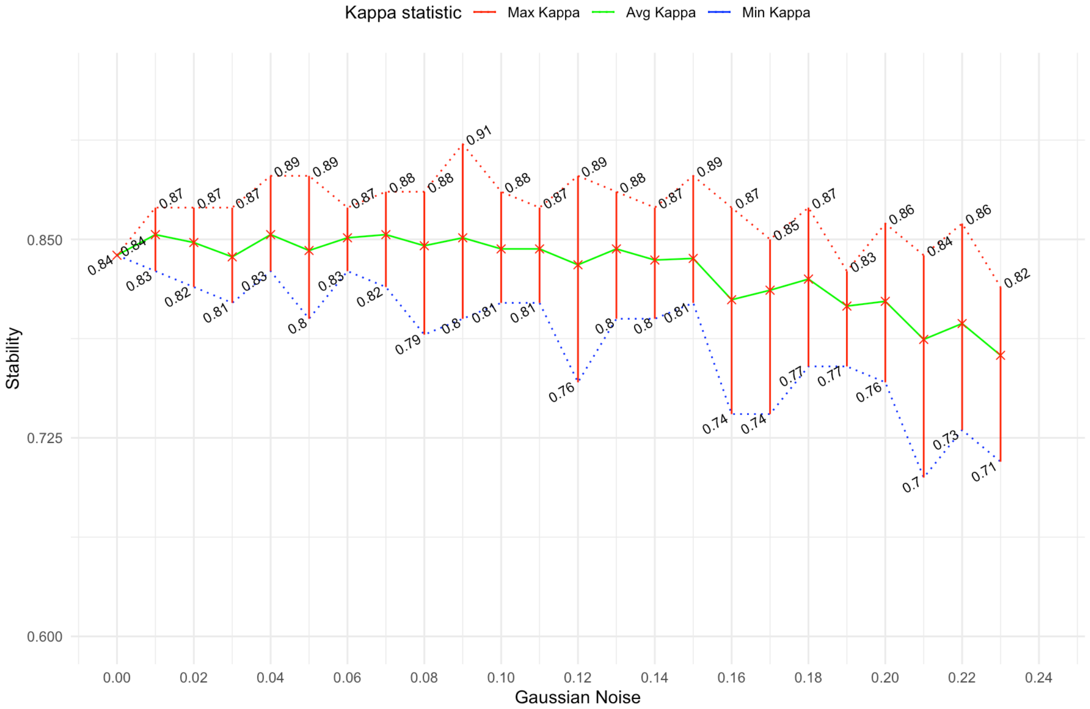


Figure 6.14: To mitigate biases, each noise level is repeatedly added multiple times within each iteration, and the Kappa statistic is calculated for each instance of noise addition. Subsequently, the minimum, average, and maximum values of the Kappa statistic are computed for each noise level over all iterations, and these values are displayed in the figure.

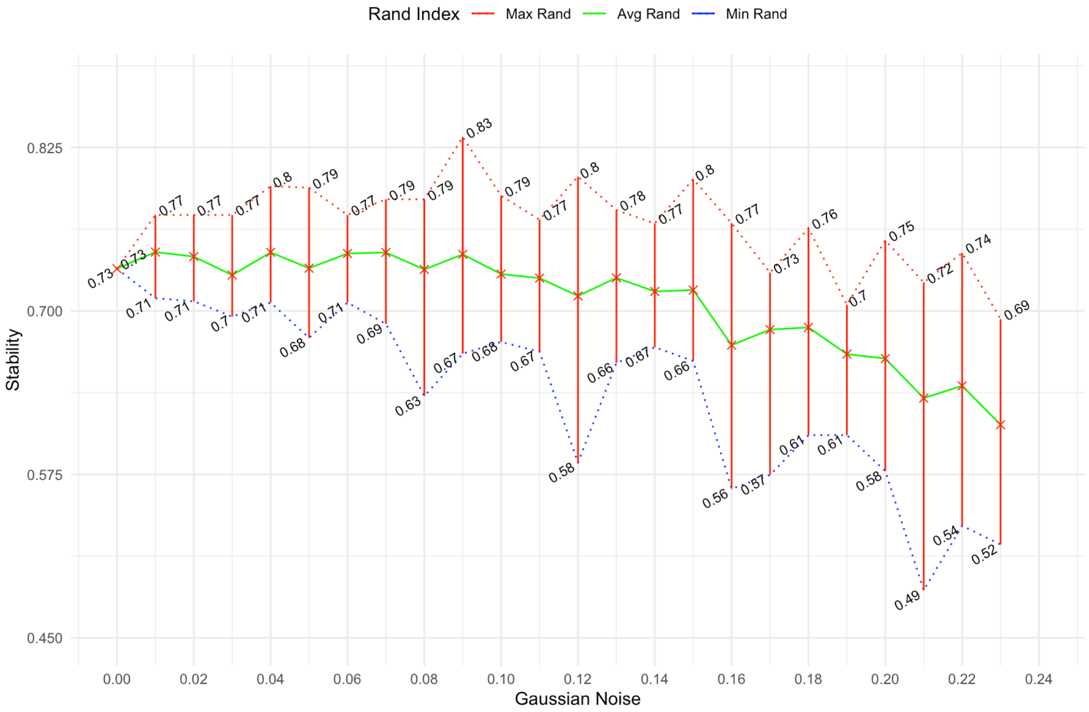


Figure 6.15: To mitigate biases, each noise level is repeatedly added multiple times within each iteration, and the Rand index is calculated for each instance of noise addition. Subsequently, the minimum, average, and maximum values of the Rand index are computed for each noise level over all iterations, and these values are displayed in the figure.

noisy dataset is clustered using the romdex clustering algorithm. The predicted labels are then compared against the ground truth labels using the evaluation metrics mentioned earlier. Therefore, for each noise level, we obtain ten sets of clustering results.

Additionally, after each iteration of the outer loop, we record the minimum, average, and maximum values of each metric across these ten clustering results to examine and visualise the range of results obtained. This provides a comprehensive analysis of the stability and performance of the romdex clustering algorithm under different noise levels.

In summary, the outer loop runs 25 times, and within each outer iteration, the inner loop iterates ten times, resulting in a total of 250 clustering results (ten clustering results for each level of Gaussian noise). These results are plotted in three separate figures: Figure (6.13) for Accuracy, Figure (6.14) for the Kappa statistic, and Figure (6.15) for the Rand index.

Each figure displays the maximum, average, and minimum values achieved for each point (corresponding to the noise level) across the ten clustering results. From these figures, we can observe that the romdex clustering algorithm exhibits robustness and stability against noise, as evidenced by the consistent average evaluation values achieved. However, it is worth noting that the Rand index values show comparatively more fluctuation between maximum and minimum values to other metrics. Nevertheless, overall, these results confirm the stability and robustness of the proposed approach against noise, as evidenced by the consistent average evaluation values achieved.

Chapter 7

Discussions

This chapter comprehensively discusses all of the key points related to the research findings raised during the peer-review process. In addition, It provides a brief conclusion that summarises the main research findings and discusses their implications for future research in this area.

7.1 A preferred distance function for clustering on high-dimensional data

Many studies have been conducted to investigate the impact of the curse of high dimensionality on clustering [109, 37]. The difficulty with clustering on such high dimensions is that data becomes significantly sparse in high-dimensional spaces [61]. As a result, the concept of similarity and distance, which are critical for clustering, loses qualitative significance. A careful investigation of the behaviour of distance functions (L_k norm) revealed that the meaningfulness is sensitive to the value k [67]. In comparison to the Euclidean distance (L_2 norm), the Manhattan distance (L_1 norm) is preferable for high-dimensional data [67]. Based on these findings, the proposed method for grouping high-dimensional omics data used the Manhattan distance (L_1 norm).

7.2 Robust statistical binning vs ranking based vs non-binning methods for clustering

Ranking-based methods usually obtain their ranks by assigning an ordered integer r_i to each value v_i of a sorted feature vector. For example, $r_i = 1$ only if v_i is the greatest value in the feature vector [21]. Rank-based approaches have limitations when it comes to taking essential information regarding data variability and the degree of proximity between values. Specifically, these approaches might not be able to provide a precise understanding of the extent to which one value is superior or inferior to another. As a result, a significant quantity of critical information included in the data is lost [21]. On the other hand, in the proposed approach, each feature vector is divided into three defined intervals (Q_1, IQR, Q_4), and the values inside each interval are classified into buckets depending on their corresponding estimated bucket width. This mechanism shifts the most influential values to the extreme buckets. These robust statistical binning approaches retain data variability and proximity information in the form of empty buckets, e.g., gaps. When the distance between two values is higher than the predicted bucket width, gaps arise. Furthermore, unlike ranking-based approaches, the bucket numbers are not always ordered consecutively, as gaps (empty buckets) may occur between distant values.

In addition, Table (6.2) compares the results of binning-based vs non-binning approaches. Similarity network fusion (SNF), in particular, is a non-binning-based strategy that is substantially connected to the proposed approach. The proposed approach differs from the SNF in that the Euclidean distance (used in the SNF) is replaced with the ROMDEX (used in the proposed). The results displayed in Table (6.2) show that the ROMDEX (using robust binning) beat the non-binning based (existing) approaches on numerous datasets.

7.3 Effect of data complexity on clustering performance

The complexity of data, such as high-dimensionality versus fever observation, influences the clustering balance. This is frequently the case with gene expression data. These are high-dimensional data types with fewer observations. These properties generate issues for both supervised ML models (which frequently leads to model overfitting) and clustering methods that employ the similarity graph to put observations into coherent groupings. The intricacy of the data-type in many circumstances leads to a sparse patient similarity graph, making it difficult for clustering algorithms to group some of the patients. As demonstrated by experiments using synthetic data with fever dimensions in Fig (6.10, 6.11), the proposed approach identified better patient groups with distinct survival.

As a summary, genomics data introduces obstacles such as the curse of dimensionality, and extreme values. Extreme values can distort distance calculations by significantly influencing overall similarity and dissimilarity measures. Consequently, clustering algorithms may exhibit bias towards these extreme values, potentially resulting in suboptimal clustering outcomes. Similarly, as dimensionality increases, the majority of distances between points tend to become similar, making it difficult for clustering algorithms to accurately group data points based on similarity [61]. Therefore, the proposed research suggested a robust statistical approach that involves grouping measurements into buckets prior to applying any distance function. This approach mitigates the impact of extreme values on the distance functions. Subsequently, the problem is modelled to enable the effective computation of distances between data points based on the constructed buckets using the Manhattan distance, which is more resilient to dimensionality compared to the Euclidean distance.

7.4 Qualitative comparison of the existing disease subtyping approaches

There are several recent methodologies that have demonstrated robustness in grouping omics data that are closely related. MRGC is one such approach, with notable results on omics and generic machine learning data [9]. On many heterogeneous actual and synthetic data, ROMDEX produces equivalent excellent clustering results. ROMDEX, on the other hand, produces more consistent and stable outcomes. Similarly, the PINS method finds stable clusters by varying the amount of noise introduced [39]. However, the clustering stability in PINS comes at the expense of high computing complexity and power. ROMDEX, on the other hand, produced comparable findings using a robust statistical technique. In S2GC, on the other hand, a supervised machine learning approach is used, which produced an optimised similarity graph and showed promising results for subtyping [11]. However, S2GC requires human intervention to provide class labels, unlike ROMDEX and other unsupervised techniques.

7.5 Graph Contrastive Learning for Clusteirng

Graph Contrastive Learning (GCL) is a prevalent self-supervised learning approach for graph-structured data. GCL methods rely on augmentation schemes for learning invariant representations across different views [110][111]. Existing approaches for multi-view contrastive learning primarily concentrate on either multiple graphs or multi-view attributes. Therefore, a generic framework called multi-view contrastive graph clustering (MCGC), which aims to cluster multi-view attributed graph data by learning a consensus graph is proposed [112]. The backbone of GCL is contrastive learning, wherein graph samples are contrasted to push similar samples together and dissimilar samples apart in an embedded space [113]. Contrastive Learning uses graph neural networks (GNN) to learn low-dimensional embeddings [114]. In recent years, it has garnered increasing popularity due to its ability to facilitate efficient training of neural networks when confronted with limited labelled data.

In contrast, the proposed approach in this thesis utilises spectral clustering which inherently performs dimensionality reduction by leveraging the eigenvectors associated with the smallest eigenvalues of the affinity matrix [16] [115]. This approach facilitates the effective representation of high-dimensional data in a lower-dimensional space, proving advantageous when working with genetics data that encompass small samples and a large number of features.

7.6 Performance advantages of the proposed research

The proposed linear approach could not outperform the non-linear CGGA model in terms of accuracy. The proposed approach, on the other hand, has its own performance advantages, such as its rapid, transparent, and explainable procedure, which is crucial in critical areas such as healthcare and finance. Furthermore, the proposed approach produces more consistent and stable outcomes. Clustering stability is achieved by a robust statistical approach that is quick and requires little processing power. The results on synthetic data further show that the proposed approach is least influenced by extreme values and data variability. Furthermore, the proposed method has the advantage of being simpler and easier to understand, which is more applicable to particular applications.

7.6.1 Generalisation of the Proposed Approach

Generic Machine Learning Vision Data

Finally, we have also included three datasets from computer vision, these datasets are related to object recognition and classification problems. These include 1)

Caltech101-7, 2) COIL20, and 3) Handwritten Digits.

The results generated on the object recognition and classification data are provided in Figure (7.1). The details and explanation for the generated results are provided in the corresponding evaluation section.

Results on object recognition and classification data

In the above section, the proposed approach was evaluated on genomics and synthetic datasets for disease sub-typing. As most of the disease sub-typing datasets lack gold standards, therefore, the cox p-value remains the most acceptable comparison metric, which has been widely adopted in comparative analysis in the field. In addition to the cox p-value, concordance statistics were performed to test the predictive ability of the proposed model.

Now, in order to test the generalisability of the proposed approach in other domains it has been evaluated on object recognition and classification data as stated in the earlier sections. We performed experiments on the vision (generic ML) datasets. As the vision dataset contained the gold standards, therefore, we computed additional metrics such as NMI, clustering purity, and clustering accuracy. We computed clustering accuracy to compare the proposed approach with existing clustering approaches on generic ML datasets. For comparison, we picked Spectral Clustering, SNF, and PINS. The comparison results are evaluated based on clustering accuracy. The source code for the proposed approach is made available on GitHub: https://github.com/bit-whacker/romdex

The results are shown in Figure (7.1) below. Before the comparison, the following basic preprocessing steps were applied which are essential for preparing the dataset for clustering.

1. Each image in each category of the ML datasets is reshaped to a fixed $width \times height$ dimensions d.

2. To extract a single channel, each image is transformed to grayscale.

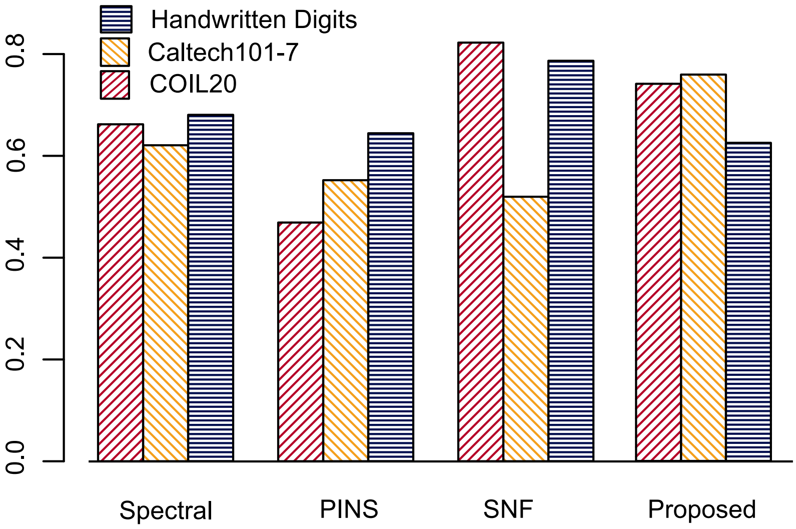


Figure 7.1: Accuracy comparison on object recognition and classification data.

3. Image data is extracted from each grayscale image to produce a d-dimensional numerical matrix.

4. The matrices are then flattened into row vectors.

5. Afterwards, all of the images from each category combined to form a single matrix with n × m dimensions, where n is the number of all images and m is equal to d (fixed *width* × *height* of the image).

6. Additionally, we added a column vector with the label (category) for each image.

As can be seen from the figure, overall the proposed approach achieved a decent clustering performance on all datasets. On the handwritten digits, Caltech, and COIL20 datasets it achieved 74%, 76%, and 62% accuracy respectively. Likewise, the proposed approach achieved the best position on the Handwritten digits and Caltech101-07 dataset. For more details please refer to Figure (7.1).

In addition, the clustering performance of the proposed approach is further evaluated through NMI, and clustering purity. These results are shown in Figure (7.2) below. Overall, the NMI, and clustering purity values on all three datasets (handwritten digits, caltech101-07, and coil20) show a decent performance. Which

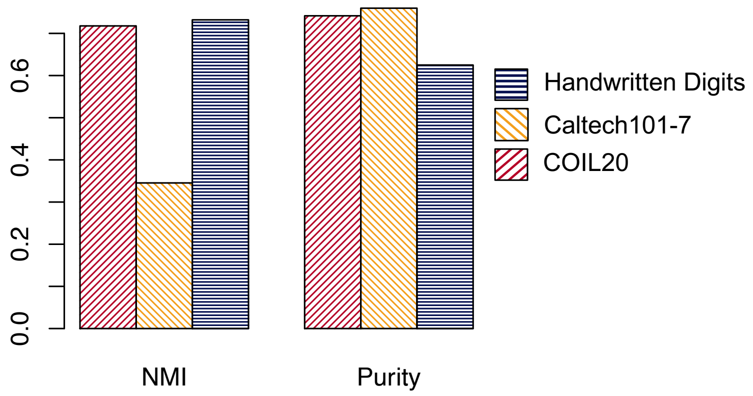


Figure 7.2: Performance evaluation of the proposed approach in terms of NMI and Clustering purity.

shows, the generalisability of the proposed approach toward other domains.

7.7 Limitations of the proposed approach and the best context in which to apply it.

Despite the best performance of the proposed approach in high-dimensional datasets with extreme values and variability, has certain limitations. Firstly, the approach may not perform optimally in datasets, where normal or nearly normal distribution and minimal data variability are present. Other models assuming normal distribution might yield better accuracy in such contexts. Secondly, the proposed approach's reliance on its internal mechanism can lead to relatively lower accuracy when applied to datasets that do not exhibit the specific characteristics it is designed to handle.

On the other hand, the proposed approach is best suited for specific contexts where its strengths can be leveraged. It excels when dealing with datasets that contain extreme values and high variability, the proposed approach's ability to handle such characteristics makes it an ideal choice. Moreover, the proposed approach is well-suited for datasets that exhibit complex patterns, non-linear relationships, and diverse data distributions, as it can capture and model these intricacies effectively.

In summary, while the proposed approach has limitations as above, it shines in high-dimensional datasets with extreme values and variability, as well as in complex and diverse datasets.

Chapter 8

Conclusion and Future Works

8.1 Conclusion

Graph-based approaches have the potential to improve overall clustering performance. These approaches exploit the underlying relationships and therefore find meaningful clusters in data. It is feasible to acquire a better knowledge of the structure inside the data and reveal better insights from previously hidden relationships by exploiting graph algorithms. Graph-based approaches for clustering have the potential to uncover hidden value in huge datasets that would otherwise remain hidden.

Another, critical factor in clustering is the notion of similarity for categorising and organising noisy knowledge. Similarity kernels are commonly used to compute it. The distance function utilised in similarity kernels, on the other hand, is not robust to extreme values, and data variability. Extreme values and data variability can have a significant impact on the distance function used in clustering algorithms. The most common distance metric is the Euclidean distance, which is simply the straight-line distance between two points. However, extreme values greatly deviate from the rest of the data than other points, causing them to have a disproportionate influence on the calculated distances. This can cause problems with clustering algorithms that are sensitive to outliers. Consequently, they impact disease subtyping by making it more difficult to identify risk factors and develop appropriate treatments.

All of these issues underscore the importance of carefully considering extreme values when conducting research on disease subtypes. Failure to do so could lead to inaccurate conclusions and potentially harmful consequences for patients. Data binning can be a helpful tool for handling them. This is because extreme values often skew results when using traditional methods, such as mean and median. With data binning, it can group data into bins, or categories, which makes it easier to see patterns and trends. This can be especially helpful when there are a lot of extreme values in the data set.

8.1.1 Research Findings

After extensive experiments, we come up with the following research findings. As the proposed approach assumes high-dimensional data with extreme values and data variability. Therefore, the proposed mechanism starts pushing the extreme values (extreme minimums, and maximums) to the far ends of the distribution. Moreover, to make the proposed approach least affected by the data variability it starts bucketing the data with variable bucket widths. Afterwards, it computes the distances between the buckets using ROMDEX (distance metric based on Manhattan distance) which is preferred for high-dimensional data compared to the L_2 norm functions as the distance in high-dimensional spaces is sensitive to the value L_K norm. Now, as the approach is explicitly proposed for the datasets with these characteristics, therefore, it works best in these scenarios as can be seen from the results section. However, if the datasets lack these characteristics then due to the internal mechanism of the proposed approach it achieves comparatively lower accuracy than the models that assume normal distribution which provide strong fitting ability. Therefore, the final takeaway is that the proposed approach is recommended for the datasets with the above-mentioned characteristics.

8.2 Future works

8.2.1 Injectable Probabilistic Graph Integration towards Explainable AI

The amount of information and data generated in healthcare is increasing exponentially. Such a large volume of data creates the need for better integration of clinical information. Therefore, in medicine a cognitive method of investigation e.g., clinical oncology, for instance, starts from a generic approach and moves towards more specialised examination by injecting evidence and information from these relevant sources until they reach a conclusion supported by facts. Therefore, in the future, the same idea will be adopted to inject context information into the graph for improvement results. A holistic view of the data leads to a better understanding of the data sets that are being analysed and can lead to more granular and insightful research. Therefore, in the future, the aim is to propose an approach that provides injectable evidence for the improvements in the integration which is a step forward towards explainable AI. The integrative approach can help to better understand the disease and its progression, identify new therapeutic targets, and improve patient care. In addition to the omics data, the proposed work will integrate clinical information and inject context into the patient graphs for better and explainable subtyping.

8.2.2 Graph Probabilistic Dependencies for Multi-view Data Integration

Wide-range of studies investigates data integration techniques for the provision of valuable insights on several business functions such as disease-subtyping, entity resolution, and clustering. Real-world datasets often exhibit intrinsic structure with possible relations between them, therefore, the integration of these datasets using graph theoretical approaches shows improved results compared to their non-graph theoretical counterparts. In graph-theoretical approaches, the individual source data is first transformed into a graph and then all these graphs are integrated to construct a single integrated target graph. The challenges exist both at the integration level and the graph construction level.

The challenges at the integration level are those that alter the state of data integrity on the integrated graph. These challenges arise from the noise and high dimensionality of the datasets, which increases the likelihood of redundancies. Consequently, this leads to poor quality of the integrated graph. The challenges at the integration level significantly reduce the accuracy of the final application. In the future, these challenges will be addressed by integrating large-scale datasets using novel graph theoretical approaches. Particularly, these challenges will be addressed with novel graph dependencies (GDs) called graph probabilistic dependencies (GPDs). GPDs provide probabilistic explanations for the issues that alter the state of data integrity on graphs.

8.2.3 Improvement Directions to the Proposed Approach

Whilst, graph-based approaches have a great potential to improve overall clustering performance, there are remaining challenges to be addressed. Some of the challenges have been addressed in this thesis by presenting a robust approach that led to considerable improvements in disease subtyping. There is space for improvement, such as determining the optimal *sigma* value for constructing similarity graphs, which is estimated empirically in this work over a few runs of the programme. As a result, in the future, an optimisation technique is required to determine the optimal *sigma* value. In addition, graph theoretical ways to compute distances on the IMG in topological space are recommended.

As the research contributed to the construction of robust similarity graphs from high-dimensional source datasets. These robust similarity graphs can be used with deep learning models as an input for clustering. Therefore, tuning the proposed approach to work the best on datasets that require non-linear fitting is another future direction to be considered.

Bibliography

[1] Z. Dlamini, F. Z. Francies, R. Hull, and R. Marima, "Artificial intelligence (ai) and big data in cancer and precision oncology," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 2300–2311, 2020.

[2] Z. Ahmed, K. Mohamed, S. Zeeshan, and X. Dong, "Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine," *Database*, vol. 2020, 2020.

[3] S.-I. Lee, S. Celik, B. A. Logsdon, S. M. Lundberg, T. J. Martins, V. G. Oehler, E. H. Estey, C. P. Miller, S. Chien, J. Dai *et al.*, "A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia," *Nature communications*, vol. 9, no. 1, pp. 1–13, 2018.

[4] G. Fatemifar, R. Lumbers, D. I. Swerdlow, and S. Denaxas, "Discovering and validating disease subtypes for heart failure using unsupervised machine learning methods," *Circulation*, vol. 136, no. suppl_1, pp. A15 862–A15 862, 2017.

[5] K. M. Boehm, P. Khosravi, R. Vanguri, J. Gao, and S. P. Shah, "Harnessing multimodal data integration to advance precision oncology," *Nature Reviews Cancer*, vol. 22, no. 2, pp. 114–126, 2022.

[6] K. M. Boehm, E. A. Aherne, L. Ellenson, I. Nikolovski, M. Alghamdi, I. Vázquez-García, D. Zamarin, K. Long Roche, Y. Liu, D. Patel *et al.*, "Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer," *Nature cancer*, vol. 3, no. 6, pp. 723–733, 2022.

[7] R. S. Vanguri, J. Luo, A. T. Aukerman, J. V. Egger, C. J. Fong, N. Horvat, A. Pagano, J. d. A. B. Araujo-Filho, L. Geneslaw, H. Rizvi *et al.*, "Multimodal integration of radiology, pathology and genomics for prediction of response to pd-(l) 1 blockade in patients with non-small cell lung cancer," *Nature cancer*, vol. 3, no. 10, pp. 1151–1164, 2022.

[8] C. Liang, M. Shang, and J. Luo, "Cancer subtype identification by consensus guided graph autoencoders," *Bioinformatics*, vol. 37, no. 24, pp. 4779–4786, 2021.

[9] X. Shi, C. Liang, and H. Wang, "Multiview robust graph-based clustering for cancer subtype identification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2022.

[10] R. A. Neff, M. Wang, S. Vatansever, L. Guo, C. Ming, Q. Wang, E. Wang, E. Horgusluoglu-Moloch, W.-m. Song, A. Li *et al.*, "Molecular subtyping of alzheimer's disease using rna sequencing data reveals novel mechanisms and targets," *Science advances*, vol. 7, no. 2, p. eabb5398, 2021.

[11] C. Liu, C. Wenming, S. Wu, W. Shen, D. Jiang, Z. Yu, and H. San Wong, “Supervised graph clustering for cancer subtyping based on survival analysis and integration of multi-omic tumor data,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2020.

[12] V. Gligorijević and N. Pržulj, “Methods for biological data integration: perspectives and challenges,” *Journal of the Royal Society Interface*, vol. 12, no. 112, p. 20150571, 2015.

[13] C. R. John, D. Watson, M. R. Barnes, C. Pitzalis, and M. J. Lewis, “Spectrum: fast density-aware spectral clustering for single and multi-omic data,” *Bioinformatics*, vol. 36, no. 4, pp. 1159–1166, 2020.

[14] S. Li, L. Jiang, J. Tang, N. Gao, and F. Guo, “Kernel fusion method for detecting cancer subtypes via selecting relevant expression data,” *Frontiers in genetics*, p. 979, 2020.

[15] M. Sinkala, N. Mulder, and D. Martin, “Machine learning and network analyses reveal disease subtypes of pancreatic cancer and their molecular characteristics,” *Scientific reports*, vol. 10, no. 1, pp. 1–14, 2020.

[16] U. Von Luxburg, “A tutorial on spectral clustering,” *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[17] M.-F. Balcan, A. Blum, and N. Srebro, “A theory of learning with similarity functions,” *Machine Learning*, vol. 72, no. 1, pp. 89–112, 2008.

[18] B. Schölkopf, K. Tsuda, and J.-P. Vert, *Kernel methods in computational biology.* MIT press, 2004.

[19] A. S. Shirkhorshidi, S. Aghabozorgi, and T. Y. Wah, “A comparison study on similarity and dissimilarity measures in clustering continuous data,” *PloS one*, vol. 10, no. 12, p. e0144059, 2015.

[20] R. Xu and D. Wunsch, “Survey of clustering algorithms,” *IEEE Transactions on neural networks*, vol. 16, no. 3, pp. 645–678, 2005.

[21] D. Jiang, C. Tang, and A. Zhang, “Cluster analysis for gene expression data: a survey,” *IEEE Transactions on knowledge and data engineering*, vol. 16, no. 11, pp. 1370–1386, 2004.

[22] J. Mao and A. K. Jain, “A self-organizing network for hyperellipsoidal clustering (hec),” *Ieee transactions on neural networks*, vol. 7, no. 1, pp. 16–29, 1996.

[23] A. K. Jain, M. N. Murty, and P. J. Flynn, “Data clustering: a review,” *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[24] R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera, *Robust statistics: theory and methods (with R).* John Wiley & Sons, 2019.

[25] L. A. Gillenwater, S. Helmi, E. Stene, K. A. Pratte, Y. Zhuang, R. P. Schuyler, L. Lange, P. J. Castaldi, C. P. Hersh, F. Banaei-Kashani *et al.*, "Multi-omics subtyping pipeline for chronic obstructive pulmonary disease," *PloS one*, vol. 16, no. 8, p. e0255337, 2021.

[26] S. Verdi, S. M. Kia, K. Yong, D. Tosun, J. M. Schott, A. F. Marquand, J. H. Cole, A. D. N. Initiative *et al.*, "Revealing individual neuroanatomical heterogeneity in alzheimer's disease," *medRxiv*, 2022.

[27] A. Nowak-Brzezińska and I. Gaibei, "How the outliers influence the quality of clustering?" *Entropy*, vol. 24, no. 7, p. 917, 2022.

[28] S. Arslanturk, S. Draghici, and T. Nguyen, "Integrated cancer subtyping using heterogeneous genome-scale molecular datasets," in *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2020.* World Scientific, 2019, pp. 551–562.

[29] S. Krishnagopal, "Multi-layer trajectory clustering: A network algorithm for disease subtyping," *Biomedical Physics & Engineering Express*, vol. 6, no. 6, p. 065003, 2020.

[30] M.-A. Schulz, M. Chapman-Rounds, M. Verma, D. Bzdok, and K. Georgatzis, "Inferring disease subtypes from clusters in explanation space," *Scientific Reports*, vol. 10, no. 1, pp. 1–6, 2020.

[31] M. Brendel, C. Su, Y. Hou, C. Henchcliffe, and F. Wang, "Comprehensive subtyping of parkinson's disease patients with similarity fusion: a case study with biofind data," *npj Parkinson's Disease*, vol. 7, no. 1, pp. 1–9, 2021.

[32] S. Paul *et al.*, "Capturing the latent space of an autoencoder for multi-omics integration and cancer subtyping," *Computers in Biology and Medicine*, vol. 148, p. 105832, 2022.

[33] A. Jasinska-Piadlo, R. Bond, P. Biglarbeigi, R. Brisk, P. Campbell, F. Browne, and D. McEneaneny, "Data-driven versus a domain-led approach to k-means clustering on an open heart failure dataset," *International Journal of Data Science and Analytics*, vol. 15, no. 1, pp. 49–66, 2023.

[34] N. Rappoport and R. Shamir, "Nemo: cancer subtyping by integration of partial multi-omic data," *Bioinformatics*, vol. 35, no. 18, pp. 3348–3356, 2019.

[35] N. D. Nguyen and D. Wang, "Multiview learning for understanding functional multiomics," *PLoS computational biology*, vol. 16, no. 4, p. e1007677, 2020.

[36] H. Yang, R. Chen, D. Li, and Z. Wang, "Subtype-gan: a deep learning approach for integrative cancer subtyping of multi-omics data," *Bioinformatics*, vol. 37, no. 16, pp. 2231–2237, 2021.

[37] M. C. Thrun and A. Ultsch, "Using projection-based clustering to find distance-and density-based clusters in high-dimensional data," *Journal of Classification*, vol. 38, no. 2, pp. 280–312, 2021.

[38] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg, "Similarity network fusion for aggregating data types on a genomic scale," *Nature methods*, vol. 11, no. 3, pp. 333–337, 2014.

[39] T. Nguyen, R. Tagett, D. Diaz, and S. Draghici, "A novel approach for data integration and disease subtyping," *Genome research*, vol. 27, no. 12, pp. 2025–2039, 2017.

[40] L. Yin, C. K. Chau, P.-C. Sham, and H.-C. So, "Integrating clinical data and imputed transcriptome from gwas to uncover complex disease subtypes: applications in psychiatry and cardiology," *The American Journal of Human Genetics*, vol. 105, no. 6, pp. 1193–1212, 2019.

[41] H. Nguyen, S. Shrestha, S. Draghici, and T. Nguyen, "Pinsplus: a tool for tumor subtype discovery in integrated genomic data," *Bioinformatics*, vol. 35, no. 16, pp. 2843–2846, 2019.

[42] M. Chierici, N. Bussola, A. Marcolini, M. Francescatto, A. Zandonà, L. Trastulla, C. Agostinelli, G. Jurman, and C. Furlanello, "Integrative network fusion: a multi-omics approach in molecular profiling," *Frontiers in oncology*, vol. 10, p. 1065, 2020.

[43] P. Govender and V. Sivakumar, "Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019)," *Atmospheric pollution research*, vol. 11, no. 1, pp. 40–56, 2020.

[44] S. Saria and A. Goldenberg, "Subtyping: What it is and its role in precision medicine," *IEEE Intelligent Systems*, vol. 30, no. 4, pp. 70–75, 2015.

[45] M. Wang, D. Spiegelman, A. Kuchiba, P. Lochhead, S. Kim, A. T. Chan, E. M. Poole, R. Tamimi, S. S. Tworoger, E. Giovannucci *et al.*, "Statistical methods for studying disease subtype heterogeneity," *Statistics in medicine*, vol. 35, no. 5, pp. 782–800, 2016.

[46] C. Sun, S. Hong, M. Song, H. Li, and Z. Wang, "Predicting covid-19 disease progression and patient outcomes based on temporal deep learning," *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, pp. 1–16, 2021.

[47] A. Prat, E. Pineda, B. Adamo, P. Galván, A. Fernández, L. Gaba, M. Díez, M. Viladot, A. Arance, and M. Muñoz, "Clinical implications of the intrinsic molecular subtypes of breast cancer," *The Breast*, vol. 24, pp. S26–S35, 2015.

[48] E. W. Forgy, "Cluster analysis of multivariate data: efficiency versus interpretability of classifications," *biometrics*, vol. 21, pp. 768–769, 1965.

[49] G. T. Huang, K. I. Cunningham, P. V. Benos, and C. S. Chennubhotla, "Spectral clustering strategies for heterogeneous disease expression data," in *Biocomputing 2013*. World Scientific, 2013, pp. 212–223.

[50] M. S. H. Zada, B. Yuan, W. A. Khan, A. Anjum, S. Reiff-Marganiec, and R. Saleem, "A unified graph model based on molecular data binning for disease subtyping," *Journal of Biomedical Informatics*, vol. 134, p. 104187, 2022.

[51] F. Nielsen, "Hierarchical clustering," in *Introduction to HPC with MPI for Data Science*. Springer, 2016, pp. 195–211.

[52] R. Thorndike, "Who belongs in the family? pyschometrika 18 (4): 267–276," 1953.

[53] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[54] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 2, pp. 411–423, 2001.

[55] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data," *Machine learning*, vol. 52, no. 1, pp. 91–118, 2003.

[56] J. Yan and W. Liu, "An ensemble clustering approach (consensus clustering) for high-dimensional data," *Security and Communication Networks*, vol. 2022, 2022.

[57] N. Nguyen and R. Caruana, "Consensus clusterings," in *Seventh IEEE international conference on data mining (ICDM 2007)*. IEEE, 2007, pp. 607–612.

[58] G. Brière, É. Darbo, P. Thébault, and R. Uricaru, "Consensus clustering applied to multi-omics disease subtyping," *BMC bioinformatics*, vol. 22, no. 1, pp. 1–29, 2021.

[59] H. Liu, R. Zhao, H. Fang, F. Cheng, Y. Fu, and Y.-Y. Liu, "Entropy-based consensus clustering for patient stratification," *Bioinformatics*, vol. 33, no. 17, pp. 2691–2698, 2017.

[60] Z. Zheng, S. S. Waikar, I. M. Schmidt, J. R. Landis, C.-y. Hsu, T. Shafi, H. I. Feldman, A. H. Anderson, F. P. Wilson, J. Chen *et al.*, "Subtyping ckd patients by consensus clustering: the chronic renal insufficiency cohort (cric) study," *Journal of the American Society of Nephrology*, vol. 32, no. 3, pp. 639–653, 2021.

[61] N. Tomašev and M. Radovanović, “Clustering evaluation in high-dimensional data,” in *Unsupervised learning algorithms.* Springer, 2016, pp. 71–107.

[62] M. Behringer, P. Hirmer, D. Tschechlov, and B. Mitschang, “Increasing explainability of clustering results for domain experts by identifying meaningful features.” in *ICEIS (2)*, 2022, pp. 364–373.

[63] D. Cohen, “Precalculus: A problems-oriented approach , cengage learning,” ISBN 978-0-534-40212-9, Tech. Rep., 2004.

[64] P. E. Black, “Manhattan distance”” dictionary of algorithms and data structures,” *http://xlinux. nist. gov/dads//.*

[65] A. Singhal *et al.*, “Modern information retrieval: A brief overview,” *IEEE Data Eng. Bull.*, vol. 24, no. 4, pp. 35–43, 2001.

[66] R. S. Strichartz, *The way of analysis.* Jones & Bartlett Learning, 2000.

[67] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, “On the surprising behavior of distance metrics in high dimensional space,” in *International conference on database theory.* Springer, 2001, pp. 420–434.

[68] B. Pfeifer and M. G. Schimek, “A hierarchical clustering and data fusion approach for disease subtype discovery,” *Journal of Biomedical Informatics*, vol. 113, p. 103636, 2021.

[69] N. Rappoport, R. Safra, and R. Shamir, “Monet: multi-omic module discovery by omic selection,” *PLoS computational biology*, vol. 16, no. 9, p. e1008182, 2020.

[70] A. Kamoun, G. Cancel-Tassin, G. Fromont, N. Elarouci, L. Armenoult, M. Ayadi, J. Irani, X. Leroy, A. Villers, G. Fournier *et al.*, “Comprehensive molecular classification of localized prostate adenocarcinoma reveals a tumour subtype predictive of non-aggressive disease,” *Annals of Oncology*, vol. 29, no. 8, pp. 1814–1821, 2018.

[71] H. Xu, L. Gao, M. Huang, and R. Duan, “A network embedding based method for partial multi-omics integration in cancer subtyping,” *Methods*, vol. 192, pp. 67–76, 2021.

[72] T. Gärtner, “A survey of kernels for structured data,” *ACM SIGKDD explorations newsletter*, vol. 5, no. 1, pp. 49–58, 2003.

[73] M. Rupp, “Machine learning for quantum mechanics in a nutshell,” *International Journal of Quantum Chemistry*, vol. 115, no. 16, pp. 1058–1073, 2015.

[74] D. Ramazzotti, A. Lal, B. Wang, S. Batzoglou, and A. Sidow, “Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival,” *Nature communications*, vol. 9, no. 1, pp. 1–14, 2018.

[75] W. Fan, C. Hu, X. Liu, and P. Lu, "Discovering graph functional dependencies," *ACM Transactions on Database Systems (TODS)*, vol. 45, no. 3, pp. 1–42, 2020.

[76] W. Fan, X. Liu, P. Lu, and C. Tian, "Catching numeric inconsistencies in graphs," *ACM Transactions on Database Systems (TODS)*, vol. 45, no. 2, pp. 1–47, 2020.

[77] Q. Chen, J. Zobel, and K. Verspoor, "Duplicates, redundancies and inconsistencies in the primary nucleotide databases: a descriptive study," *Database*, vol. 2017, 2017.

[78] T. Ma and A. Zhang, "Affinity network fusion and semi-supervised learning for cancer patient clustering," *Methods*, vol. 145, pp. 16–24, 2018.

[79] X. Chen, N. Garcelon, A. Neuraz, K. Billot, M. Lelarge, T. Bonald, H. Garcia, Y. Martin, V. Benoit, M. Vincent *et al.*, "Phenotypic similarity for rare disease: ciliopathy diagnoses and subtyping," *Journal of Biomedical Informatics*, vol. 100, p. 103308, 2019.

[80] E. Parimbelli, S. Marini, L. Sacchi, and R. Bellazzi, "Patient similarity for precision medicine: A systematic review," *Journal of biomedical informatics*, vol. 83, pp. 87–96, 2018.

[81] K. K. Sharma and A. Seal, "Multi-view spectral clustering for uncertain objects," *Information Sciences*, vol. 547, pp. 723–745, 2021.

[82] O. Rafique and A. H. Mir, "Weighted dimensionality reduction and robust gaussian mixture model based cancer patient subtyping from gene expression data," *Journal of Biomedical Informatics*, vol. 112, p. 103620, 2020.

[83] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou, "Patient subtyping via time-aware lstm networks," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 65–74.

[84] C. Lee and M. van der Schaar, "A variational information bottleneck approach to multi-omics data integration," in *International Conference on Artificial Intelligence and Statistics.* PMLR, 2021, pp. 1513–1521.

[85] G. Zhang, Z. Peng, C. Yan, J. Wang, J. Luo, and H. Luo, "Multigatae: A novel cancer subtype identification method based on multi-omics and attention mechanism," *Frontiers in Genetics*, vol. 13, 2022.

[86] D.-J. Zhang, Y.-L. Gao, J.-X. Zhao, C.-H. Zheng, and J.-X. Liu, "A new graph autoencoder-based consensus-guided model for scrna-seq cell type detection," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[87] P. Kokol, M. Kokol, and S. Zagoranski, "Machine learning on small size samples: A synthetic knowledge synthesis," *Science Progress*, vol. 105, no. 1, p. 00368504211029777, 2022.

[88] E. Li, L. Wang, Q. Xie, R. Gao, Z. Su, and Y. Li, "A novel deep learning method for maize disease identification based on small sample-size and complex background datasets," *Ecological Informatics*, vol. 75, p. 102011, 2023.

[89] R. Keshari, M. Vatsa, R. Singh, and A. Noore, "Learning structure and strength of cnn filters for small sample size training," in *proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9349–9358.

[90] J.-Y. Park, H. K. Na, S. Kim, H. Kim, H. J. Kim, S. W. Seo, D. L. Na, C. E. Han, and J.-K. Seong, "Robust identification of alzheimer's disease subtypes based on cortical atrophy patterns," *Scientific reports*, vol. 7, no. 1, pp. 1–14, 2017.

[91] C. R. Planey and O. Gevaert, "Coincide: A framework for discovery of patient subtypes across multiple datasets," *Genome medicine*, vol. 8, no. 1, pp. 1–17, 2016.

[92] A. S. Herbert, "The choice of a class interval," *Journal of The American Statistical Association*, vol. 21, pp. 65–66, 1926.

[93] D. W. Scott, "On optimal and data-based histograms," *Biometrika*, vol. 66, no. 3, pp. 605–610, 1979.

[94] D. M. Lane *et al.*, "Online statistics education: a multimedia course of study (http://onlinestatbook. com/)," *Rice University*, 2006.

[95] D. Freedman and P. Diaconis, "On the histogram as a density estimator: L 2 theory," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 57, no. 4, pp. 453–476, 1981.

[96] J.-P. Vert, K. Tsuda, and B. Schölkopf, "A primer on kernel methods," *Kernel methods in computational biology*, vol. 47, pp. 35–70, 2004.

[97] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, "Evaluating the yield of medical tests," *Jama*, vol. 247, no. 18, pp. 2543–2546, 1982.

[98] M. Schmid, M. N. Wright, and A. Ziegler, "On the use of harrell's c for clinical risk prediction via random survival forests," *Expert Systems with Applications*, vol. 63, pp. 450–459, 2016.

[99] R. A. Fisher, "Iris," UCI Machine Learning Repository, 1988, DOI: https://doi.org/10.24432/C56C76.

[100] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[101] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical association*, vol. 66, no. 336, pp. 846–850, 1971.

[102] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 14, 2001.

[103] Y.-C. Wei and C.-K. Cheng, "Towards efficient hierarchical designs by ratio cut partitioning," in *1989 IEEE International Conference on Computer-Aided Design. Digest of Technical Papers.* IEEE, 1989, pp. 298–301.

[104] B. Nowok, G. M. Raab, and C. Dibben, "synthpop: Bespoke creation of synthetic data in r," *Journal of statistical software*, vol. 74, pp. 1–26, 2016.

[105] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[106] J. Emmerson and J. Brown, "Understanding survival analysis in clinical trials," *Clinical Oncology*, vol. 33, no. 1, pp. 12–14, 2021.

[107] Q. Mo, S. Wang, V. E. Seshan, A. B. Olshen, N. Schultz, C. Sander, R. S. Powers, M. Ladanyi, and R. Shen, "Pattern discovery and cancer gene identification in integrated cancer genomic data," *Proceedings of the National Academy of Sciences*, vol. 110, no. 11, pp. 4245–4250, 2013.

[108] T. M. Therneau and D. A. Watson, "The concordance statistic and the cox model," *Department of Health Science Research Mayo Clinic Technical Report*, vol. 85, pp. 1–18, 2017.

[109] Z. T. Kosztyán, A. Telcs, and J. Abonyi, "A multi-block clustering algorithm for high dimensional binarized sparse data," *Expert Systems with Applications*, vol. 191, p. 116219, 2022.

[110] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen, "Graph contrastive learning with augmentations," *Advances in neural information processing systems*, vol. 33, pp. 5812–5823, 2020.

[111] N. Liu, X. Wang, D. Bo, C. Shi, and J. Pei, "Revisiting graph contrastive learning from the perspective of graph spectrum," *arXiv preprint arXiv:2210.02330*, 2022.

[112] E. Pan and Z. Kang, "Multi-view contrastive graph clustering," *Advances in neural information processing systems*, vol. 34, pp. 2148–2159, 2021.

[113] Y. Zhu, Y. Xu, F. Yu, Q. Liu, S. Wu, and L. Wang, "Graph contrastive learning with adaptive augmentation," in *Proceedings of the Web Conference 2021*, 2021, pp. 2069–2080.

[114] K. Hassani and A. H. Khasahmadi, "Contrastive multi-view representation learning on graphs," in *International conference on machine learning.* PMLR, 2020, pp. 4116–4126.

[115] D. Verma and M. Meila, "A comparison of spectral clustering algorithms," *University of Washington Tech Rep UWCSE030501*, vol. 1, pp. 1–18, 2003.