Discovering Gene Associations Across Diseases Using a Knowledge-based Machine Learning Approach

Prepared by: Emma Mamdouh Jeries Qumsiyeh

Ph.D. Thesis

Jerusalem - Palestine

1446/ 2024

Discovering Gene Associations Across Diseases Using a knowledge-based Machine Learning Approach

Prepared by:
Emma Mamdouh Jeries Qumsiyeh

Ph.D. Joint Program in Information Technology Engineering, Al- Quds University, Palestine

Supervisor: Prof. Malik Yousef
Co-supervisor: Dr. Rashid Jayousi
Co-supervisor: Prof. Zaidoun Salah

A Thesis Submitted in Partial Fulfillment of Requirements for the Degree of Doctor of Philosophy in Information Technology Engineering – Joint program - Palestine.

1446/ 2024

Al Quds University
Deanship of Graduate Studies
Ph.D. Joint Program in Information Technology Engineering

Thesis Approval

Discovering Gene Associations Across Diseases Using a knowledge-based Machine Learning Approach

Prepared by: Emma Mamdouh Jeries Qumsiyeh
Registration No.: 22012320

Supervisor: Prof. Malik Yousef
Co-supervisor: Assistant Professor Rashid Jayousi
Co-supervisor: Assoc. Prof. Zaidoun Salah

Ph.D. thesis submitted and accepted, Date: 4/11/2024
The names and signatures of the examining committee members are as follows:

1. **Head of Committee: Prof. Malik Yousef. Signature**
2. **Internal Examiner: Prof. Radwan Qasrawi. Signature**
3. **External Examiner: Prof Mourad Elloumi. Signature**
4. **External Examiner: Prof Burcu Bakir-Gungor. Signature**

Jerusalem, Palestine

1446/ 2024

Dedication

To my husband Issa, your unwavering support and love have been my guiding light. Your belief in me has fueled my determination to succeed. This achievement is a reflection of our partnership and your constant encouragement.

To my kids, Carla, Elias, and Laura. You are the heart and soul of my journey. Your innocence and curiosity have constantly reminded me why I strive for excellence. May this achievement inspire you to dream big and chase your aspirations fearlessly.

To my parents, whose unwavering love, sacrifice, and unending belief in my potential have laid the foundation for my journey.

To my sister Dina and brother Johny, your unwavering faith in my abilities and continuous motivation have been a driving force.

To Jesus Christ and Virgin Mary, your grace and blessings have given me strength and resilience. Your unwavering presence has carried me through moments of doubt and has been a source of hope.

To all those who stood by me, your guidance, words of wisdom, and support have illuminated my path. This achievement is dedicated to all of you. Your love, support, and presence have shaped my success and enriched every step of this remarkable journey.

Declaration

I certify that this thesis submitted for the degree of Ph.D. is the result of my own research, except where otherwise acknowledged, and that this study (or any part of the same) has not been submitted for a higher degree to any other university or institution.

Signed: 

Emma Mamdouh Jeries Qumsiyeh

Date: 4/11/2024

Acknowledgments

I extend my heartfelt gratitude to my advisor, Prof. Malik Yousef, whose unwavering support and guidance have been instrumental in shaping my doctoral journey. His mentorship at every stage of my thesis work and his unwavering trust and encouragement have been the driving force behind my accomplishments. I consider him an inspiration and a role model for my success.

Special thanks go to my second advisor, Dr Rashid Jayousi, for his exceptional patience, unwavering faith, and invaluable insights. His trust in me and valuable ideas and suggestions have been invaluable assets throughout my journey.

I also express my sincere appreciation to Assoc. Prof. Zaidoun Salah's support and contributions have greatly enriched my research experience.

Your contributions have made this achievement possible, and I am deeply grateful for your unwavering support and guidance.

ABSTRACT

Complex diseases such as diabetes, Alzheimer's, and cancer are influenced by a combination of genetic, lifestyle, and environmental factors that do not follow straightforward inheritance patterns. Biological systems are immensely complex and heterogeneous. To resolve the enigmas surrounding these systems, extensive research provides huge amounts of biological data. In this thesis and in our first study, a novel approach called GediNET was developed to integrate prior biological knowledge into disease-associated gene groups. GediNET employs a Grouping, Scoring, and Modeling (G-S-M) approach to identify top-performing gene groups, which are then used to train a machine-learning model. Following the data exploration and preprocessing steps, various classification models were built with 100-fold Monte Carlo Cross-Validation, and the performance of these models was evaluated. By applying Disease-Disease Association (DDA) based machine learning, GediNET uncovered new relationships between diseases, improving diagnosis, prognosis, and treatment approaches. In the second study, GediNETPro, an advanced version of GediNET, was developed. This version utilizes Cross-Validation (CV) information and clustering techniques, such as K-means, to identify patterns of disease group associations. GediNETPro provides visualization tools, like heatmaps and in-depth analysis of disease group clusters, offering insights for developing effective diagnostic interventions. The third study leveraged molecular-level data to develop effective methods for predicting Disease-Disease Associations (DDAs). A statistical technique was developed by employing the G-S-M-P model of GediNETPro to compute semantic similarity metrics between diseases. The semantic approach detects representative diseases within clusters and establishes a semantic relationship between the disease under investigation and other diseases. The studies presented in this thesis contribute to understanding disease complexity, uncovering disease associations, and identifying potential biomarkers and drug targets.

Keywords: Biological Knowledge Integration, Gene Expression, Disease–Disease Association (DDA), Machine Learning, Feature Selection, Grouping, Scoring, and Modeling (G-S-M) approach, Cross-Validation, K-means clustering, Heatmap, Heterogeneity, Monte-Carlo Cross-Validation, Semantic Relationship.

Table of Contents

List of Tables

List of Figures

CHAPTER ONE

INTRODUCTION

This study aims to uncover disease associations using novel approaches by integrating pre-existing biological knowledge, feature selection, and machine learning algorithms. It also focuses on leveraging molecular-level data to predict disease-disease associations and establish semantic relationships, contributing to advancements in disease understanding, biomarker identification, and improved diagnosis and treatment approaches. This chapter provides an overview of the study's background, research challenge, aims, significance, and scope.

Detecting disease-disease associations (DDAs) is essential for advancing research in medicine and systems biology, as it provides valuable insights into the complex relationships between different diseases. However, the heterogeneous nature of biological data sources presents challenges for identifying meaningful disease associations. Traditional approaches often focus on individual genes or biomarkers, overlooking the broader context of disease interactions and shared biological mechanisms. This traditional approach neglects the intricate biological context surrounding diseases, hindering the identification of potential therapeutic targets and interventions. A computational tool is needed for integrating pre-existing biological knowledge and molecular-level data, while also addressing semantic associations and capturing the variation within disease groups. Such a tool would provide a more comprehensive understanding of disease associations and facilitate the discovery of novel insights.

Machine learning algorithms and feature selection methods have been frequently used in recent years to pinpoint genes associated with particular diseases. While these approaches have shown promise in biomarker discovery, there is an increasing acknowledgment of the significance of an integrative approach incorporating prior biological knowledge into machine learning to enhance the discovery process. By incorporating existing knowledge, such as gene-disease associations, pathways, and functional annotations, researchers can leverage the collective information to gain a more holistic understanding of disease associations.

This research is important because it has the potential to reveal hidden associations across diseases, provide insights into common biological processes, and facilitate the creation of improved diagnostic and treatment approaches. This thesis can significantly advance our understanding of disease associations and contribute to translational research efforts by considering the broader context of disease interactions and incorporating prior biological knowledge.

1.1 Research Problem

Genetic, lifestyle, and environmental factors exert a significant influence on the development of complex diseases that display a deviation from straightforward inheritance patterns. Biological systems, known for their immense complexity and heterogeneity, pose substantial challenges in unraveling their mysteries. Extensive research efforts are undertaken to tackle these challenges, generating vast volumes of valuable biological data. Gene expression data-targeted research tries to find disease-associated genes that may be used to identify new biomarkers (X. Wang et al., 2011). These gene expression patterns can have diagnostic, prognostic, or therapeutic values in complex diseases like cancer (B. Chen et al., 2016).

As high-throughput technologies progress, large transcriptome datasets become available, making it increasingly challenging to glean insights from extensive lists of differentially expressed genes. Besides, the cost of acquiring a sample's gene expression profile is decreasing rapidly (Hasin et al., 2017). As a result, expression profiling, the subject of this thesis, has become a standard practice in biological laboratories. As these data are frequently acquired from a small number of samples, dealing with multidimensional biological data is a hard problem. The dimensions of a gene's characteristics are crucial input factors, and a predefined number of samples is necessary to estimate or learn various functions. Of course, a larger sample size enhances the accuracy of a prediction (Bellman, 1961). Gene selection and removal of duplicated or unnecessary genes are crucial for addressing this issue.

The majority of feature selection techniques for the analysis of gene expression data choose genes only based on expression values. Their foundation lies on statistics and machine learning. Biological knowledge is utilized at the conclusion of the research to gain biological insights, confirm the initial findings, or do enrichment analysis (Ben-dor, 2002; Bittner et al., 2000). The traditional method of gene selection has various limitations. They assess the importance of each gene individually without taking into account its interconnections and interactions. They assess the importance of each gene individually without taking into account its interconnections and interactions. The primary drawbacks of such approaches are their biological interpretation challenges and their inability to generate new biological information (Fang et al., 2014).

The incorporation of pre-existing biological knowledge into gene selection has replaced data-driven approaches in recent times. Integrative gene selection methods include pre-existing knowledge from external biological resources, enhancing interpretability and predictive performance (Bellazzi & Zupan, 2007; Fang et al., 2014). Although specific integrative systems already combine different biological data sources, most use unsupervised learning techniques to identify regulatory modules and regulatory motifs instead of feature selection or disease classification (Elati & Rouveirol, 2010).

Since biological systems are complex and interrelated, a model trained on a specific dataset can only utilize a small portion of all available scientific information. To get a comprehensive understanding of molecular biology and health care, it is necessary to integrate multiple biological resources, such as pre-existing biological knowledge and multi-omics data. Thus, we can formulate the research questions as follows:

How to integrate pre-existing biological knowledge associated with diseases (disease genes are a group) into a machine learning framework and perform feature selections as groups to detect the most important disease genes?

1.2 Research Objectives

In this section, the study outlines its general and specific objectives, which were formulated based on the problem statement provided earlier. The study objectives are outlined as following:

This project aims to develop a platform that integrates gene-disease associations to improve the detection of disease-disease associations, molecular disease biomarkers, and the discovery of targeted therapies. Our tool, GediNET, is developed to accomplish this purpose. This novel method integrates pre-existing biological knowledge about the disease to enhance feature selection in classification tasks and reveal hidden patterns related to the gene-disease association. GediNET is a machine learning method that utilizes the Grouping, Scoring, and Modelling (G-S-M) approach, developed by Yousef and his colleagues (Yousef et al., 2020). To analyze gene expression data, the method utilizes DisGeNET (Piñero et al., 2017), an external biological knowledge database that contains associations between genes and diseases. The GediNET analysis promotes a global picture by enabling a deeper comprehension of the information flow underlying the data under study. Besides, GediNET is unique because it makes finding meaningful relationships between the specific disease under research and other diseases possible. GediNET identifies these relationships through Disease–Disease Association (DDA) based machine learning. The necessity of examining new disease associations and expanding our understanding of disease connections has led to a recent increase in interest in various initiatives aimed at establishing DDA. These studies have the potential to enhance methods for diagnosing, treating, and prognosing diseases. There are very few reliable and recognized DDA. As a result, it implies that more effort is needed to detect DDAs (Suratanee & Plaimas, 2015).

The specific objectives of the study are:

1. To develop a machine learning approach, named GediNET, based on the Grouping, Scoring, and Modelling (G-S-M) framework, to integrate pre-existing biological knowledge, and enhance feature selection in classification tasks.
2. To use DisGeNET, an external biological knowledge database containing gene-disease associations, to analyze gene expression data within the GediNET framework.
3. To promote a global perspective by enabling a deeper comprehension of the information flow underlying the analyzed data and uncovering hidden patterns related to gene-disease associations.
4. To enhance the detection of molecular disease markers by integrating gene-disease associations into the GediNET platform.
5. To identify meaningful relationships between the specific disease under research and other diseases through Disease-Disease Association (DDA) based machine learning within the GediNET framework.
6. To contribute to the disease research field by exploring novel associations and enhancing the knowledge of disease relationships through the detection of DDAs.
7. To enhance disease diagnosis, and treatment methods by broadening the knowledge of disease connections and contributing in the discovery of target therapies.

These specific objectives collectively contribute to achieving the general objective of developing GediNET and its application in enhancing the understanding and management of complex diseases.

1.3 Research Relevance

The relevance of this research covers multiple sectors, offering significant benefits to the biomedical and healthcare community by providing GediNET, a tool that integrates existing biological knowledge with machine learning to enhance gene-disease association detection. This occurrence can facilitate the identification of innovative biomarkers and therapeutic targets, thereby promoting the progress of disease diagnosis, treatment, and enhancing the quality of patient care. GediNET's capacity to uncover hidden patterns in diseases can help with the creation of customized treatment regimens, thereby personalizing healthcare in the context of precision medicine and personalized treatment. Within the pharmaceutical sector, the utilization of GediNET holds the potential to optimize drug development by facilitating the identification of novel therapeutic targets. This, in turn, allows pharmaceutical companies to concentrate their efforts on promising drug candidates and accelerate the whole process of drug discovery.
Moreover, the GediNET method has the potential to yield significant advantages for the academic and research domains. It can provide vital insights into the associations between genes and diseases, hence stimulating further investigations into the intricacies of various diseases. The insights offered by GediNET can additionally inform healthcare policy and decision-making, thereby assuring the effective allocation of resources towards addressing specific disease correlations. Enhancements in diagnostic tools, risk assessment, and treatment options are anticipated to result in improved healthcare outcomes and enhanced patient well-being for both patients and healthcare professionals. GediNET results can be used by disease foundations and advocacy groups to determine possible therapeutic targets associated with particular disease priority areas.

This study has important consequences for understanding diseases and encourages new approaches for disease diagnosis, treatment, and management. The technology is strategically situated to provide a beneficial influence within the realms of science, medicine, pharmaceuticals, and healthcare, representing notable advancements in comprehending, diagnosing, and treating diseases. The research holds the potential to provide positive outcomes for various stakeholders, ultimately contributing to the well-being of patients and society as a whole.

1.4 Significance of the study

The thesis has significance as it helps to enhance our comprehension of intricate diseases and their associations with each other. The following points highlight the significance of the thesis:

1. Search for Significant Biomarkers/Genes: GediNET's emphasis on gene groups rather than individual genes in the search for significant biomarkers brings several advantages. By considering gene groups, GediNET considers the intricate interactions and synergistic effects among genes, providing a more comprehensive understanding of their collective contribution to disease development and progression. This approach can uncover hidden associations and reveal novel biomarkers that may have been overlooked in traditional single-gene analyses. GediNET enhances the accuracy and reliability of biomarker identification by focusing on gene groups, paving the way for more targeted and effective diagnostic and therapeutic strategies.

The application of GediNET in biomarker identification is especially important for diseases like cancer, where early detection and accurate therapy are essential. In Alzheimer's disease, GediNET may discover groups of genes linked to early-stage pathophysiological alterations, thereby facilitating the creation of early intervention strategies. In cancer research, GediNET may identify novel biomarkers for cancer subtypes, perhaps resulting in more targeted and effective therapy alternatives. By finding these biomarkers, GediNET enhances the comprehension of disease mechanisms and paves the way for the advancement of diagnostic tests and therapeutic strategies.

2. Defining New Disease-Disease Associations: The final list of genes generated by GediNET not only serves as potential biomarkers but also enables the exploration of new disease-disease associations. By leveraging DDA-based machine learning, GediNET uncovers meaningful relationships between previously unknown or not extensively studied diseases. The finding of new disease association improves our comprehension of disease relationships, making it easier to identify common biological mechanisms and shared pathways. These findings can lead to major implications, such as creating comprehensive treatment strategies, discovering new targets for therapy, and the possibility of using current drugs for other diseases.

3. Revealing Marker Impact and Interactions: GediNET's scoring algorithm creates high and low-scoring groups, offering valuable insights into the impact of markers on disease mechanisms. At the single omics level, the analysis of high-scoring groups highlights the individual marker's significance and its potential role in driving disease processes. Conversely, low-scoring groups shed light on markers that may play a lesser role or have a minimal impact. Moreover, at the multi-omics level, the interaction of markers within high-scoring groups provides a deeper understanding of how these markers collaboratively contribute to disease mechanisms, offering valuable clues for designing targeted interventions and personalized treatment strategies.

4. Prognostic Analysis: GediNET enables the evaluation of the predictive value of top gene groups and individual genes by utilizing survival time and clinical status data from gene expression datasets. By integrating clinical outcomes, GediNET facilitates the identification of gene signatures and gene groups that correlate with disease prognosis. This information can be used to develop robust predictive models, enabling clinicians to assess patient outcomes, predict disease progression, and make informed decisions regarding treatment strategies. The predictive insights gained from GediNET's analysis have the potential to improve patient management, enhance risk stratification, and guide personalized therapeutic interventions.

5. Hub Genes and Modules in Network Analysis: GediNET's identification of top-scored genes provides a valuable resource for network analysis. With their high significance in the context of disease, these genes can be considered hub genes within biological networks. Their central position in the network suggests their crucial roles in disease-related pathways and molecular interactions. The top-scored gene groups can also be treated as hub modules, representing functionally related sets of genes intricately involved in disease mechanisms. The p-values generated by GediNET offer an opportunity to incorporate them as weights in weighted correlation network analysis, enabling a more refined and accurate assessment of gene co-expression patterns and network properties. This comprehensive network analysis can unravel complex molecular interactions, identify critical regulators, and provide insights into the underlying biological processes driving disease progression.

6. Versatile Tool for System-Level Understanding: GediNET, recommended as a versatile tool in this study, significantly contributes to the system-level understanding of cellular behaviors by integrating pre-existing biological knowledge. By incorporating external databases such as DisGeNET, GediNET harnesses the wealth of biological knowledge and integrates it with molecular-level data. This integration allows researchers to bridge the gap between individual genes and their biological context. GediNET's ability to enhance the comprehension of cellular behaviors facilitates the identification of critical biological pathways, promotes the discovery of new therapeutic targets, and opens avenues for developing innovative treatment approaches.

7. Current Limitations in Personalized Medicine: Personalized medicine presently encounters considerable obstacles, especially regarding biomarker validity, the intricacy of disease causes, and the tailoring of therapeutic approaches. GediNET mitigates these constraints by amalgamating extensive biological information with sophisticated machine learning methodologies to improve the identification and confirmation of disease-related gene groups. This method enhances the precision of biomarker discovery and aids in comprehending intricate disease interactions at the molecular level. GediNET's capacity to discern interrelated biomarkers across many ailments can enhance the comprehension of comorbidities in diseases such as diabetes and cardiovascular disorders, resulting in more efficacious multi-target therapies.

In summary, the significance of the thesis extends to various aspects, including identifying biomarkers, discovering disease associations, understanding marker impact and interactions, predictive analysis, network analysis, and enhancing system-level understanding. The multidimensional contributions of GediNET and the research presented in this thesis have the potential to revolutionize disease research, diagnosis, and treatment, ultimately benefiting patients and improving healthcare outcomes.

1.5 Motivation

The primary motivation behind this study is to comprehensively understand the intricate networks of gene-disease associations and the interconnected links between diseases. These aspects play a crucial role in the advancement of biomedical science. As we explore the enormous amount of genetic data made available by high-throughput gene expression technologies, we are faced with the task and potential to extract profound discoveries that have the capacity to revolutionize our comprehension of human health and disease.

The intellectual motivation behind the pursuit of this thesis stems from its potential to make novel contributions in medicine and customized healthcare. The exploration of complex relationships between genes and diseases is a primary focus of this research, offering the potential to provide personalized treatment approaches and diagnostic methods. This undertaking is in accordance with the overarching goal of shifting away from standardized medical interventions towards personalized treatment strategies that are closely influenced by the patient's own genetic composition.
Concurrently, our study is driven by the prospect of revolutionary findings in the domain of disease-disease associations. These findings can reshape our understanding of comorbidities and the interconnected mechanisms underlying diseases, providing a fresh perspective for examining and tackling the complex nature of human ailments. The motivation to offer innovative approaches and knowledge to the scientific community in the fields of

bioinformatics and systems biology enhances the academic drive, consequently improving the continuing discourse.

Furthermore, the foundation of this concept is rooted in a social drive to effectively influence healthcare outcomes in a good manner. This research aims to address the gap between huge genetic databases and practical healthcare insights, with the goal of informing public health policy, improving clinical decision-making, and promoting a predictive, preventative, and participative approach to medicine.

The G-S-M (Grouping, Scoring, and Modeling) approach of GediNET presents a novel way for detecting disease-disease associations (DDAs), setting it apart from conventional approaches that mostly emphasize direct gene-disease relationships. The G-S-M framework creates grouped gene sets across different diseases, which makes it easier to find intricate inter-disease interactions than traditional methods that examine individual genes in isolation. This not only improves the model's capacity to detect possible biomarkers but also reveals novel opportunities for therapeutic intervention. The 'Grouping' stage categorizes genes according to common disease associations, 'Scoring' assesses these clusters to ascertain their significance and impact on disease outcomes, and 'Modeling' utilizes the highest-ranked gene groups to develop predictive models, combining machine learning with comprehensive biological insights.

Recent studies, such as Yao et al. (Yao et al., 2024), provide empirical evidence for the integration of biological insights with machine learning, demonstrating that machine learning techniques applied to integrated gene-disease networks revealed possible novel biomarkers for Alzheimer’s disease. Wang et al. (Z. Wang et al., 2023) employed a comparable methodology to elucidate critical pathological pathways in cancer progression, highlighting the capacity of these techniques to improve diagnostic precision and inform the creation of targeted therapeutics. These examples offer empirical validation for the G-S-M framework's capacity to enhance our comprehension of intricate disease relationships and treatment prospects.

The primary objective of this academic investigation is two-fold: firstly, to advance the scientific comprehension of gene-disease and disease-disease connections to unprecedented levels, and secondly, to use these discoveries in a practical manner that yields tangible advantages for both individual patients and the wider healthcare domain.

1.6 Scope of the Study

Through this study, we aim to create and test GediNET, a new computational tool for combining biological knowledge and molecular-level data to improve feature selection in classification tasks and find hidden patterns in the gene-disease association. The study analyzes gene expression data of complex diseases such as diabetes, Alzheimer's, and cancer. Various topics and theories will be discussed, including the Grouping, Scoring, and Modeling (G-S-M) approach employed by GediNET, using DisGeNET (Piñero et al., 2017) as an external biological knowledge database, Disease-Disease Association (DDA) based machine learning, and the impact and interaction of markers involved in disease mechanisms at the single omics and multi-omics levels.

The study carefully identifies gene-disease associations with an emphasis on complicated disorders like diabetes and Alzheimer's. These diseases are chosen because of their

prevalence and genetic influence on their progression and therapy. The use of public gene expression databases like GEO (Barrett et al., 2013), TCGA (Tomczak et al., 2015a), and DisGeNET (Piñero et al., 2017) limits the coverage. These databases are rich in data, yet they have constraints including data quality, completeness, and sample representativeness. These considerations can affect findings generalizability.

Sample size and processing demands make high-dimensional data handling difficult. The study used datasets with a high number of features (genes) relative to samples, requiring sophisticated statistical methods and computational algorithms for meaningful analysis. Dimensionality reduction, robust statistical approaches, and huge dataset-handling machine learning algorithms are used to address these issues. The computational intensity needed to handle and model these data is high, thus the results must be recognized within these computational and methodological limits.

1.7 Methodology

Our methodology progresses from the development of the original GediNET framework to its advanced version, GediNETPro. GediNET lays the foundation by sourcing gene expression data and disease-gene associations to identify potential biomarkers and infer disease-disease associations. This process is critical for understanding disease mechanisms and guiding personalized medicine.

GediNET employs the G-S-M approach, starting with the "G" component to group genes based on associations from the DisGeNET database. The "S" component then scores these groups based on differential expression, utilizing statistical measures and machine learning classifiers like Random Forest within a cross-validation framework. The "M" component consolidates these efforts by training a model on the top-scored gene groups, with a subsequent evaluation on a testing set.

The Grouping-Scoring-Modeling (G-S-M) approach integrates gene-disease associations to improve disease network identification and analysis, making it an innovative bioinformatics approach. The G-S-M framework groups genes based on shared disease associations from comprehensive databases like DisGeNET using a sophisticated grouping algorithm. Traditional techniques frequently investigate genes in isolation or use simplistic disease relationships. This explores intricate gene relationships and makes gene groupings more biologically relevant.

The scoring component of GediNET uses powerful statistical models to assess each gene group's impact on certain diseases. Prioritizing gene groups by predictive power and disease phenotype relevance improves the model's accuracy. The scoring process uses established and unique measures to quantify each gene group's contribution before machine learning modeling.

Monte Carlo cross-validation (Xu & Liang, 2001) is used throughout the investigation to verify the GediNET framework's efficacy and resilience. The model is validated using numerous random subsamples, guaranteeing that the results are reproducible and resilient to data sampling and distribution. In high-dimensional data, this validation technique reduces overfitting and improves reliability and generalizability. Monte Carlo cross-validation shows our dedication to rigorous scientific methods and boosts trust in the G-S-M framework's robustness and applicability.

Building on this, GediNETPro (Qumsiyeh et al., 2023a) introduces significant enhancements, particularly the "P" component, which employs rank aggregation and K-means clustering within a Monte Carlo cross-validation loop to uncover hidden patterns and deeper insights into disease associations. This sophisticated approach involves repeated random sub-sampling to prevent overfitting, managing the imbalanced datasets through under-sampling. Model performance is meticulously evaluated using accuracy, sensitivity, and specificity metrics across multiple iterations. The GediNETPro (Qumsiyeh et al., 2023a) workflow is thus a multi-faceted approach that integrates data acquisition, preprocessing, scoring, and machine learning to facilitate a comprehensive analysis of disease associations, ultimately aiming to propel the field of bioinformatics into new frontiers of disease understanding and treatment discovery.

Later, we introduce a novel statistical method within a Grouping-Scoring-Modeling framework to calculate semantic similarity between a target disease and other diseases, supplemented by Jaccard similarity for comparing disease groups. Leveraging GediNETPro (Qumsiyeh et al., 2023a), the research captures the relative importance of disease groups through a ranking system over numerous Monte Carlo cross-validation iterations. Clusters of diseases are then formed using K-means clustering, with the most significant clusters identified by the lowest average rank scores. These clusters are further analyzed using a semantic approach to uncover deeper disease relationships, and diversity measures like the Jaccard index to validate these semantic connections.

The use of databases like DisGeNET(Piñero et al., 2017), GEO (Barrett et al., 2013), and TCGA (Tomczak et al., 2015a) enhances our research with comprehensive genomic and clinical data; nonetheless, it is essential to recognize the inherent limits resulting from the possible overrepresentation of specific genes and diseases. This bias may distort predictive modeling and impact the generalizability of the study results. To address these biases, our methodology incorporates data normalization and the integration of supplementary datasets to ensure balanced representation of genes and diseases. Moreover, sophisticated statistical methods like robust rank aggregation (Kolde et al., 2012) are utilized to prevent the results from being skewed by overrepresented data points, hence improving the reliability and applicability of our conclusions. This methodology enhances our ability to unravel disease networks, potentially informing more effective treatment approaches, particularly when considering non-genetic disease factors.

CHAPTER TWO

BACKGROUND

The background chapter provides a comprehensive exploration of the datasets, methodologies for data processing, approaches for feature selection, and the machine learning classifiers that form the foundation of our research. It begins with a detailed presentation of the gene expression databases retrieved from the GEO database, laying out the process of data selection, normalization, and preparation that makes them fit for analysis. The chapter proceeds to explore the DisGeNET disease-gene association dataset, a pivotal resource for understanding the genetic underpinnings of disease. In addition, we conduct a thorough examination of the TCGA dataset, specifically emphasizing its gene expression data in the context of breast cancer research. In addition to data curation, this chapter also presents advanced feature selection approaches that are used to enhance our dataset, ensuring that only the most important features are utilized to inform the prediction models. The final section of the chapter focuses on providing an exposition of the several machine learning classifiers utilized in this study. It elaborates on their functioning, the reasoning behind their selection, and their contribution to the advancement of our investigation. This chapter serves the dual purpose of establishing the foundation for our investigation and outlining the methodological framework that directs our journey towards knowledge acquisition.

2.1 Datasets and Data Processing

2.1.1 GEO Gene Expression Datasets:

This thesis examined various gene expression databases within the "GEO database," which stands for the Gene Expression Omnibus (GEO) database. The Gene Expression Omnibus is a public archive for functional genomics data managed by the National Center for Biotechnology Information (NCBI), which is part of the National Library of Medicine (NLM). The GEO database stores high-throughput gene expression data such as microarray and next-generation sequencing data, along with additional functional genomics data including ChIP-seq and SNP array data.

Researchers and scientists use GEO to store and retrieve a vast array of experimental data related to gene expression and molecular biology. This vast and varied accumulation of data enables researchers to perform data analysis, compare gene expression patterns, and make discoveries regarding gene function and regulation across a spectrum of biological conditions and disease states. Through the NCBI website, we have accessed the GEO

database and its associated tools, allowing us to search, obtain, and analyze the data for our research projects. Usually, researchers process and normalize their data before making it accessible to the public and comparable to other datasets.

In this thesis, we downloaded 10 human gene expression datasets for different types of complex diseases. For each dataset, the name of the disease and the number of samples were defined. Moreover, positive and negative samples were available. Table 2.1 describes the 10 datasets in more detail.

Table 2.1: Description of the 10 datasets used in the study. Each entry has the GEO accession, the name of the disease, the number of samples, and the data classes.

GEO accession	Title	Disease	#Samples	Classes
GDS1962	Glioma-derived stem cell factor effect on angiogenesis in the brain	Glioma	180	negative = 23 positive = 157
GDS2545	Metastatic prostate cancer (HG-U95A)	Prostate cancer	171	negative = 81 positive = 90
GDS2771	Large airway epithelial cells from cigarette smokers with suspected lung cancer	Lung cancer	192	negative = 90 positive = 102
GDS3257	Cigarette smoking's effect on lung adenocarcinoma	Lung adenocarcinoma	107	negative = 49 positive = 58
GDS4206	Pediatric acute leukemia patients with early relapse: white blood cells	Leukemia	197	negative = 157 positive = 40
GDS5499	Pulmonary hypertension: PBMCs	Pulmonary hypertension	140	negative = 41 positive = 99
GDS3837	Non-small cell lung carcinoma in female nonsmokers	Lung cancer	120	negative = 60 positive = 60
GDS4516_4718	Colorectal cancer: laser microdissected tumor tissues	Colorectal cancer	148	negative = 44 positive = 104
GDS2547	Metastatic prostate cancer (HG-U95C)	Prostate cancer	164	negative = 75 positive = 89
GDS3268	Colon epithelial biopsies of ulcerative colitis patients	Colitis	202	negative = 73 positive = 129

2.1.2 DisGeNET disease-gene association dataset:

DisGeNET is a comprehensive database that contains information on the relationships between genes and human disorders. This is a helpful resource for academics and medical professionals who want to comprehend the genetic foundation of disorders and investigate potential disease-related genes. DisGeNET collects and integrates information from various sources, including scientific literature, genome-wide association studies and other databases. The data in DisGeNET are curated and mapped to standardized identifiers, enabling easy cross-referencing and integration with other genomic and biomedical databases (Piñero et al., 2017).

The database includes information on:

1. Gene-Disease associations: DisGeNET provides information about known associations between specific genes and various human diseases. Evidence from the scientific literature and other sources of experimental data supports these associations.
2. Disease-Disease Associations: DisGeNET also catalogs relationships between different diseases. This information helps researchers identify potential comorbidities or shared genetic factors among different diseases.
3. Variant-Disease Associations: DisGeNET contains information on the connections between genetic variations (such as single nucleotide polymorphisms, or SNPs) and diseases, offering understanding into the genetic variables that increase the risk of certain diseases.
4. Drug-Disease Associations: DisGeNET also contains information on drug-disease relationships, helping researchers identify potential drug candidates or repurposing opportunities for specific diseases.

The data in DisGeNET is regularly updated and curated to ensure accuracy and relevance. It is accessible through a user-friendly web interface, allowing researchers to search, browse, and download gene-disease association data.

In this thesis, the dataset containing genes and their associated diseases was downloaded from DisGeNET version 7.011. The dataset contains 30,170 diseases and 21,666 genes that form 3,241,576 gene-disease connections. Given the massive dataset size, two filters were used to reduce the number of associations in terms of practicality and computational complexity. The filters were set on the columns diseaseType and diseaseSemanticType in the DisGeNET dataset. The diseaseType column divided the data into three categories—disease, phenotype, and group—and we only chose disease as concerning for our study. On the column diseaseSemanticType, we only chose those rows categorized as Neoplastic Process and Disease. This was done to increase compatibility and better understand the workflow results. After filtering, only 15,991 genes and 3929 diseases remained for further analysis, which accounted for 329,936 gene-disease associations. Figure 2.1 illustrates a portion of the disease distribution over the number of genes for associated with each disease.

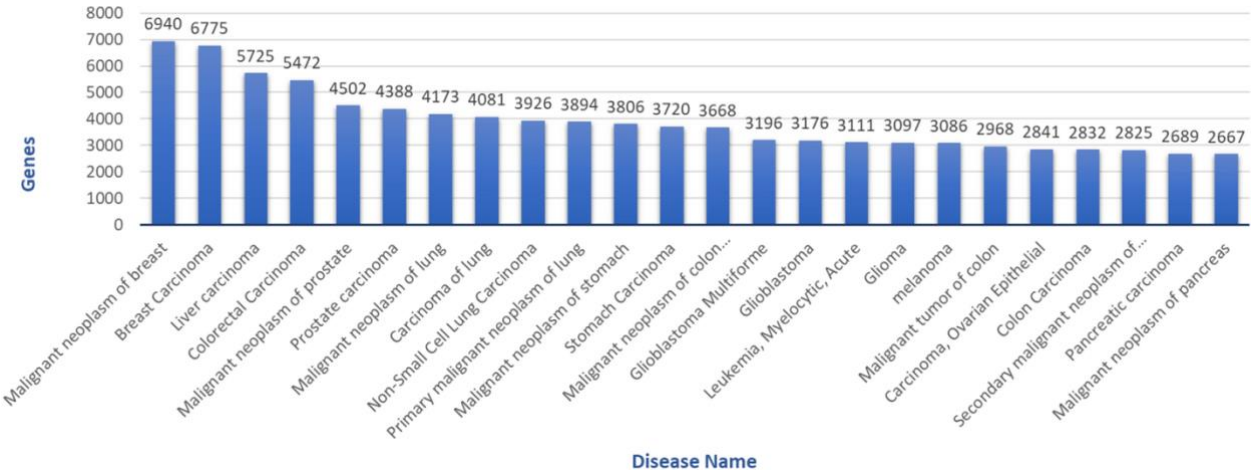


Figure 2.1: A part of the DisGeNET dataset histogram frequency plot. It shows the number of genes associated with each disease, where the X-axis is the disease name, and Y-axis is the number of genes.

2.1.3 TCGA dataset:

We utilized the Cancer Genome Atlas - Breast Invasive Carcinoma (TCGA-BRCA) (Tomczak et al., 2015b) dataset from the Genomic Data Commons, which is managed by the National Cancer Institute. We analyzed gene expression (mRNA) datasets with reads aligned to GRCh38, obtained from Xena Public Data Hubs for analysis (Goldman et al., 2020). The dataset included 302 positive class samples and 247 negative class samples, and was refined for Luminal A, Luminal B, HER2-enriched, and Basal-like intrinsic subtypes. The molecular intrinsic subtype classes in the BRCA dataset were identified by the utilization of the PAM50 assay (Prediction Analysis of Microarray 50). This test relies on 50 gene signatures to categorize samples into various molecular subtypes (Parker et al., 2009). The gene expression raw counts were normalized using the Trimmed Mean of M-Values approach from the edgeR Bioconductor package (Robinson et al., 2010). The dataset contained 21,839 genes.

Uniform normalization methods are employed to guarantee consistent and precise comparisons across various datasets, including those from GEO (Barrett et al., 2013), TCGA (Tomczak et al., 2015b), and DisGeNET (Piñero et al., 2017). The Trimmed Mean of M-Values (TMM) (Robinson et al., 2010) normalization for TCGA data compensates for compositional differences among libraries. Normalization of GEO datasets is performed during preprocessing to equalize gene expression levels across samples, enabling precise cross-sample comparisons and analyses.

For DisGeNET(Piñero et al., 2017), which offers gene-disease association data instead of gene expression levels, we standardize gene-disease linkages by normalizing the frequency of correlations in relation to the incidence of diseases and genes within the dataset. This normalization facilitates fair comparison of gene-disease association intensities across various situations and studies.

Mitigating class imbalance is essential in disease classification research, particularly in datasets characterized by unequal class distribution. We employed under-sampling (Bach et al., 2019) to equilibrate the class representation in the GEO dataset. This approach preserves all instances in the minority class while diminishing the instances in the majority class, so averting the model from acquiring a bias towards the more dominant class.

Alongside under-sampling, we employed the Random Forest Classifier, allocating 90% of the data for training and 10% for testing. To guarantee the robustness and dependability of the model, 100-fold Monte Carlo Cross-Validation (MCCV) (Xu & Liang, 2001) was utilized. MCCV entails the random selection of sample fractions for training data, while the remainder is designated as test data, with performance measures computed by averaging outcomes across 100 iterations. This method was selected over conventional cross-validation because of its superior repeatability and reduced variance, rendering it especially appropriate for datasets with class imbalances and guaranteeing the consistency and precision of our results.

2.2 Feature Selection Techniques

Feature selection tries to reduce computing costs by removing redundant or irrelevant features from the input data. This methodology allows for the refinement of the model by focusing on important attributes and helps to improve understanding of the resulting model.

To do this, it is necessary to rate or score features according to how well they can predict the outcome. Numerous methods exist for ranking features, such as statistical evaluations and machine learning-based wrapper methods (Jovic et al., 2015). Moreover, more sophisticated strategies incorporating biological knowledge into machine learning algorithms have been employed for feature selection or the selection of feature groups, as evidenced by a number of recent tools. Several tools, such as SVM RCE, SVM-RCE-R (Yousef, Bakir-Gungor, et al., 2021; Yousef et al., 2007; Yousef, Jabeer, et al., 2021), maTE (Yousef et al., 2019), CogNet (Yousef, Ülgen, et al., 2021), miRcorrNet (Yousef, Goy, et al., 2021), miRModuleNet (Yousef, Goy, et al., 2022), and Gene Ontology (Yousef, Sayıcı, et al., 2021), have embraced such an approach. A recent review of these tools and their competitors is available in (Yousef et al., 2020).

2.2.1 Maximizing Relevance while Minimizing Redundancy (mRMR):

Maximum Relevance—Minimum Redundancy (mRMR) is a feature selection method employed in machine learning and data analysis to identify a subset of pertinent and informative features from a broader range of input variables. mRMR aims to enhance the performance of a predictive model by selecting features that have high relevance to the target output while also minimizing redundancy between the selected features.

Two primary phases comprise the mRMR algorithm. First, features are ranked based on their relevance to the output variable using a relevance metric such as mutual information. Second, it selects unique, minimally redundant features to create a cohesive set with complementary information, making the model more robust and understandable (Hanchuan Peng et al., 2005) .

2.2.2 Maximization of Conditional Mutual Information (CMIM):

In CMIM, candidate features and target variables are measured mutually in order to quantify the amount of information that each candidate feature contributes to the target. Contrary to other methods, CMIM takes conditional mutual information into account. By considering other features that have been selected, this additional level of analysis examines the correlation between a particular feature and the target variable. In other words, CMIM evaluates how much new information a feature adds to the prediction while taking into account the existing feature set.

By incorporating this conditional perspective, CMIM seeks to achieve a balance between relevance and duplication. It favors features with high predictive power for the target variable as well as those that provide distinctive insights that are less likely to be redundant with other chosen features (G. Brown et al., 2012).

2.2.3 Extreme Gradient Boosting (XGB):

Extreme Gradient Boosting (XGB) is a well-known feature selection technique as well as a highly esteemed machine learning algorithm. In XGB, feature importance is assessed using a scoring mechanism that takes into account the importance of each attribute when building boosted decision trees inside the model. These decision trees' importance scores rise as they depend more heavily on particular attributes for critical decisions, indicating that they are becoming more and more important to the modeling process. This approach fits with XGB's reputation for effectiveness and predictive abilities, as well as helping to identify key

features needed to improve the model's predictive capacity and decision-making prowess. It builds a series of decision trees that iteratively correct errors, improving overall prediction accuracy. Additionally, by integrating regularization, XGB makes it simple to create unique loss functions, handles missing data, and excels at parallel processing, cross-validation, and hyperparameter tuning (T. Chen & Guestrin, 2016).

2.2.4 Information Gain (IG):

Information gain (IG) is a crucial technique for feature selection, analyzing the contribution of each variable to the target variable. The procedure involves evaluating the information gain of each independent feature and quantifying their individual effects. The computed information gain value's subsequent arrangement in descending order then ranks the features according to their individual relevance. A threshold is established and used as a cutoff point to speed up the feature inclusion procedure in machine learning models. All features that surpass this cutoff in terms of information gain are then seamlessly incorporated into the machine learning algorithms, ensuring that the features chosen have the greatest influence on the model's predictive performance and insights. This systematic approach improves the model's efficiency, interpretability, and overall efficacy in identifying meaningful patterns in the data (Kent, 1983).

2.3 Machine Learning Classifiers

Machine learning algorithms play a critical role in predicting gene-disease relationships and uncovering complex interconnections between genes and diseases using a knowledge-driven approach. The creation of predictive systems using computerized statistical and mathematical procedures and data-driven inferences is the focus of this field of study. Machine learning includes the development of complex models that can analyze data and produce insightful results. This method stands out for its ability to learn from data to increase the accuracy of predictions. Machine learning algorithms in this case learn from datasets, adapting their decision-making to the complex details of the data, in contrast to rigid rule-based techniques. In fields like bioinformatics, where it facilitates the investigation of the genetic factors underlying diverse diseases, this flexibility has enormous promise. Machine learning is guiding improvements in gene-disease association predictions by expanding its capabilities beyond conventional constraints, fostering discoveries that deepen our understanding of complex biological systems. Based on this purpose, we used traditional machine learning classifiers in this thesis.

2.3.1 Random Forest Classifier (RF):

During the training phase, the Random Forests (RF) ensemble learning approach builds an enormous quantity of decision trees and estimates the class based on the type of task. It can be applied to classification, regression, and other applications (Ho, 1995). In basic terms, the algorithm builds a decision tree for each sample, from which it derives the estimated value result. For each value resulting from the prediction, a vote is conducted. The algorithm then creates the outcome by selecting the last guess with the highest number of votes. Random forest parameters may increase the model's prediction ability or simplify the training procedure. Numerous parameters are employed and improved to boost the algorithm's performance. The "Max_features" option specifies the maximum number of features that Random Forest is able to test in each tree. The "n_estimator" parameter refers to the number of trees you want to construct prior to taking the most possible predictions into consideration.

The "max_depth" parameter of a tree in Random Forest is the longest path from the root to the leaf node.

2.3.2 Support Vector Machines (SVM):

The support vector machine is one of the discriminative classifiers used in machine learning. Finding a hyperplane that discriminates between two or more classes is the main goal of SVM. SVM can be used to classify datasets that are both linear and nonlinear. There are numerous ways to use linear separation to separate data from different classes. A non-linear mapping is utilized because a linear decision boundary cannot separate the data in a nonlinear way. A higher-dimensional space is used to transfer the data samples that were taken from the input feature space, and a linear hyperplane that separates the data samples in the new space is identified. By expressing the optimization issue in dual space and applying the kernel approach, it can be solved without explicitly transferring the data points to the new space (Cortes & Vapnik, 1995).

The two most crucial hyperparameters for SVM are The C and gamma parameters. The C determines the classifier's penalty. If the C is really large, the margin will be very small because incorrect training would carry a heavy cost. The penalty will be minimal, and the margin will be high if the C is low. The gamma parameter sets the range within which a single training point influences the model. Small gamma values indicate a wide similarity radius, resulting in the clustering of several points. The points must be relatively close to one another in order to be grouped together (or placed in the same class) with high gamma values.

2.3.3 AdaBoost:

Boosting techniques are intelligent ways of creating a strong learner by combining many weaker ones. It works by training these learners step by step. First, a weak learner is used to start training. Then, the algorithm pays special attention to the mistakes it made in the first round and trains again, focusing more on fixing those mistakes (R. Wang, 2012). This makes the learner smarter over time. In the AdaBoost Algorithm, we have some important settings. The "base_estimator" setting helps pick the type of weak learner we use. The "n_estimators" setting decides how many of these weak learners we use. And there's a "learning_rate" setting that lets us control how much each learner's contribution matters. By adjusting these settings, we guide AdaBoost to learn and predict better, making it a strong and accurate tool for various tasks.

2.3.4 LogitBoost:

Among the algorithms developed to address AdaBoost's overfitting problem is the boosting classification algorithm LogitBoost. It provides a solution by gradually reducing training errors in a linear fashion. Both LogitBoost and AdaBoost utilize an additive logistic regression technique. The type of loss they seek to minimize, however, is where they differ from one another. While LogitBoost concentrates on reducing logistic loss, AdaBoost concentrates on reducing exponential loss (Friedman et al., 2000).

The hyperparameters of the LogitBoost Algorithm match those of the AdaBoost Algorithm. As a result, their primary settings are the same. By utilizing the well-known control mechanisms in this way, LogitBoost can improve the accuracy of its predictions and fine-tune its learning algorithm. This synchrony in hyperparameters makes parameter selection

easier and emphasizes how both algorithms are flexible and adaptable to different problem domains. The switch in LogitBoost from exponential to logistic loss strengthens its ability to address overfitting issues, resulting in a useful tool for developing reliable and accurate classification models.

2.3.5 Decision Tree:

The decision tree algorithm constructs a model for classification or regression by utilizing a tree-like structure. An interconnected decision tree grows in parallel when the program splits the dataset into segments that are smaller. Both classification and regression are accomplished using this machine learning technique. The decision tree is a graphic representation of a series of options and the possible outcomes connected to those options. It starts at the root node and moves up the tree, using the property values at each node to direct its choices (Breiman et al., 2017). This strategy is advantageous in the field of data science due to its clarity and interpretability. Moreover, decision trees can manage both continuous and categorical data types proficiently.

The parameter termed "criterion" dictates the manner in which the quality of a split within a decision tree is assessed. Conversely, the "Max_Depth" parameter sets a limit on the deepest level that the tree can reach. The "Min_Samples_Split" parameter sets a minimum threshold for the number of samples required to start a division at an internal node. The "Min_Samples_Leaf" argument specifies the minimum number of samples needed to create a leaf node. The "Max_Features" parameter affects the number of features used when determining the best split. All of these settings have predetermined values.

2.3.6 k-Nearest Neighbor:

The k-nearest neighbor (kNN) algorithm is a leading contender among supervised learning techniques, as it effectively addresses both classification and regression problems. This methodology organizes data classification by employing a democratic process, taking into account the consensus of the "k" closest points to an unmarked data point. Its mode of operation involves examining latent data and sifting through the training dataset to identify the "k" most similar instances. Several metrics, including but not limited to the Euclidean distance and the Hamming distance, can be used to measure the spatial separation between two data points (Fix & Hodges, 1989).

In the realm of parameterization, the "n_neighbors" parameter plays the role of specifying the count of neighboring points to be considered. Meanwhile, the "metric" parameter assumes the task of dictating the yardstick for distance measurement when assessing similarity. This ensures that the algorithm adeptly determines the proximity between instances while bearing in mind the chosen distance metric.

2.3.7 Stacking:

Stacking, a prevalent technique in the field of ensemble machine learning, aims to improve model performance by coordinating a multitude of predictors. This methodology entails assembling a variety of distinct models, combining their insights, and developing a completely novel model with improved performance. Utilizing stacking to address interconnected problems results in the development of a more potent and adaptable predictive model (Lu et al., 2023).

CHAPTER THREE

LITERATURE REVIEW

Gene expression datasets offer useful insights into the intricate molecular pathways involved in many biological processes and diseases. Examining these datasets to pinpoint important genes or biomarkers is crucial in the fields of bioinformatics and biomedical research (van ’t Veer et al., 2002). Gene selection approaches can be classified into traditional gene selection and integrated gene selection. Traditional approaches use statistical and computational analysis of gene expression levels, whereas integrative methods use domain knowledge from external biological sources to assist in gene selection (Perscheid, 2021; Yousef et al., 2020).

To identify genes that exhibit statistically significant changes in expression levels across diverse conditions, the traditional gene selection methods rely heavily on a computational and statistical analysis. Differentially expressed genes are a potential indicator of biological significance, which is why these methods seek to identify them. Lazar et al. (Lazar et al., 2012) provided a brief overview of filter strategies used for selecting features in gene expression microarray analysis, considering factors like accuracy, scalability, robustness, and interpretability. Different approaches, such as statistical tests, correlation-based methods, and information theory, were discussed, highlighting the importance of selecting relevant features and avoiding overfitting. The authors concluded that choosing an appropriate feature selection method depends on the dataset's characteristics and the research question being addressed.

In a comparison between two different feature selection methods in DNA microarray analysis, filter and wrapper methods, Inza et al. (Inza et al., 2004) examined the performance of these methods on several publicly available DNA microarray datasets, comparing their accuracy, robustness, and computational efficiency. The authors determined that the selection between filter and wrapper approaches is contingent upon the specific attributes of the dataset and the research question being investigated, with filter methods generally being faster and less prone to overfitting, while wrapper methods may provide better predictive accuracy. Meanwhile, Anggraeni et al. (Anggraeni et al., 2021) evaluated the performance of filter and wrapper feature selection methods in the context of spam comment classification. The authors used three different datasets of comments collected from social media platforms and applied filter and wrapper methods to select the most relevant features. The results showed that both the filter and wrapper methods effectively reduced the dimensionality of the data and improved the classification accuracy. However, the wrapper method outperformed the filter method regarding accuracy and stability across different datasets.

The traditional approach to gene selection has various limitations. For instance, the filtering method examines the relevance of each gene separately without considering their interrelationships and connections. The primary drawbacks of such approaches are their biological interpretation challenges, low robustness, and inability to generate new biological information, as confirmed by multiple studies (Mungloo-Dilmohamud et al., 2020; Pes et al., 2017; Zhang et al., 2013).

Multiple studies finding genes associated with human disorders have led to the development of diagnostic tools and, in some instances, the creation of innovative medications. Numerous computational tools that vary in their techniques and resource utilization have been developed, including ones that integrate diverse biological data into machine learning. Liekens et al. presented BioGraph, a computational tool for unsupervised biomedical knowledge discovery. The tool uses automated hypothesis generation to extract new insights from large-scale biological data, such as gene expression profiles and protein interactions. They inferred that BioGraph is a valuable tool for hypothesis generation and data-driven discovery in biomedical research. Qi and Tang (Qi & Tang, 2007) proposed a novel approach for feature selection in microarray data analysis by integrating Gene Ontology (GO) information. They calculated the discriminative power of genes using Information Gain (IG) based on gene expression and class labels. They weighed it based on GO annotations, demonstrating improved accuracy and biological relevance compared to other methods and emphasizing the enhanced interpretability of analysis results.

Papachristoudis (Papachristoudis et al., 2010) introduced a feature selection method for microarray data using Gene Ontology (GO) annotations. By constructing a gene-GO bipartite graph and calculating the relevance of GO terms using either the hypergeometric distribution or mutual information, the authors showed better performance than previous techniques for selecting features in classification examinations.

Fang et al. (Fang et al., 2014) introduced an integrated approach machine learning approach for gene selection in microarray analysis, combining Relief F for feature selection, the chi-square test for association analysis, and SVM for classification. The approach outperformed existing methods in terms of accuracy, sensitivity, and specificity while utilizing a smaller number of genes, emphasizing the importance of integrating multiple techniques for improved classification accuracy. The authors in (Raghu et al., 2017) proposed a gene selection method incorporating genetic metainformation from KEGG and DisGeNET. The method calculates importance scores based on gene-disease associations and gene expression levels. It then uses a gene distance metric to choose relevant and diverse gene sets, showing better predictive modeling performance than variance-based gene selection techniques. Wang et al. (J. Wang et al., 2018) proposed an integrative approach to identify gene-disease associations by combining GWAS, eQTL, and PPI network data. The study demonstrated improved accuracy in predicting gene-disease associations, revealing potential associations, such as CD40, CCL5, and CDK6, with rheumatoid arthritis (RA), showcasing the approach's effectiveness in discovering novel gene-disease relationships.

In their study, Giambartolomei et al. (Giambartolomei et al., 2014) presented a Bayesian method that utilizes GWAS summary statistics to identify the colocalization of genetic association signals between study pairs. By considering prior knowledge and expected association patterns, the method revealed genomic regions where the same variant showed likely associations with both multiple sclerosis and ulcerative colitis, indicating potential shared genetic pathways underlying the two diseases. In another approach to improving

understanding of complex diseases. Greene et al. (Greene et al., 2015) devised a technique for creating tissue-specific networks that combines gene expression data and protein-protein interaction information. They found functional connections between genes and proteins in many tissues and discovered tissue-specific networks enriched in disorders like breast cancer, Alzheimer's disease, and type 2 diabetes enriched with known disease-related genes and pathways, as well as novel candidates not previously associated with the diseases.

Peng et al. (Peng et al., 2017) introduced SLN-SRW, a network-based disease gene prediction method that integrates heterogeneous biomedical data to identify disease-related genes. They applied their approach to diseases like Alzheimer's, breast cancer, and type 2 diabetes, revealing novel genes and potential therapeutic targets. In a study by Asif et al. (Asif et al., 2018), machine learning classifiers were trained on gene similarities computed by Gene Ontology (GO) demonstrated effectiveness in identifying genes associated with complex diseases, specifically autism spectrum disorder (ASD).

Luo et al. (Luo et al., 2020) propose EdgCSN, an ensemble learning algorithm that integrates clinical sample-based networks with protein-protein interaction networks to predict disease-associated genes. They identified potential biomarkers for breast cancer and found shared genes across different diseases, suggesting common molecular mechanisms. Hamzeh and Rueda (Hamzeh & Rueda, 2019) developed a machine learning method utilizing the DisGeNET database to detect biomarkers in prostate cancer, using wrapper-based feature selection to group genes based on classification accuracy. Nikhil et al. (Acharya et al., 2017) presented an unsupervised gene selection technique for sample clustering using biological knowledge. Their approach involved constructing a gene co-expression network, identifying gene clusters enriched in specific biological processes, and selecting differentially expressed genes for improved sample clustering accuracy.

None of the above computational tools suggests a framework to embed the process of grouping selections into the core of the machine learning algorithms. Most of those algorithms integrate the knowledge as part of or prior to applying the ML algorithm, in some cases, to perform enriched analyses. Various methods that incorporate biological structure or information for selecting features are examined in the context of machine learning methodologies based on biological domain knowledge.

Yousef et al.(Yousef et al., 2007) introduced the Recursive Cluster Elimination (RCE) algorithm-based feature selection for gene expression data classification. They demonstrated the effectiveness of RCE in identifying informative gene clusters associated with sample phenotypes. In another study by Yousef et al. (Yousef et al., 2019), they presented maTE, a computational method for discovering expressed interactions between microRNAs (miRNAs) and target genes. maTE utilized both miRNA and gene expression data to identify biologically relevant and dynamically regulated miRNA-gene interactions, showcasing its effectiveness in breast cancer research. Besides, Yousef et al. (Yousef, Ülgen, et al., 2021) introduced CogNet, a computational method for classifying gene expression data based on ranked active-subnetwork-oriented KEGG pathway enrichment analysis. CogNet effectively classified breast cancer and glioma subtypes by utilizing gene expression data and pathway enrichment scores. In a different study (Yousef, Goy, et al., 2021), the authors proposed miRcorrNet, which performs machine learning-based integration of miRNA and mRNA gene expression profiles. The tool accurately prioritizes pan-cancer-regulating high-confidence miRNAs, showcasing its potential to uncover functional effects in complex diseases like cancer.

GediNET is explicitly formulated to address major shortcomings in the existing framework of gene-disease association research, notably the absence of group-based gene selection and the necessity for comprehensive integration of disease associations. Traditional methods frequently emphasize singular genes, neglecting the valuable insights derived from interactions among groups and the complex connections of diseases. The G-S-M (Grouping, Scoring, Modeling) structure of GediNET effectively mitigates these constraints by categorizing genes according to common disease connections, hence augmenting the biological significance and interpretative capacity of the analysis. This classification facilitates a more refined comprehension of gene interactions and their cumulative effects on diseases, establishing a robust basis for predictive modeling and the identification of therapeutic targets.

Current methodologies in gene-disease association research encounter substantial obstacles with feature selection, including overfitting and difficulties in generalizability across many biological contexts. Numerous contemporary approaches lack strong mechanisms to guarantee that the picked features are both statistically significant and biologically pertinent, thereby compromising their efficacy in clinical applications. GediNET's methodology improves the generalizability and applicability of its conclusions by incorporating a more rigorous scoring system and utilizing empirical evidence from extensive datasets. In-depth examination of pivotal findings in the field indicates an urgent want for models such as GediNET that merge methodological precision with profound biological understanding, thus reinforcing the argument for sophisticated integrative approaches in genomic research.

CHAPTER FOUR

GEDINET FOR DISCOVERING GENE ASSOCIATIONS ACROSS DISEASES USING KNOWLEDGE-BASED MACHINE LEARNING APPROACH

4.1 Introduction

Common methods for identifying disease-associated genes involve machine learning and various feature selection techniques to pinpoint important genes that can act as biomarkers for a particular disease. Recently, incorporating prior knowledge-based methods into this process has demonstrated considerable potential for identifying new biomarkers with practical applications.

In this research, a novel approach called GediNET is developed that integrates prior biological knowledge to Group genes associated with diseases such as cancer. The novelty of GediNET is that it discovers gene associations across diseases rather than single disease-gene associations. The selected gene groups are subjected to a Scoring component to identify the top performing groups rather than single feature selections. The highest-ranked gene groups are subsequently utilized to train the Machine Learning model. The Grouping, Scoring, and Modeling (G-S-M) model is used to discover groups of disease associated genes or biomarkers for specific groups/diseases. One of the outputs of GediNET is a list of disease groups that can combine their gene signatures to identify new biomarkers and potential drug targets. GediNET identifies these relationships through Disease–Disease Association (DDA) based machine learning. DDA explores novel associations between diseases and identifies relationships that could be used to further improve approaches to treat cancers.

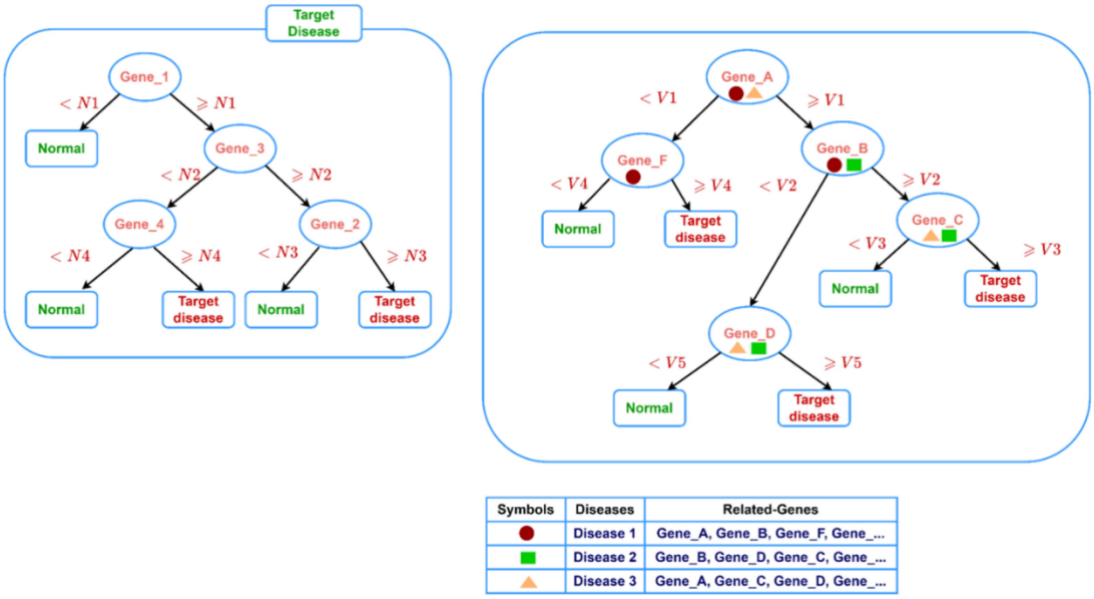
4.2 The merit of GediNET in the discovery of Disease -Disease Associations

Let *D* be a two-class gene expression dataset designed to study a specific disease (for example, Lung Cancer or Breast cancer) in order to detect significant genes that will serve as biomarkers. The traditional approach to the classification model suggests a list of *k* genes that can serve as biomarkers for predicting the number of patients with the disease. In other words, identifying disease-gene associations. One solution could be a linear function F(X) that might be expressed as:
F(X) = $w_1 g_1 + w_2 g_2 + \ldots + w_k g_k$, where w_i are the weights (scores) while the g_i are the gene expression values. The weights indicate the importance (significantness) of each gene expression for the linear model. For instance, a value with a weight close to zero indicates

that the associated genes contribute less to the equation model. In other words, F(X) describes the biological interaction between those *k* genes to form biomarkers.

GediNET differs from traditional approaches by considering groups of genes, not individual ones. The group represents a biological pre-existing knowledge about the association between genes and disease. In GediNET, a group is represented as a set of genes associated with a disease. GediNET scores those groups and their contribution to the classification task by applying the S component of GediNET (see Section (The S component)). The top j-scored groups will be used for training the final model. In other words, the genes that appear on those *j* groups will be used to train the machine learning model. The S component relies on representing the group of genes as a subdataset of the original dataset *D*, preserving the class labels, as described in detail in the two following sections (Grouping Genes Based on Disease (The G component) and Creating a Sub-dataset).

For simplicity, the final model might be visualized as a decision tree, as illustrated Figure 4.1 (Right panel). The left panel of Figure 4.1 illustrates the decision tree model of the significant genes selected by the traditional approach. The right panel of Figure 4.1 shows



that the decision tree model consists of genes associated with the top three GediNET ranked diseases (groups). This model contains information about the biological knowledge of the diseases and shows the disease-disease associations.

Figure 4. 1: Decision Tree model. The left panel illustrates the traditional approach that detects gene-disease associations, while the right panel illustrates the disease-disease association as the output of GediNET.

For example, considering the dataset GDS1962 that studies the Glioma disease, GediNET suggests a model that is based on the top three significant groups/diseases, as follows: Grp1_disease= {PAPILLARY RENAL CELL CARCINOMA}, Grp2_disease= {PLASMA CELL}, and Grp3_disease= {NEOPLASM and ADULT GLIOBLASTOMA}.

The following are the sets of genes associated with each disease:
Grp1_genes= {SLC16A1, TAGLN2, TIMP3, IGFBP7, TOP2A, TP53, RRM2…}, Grp2_ genes = {CD99, TP53, LPL, CD40, CD38, NCAM1, MYC, CSF3, CDKN2A, FGFR3, CCND1}, and Grp3_genes= {EDNRA, CSPG4, MELK, ENPEP, …}.

Applying GediNET will compute F*(x) that describes the association between the Grp1, 2 and 3_diseases with the disease under study (in this case, Glioma disease). This might lead to new discoveries that have not been observed before by traditional approaches.

4.3 The G-S-M components of GediNET

GediNET utilizes the generic technique G-S-M, which is also used by several tools like SVM-RCE (Yousef et al., 2007), SVM-RCE-R (Yousef, Jabeer, et al., 2021), SVM-RNE (Yousef et al., 2009), maTE (Yousef et al., 2019), CogNet (Yousef, Ülgen, et al., 2021), miRcorrNet (Yousef, Goy, et al., 2021), Integrating Gene Ontology (Yousef, Sayıcı, et al., 2021), miRModuleNet (Yousef, Goy, et al., 2022), and PriPath (Yousef, Ozdemir, et al., 2022), and was recently reviewed in Yousef et al. (Yousef et al., 2020). The main workflow of GediNET is illustrated in Figure 4.2, where the G-S-M approach is presented in the three main sections labeled with the orange section (G), the yellow section (S), and the green section (M), which represent:

1. The G Component (Grouping): where the genes are grouped according to the biological pre-existing knowledge of disease. Each group is represented by an extracted two-class subdataset from the main dataset.
2. The S Component (Scoring): where the groups are scored and ranked by considering the related two-class subdatasets.
3. The M Component (Machine Learning model): where a model is developed by training a classifier (Random Forest) on the top ranked groups' genes.

GediNET requires a two-class gene expression dataset and a table containing biological pre-existing knowledge of the diseases as inputs. The dataset has two categories of samples: control (negative) and disease (positive). The dataset is divided into training and testing subsets. The training dataset is utilized for the G-S-M components, while the testing dataset is employed to assess the model's performance. The entire process is iterated 100 times using cross-validation, with the input randomly divided into 90% for training and 10% for testing in each iteration. A statistical t-test, specifically Levene's test for equality of variances, is conducted on the training dataset to identify the most significantly expressed genes (M. B. Brown & Forsythe, 1974). 2000 genes that show significant differential expression with a P-value below 0.05 are chosen. The primary contribution of the generic approach and the functions of each component are elaborated in the subsequent sections.

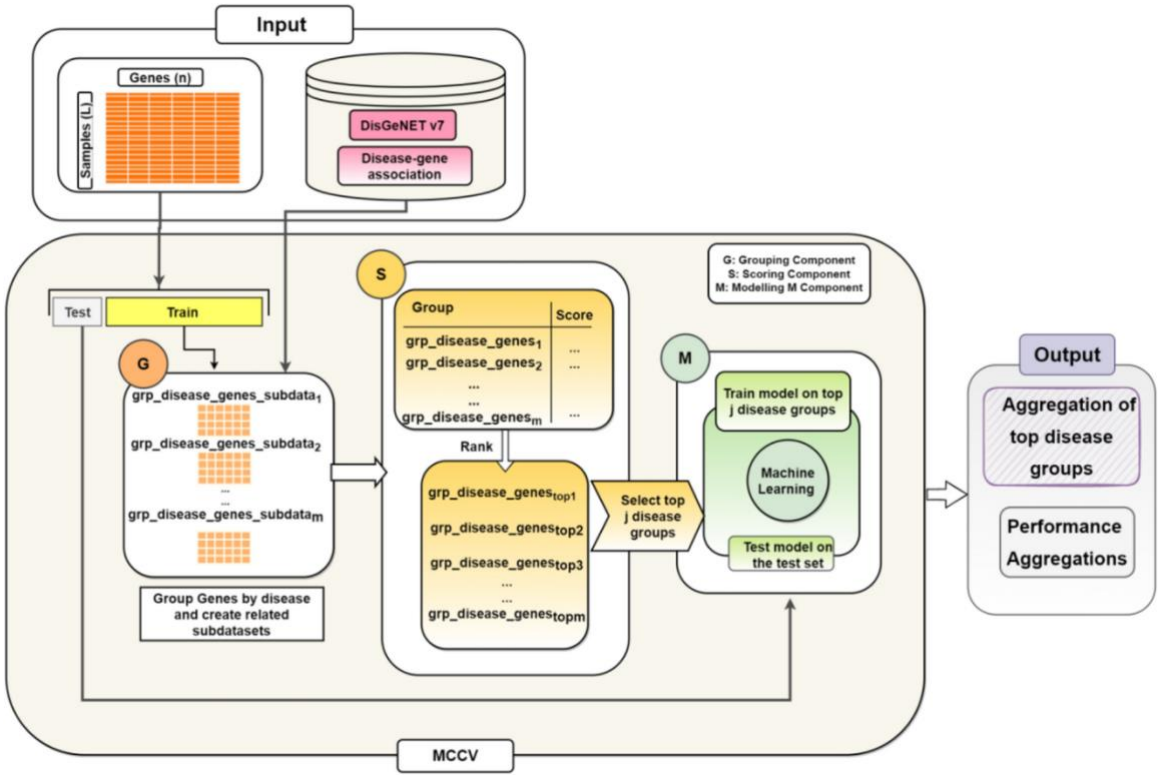


Figure 4. 2: GediNET workflow. The main workflow of G-S-M that integrates pre-existing biological knowledge for grouping genes based on disease-gene association, which is derived from the DisGeNET v7 database.

4.3.1 G Component: Grouping Genes based on Disease:

The first component of GediNET is the grouping component G (the orange section in Figure 4.2), which groups genes into groups. The G component might be based on any pre-existing biological knowledge, such as miRTarBase, KEGG pathway, etc., for creating groups of genes. In this tool, the G component group genes are based on the DisGeNET v7 database (Piñero et al., 2017), which is a gene-disease association database. Table 4.1 is an example of such a group that includes the disease name (group name), the $_{set}$ of genes associated with this disease, and the number of genes in the associated group.

Table 4.1-A: An example of groups of diseases with their associated genes. The last column represents the group size.

Group Name	Genes	#Genes
Small Cell Carcinoma of Lung	VPS13B, SLC16A1, ANXA1, CD99, SMARCC1, PCNA…	41
Leukemia, B-Cell	TP53, LAMA4, STK11, CSPG4, CD40, TNFRSF1A…	43
Stage III Breast Cancer Ajcc V6	TP53, BRCA2	2

Table 4.1-B: An example of groups of diseases with their associated genes. The last column represents the group size.

Group Name	Genes	#Genes
Head And Neck Carcinoma	PRMT5, ANXA1, LGALS1, TIMP3, IGFBP7, PCNA, TNC, TP53…	149
Secondary Malignant Neoplasm of Bone	ADAM9, SLC16A1, CD99, NME1-NME2, DPYSL3, TNC, TP53, NRAS…	145
Malignant Glioma	TK1, NPAS3, CD63, HMGB1, TAGLN2, TXNIP…	162
Adenocarcinoma, Tubular	PCNA, TP53, EFEMP1, APOE, STK11, PRKD1…	31
Childhood Brain Neoplasm	TP53, NRAS, SOX9, MYC, TNFRSF11B	5
Adult Myelodysplastic Syndrome	CSNK1A1, CTNNA1, HMGB1, PCNA, TOP2A, TP53…	58
Non-Small Cell Lung Cancer Stage I	TP53, PRRX1, IGFBP3, VEGFA, S100A6, GSTK1…	22

4.3.1.1 G Component: Creating Two-class Subdataset:

We assume that *D* consists of columns that represent the gene expressions, while the rows represent the samples. *D* also has a class label column with information about each sample, as illustrated in Figure 4.3 at the Input panel (labeled by I).
To score each group, we have created a two-class subdataset related to each group/disease. Each subdataset is specific to one group/disease and contains the genes belonging to this group/disease. This is achieved by extracting the gene columns belonging to the specific group and their original class label from the original dataset *D*. Let *m* be the number of groups. In this stage, we will extract or create m two-class subdatasets that will be input to the S (Scoring) component. In Figure 0.3, the I panel (input panel) contains two matrices. The left one is an example of the gene expression matrix *D*, with the class label appearing in column "Class". The right one is the pre-existing biological knowledge containing the disease name (group name) and its set of genes. In our example, the right matrix contains four group diseases labeled with group_diseasei, *i=1...,4*. For example, group_disease1 represents the disease named "Well Differentiated Pancreatic Endocrine Tumor," along with three genes associated with this specific disease. The genes are RBMS3, TFE3, and NTRK1.

Within the G panel, the extraction of two-class subdatasets is performed. As evident in Figure 4.3, four subdatasets are created. For each subdataset, the gene columns belonging to each disease group are extracted from the *D* dataset with the original class label, where pos is for the positive class and neg is for the negative class. The four subdatasets serve as input to the following component, S, to be scored and ranked.

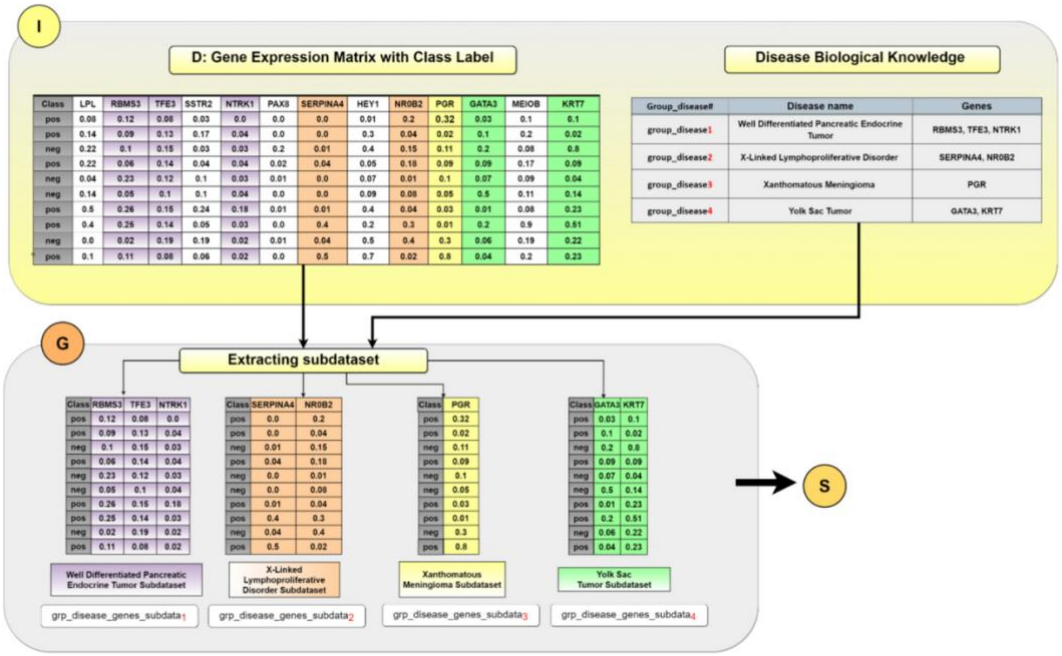


Figure 4. 3: An example of creating two-class subdatasets extracted according to disease-group names. These subdatasets will be subject to the S component for scoring.

4.3.2 S Component: Scoring the Groups:

As a result of the G component, *m* two-class subdatasets are created, each representing one group. The task of the S component is to compute a score that measures to what extent it is differentially expressed considering the given two classes. The group is a set of genes; one way of computing a group-score is by computing each individual gene's t statistic and then averaging those scores to be the final score of the group, as suggested in (Nacu et al., 2007). The following equations might be used to compute this score for a given gene *i*:

$$T_i = (\mu_{i\_pos} - \mu_{i\_neg}) / \sqrt{\sigma^2_{i\_pos}/n_1 + \sigma^2_{i\_neg}/n_0} \qquad (1)$$

Where μ_{i_pos} and μ_{i_neg} are the average expressions over the pos and neg classes, respectively. σ_{ipos} and σ_{ineg} are the standard deviations over both classes, while n_1 is the number of positive samples and n_0 is the negative samples.

Based on equation number 1, one might compute a score for a given group that consists of k genes as follows:

$$S(group) = \frac{1}{k}\sum_{i=1}^{k} T_i \qquad (2)$$

However, GediNET uses a more progressive approach based on machine learning to compute such scores. Figure 4.4 illustrates the steps of the S component that end by assigning the performance measurement as the group score. In our case, we consider accuracy. Each two-class subdataset is randomly split into training and testing (90% of the data is allocated for training, while 10% is reserved for testing), as shown in Figure 4.4, Panel S-Splitting, where this procedure is repeated *r* times. The training is used to train the machine learning algorithm (we have used Random Forest), and the performance is evaluated on the test split as seen in the Panel, S-FitTestModel. The accuracy average of the *r* splits is computed to form the group score. All of the group scores are collected to form a table of *m* scores. For the M component, we perform a ranking step by ordering the table in descending order. An

example of the Scoring component applied to the GDS2545 dataset is presented in Table 4.2.

Table 4.2: An example of the output of the Scoring S component. The first column is the name of the group disease, the Score column is the computed score computed by the S component, and the rank column is the rank of the group.

Disease	**Score**	**Rank**
PAPILLARY RENAL CELL CARCINOMA	0.98	1
PLASMA CELL NEOPLASM	0.96	2
ADULT GLIOBLASTOMA	0.94	3
INTESTINAL CANCER	0.91	4
MALIGNANT NEOPLASM OF COLON STAGE IV	0.89	5
DERMATOFIBROSARCOMA	0.87	6

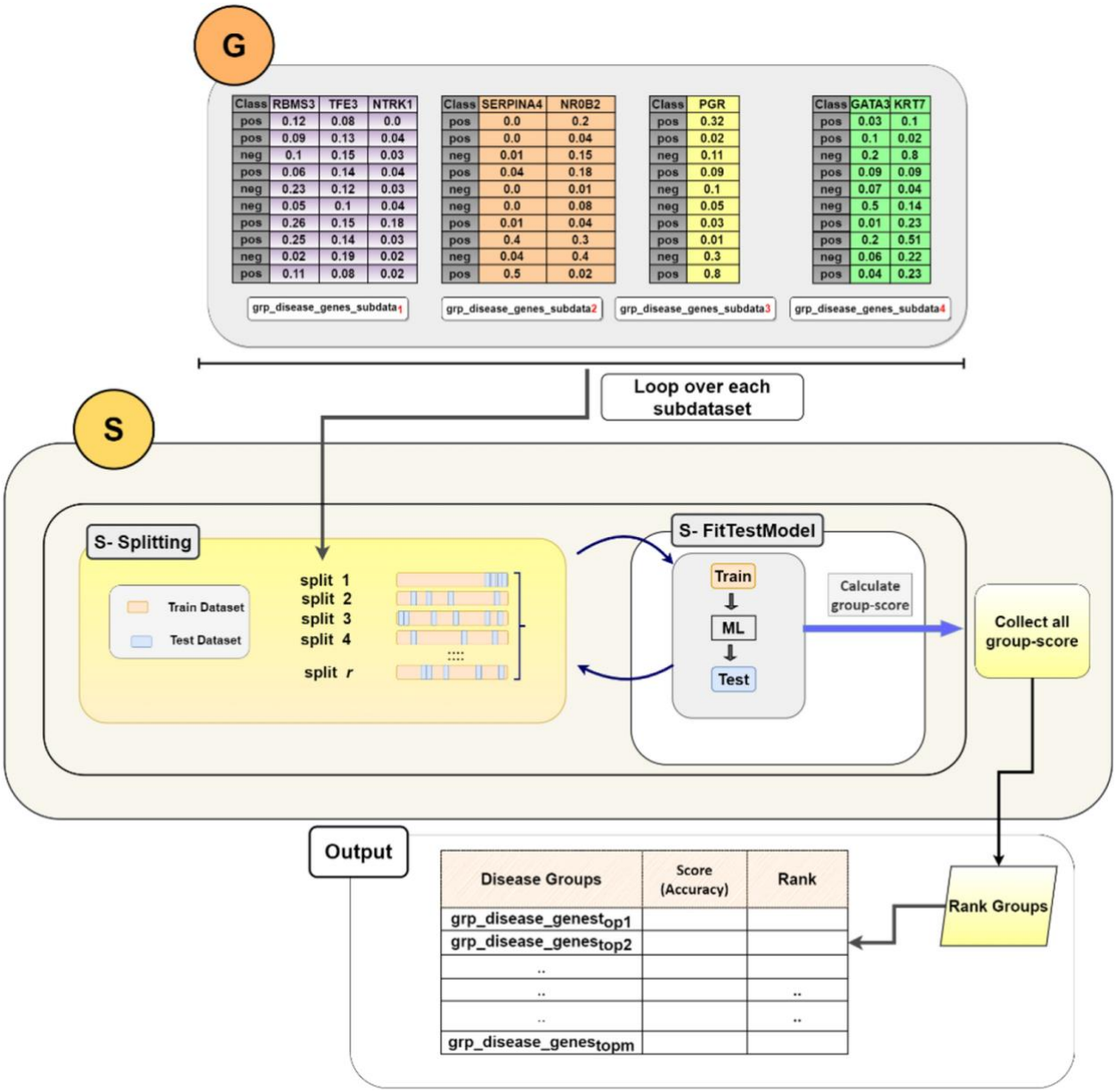


Figure 4. 4: The details of the S component. The G panel contains all the two-class sub-datasets that each one is subject to the S component.

GediNET uses the accuracy measurement to assign a score; one might use a different measurement or a combination of measurements (such as sensitivity, specificity, the AUC, etc.). For more information on such an option, we refer to (Yousef, Jabeer, et al., 2021).

4.3.3 M component: Fitting the Model:

The M component considers the top-ranked j groups of disease, and their genes are merged to form the top-ranked associated genes (as seen in Figure 4.4, the output panel). A subdataset is extracted considering the top-ranked associated genes from the training part of the dataset (90% training, 10% testing, as mentioned before). The retrieved subdataset is used to train an RF model. The model is assessed on the testing dataset consisting of those

genes, and the performance statistics are recorded. We have reported the performance of $j = 1...,10$.

Our approach involves training Random Forest classifier on randomly selected data, with 90% used for training and 10% for testing. Settings can be modified in our KNIME version of GediNET.

4.4 Implementation of GediNET

We have utilized the GediNET tool by integrating it with the KNIME platform, which is free and open-source, due to its straightforward and user-friendly graphical interface. KNIME is a versatile platform that allows us to combine Python and R scripts to create a KNIME workflow (Berthold et al., 2008). The workflow created on KNIME comprises several nodes with separate functions.

Figure 4.5 displays the KNIME workflow for GediNET. The process begins by uploading a list of the dataset names using the "List Files/Folders" node. A loop iterates across the datasets to be read by the "Table Reader" node and then processed by the meta-node "FilterMissingValues" to eliminate or filter rows with missing values. The filtered data is sent as input to the GediNET meta-node. Use the "Integer Input" node to adjust the amount of iterations during model training.

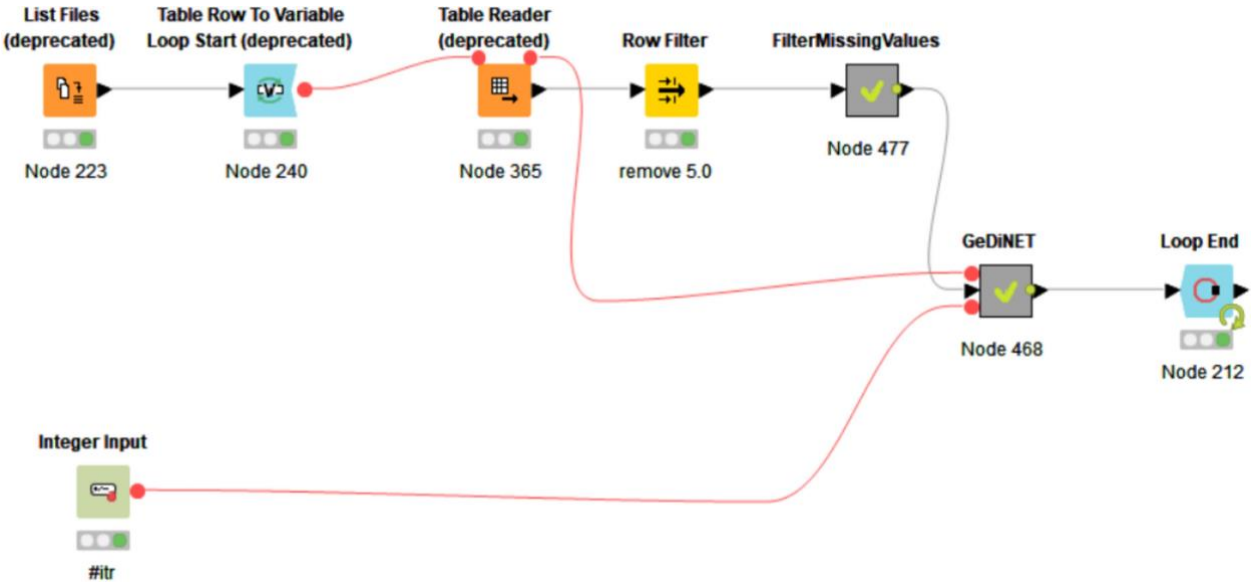


Figure 4. 5: GediNET workflow in KNIME.

You can download the GediNET KNIME workflow from: https://github.com/malikyousef/GediNET.git or https://kni.me/w/3kH1SQV_mMUsMTS-

4.5 Model Performance Evaluation

We utilized the Random Forest Classifier to partition the data into 90% for training and 10% for testing. We used the under-sampling method to address the imbalance in the datasets, where the class label distribution is uneven. This strategy addresses imbalanced datasets by retaining all samples in the minority class while reducing the size of the majority class. We utilized 100-fold Monte Carlo cross-validation (MCCV) for model training. Monte Carlo

cross-validation (MCCV) involves randomly selecting fractions of the samples as training data and assigning the remaining as test data. The performance metrics are calculated by averaging the results of 100-fold MCCV. We choose for MCCV over standard CV due to its higher repeatability resulting from lower variance.

Quantitative measures were calculated to assess the performance of the RF model, such as accuracy, sensitivity, and specificity (El-Hadj Imorou, 2020), using the following formulations:

$$\text{Sensitivity (SEN)} = TP/ (TP + FN). \quad (3)$$

$$\text{Specificity (SPE)} = TN/ (TN + FP). \quad (4)$$

$$\text{Accuracy (ACC)} = (TP + TN) / (TP + TN + FP + FN). \quad (5)$$

Where TP=true positive; FP = false positive; TN = true negative; and FN = false negative. The Area Under the Curve (AUC) quantifies a classifier's capability to differentiate between classes and serves as a condensed representation of the ROC curve. We utilized the AUC metric to assess the performance outcomes (Hand & Till, 2004).

Our method produces lists of disease groups and their corresponding genes that vary slightly in each iteration. Therefore, a prioritization strategy should be applied to those lists. We employed rank aggregation methods in miRcorrNet. We have incorporated the RobustRankAggreg R package, created by Kolde and colleagues in 2012, into the GediNET workflow (Kolde et al., 2012). RobustRankAggreg assigns a p-Value to each element in the aggregated list, indicating how well each element/entity was rated in comparison to the expected value

4.6 Results

4.6.1 Performance Evaluation of GediNET:

Table 4.3 displays the average 100-fold MCCV performance of GediNET for the top 10 groups in the GDS1962 dataset. The final row displays the performance of the highest-ranked group (#Groups = 1). An AUC of 97% was achieved by utilizing an average of 21.61 genes. The row (#Groups = 2) displays performance indicators for the top 2 groups by aggregating the genes of the first and second-ranked groups. GediNET reports the cumulative performance results for the top 10 groups.

Table 4.3: An example average of 100 MCCV performance table for GediNET for the top-ranked 10 groups for the GDS1962 dataset cumulatively.

#Groups	#Genes	Accuracy	Sensitivity	Specificity	AUC
10	136.74	0.928	0.93	0.92	0.98
9	127.68	0.93	0.93	0.92	0.98
8	116.02	0.93	0.94	0.92	0.98
7	111.16	0.93	0.93	0.91	0.98
6	102.02	0.93	0.9	0.92	0.98
5	92.88	0.93	0.93	0.93	0.98
4	78.37	0.93	0.93	0.92	0.98
3	62.47	0.93	0.94	0.92	0.98
2	45.57	0.93	0.93	0.93	0.97
1	21.61	0.92	0.93	0.92	0.97

Table 4.4: Performance outcomes of GediNET compared to the highest-ranked group. ACC represents Accuracy, SEN represents Sensitivity, SPE represents Specificity, FM represents F-Measure, and AUC represents Area Under the ROC Curve.

GEO Accession	#Genes	ACC	SEN	SPE
GDS1962	45.57	0.93	0.93	0.93
GDS2545	113.76	0.73	0.72	0.74
GDS2771	97.83	0.64	0.69	0.59
GDS3257	74.81	0.97	0.99	0.94
GDS3837	21	0.92	0.83	1
GDS4206	83	0.66	0.3	0.82
GDS4516_4718	40.72	0.99	0.99	0.99
GDS2574	102.49	0.76	0.77	0.76
GDS3268	115.7	0.67	0.7	0.63
GDS5499	80.23	0.9	0.96	0.77

Table 4.4 shows the GediNET performance over 10 datasets for the top 2 groups. All values are the results of an average of 100-MCCV iterations while considering the AUC for presenting the performance. The complete performance results are attached in the

supplementary data. The table shows the GEO accession in the first column, the number of genes in column #Genes while ACC is the accuracy, SEN is the sensitivity, SPE is the specificity, and the AUC is the area under the curve. We see only one unsuccessful result for the dataset GDS4206. However, a similar observation was made when applying other tools to this specific dataset, as illustrated in Figure 4.6.

The average number of genes associated with the top 2 groups is slightly high because the distribution of genes over the disease is slightly high compared, for example, to other biological knowledge such as microRNA target or KEGG pathways. Moreover, this number of genes could be reduced by removing the least contributed genes when processing each group. This step will be considered in the future version of the algorithm. Also, one can use additional biological knowledge to filter out more genes from the group by, for example, leaving the most associated genes with the disease. The last suggestion requires other biological resources to be embedded into GediNET.

4.7 Comparative Evaluation with other biological G-S-M

We have examined similar methods like CogNet, maTE, and PriPath, which utilize the G-S-M strategy to incorporate biological information for gene grouping and scoring within the group. We have calculated the Area Under the Curve (AUC) values for the top 1-10 groups based on the scoring component of each tool using 100-fold Monte Carlo Cross-Validation (MCCV). We carefully examined the top two groups for comparison.

Figure 4.6 shows the average AUC values of the four tools over 10 datasets. Figure 4.7 displays the average number of genes for the four tools. The AUC values of GediNET, CogNet, maTE, and PriPath for the top two groups across 10 distinct datasets are almost identical, as shown in Figure 4.6. The performance of the tools is similar. This close performance suggests that the developed tool, GediNET, is reliable and stable. However, each tool produces a unique conclusion based on its individual strengths and is designed to identify important groups associated with particular pre-biological information.

Figure 4.7 indicates that GediNET utilizes, on average, ten times more genes than the other tools. The reason for this is that the number of gene groups associated to the diseases are extensive.

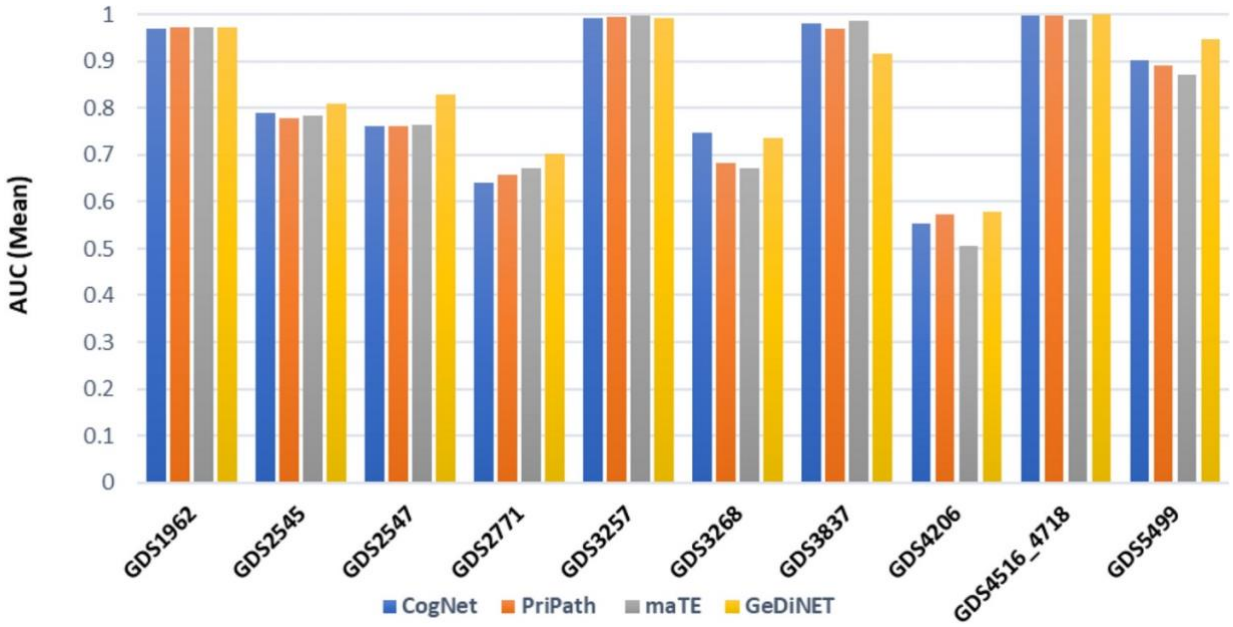


Figure 4. 6: The average AUC values of GediNET, CogNet, maTE, and PriPath for the top two groups across ten different datasets.

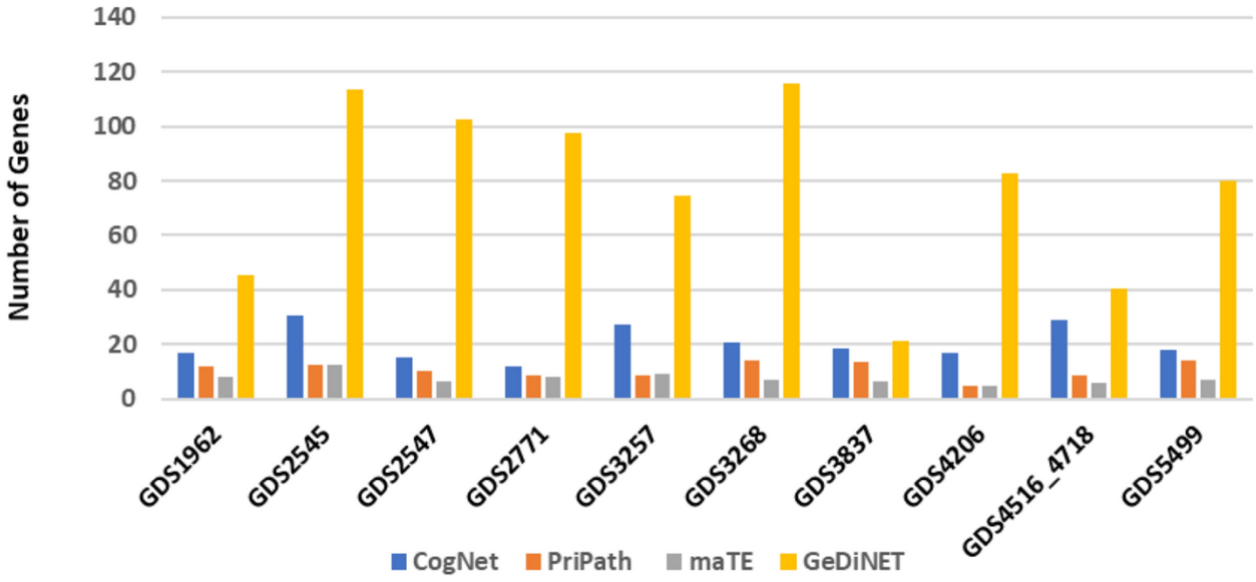


Figure 4.7: The mean number of genes of GediNET, CogNet, maTE and PriPath tools over the ten datasets for the top two groups.

One of the tool's outputs is a list of ranked disease groups that were assigned a p-value by the robust rank aggregation package (Kolde et al., 2012). Table 4.5 is an example of this tool for the GDS1962 dataset.

Table 4.5: An output of the RobustRankAggreg tool for the GDS1962

GDS1962			
Disease Name	**p-value**	**#Genes**	**List of genes**
PAPILLARY RENAL CELL CARCINOMA	0.00052	22	SLC16A1, TAGLN2, TIMP3, IGFBP7...
PLASMA CELL NEOPLASM	0.0010	11	CD99, TP53, LPL, CD40...
COMMON ACUTE LYMPHOBLASTIC LEUKEMIA	0.001772	3	KNG1, MME, BCL2
DUCTAL BREAST CARCINOMA	0.002363	13	TCF21, AFAP1L2, PLG...
GASTRIC MUCOSA-ASSOCIATED LYMPHOID TISSUE LYMPHOMA	0.002953	2	BCL2, EPCAM
NTRAHEPATIC CHOLANGIOCARCINOMA	0.003544	27	SHBG, BAX, TYMS, GPC3...
LYMPHOMA, NON-HODGKIN	0.004135	44	BAX, SLC23A1, MME, TYMS, …
MALIGNANT NEOPLASM OF COLON STAGE IV	0.004725	7	TYMS, MYCN, KLK6, NDRG1, …
NEUROECTODERMAL TUMOR, PRIMITIVE	0.005316	14	SFRP1, PCSK2, MYCN, CAPS...
PAPILLARY THYROID CARCINOMA	0.005907	75	BAX, PKHD1L1, MME, GPC3…

This is a novel output of the feature selection techniques that GediNET is providing. This table will be used to analyze the relationship between the diseases further. For example, Table 4.5 raises a biological question about the association between the top-ranked disease (PAPILLARY RENAL CELL CARCINOMA, PLASMA CELL NEOPLASM, ...) and the target disease of the study (dataset GDS1962 with target disease Glioma). Additionally, GediNET provides a list of significant genes that were also aggregated by the Robust Rank Aggregation tool. While scoring each group, the genes associated with the group is scored with the same score as the group. This list with its scores is aggregated at the end to compile and report a list of significant genes. Table 4.6 provides an example of such a list.

Table 4.6-A: Top 10 significant genes that were aggregated by the RobustRankAggreg tool for the GDS2545 dataset.

Genes	**p-value**
MYL1	0.003
RNF44	0.016
UBN1	0.051

Table 4.6-B: Top 10 significant genes that were aggregated by the RobustRankAggreg tool for the GDS2545 dataset.

Genes	p-value
N4BP2L1	0.060
GDI1	0.066
ARL17B	0.093
MYLPF	0.133

The user can consider the list of significant genes for functional and enrichment analysis as was done in similar studies such as PriPath and miRmodulnet using different tools such as David (*DAVID: Functional Annotation Tools*, n.d.), EnrichR (Kuleshov et al., 2016), and GeneMANIA (*GeneMANIA*, n.d.).

4.8 Biological Interpretations

One of the outputs of GediNET is a list of significant diseases that have been scored by the S component, as illustrated in Table 4.5. RobustRankAggreg has ranked this list according to p-value. For all 10 GEO datasets, the top 2 diseases and their set of genes were considered for pathway enrichment analysis. Their total number of distinct genes is 1184.
The web tool EnrichR (Kuleshov et al., 2016) was used to perform the pathway enrichment analysis. The tool was run to collect the top enriched pathways for each disease-gene group per dataset, and the top pathways (with the least p-values) were selected. The WikiPathway database (Martens et al., 2021) version 2021 for human genes was used to select our results. The top cell signaling pathways' names for the 10 GEO datasets, p-values, adjusted p-values, and associated genes are illustrated in Table 4.7. Evidence from the literature was then gathered for the dataset on cancer and the top-performing disease, along with the enriched genes and pathways found from the enrichment analysis.

Table 4.7- A: The top cell signaling pathways' names for the 10 GEO datasets. The first column is the name of the cell signaling pathway, the second column is the p-values, the third column is the adjusted p-value, the Genes column represents an example of the associated genes, and finally, the last column is the total number of associated genes.

Cell signaling pathways term	P-value	Adjusted P-value	List of Genes	#Genes
Head and Neck Squamous Cell Carcinoma WP4674	2.24E-13	6.31E-11	CCND1;CDKN2A; AKT1…	9
DNA damage response (only ATM dependent) WP710	2.95E-16	1.08E-13	GSK3B;SMAD4;CDKN1A,...	14

Table 0.7- B: The top cell signaling pathways' names for the 10 GEO datasets. The first column is the name of the cell signaling pathway, the second column is the p-values, the third column is the adjusted p-value, the Genes column represents an example of the associated genes, and finally, the last column is the total number of associated genes.

Cell signaling pathways term	P-value	Adjusted P-value	List of Genes	#Genes
Apoptosis WP254	1.88E-06	4.25E-04	CASP10;MYC;PMAIP1;...	6
Hepatitis C and Hepatocellular Carcinoma WP3646	5.41E-12	2.07E-09	CDKN1A;IL6;CXCL8;...	10
VEGFA-VEGFR2 Signaling Pathway WP3888	1.66E-10	6.37E-08	LRRC59;NRP2;PRKAA2;...	27
VEGFA-VEGFR2 Signaling Pathway WP3888	1.05E-11	2.59E-09	HSP90AA1;ANXA1;...	18
Lung fibrosis WP3624	6.32E-09	1.73E-06	GREM1;CSF3;IL6;PLAU;EGF; MUC5B;MMP9	7
IL-18 signaling pathway WP4754	2.33E-17	1.05E-14	GSK3B;CEBPB;CXCL8;...	29
Effects of nitric oxide WP1995	2.93E-05	0.00310457	NOS1;XDH	2
TP53 network WP1742	2.14E-13	9.13E-11	CDKN1A;CDKN2A;MYC;...	9

Next, we used the Cytoscape tool (Franz et al., 2016) to visualize the correlation network between the cell signaling pathways and the overlapping genes for all the top enriched pathways from the previous step. In total, we took the 10 significant pathways that were enriched among the 20 disease-gene group pairs to visualize. Figure 4.8 represents the signaling pathway networks with overlapping genes across different GEO datasets.

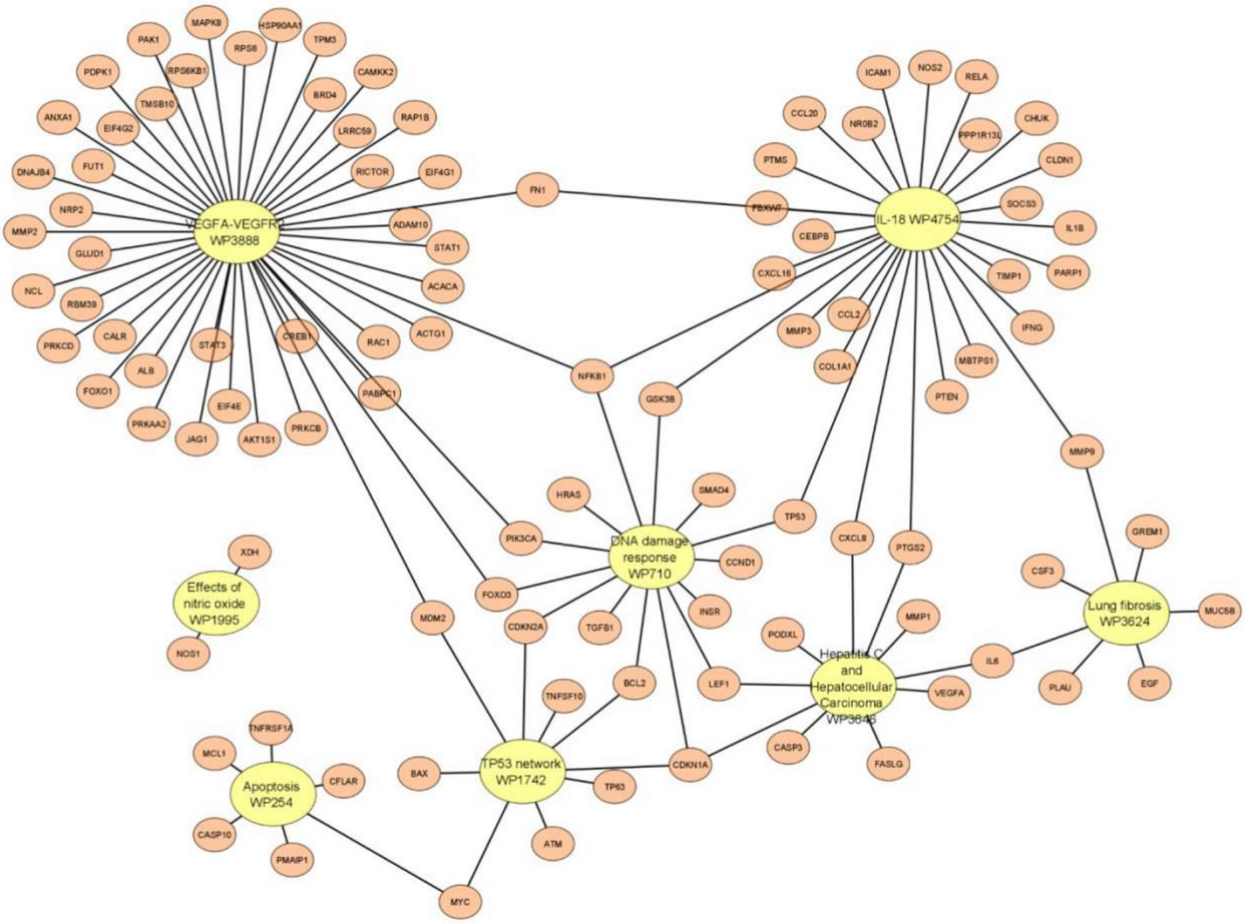


Figure 4. 8: Network visualization of the gene interaction for the cell signaling pathway with overlapping genes for the ten GEO datasets using the cytoscape tool.

As we have stated, we examine 10 different GEO gene expression datasets, studying mostly different diseases. Figure 4.8 illustrates the most significant pathways related to all given datasets, indicating that disease genes are correlated and associated even when studying different diseases. The network in Figure 4.8 shows that GediNET discovers important biological information related to various diseases. Moreover, we have studied the significance of GediNET on the GDS3257 data by considering the top 2 significant diseases having 12 distinct genes. Figure 4.9 illustrates the network of the most significant pathways and their related genes.

Figure 4. 9: Network visualization of the cell signaling pathway with overlapping genes for the GDS3257 dataset using the cystoscope tool.

4.9 Disease-Disease Associations

We suggest that a disease is characterized by a particular set of genes. To find a disease-disease relationship, one can use various association indices that take into account the shared genes between the two diseases. One could utilize metrics such as Jaccard, Simpson, Geometric, Cosine, and Pearson correlation coefficient (PCC) (Yousef et al., 2009; Yousef, Sayıcı, et al., 2021) .

Efforts towards Disease-Disease associations (DDA) have been recognized for their significance in uncovering new disease associations and expanding understanding of disease interactions, potentially leading to advancements in disease diagnosis, prognosis, and treatment. However, shared genes provide only restricted insight into the connection

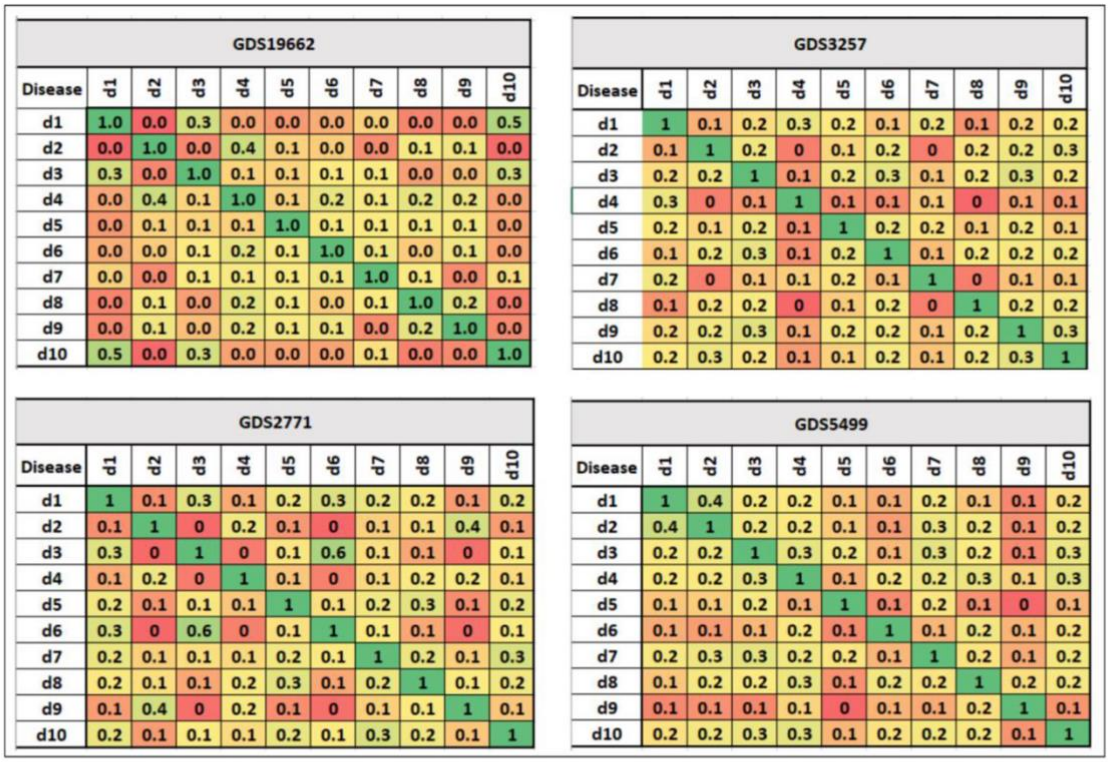
between two disorders. There are extremely few known DDAs and reliable relationships. Therefore, it indicates that additional efforts are needed for DDA detection.

Computational methods establish mathematical criteria for identifying disease modules within the incomplete human interactome. These methods demonstrate that the network-based position of each disease module dictates its pathobiological connection to other diseases (Menche et al., 2015a). Suratanee A. and Plaimas K. (Suratanee & Plaimas, 2015) created a new network-based scoring system named DDA to detect connections between diseases in a comprehensive research project. They have devised a technique that relies on random walk prioritizing within a protein-protein interaction network.

DisGeNET's API offers disease-disease associations based on shared genes and variations calculated between pairs of diseases by source. DisGeNET uses two metrics to compute the DDA. The first one is the Jaccard Index (JI) $Jaccard_G = \frac{G_1 \cap G_2}{G_1 \cup G_2}$, G1 represents the genes related to Disease 1, while G2 represents the genes associated with Disease 2.

The second one is Jaccard variance $Jaccard_V = \frac{V_1 \cap V_2}{V_1 \cup V_2}$, V_1 represents the variants linked to Disease 1, while V_2 represents the variants linked to Disease 2.

In order to compute, for each dataset, the standard DDA in GediNET, we have computed the fraction of the number of shared genes for each pair of the top-scored disease group for 4 datasets, as illustrated in Figure 4.10.



GDS19662

Disease	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
d1	1.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.5
d2	0.0	1.0	0.0	0.4	0.1	0.0	0.0	0.1	0.1	0.0
d3	0.3	0.0	1.0	0.1	0.1	0.1	0.1	0.0	0.0	0.3
d4	0.0	0.4	0.1	1.0	0.1	0.2	0.1	0.2	0.2	0.0
d5	0.0	0.1	0.1	0.1	1.0	0.1	0.1	0.1	0.1	0.0
d6	0.0	0.0	0.1	0.2	0.1	1.0	0.1	0.0	0.1	0.0
d7	0.0	0.0	0.1	0.1	0.1	0.1	1.0	0.1	0.0	0.1
d8	0.0	0.1	0.0	0.2	0.1	0.0	0.1	1.0	0.2	0.0
d9	0.0	0.1	0.0	0.2	0.1	0.1	0.0	0.2	1.0	0.0
d10	0.5	0.0	0.3	0.0	0.0	0.0	0.1	0.0	0.0	1.0

GDS3257

Disease	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
d1	1	0.1	0.2	0.3	0.2	0.1	0.2	0.1	0.2	0.2
d2	0.1	1	0.2	0	0.1	0.2	0	0.2	0.2	0.3
d3	0.2	0.2	1	0.1	0.2	0.3	0.1	0.2	0.3	0.2
d4	0.3	0	0.1	1	0.1	0.1	0.1	0	0.1	0.1
d5	0.2	0.1	0.2	0.1	1	0.2	0.2	0.1	0.2	0.1
d6	0.1	0.2	0.3	0.1	0.2	1	0.1	0.2	0.2	0.2
d7	0.2	0	0.1	0.1	0.2	0.1	1	0	0.1	0.1
d8	0.1	0.2	0.2	0	0.1	0.2	0	1	0.2	0.2
d9	0.2	0.2	0.3	0.1	0.2	0.2	0.1	0.2	1	0.3
d10	0.2	0.3	0.2	0.1	0.1	0.2	0.1	0.2	0.3	1

GDS2771

Disease	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
d1	1	0.1	0.3	0.1	0.2	0.3	0.2	0.2	0.1	0.2
d2	0.1	1	0	0.2	0.1	0	0.1	0.1	0.4	0.1
d3	0.3	0	1	0	0.1	0.6	0.1	0.1	0	0.1
d4	0.1	0.2	0	1	0.1	0	0.1	0.2	0.2	0.1
d5	0.2	0.1	0.1	0.1	1	0.1	0.2	0.3	0.1	0.2
d6	0.3	0	0.6	0	0.1	1	0.1	0.1	0	0.1
d7	0.2	0.1	0.1	0.1	0.2	0.1	1	0.2	0.1	0.3
d8	0.2	0.1	0.1	0.2	0.3	0.1	0.2	1	0.1	0.2
d9	0.1	0.4	0	0.2	0.1	0	0.1	0.1	1	0.1
d10	0.2	0.1	0.1	0.1	0.2	0.1	0.3	0.2	0.1	1

GDS5499

Disease	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
d1	1	0.4	0.2	0.2	0.1	0.1	0.2	0.1	0.1	0.2
d2	0.4	1	0.2	0.2	0.1	0.1	0.3	0.2	0.1	0.2
d3	0.2	0.2	1	0.3	0.2	0.1	0.3	0.2	0.1	0.3
d4	0.2	0.2	0.3	1	0.1	0.2	0.2	0.3	0.1	0.3
d5	0.1	0.1	0.2	0.1	1	0.1	0.2	0.1	0	0.1
d6	0.1	0.1	0.1	0.2	0.1	1	0.1	0.2	0.1	0.2
d7	0.2	0.3	0.3	0.2	0.2	0.1	1	0.2	0.1	0.2
d8	0.1	0.2	0.2	0.3	0.1	0.2	0.2	1	0.2	0.2
d9	0.1	0.1	0.1	0.1	0	0.1	0.1	0.2	1	0.1
d10	0.2	0.2	0.3	0.3	0.1	0.2	0.2	0.2	0.1	1

	GDS19622	GDS3257	GDS2771	GDS5499
d1	ADULT GLIOBLASTOMA	ACOUSTIC NEUROMA	BLADDER NEOPLASM	ACUTE LEUKEMIA
d2	DERMATOFIBROSARCOMA	ADENOCARCINOMA OF COLON	CARCINOMA, SMALL CELL	ACUTE MONOCYTIC LEUKEMIA
d3	EPITHELIAL OVARIAN CANCER	ADENOCARCINOMA OF ESOPHAGUS	COLORECTAL CARCINOMA	CHOLANGIOCARCINOMA
d4	GIANT CELL FIBROBLASTOMA	ADENOCARCINOMA OF LUNG, STAGE I	GASTROINTESTINAL CARCINOID TUMOR	GALLBLADDER CARCINOMA
d5	INTESTINAL CANCER	ADENOCARCINOMA OF PROSTATE	LEUKEMOGENESIS	GLIOBLASTOMA MULTIFORME
d6	MALIGNANT NEOPLASM OF COLON STAGE IV	ADENOMATOUS POLYPOSIS COLI	LIVER CARCINOMA	HAMARTOMA SYNDROME, MULTIPLE
d7	PAPILLARY RENAL CELL CARCINOMA	ADENOSQUAMOUS CARCINOMA	MALIGNANT HEAD AND NECK NEOPLASM	HEPATOCARCINOGENESIS
d8	PLASMA CELL NEOPLASM	ADULT ACUTE LYMPHOCYTIC LEUKEMIA	MANTLE CELL LYMPHOMA	PAPILLOMA
d9	RHABDOID TUMOR OF THE KIDNEY	ADULT RHABDOMYOSARCOMA	MUCINOUS ADENOCARCINOMA	PROGRESSION OF NON-SMALL CELL LUNG CANCER
d10	SECONDARY MALIGNANT NEOPLASM OF LYMPH NODE	ADULT T-CELL LYMPHOMA/LEUKEMIA	PANCREATIC NEOPLASM	RECTAL CARCINOMA

Figure 4. 10: An example of the DDA for four datasets in GediNET. The number of shared genes for the top-scored disease group is represented. The upper panel shows the DDA for GDS1962, GDS3257, GDS2771 and GDS5499 datasets. The lower panel shows the annotations used

GediNET differs from the tools mentioned above in that it is based on machine learning for detecting the relationships between diseases and DDAs, which detect novel and previously unknown associations. We conducted a further analysis to explore if GediNET can identify novel relationships between diseases using the DisGeNET API.

Table 4.8 A, B illustrates, for each data set, the three top diseases detected by the DisGeNET API and the top 3 ranked diseases by GediNET. For each detected disease by DisGeNET, we have looked up the disease in the list of ranked diseases by GediNET to examine the two tools.

Table 4.8- A: illustrates the three top detected diseases by DisGeNET API and the top 3 ranked diseases by GediNET for each GEO dataset. For each detected disease by DisGeNET, we have looked up the disease in the list of robust ranked aggregated disease results by

GediNET. The values in parenthesis for the rows of DisGeNET are the position of the disease and the p-value assigned by GediNET.

GEO Data Set/ Target Disease	The Data Disease	Top 1 Disease name	Top 2 Disease name	Top 3 Disease name
GDS1962/ BrainStem Glioblastoma	DisGeNET	Recurrent Endometrial Cancer (#193, pv=0.16)	Adult Astrocytic Tumor (#253, pv=0.22)	ALPHA-THALASSEMIA/MENTAL RETARDATION SYNDROME, NONDELETION TYPE, X-LINKED
	GediNET	PAPILLARY RENAL CELL CARCINOMA	PLASMA CELL NEOPLASM	ADULT GLIOBLASTOMA
GDS2545/ Metastatic prostate cancer	DisGeNET	Metastasis from malignant tumor of prostate (#25, pv=0.01)	Hormone refractory prostate cancer (#274, pv=0.34)	Secondary malignant neoplasm of bone (#62, pv=0.04)
	GediNET	CHILDHOOD RHABDOMYOSARCOMA	RHABDOMYOSARCOMA	SECONDARY MALIGNANT NEOPLASM OF LIVER
GDS2771/ Lung Cancer	DisGeNET	Primary malignant neoplasm of lung (#50, pv=0.03)	Carcinoma of lung (#97, pv=0.08)	Non-Small Cell Lung Carcinoma (#141, pv=0.14)
	GediNET	MANTLE CELL LYMPHOMA	GASTROINTESTINAL CARCINOID TUMOR	MUCINOUS ADENOCARCINOMA
GDS3257/ Lung Adenocarcinoma	DisGeNET	Non-small cell lung cancer recurrent (#116, pv=0.11)	Adenosquamous cell lung cancer (#137, pv=0.15)	Adenocarcinoma, metastatic (#200, 0.22)
	GediNET	ACOUSTIC NEUROMA	ADENOCARCINOMA OF COLON	ADENOCARCINOMA OF ESOPHAGUS

Table 4.8- B: illustrates the three top detected diseases by DisGeNET API and the top 3 ranked diseases by GediNET for each GEO dataset. For each detected disease by DisGeNet, we have looked up the disease in the list of robust ranked aggregated disease results by GediNET. The values in parenthesis for the rows of DisGeNET are the position of the disease and the p-value assigned by GediNET.

GEO Data Set/ Target Disease	**The Data Disease**	**Top 1 Disease name**	**Top 2 Disease name**	**Top 3 Disease name**
GDS4206/ Pediatric acute leukemia patients with early relapse: white blood cells	DisGeNET	Childhood Leukemia (#96, pv=0.13)	Melanoma (#29, pv=0.03)	Glioblastoma Multiforme (#115, pv=0.18)
	GediNET	ACUTE LEUKEMIA	ADULT DIFFUSE LARGE B-CELL LYMPHOMA	ESOPHAGEAL CARCINOMA
GDS5499/ Pulmonary hypertension	DisGeNET	Idiopathic pulmonary hypertension	Vascular Diseases	Endothelial dysfunction
	GediNET	CHOLANGIOCARCINOMA	HEPATOCARCINOGENESIS	PAPILLOMA
GDS3837/ Non-small cell lung carcinoma in female nonsmokers	DisGeNET	Primary malignant neoplasm of lung	Carcinoma of lung (#10, pv=0.009)	Neoplasm Metastasis
	GediNET	EARLY-STAGE BREAST CARCINOMA	MENINGIOMA, BENIGN, NO ICD-O SUBTYPE	COLORECTAL CARCINOMA
GDS4516_4718/ Colorectal Carcinoma	DisGeNET	Malignant neoplasm of colon and/or rectum (#3, pv=0.002)	Carcinogenesis	Neoplasm Metastasis
	GediNET	ACUTE LEUKEMIA	ACUTE LYMPHOCYTIC LEUKEMIA	Malignant neoplasm of colon and/or rectum

In Table 4.8, we have included additional information. The values in parenthesis for the rows of DisGeNET are the position of the disease and the p-value assigned by GediNET. Interestingly, excluding just one disease, all the top three significant diseases detected by GediNET are novel. This suggests that the tool detects new biological knowledge that the biology researcher should consider.

4.10 Discussion

This study presents an innovative approach for identifying associations between diseases and identifying the genes/biomarkers linked to those diseases. The method involves categorizing genes based on disease connections and evaluating these groups for their categorization significance in training the machine learning model. For example, if a model created from the given data associated with a specific disease, such as lung cancer, is also found to apply to a subset of different diseases, this could suggest a previously undetected biological relationship with those other diseases that could inform clinical approaches not previously considered. The traditional approach of searching for genes that could be used as a biomarker in most cases yields a list of significant genes that solve the computational problem and does not take into account any prior knowledge about those genes, as such, their association with other diseases or even with other biological knowledge such as microRNA targets (see maTE (Yousef et al., 2019)), or Pathways (See CogNet (Yousef, Ülgen, et al., 2021)), GeneOntology (See (Ersoz et al., 2023)).

Our GediNET tool is unique in that: (1) the search for the significant biomarkers/genes focuses on gene groups rather than single genes associated with the disease and (2) the final list of genes can be used to define new disease-disease associations. GediNET identifies important relationships between diseases, using DDA based machine learning, which investigates new associations that expand our understanding of disease interactions and enhance disease diagnosis, prognosis, and treatment by identifying novel disease associations.

CHAPTER FIVE

GEDINETPRO: DISCOVERING PATTERNS OF DISEASE GROUPS

5.1 Introduction

The GediNET tool is based on the Grouping, Scoring, and Modeling (G-S-M) approach for detecting disease-disease association (DDA). In this study, we have developed the pro version, GediNETPro, that utilizes the Cross-Validation (CV) information to detect patterns of disease group association by applying clustering approaches, such as K-means, extracted from the groups' ranks over the CV iterations. Additionally, a cluster score is computed to measure its significance and provide a deep analysis of the output of GediNET, yielding new biological knowledge that GediNET did not detect. Further, GediNETPro utilizes a visualization approach, such as a heatmap, to get novel insights and in-depth analysis of the disease clusters, revealing the relationship between diseases that might be used for developing effective interventions for diagnosing. We have tested GediNETPro on the Breast cancer dataset downloaded from the TCGA database. Results showed deeper insight into the interaction and collective behavior of the DDA, facilitating the identification and association of potential biomarkers.

5.2 GediNETPro

In the field of Machine Learning, one is required to evaluate the model created after training the classifier. Different approaches for evaluating the performance are used, such as k-fold cross-validations, repeated k-folds, and leave-one-out (Wong, 2015).

Monte Carlo Cross-Validation (MCCV), often referred to as repeated random sub-sampling CV, is a reliable technique for dividing a dataset into training and testing sets. As the name suggests, it randomly chooses the percentage of each split in each iteration, meaning no defined percentage of the dataset is left out in each iteration. MCCV is preferred over leave-one-out CV as the splits' proportion is independent of the number of iterations, which avoids the cause of over-fitting in prediction (Xu & Liang, 2001).

To perform the MCCV, the dataset is first randomly divided into training and testing parts. In each iteration, the percentage of splits is different; for example, it might be 80% training and 20% testing or 75% training and 25% testing. Some data splits are never selected in training, and others are chosen more than once. Second, the model is computed by fitting the ML using the training part, and the performance is measured with the testing dataset. The performance metrics are calculated through cross-validation iterations.

Our recently developed integrative machine learning-based tool, GediNET (Qumsiyeh et al., 2022), detects disease-disease associations and gene biomarkers for the disease under study.

The tool based on the G-S-M approach initially incorporates gene-disease associations from the DisGeNET database (Piñero et al., 2017) and gene expression datasets in the G component. A set of groups is given as input. Each group has a unique disease name and associated genes with the disease. Further, the task of the S component is to compute a score that measures to what extent it is differentially expressed considering the given two classes. This is performed after training each group with its associated sub_data using a Random Forest (RF) classifier. The GediNET tool was implemented in KNIME (Berthold et al., 2008).

GediNET provides a unique output of a list of groups ranked by a score, while the traditional approach output is a list of genes ranked by a score. Additionally, it provides a relationship between the top detected significant disease groups among those groups and their association with the main disease under study. Besides, in the original version of GediNET, a Monte Carlo CV (MCCV) is applied to estimate the tool's performance. We have applied 100 iterations of splitting the data into training and testing, where 90% of the data is used for training the classifier in the (the M component). While the remaining 10% is for testing to evaluate its performance. The aggregation of all those splits is collected, while the means and standard deviations are reported for each performance measurement. However, the tool did not exploit the knowledge that can be extracted from the MCCV iterations. Therefore, we have developed a new component, P, to extract the hidden patterns in MCCV using the new Pro version. Figure 5.1 illustrates the GediNETPro version that utilizes the MCCV to reveal hidden patterns and additional biological knowledge.

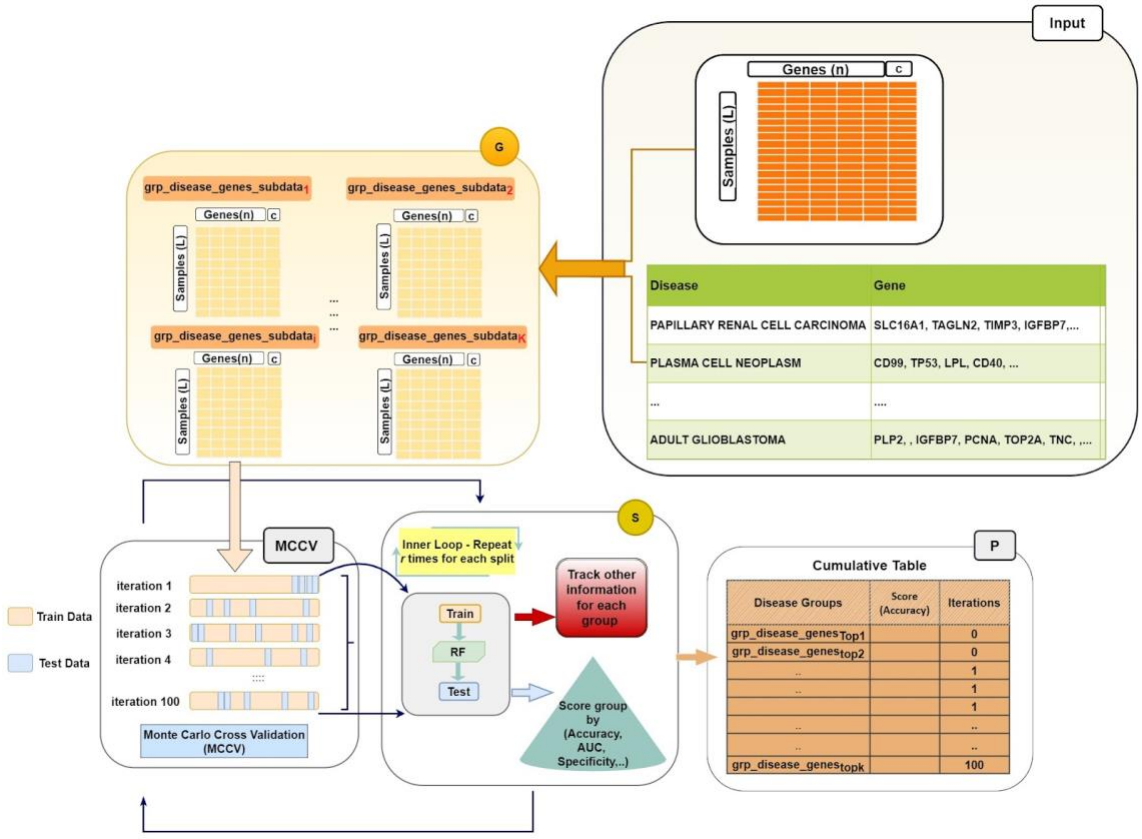


Figure 5. 1: Illustration of the GediNETPro. The input panel contains the gene expression data and the grouping table. Component G creates the sub_datasets based on the Input

panel. The MCCV panel uses the S component to perform the looping. The P component tracks the output of MCVV and S to be stored as a cumulative table.

The "Input" panel in Figure 5.1 contains the two-class gene expression table and the grouping table. Both tables will serve as input to the G component. The G component creates for each disease group its related two-class sub_datasets (G panel, Figure 5.1) by extracting the related columns (genes) from the original data with the class label (the c column in G panel, Figure 5.1) based on the Input panel. The "MCCV" panel cooperates with the S component to perform looping of *r* iterations. The P component collects cumulative information from each disease group, including gene sets, scores, and ranks. All the information is collected under the "Cumulative Table" in Figure 5.1, P component, whereas the "Cumulative Table" is summarized (See an example in Table 5.5). In the current version, we have redefined the rank according to Table 5.1, utilizing the score (accuracy) computed by the S component (See Figure 5.1). This system of ranks allows us to explore the patterns of the groups in more depth.

Table 5.1: The Rank scale is based on the score values.

Rank	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Score, ACC	>0.95	[0.90 - 0.95)	[0.85 -0.9)	[0.8-0.85)	[0.75 -0.8)	[0.7-0.75)	[0.65-0.7)	[0.6-0.65)	[0.55 -0.6)	<0.55	Absent of group

Table 5.1 shows that the value of 11 will be assigned to the group that failed to extract its associated subdataset due to filtering out genes with low signal.

5.3 P Component: Detect Patterns of Diseases Associations

Let's assume we have *m* groups of diseases. The S component assigns each group a score and a rank over the *r* iterations (We have set r to be 100). As a result, a matrix R with *m* rows and 100 columns is computed, where each row represents one disease group and the columns are the iterations ranks. R(i,j) is the rank assigned by the S component for group *i* in iteration *j*. Table 5.5 is an example of such output, where the ranks are stored in the column " Ranks list."

Let R_p be row *p* of matrix R representing all rank values over the 100 iterations. Each R_p is a point in 100 dimensions. One way to detect patterns of group ranks is by computing the similarity between R_p, p = 1,...,*m*. Clusters of those rows (points) would serve to find associations between diseases (groups). We have used K-means to detect such clusters. Then, for each cluster, a cluster score is assigned by averaging all the scores of its members. We have used K-means to estimate the number of clusters. The cluster with the least value is the most significant cluster that contains the top-ranked groups. The pseudo-code of the new P component is presented in Table 5.2.

Table 5.2: Pseudo code for detecting patterns of ranks of disease groups over 100 iterations.

P component

R is the diseases group ranks matrix over 100 iterations

Let *k* be the estimated number of clusters over R
clusters = K-means (R,k) //Apply clustering approach
for i = 1 to k
    c_score{i} = mean (clusters{i}) //compute the average ranks of each cluster
sort(c_score, "increasing order")

We have implemented the P component in KNIME (Berthold et al., 2008) using H2O (Aiello et al., n.d.). The H2O k-means node has the option of estimating the number of clusters that were used in P.

5.4 Results

GediNETPro is executed on the BRCA-TCGA data explained in section 2.1.3 with a 100-fold MCCV. The performance measures of accuracy, sensitivity, specificity, and AUC are reported in Table 5.3. The performance of the top-10 ranked groups is cumulatively presented in Table 5.3. The last row presents the results of Group number 1, the top-ranked cumulative group, with an AUC of 0.91, specificity of 0.83, sensitivity of 0.83, and accuracy of 0.83, obtained by an average of 6.17 genes. The last second row, Group number 2, presents the performance results of the top cumulative two groups.

Table 5.3: The performance table of GediNETPro for the top-ranked 10 groups averaged for 100 MCCV.

#Groups	#Genes	Accuracy	Sensitivity	Specificity	AUC
10	7.78	0.84	0.84	0.84	0.92
9	7.59	0.84	0.84	0.84	0.92
8	7.55	0.84	0.84	0.84	0.92
7	7.43	0.84	0.84	0.84	0.92
6	7.25	0.84	0.84	0.84	0.92
5	7.09	0.84	0.84	0.84	0.92
4	7	0.84	0.84	0.84	0.92
3	6.67	0.84	0.83	0.84	0.91
2	6.52	0.84	0.83	0.84	0.91
1	6.17	0.83	0.83	0.83	0.91

As seen from Table 5.3, there was no improvement in the AUC after the level of 4 accumulative groups. However, the user might be interested in examining more than 4 top groups to explore the association between disease groups. Since there is no change in the value of the AUC, one might use this level as the optimal threshold for the number of groups.

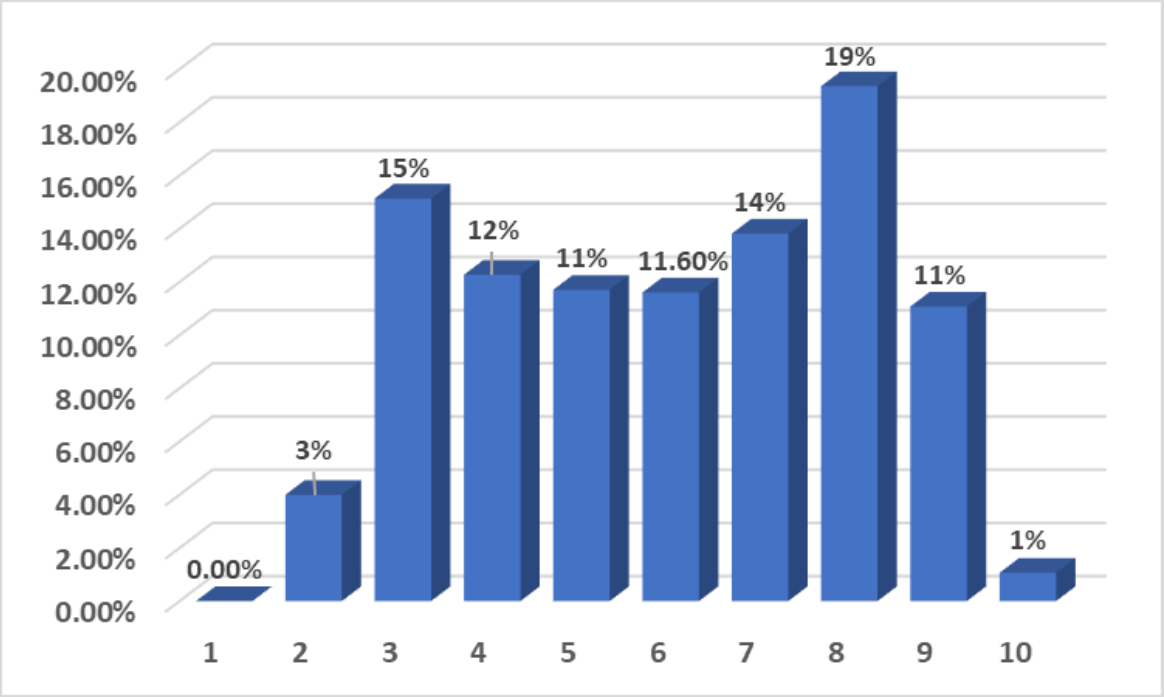


Figure 5. 2: The frequency of the groups ranks over all the iterations.

Figure 5.2 shows that none of the disease groups (1 out of 207,565) reaches the highest rank of 1, which impacts the performance of GediNETPro which has an accuracy of about 84%, as shown in 0.4. We have also seen that about 50% of the groups are ranked above the average in the range [1-6]. However, researchers would be mostly interested in the groups that are highly ranked. We might consider the range [1-4] for that purpose. Moreover, just 1% of the groups ranked with the lowest rank of 10. This is the impact of the filter step we applied using the statistics t-test, as explained in more detail in (Qumsiyeh et al., 2022).

Table 5.4-A: Cumulative Table of GediNETPro analysis of the molecular subtype datasets of BRCA. The table summarizes frequency, average score and rank, number of associated genes, and corresponding gene list over 100 iterations

Group	Average Score	Average rank	#Associated Gene	Associated Genes	Ranks list
ACINAR CELL CARCINOMA	0.85	3.28	10	MSANTD3, CENPF, PSAT1,…	3, 4, 3, 3, 4, 3, 3, 4,...
ACINAR CELL CARCINOMA OF PANCREAS	0.72	3.4	4	BRCA2, TP53, CDKN2A,...	6, 7, 5, 5, 8, 5,5, 8,..
ACINIC CELL CARCINOMA OF SALIVARY GLAND	0.66	7.29	1	MSANTD3	7, 8, 7, 7, 7, 7 8, 8, …

Table 5.4-B: Cumulative Table of GediNETPro analysis of the molecular subtype datasets of BRCA. The table summarizes frequency, average score and rank, number of associated genes, and corresponding gene list over 100 iterations

Group	Average Score	Average rank	#Associated Gene	Associated Genes	Ranks list
ACOUSTIC NEUROMA	0.85	3.4	18	GSTP1, MET, TP53,...	3, 4, 3, 3,4, 3, 4, 4, …
ACRAL LENTIGINOUS MALIGNANT MELANOMA	0.64	7.54	3	CD38, KDM1A, PROM1,...	8, 7, 7, 7, 7, 8, 7, 8,...
ACROSPIROMA	0.68	6.88	3	SLC6A2, ERBB2, EPHB1	8, 7, 7, 7, 7, 7, 8, 8, …
ACTH-SECRETING PITUITARY ADENOMA	0.76	5.28	12	IFNG, GAPDH, TP53, …	8, 7, 7, 7, 7, 8, 7, …

Table 5.4 is an example of a “Cumulative Table” that appears in Figure 5.1, with summary statistics. The average score and rank over the 100 iterations for each disease group are calculated in the S component (S panel, Figure 5.2) and presented correspondingly under the “Average Score” and “Average Rank” columns. The number of associated genes for each disease group and their unique associated genes are listed in the “#Associated Gene” and “Associated Genes” columns, respectively.

5.5 Detect Clusters of Groups by P component

The output creates 2414 groups, thus a rank matrix R with dimensions of 2414 rows and 100 columns.

Table 5.5: The summary output of component P describes 8 detected clusters of disease groups.

Cluster name	Number of Groups	Group Score
cluster_0	316	2.79
cluster_1	379	3.73
cluster_2	257	4.76
cluster_3	498	5.83
cluster_4	279	6.72
cluster_5	247	7.56
cluster_6	218	8.47

Applying the P component detects 8 clusters of groups, while the top-ranked cluster gets a score of 2.79, and has 316 disease groups, as illustrated in Table 5.6. All the disease groups belonging to cluster_0 have similar high ranks over the iterations.

Table 5.6: The top 10 ranked disease groups detected by component P.

Disease Group Name	Score
ADENOMA_OF_LARGE_INTESTINE	2.29
MALIGNANT_GLIOMA	2.3
CONVENTIONAL_(CLEAR_CELL)_RENAL_CELL_CARCINOMA	2.3
PAPILLARY_THYROID_CARCINOMA	2.31
MALIGNANT_NEOPLASM_OF_THYROID	2.31
ASTROCYTOMA	2.33
NON-SMALL_CELL_LUNG_CARCINOMA	2.33
SECONDARY_MALIGNANT_NEOPLASM_OF_LYMPH_NODE	2.35
EPITHELIAL_OVARIAN_CANCER	2.36
CARCINOMA_OF_URINARY_BLADDER, _INVASIVE	2.36

Table 5.6 shows the top-ranked 10 disease groups that belong to cluster_0, with their score, as suggested by the pseudocode, being the mean of all the ranks over the 100 iterations.

5.6 Detect Clusters of Groups by Visualization

One of the outputs of GediNETPro is the heatmap in Figure 5.3, which illustrates the clusters of diseases over the 100 iterations. Random groups of diseases with their average rank and iteration information are visualized in the heatmap in Figure 5.3. The rank scale is also apparent in Figure 5.3. The top-ranked groups are colored dark red, whereas low-ranked groups rarely detected within the 100 iterations are colored blue and dark purple. Therefore, while analyzing the heatmap, significant diseases that have a red color are essential to be analyzed. Once analyzed, new information would reveal hidden patterns with new biological meanings. For example, as seen in Figure 5.3, the MALIGNANT NEOPLASM OF GALLBLADDER and MALIGNANT NEOPLASM OF STOMACH co-occurred with a very high rank. Thus, these two diseases might be associated with the BRCA disease. Moreover, for validation, according to the literature, we have found a strong connection between the two diseases and BRCA. Missori, Giulia, et al. (Missori et al., 2020) have reported that breast cancer's potential for secondary malignant growth within gallbladder tissues is very high. The growth of small, flat nodules on the inner surface of the gallbladder mucous cells in patients with breast cancer is also expected. Their findings reported the significance of carefully examining the Gallbladder postoperatively for older patients with breast cancer. They also confirmed a high risk of getting Gallbladder cancer from Stomach cancer.

From Figure 5.3, HEREDITARY NON-POLYPOSIS COLON CANCER TYPE 2 AND HYPERPLASTIC POLYP diseases are two complementary pairs. This means that when one group appears highly ranked in a specific iteration, the second complementary one appears with a lower rank. This has been true for these two disease groups over the 100 iterations.

Furthermore, Figure 5.3 shows 6 significant disease groups that are highly ranked and appear in all iterations. These groups are MALIGNANT NEOPLASM OF GALLBLADDER, MALIGNANT NEOPLASM OF STOMACH, NON-SMALL CELL LUNG CARCINOMA, RENAL CARCINOMA, THYROID NEOPLASM AND TRANSITIONAL CELL CARCINOMA OF BLADDER. Their average rank is reported to be 3.01, 2.41, 2.33, 2.66, 2.38, and 2.74, respectively. Such behavior invites and suggests more investigations are needed to find hidden patterns and possible correlations between these diseases and BRCA at the molecular-basis cell level.

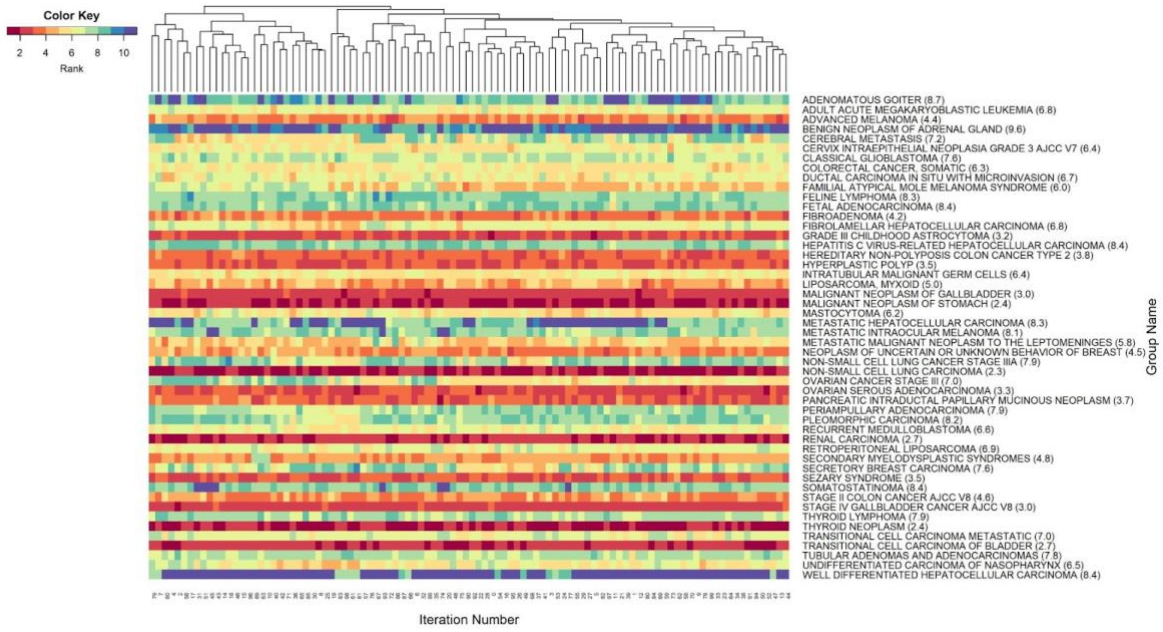


Figure 5. 3: Heatmap of groups with rank information over 100 iterations.

The low ranks, such as 9 and 10, would also provide biological knowledge. For example, Figure 5.3 shows that the disease WELL DIFFERENTIATED HEPATOCELLULAR CARCINOMA was scored all over the iterations with a very low rank, suggesting that this disease is not associated with the BRCA disease.

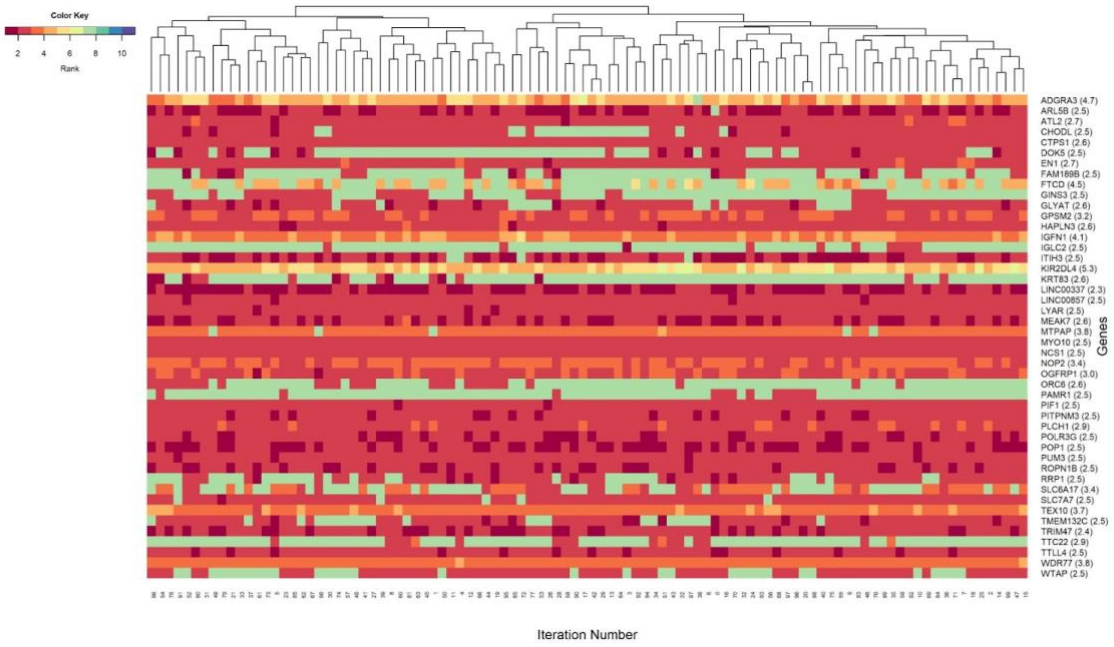


Figure 5. 4: Heatmap of the genes ranks over iterations.

The S component assigns each group a score, which is also assigned to the genes that are members of this group. Thus, at the end, we will also have information about the ranks of the genes. The robustrank aggregation is applied to those 100 lists to assign a p-value for each gene. For visualization, genes that appear less than 5 times out of 100 iterations are filtered out. Genes with a p-value less than 0.05 are selected. Then we randomly selected 50 genes, which are presented in Figure 5.4 as a heatmap. Figure 5.4 shows that most of those genes belong to groups that are also highly ranked.

CHAPTER SIX

DETECTING SEMANTIC SIMILARITY OF DISEASES BASED MACHINE LEARNING

6.1 Introduction

Identifying disease-disease associations (DDAs) is critical in medicine and systems biology, providing valuable insights into the complex relationships among different diseases. Recent advances in systems biology and the increasing availability of diverse biological data have provided opportunities for such research. However, the heterogeneity of the data makes it difficult to identify disease associations. To develop effective methods for predicting DDAs, it is necessary to leverage data at the molecular level. In this study, utilizing the Grouping-Scoring-Modeling approach, we have developed a statistical technique to compute a semantic similarity metric between the disease under study and other diseases. We have also used Jaccard similarity to compute similarity among disease groups. We have considered GediNETPro to record, over the Monte Carlo cross-validation, all the ranks based on the scores assigned by the S component to each disease group. The K-means clustering algorithm is employed to these recorded ranks for each disease group to detect patterns of similar diseases by placing them into coherent clusters. Each of the clusters is given a cluster score by averaging the ranks of its members. Notably, the lowest cluster score is the more significant one. The semantic approach is applied to the top-ranked clusters to detect a semantic relationship between the disease under investigation and other diseases (groups). In addition, we have applied diversity metrics based on the Jaccard index to each disease cluster to support the semantic findings. A high degree of diversity indicates a greater likelihood of detecting distinct groups with semantic relationships. This study enhances our understanding of disease associations and increases our ability to improve treatment strategies. This is especially important in a context where disease associations go beyond the scope of genetic factors.

Disease-disease association (DDA) detection has become an emerging topic in the field of bioinformatics and medical research. It involves identifying the underlying relationships and associations between different diseases, which could provide crucial understanding of the biological pathways and mechanisms underlying the onset and progression of disease. Furthermore, DDA detection can also facilitate drug repurposing studies, accelerate the identification of novel therapeutic targets, and aid in developing personalized treatment strategies (Xiang et al., 2022). Recent studies have demonstrated that similar diseases frequently have similar molecular causes, similar markers or traits for diagnosis, and similar medications for treatment. This suggests that we could potentially discover new effective molecules for a disease by looking at similar diseases we already know about. As a result, there's an increasing focus on researching how diseases are similar (Cheng et al., 2019) .

Despite the potential benefits of DDA detection, this area remains relatively unexplored compared to other disease association studies, such as disease-gene or disease-protein associations. The limited amount of research on DDA detection can be due to several factors, such as the complexity and heterogeneity of diseases, the lack of comprehensive and reliable data sources, and the difficulty in analyzing large and diverse datasets (Menche et al., 2015b).

The similarity between sets or groups is mainly based on the content of the groups; by this, we mean applying some distance metrics between the sets to compute their similarity or dissimilarity. Different distance metrics, including the Euclidean distance, Manhattan distance, cosine similarity, Jaccard similarity, etc., can be used to measure how similar two sets are. These metrics compare the elements in the sets or groups and provide a quantitative measure of how similar or dissimilar they are (Irani et al., 2016). On the other hand, semantic similarity measures how similar two pieces of text are in terms of their meaning and context. It is based on the concept of word embeddings, in which words are mapped to high-dimensional vectors that capture their semantic meaning. Semantic similarity can be computed using various techniques, such as word embeddings (Ye et al., 2016), Latent Semantic Analysis (LSA) (Suleman & Korkontzelos, 2021), and WordNet (Meng et al., 2013). The application of semantic similarity in text mining allows for more accurate analysis of text data, such as clustering and classification of documents, recommendation systems, and sentiment analysis. It helps identify patterns and relationships between pieces of text that may not be apparent from just the content alone (Hadj Taieb et al., 2013).

This study extends the concept of semantic similarity into the biological domain to detect a relationship between groups of diseases and the disease under study. Each group represents one disease by containing its associated set of genes. We have proposed a novel semantic similarity metric that is based on machine learning and adopts the G-S-M-P model of our earlier tool, GediNETPro (Qumsiyeh et al., 2023b).

The unique semantic similarity metric developed inside the G-S-M-P framework aims to address the shortcomings of conventional similarity measures that frequently do not adequately represent the intricate links across diseases in multi-omics data. Conventional techniques like cosine similarity or Pearson correlation emphasize statistical correlations while neglecting the biological context of these linkages. Our semantic similarity measure incorporates biological knowledge, enabling a more contextually informed evaluation of disease associations. This metric is crucial since it utilizes both quantitative data and qualitative biological insights, facilitating a more profound comprehension of disease interactions at the molecular level.

6.2 Method

The method proposed is depicted in Figure 6.1. The suggested method uses GediNETPro, adding two new components to the workflow, Rn Component and the Semantic Component (SC). The tool requires two inputs: the first is the two-class gene expression data for a specific disease, which we refer to as the target disease *D* (See Figure 6.1, component Input, Dataset *D*). The second input is the pre-existing biological knowledge obtained from the database DisGeNET. In this study, we follow the same grouping procedure proposed in GediNETPro (Qumsiyeh et al., 2023b). In other words, in the grouping component G, genes are grouped based on the disease information obtained from the DisGeNET. See Table 6.1 for an example of those groups.

Table 6.1: The Disease groups and their associated genes are formed based on the DisGeNET database.

Disease Groups	Genes
adult_anaplastic_ependymoma	CD151, DKC1, GSTP1,...
adult_diffuse_astrocytoma	ATP2A2, BRAF, CDKN2A, …
adult_gliosarcoma	AIF1, AKT1, BRAF, …
adult_undifferentiated_pleomorphic_sarcoma	AMPD2, CDKN2A, CTLA4, …
barrett's_adenocarcinoma	APC, BCL2, BRAF, …
benign_mastocytoma	CASP9, CSF2, CTLA4, …
gliomatosis_cerebri	BRAF, CD44, FGFR4, …
stage_i_lung_cancer_ajcc_v6	CDK8, EGFR, ENO1, …
stage_ii_colorectal_cancer	BRAF, CASP10, CDK1, ….
stage_iii_colorectal_cancer	SERPINA3, ACTG1, ACTG2, …
unilateral_breast_carcinoma	ATM, BRCA1, BRCA2, …

For each group *g* we extracted its associated two-class subdataset from the dataset *D* (specifically the train part of *D*). In the scoring component S, a *score* is assigned by applying an internal Monte-Carlo cross-validation on the two-class subdataset (see Figure 6.1, Component S).

A new component, Rn, is added to the workflow to obtain ranks for each group. The Rn component sorts the groups in descending order based on their scores. Groups with higher scores appear at the top of the list, while those with lower scores appear toward the bottom. Next, a rank is assigned to each group based on its position in the sorted list. The regular ranks of the group with the highest score receive rank 1, the next-highest score receives rank 2, and so on. In our tool, we have provided a transformed rank based on the following table.

Table 6.2: The Rank scale is based on the score values.

Rank	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Score, ACC	>0.95	[0.90 - 0.95)	[0.85 -0.9)	[0.8-0.85)	[0.75 -0.8)	[0.7-0.75)	[0.65-0.7)	[0.6-0.65)	[0.55 -0.6)	<0.55	Absent of group

Considering the scores assigned by the S component, Table 6.2 describes the ranking method that the Rn component adopted. The Rn component categorizes disease groups into 11 ranks, each associated with specific ranges of scores. The ranks range from the highest accuracy (Rank 1 with a score above 0.95) to the lowest (Rank 10 with a score below 0.55). An additional Rank 11 represents a group that couldn't extract its subdataset due to filtering out genes with low signal.

In GediNETPro, the dataset *D* is split into two parts, 90% of *D* for training and 10% for testing. This process is repeated 100 times (we refer to it as the Monte Carlo cross-validation

external loop). In each loop, the S score is implemented to score all the groups. The Rn component uses this output to create the R matrix. The dimensions of R are m rows, which is the number of groups, and 100 columns. Each column corresponds to an iteration (R1, R2, R3,..., R100) and contains the rank value assigned to the respective group for that iteration. In the Rn component, these scores are converted to ranks based on Table 6.2. Table 6.3 represents a part of the R matrix. Besides, See Figure 6.1, component Rn, and matrix R.

Table 6.3: A part of the R matrix

Group names	**R1**	**R2**	**R3**	**R4**	**R5**	**R6**	**..**	**..**	**R97**	**R98**	**R99**	**R100**
ACINAR_CELL_CARCINOMA	3	4	3	3	4	3	..	..	2	4	3	3
ACINAR_CELL_CARCINOMA_OF_PANCREAS	6	7	5	5	8	5	..	..	7	5	8	5
ACINIC_CELL_CARCINOMA_OF_SALIVARY_GLAND	7	8	7	7	7	7	..	..	7	8	8	7
ACOUSTIC_NEUROMA	3	4	3	3	3	3	..	..	3	4	4	4
ACRAL_LENTIGINOUS_MALIGNANT_MELANOMA	8	7	7	7	7	8	..	..	8	7	8	8
ACROSPIROMA	7	7	7	7	6	6	..	..	6	7	7	7
ACTH-SECRETING_PITUITARY_ADENOMA	5	5	6	6	6	4	..	..	4	5	5	5
ACTINIC_KERATOSIS	3	4	2	3	3	3	..	..	3	4	3	2
ACTIVATED_B-CELL_TYPE_DIFFUSE_LARGE_B-CELL_LYMPHOMA	5	5	6	6	5	5	..	..	5	5	5	5
ACUTE_BILINEAL_LEUKEMIA	9	9	9	9	9	9	..	..	9	9	9	9
ACUTE_BIPHENOTYPIC_LEUKEMIA	9	11	10	10	11	11	..	..	8	11	11	11
ACUTE_ERYTHROBLASTIC_LEUKEMIA	3	3	3	3	3	3	..	..	3	3	3	3

In the SC component, the K-means clustering algorithm is employed to the R matrix to detect patterns of similar groups (diseases) by placing similar groups into clusters. A cluster_score is computed for each cluster by averaging the ranks of its members. The cluster with the lowest value is considered the most significant cluster. For more details, see Table 6.6.

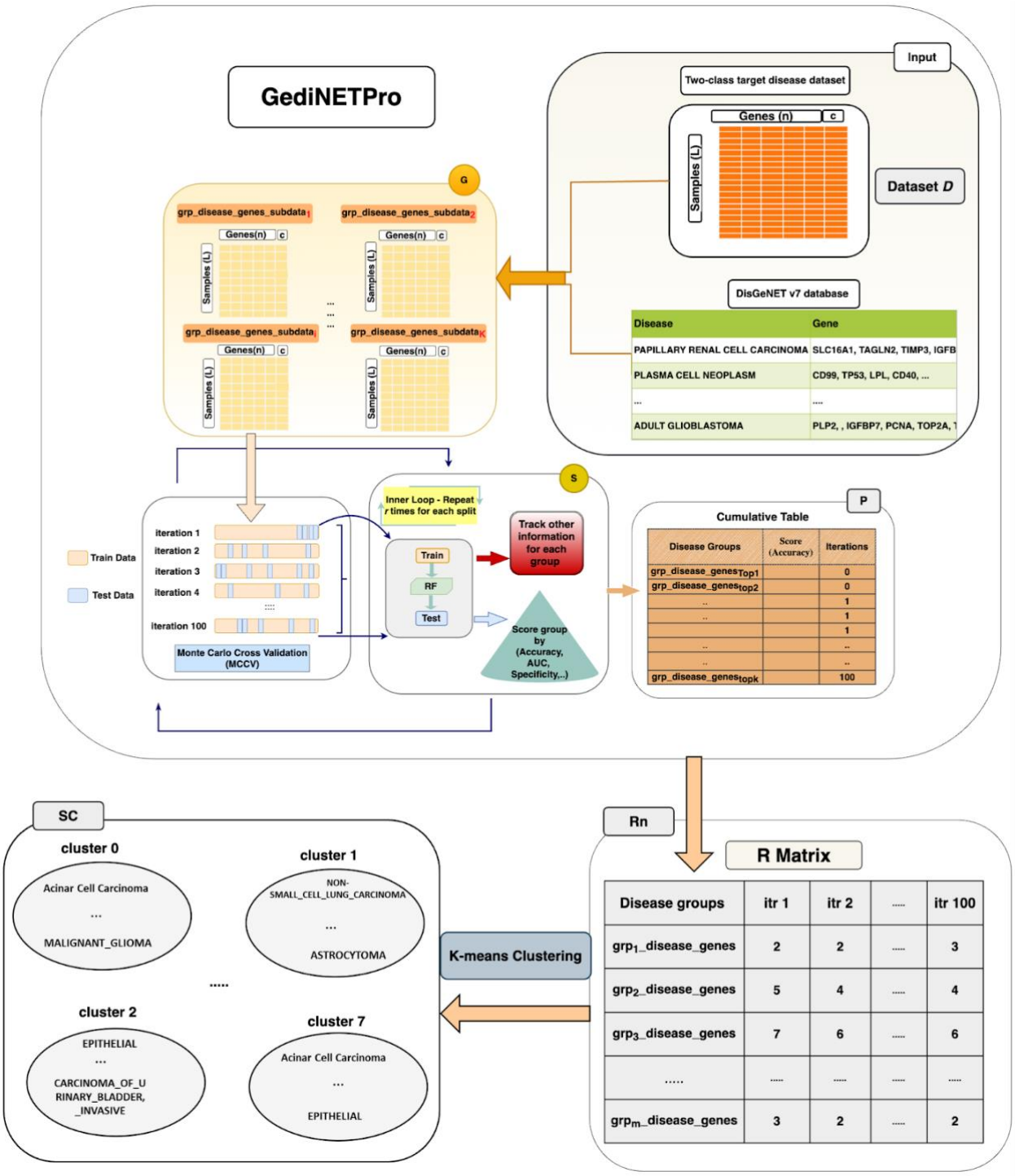


Figure 6. 1: The main workflow for the GediNETPro similarity approach. The Semantic Component (SC) computes the semantic similarities.

6.3 Similarities between groups

We have defined a group as a set of genes that are related to a particular disease. For a given two sets or groups, there are several metrics to calculate the similarity between sets of items or genes. For example, the Jaccard similarity is calculated by dividing the size of the intersection of two sets by the size of their union (Real & Vargas, 1996). The definition of the Jaccard Similarity, given two sets, A and B, is defined as:

$$J(A,B) = \left|\frac{A \cap B}{A \cup B}\right| = \frac{|A \cap B|}{|A|+|B|-|A \cap B|} \qquad (6)$$

6.3.1 Similarity Measurements for the cluster of disease groups:

Let a *group* be a set of genes that are related to a particular disease. We will refer to it as a disease group. We define a *cluster* of diseases as a group of diseases detected by the k-means applied to the R matrix, as appears in Figure 6.1, components Rn and SC. In this study, we define two kinds of similarities; the first is the similarity that is based on the common genes between groups (diseases), utilizing some known measurements such as the Jaccard similarity that is used to compute a diversity score for each cluster. The second is a novel semantic similarity, which is based on machine learning and specifically the G-S-M-P model of GediNETPro, which detects the association between a cluster of diseases and the target disease.

6.3.2 Diversity:

To calculate the diversity of sets, one can use a metric that quantifies how different the groups /sets are from each other. The Jaccard distance is a frequently used metric for measuring diversity, calculated as 1 minus the Jaccard similarity. The Jaccard distance quantifies the difference or dissimilarity between two sets, where a score of 0 signifies identical sets and a score of 1 signifies completely divergent sets.

To compute the diversity of *k* groups, we calculate the pairwise Jaccard distances between all groups and then take the average or maximum distance as the diversity of the *k* groups. The pseudocode of the diversity measurement is presented in Table 6.4.

Table 6.4: The pseudocode of the diversity measurement.

```
Input: list_sets is a k of sets (groups)
distances = []
for each combination (s1, s2) in list_sets:
    intersection = intersect(s1, s2)
    union = union(s1, s2)
    jaccard_distance = 1 - length(intersection) / length(union)
    add jaccard_distance to distances

diversity = sum(distances) / length(distances)
```

This diversity is used to compute a diversity score for each cluster that is created in component SC (see Figure 6.1, SC component). The cluster contains a number of groups detected by the k-means algorithm. A higher value of this diversity measure indicates greater dissimilarity between the groups. A diversity value of 1 would indicate that the sets are completely dissimilar, while a diversity value closer to 0 would indicate that the groups are more similar.

6.3.3 Semantic Similarity:

In the context of a cluster comprising *r* groups identified through the K-means algorithm (Figure 6.1, Component SC), it becomes a question of why these groups consistently maintain similar rankings across the 100 iterations in GediNETPro. There are two possible explanations for this phenomenon.

Firstly, this consistency could arise because the *r* groups possess a degree of inherent similarity in terms of their genes. This form of similarity can be regarded as a standard measure. On the other hand, a different perspective exists where these groups exhibit a biological connection with the target disease of the specific two-class dataset *D* (as depicted in Figure 6.1's input panel). This perspective introduces what we term "semantic similarity," a novel relationship grounded in machine learning principles.

The standard similarity can be calculated using a variety of similarity metrics, as shown in Equation (6). One could compute the pairwise similarity between all groups and compute the average as the final similarity score. On the other hand, our semantic similarity algorithm is illustrated in Table 6.5.

Table 6.5-A: Our novel semantic cluster similarity algorithm.

Algorithm: Semantic Cluster Analysis

Input:
- (R) matrix: Each row represents a group, and each column represents the rank of the group over an iteration.

Output:
- SortedClusters: List of clusters sorted by their semantic scores.

Steps:

1. Compute the (R) matrix: Rows represent groups, and columns represent ranks.

2. Apply K-means clustering on (R) matrix to detect (k) clusters:$(C_1, C_2, \ldots, C_k)$.
3. For each cluster (C_i):
   - Calculate average ranks for each group member:
   - For each group $(g) in (C_i)$, compute average rank for each value:
   - $$\text{avg\_rank}_{gj} = \frac{1}{\text{num\_iterations}} \sum_{k=1}^{\text{num\_iterations}} R_{gjk}$$
   - Calculate the overall average rank for group (g):
   - $$\text{overall\_avg\_rank}_g = \frac{1}{\text{num\_values}} \sum_{j=1}^{\text{num\_values}} \text{avg\_rank}_{gj}$$
   - Compute cluster score as the average of the overall average ranks for groups in cluster (C_i):
   - $$\text{cluster\_score}_i = \frac{1}{\text{num\_groups\_in\_cluster}} \sum_{g \in C_i} \text{overall\_avg\_rank}_g$$

4. Sort clusters based on cluster scores:

Table 6.5-B: Our novel semantic cluster similarity algorithm.

$\text{SortedClusters} = \text{SortByScore}(C_1, C_2, \ldots, C_k)$.

5. For each cluster (C_i):
- Compute diversity score:
  - Calculate Jaccard similarity coefficient for every pair of groups within (C_i) using ranks.
  - Compute average Jaccard similarity coefficient for every pairs of groups in (C_i).
  - Store diversity score for cluster (C_i).

6. Return SortedClusters: List of clusters sorted by cluster scores along with diversity scores.

6.4 Results

We have applied our approach to the Breast Invasive Carcinoma dataset in section 2.1.3. GediNETPro generated the R matrix as part of its output tables, which is gathered within the P component (refer to Figure 6.2). This matrix is then employed as an input for the k-means algorithm, facilitating the identification of 8 clusters. Our implementation involves utilizing H2O.ai (KNIME version (Berthold et al., 2008)) to determine the optimal number of clusters.

The SC component (shown in Figure 6.1) is applied to each cluster to compute the cluster_score where lower scores indicate a significant group for the classification task, and the groups that are in the cluster are more related to the disease understudy than other groups. The diversity score is also computed to validate the relationship of the cluster to the target disease. A diversity value close to 1 indicates that the groups are completely dissimilar and are placed in the cluster due to their relationship to the disease under study. Table 6.6 is ordered in descending order according to the values in column "Cluster_Score". The count column indicates the number of groups in each cluster.

Table 6.6: The diversity score and Cluster_Score for each detected cluster.

Cluster	Diversity	Count	Cluster_Score
cluster_0	0.91	316	2.79
cluster_2	0.96	257	3.74
cluster_5	0.97	247	4.77
cluster_7	0.97	220	5.84
cluster_6	0.97	218	6.72
cluster_4	0.96	279	7.56
cluster_3	0.95	498	8.48
cluster_1	0.99	379	10.61

In Table 6.6, it is evident that cluster_0 comprises 316 groups (diseases) that consistently achieved high scores (lower rank indicates better scores) throughout the 100 iterations. The average score, derived from the data presented in the table, stands at 2.79. Upon closer examination of these 316 groups, it became apparent that their ranks ranged between 2 and 4, with only one instance of rank 1 being achieved.

Cluster_0 demonstrates a diversity score of 0.91, signifying that most of the groups within it exhibit dissimilarity. This dissimilarity becomes noteworthy since the similarity detected by the k-means algorithm is primarily attributed to their comparable scores (ranks) rather than the genetic content they contain. This concept aligns with what we term "semantic similarity," where the relationship to the disease under study is akin, even if the content itself does not mirror this similarity.

An important inquiry arises regarding the interconnection among those diseases (groups) that find themselves placed within the same cluster.

6.5 Biological Findings

Table 6.7-A: The disease groups of Cluster_0 with relation to the main Breast cancer genes.

Disease	ESR1	ESR2	PGR	ERBB2	BRCA type
adenocarcinoma_basal_cell	Yes	Yes	Yes	Yes	Luminal B
adult_anaplastic_astrocytoma	NO	NO	NO	NO	Basal-like
adult_glioblastoma	NO	NO	NO	Yes	HER2-enriched
adult_liposarcoma	NO	NO	NO	NO	Basal-like
adult_meningioma	Yes	NO	NO	NO	Luminal A
adult_oligodendroglioma	NO	NO	NO	NO	Basal-like
adult_synovial_sarcoma	NO	NO	NO	Yes	HER2-enriched
carcinoma_cribriform	Yes	Yes	Yes	Yes	Luminal B
carcinoma_of_lung	Yes	Yes	Yes	Yes	Luminal B
cervical_cancer	Yes	Yes	Yes	Yes	Luminal B
cervix_carcinoma	Yes	Yes	Yes	Yes	Luminal B
childhood_burkitt_lymphoma	Yes	NO	NO	NO	Luinal A
childhood_lymphoma	Yes	Yes	NO	Yes	Luminal B
childhood_myelodysplastic_syndrome	NO	NO	NO	NO	Basal-like
childhood_neuroblastoma	Yes	Yes	NO	Yes	Luminal B

Table 6.7-B: The disease groups of Cluster_0 with relation to the main Breast cancer genes.

Disease	ESR1	ESR2	PGR	ERBB2	BRCA type
childhood_non-hodgkin_lymphoma	NO	NO	NO	Yes	Luminal B
childhood_rhabdomyosarcoma	Yes	Yes	NO	Yes	Luminal B
fibrosarcoma	NO	NO	NO	Yes	HER2-enriched
liver_and_intrahepatic_biliary_tract_carcinoma	Yes	Yes	NO	NO	Luminal A
malignant_neoplasm_of_lung	Yes	Yes	Yes	NO	Luminal A
mammary_carcinoma_human	Yes	Yes	Yes	NO	Luminal A
metastatic_prostate_carcinoma	Yes	Yes	NO	NO	Luminal A
neuroblastoma	Yes	Yes	NO	Yes	Luminal B
rhabdomyosarcoma	Yes	Yes	NO	Yes	Luminal B
stage_0_gallbladder_cancer_ajcc_v8	Yes	NO	Yes	Yes	Luminal B

We have validated the results obtained on cluster_0 by associating, for each group, the known breast cancer gene markers ESR, ESR2, ESR2, PGR, and ERBB2. The Breast Cancer type is defined based on the presence (indicated by "Yes") of the marker genes. "No" is an indication of the absence of the marker.

Most of the time, the three biomarkers estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) are utilized for categorizing breast cancer into subtypes. Different algorithms are used to categorize breast cancer subtypes based on these biomarkers. An often utilized method involves determining if the tumor cells exhibit ER and/or PR expression, as well as HER2 overexpression. Tumor cells that express ER and/or PR but do not overexpress HER2 (ESR+, PGR+/-, ERBB2-) are classified as Luminal A subtype. Tumor cells expressing ER and/or PR and overexpressing HER2 (ESR+, PGR+/-, ERBB2+) are categorized as Luminal B subtype. Tumor cells lacking ER and PR expression but showing high levels of HER2 expression (ESR-, PGR-, ERBB2+) are classified as HER2-enriched subtype. If the tumor cells do not express any of these three biomarkers, then they are categorized as a basal-like subtype (Perou et al., 2000).

The validation of Cluster_0 using known breast cancer gene markers ESR1, ESR4, PGR, and ERBB2 revealed intriguing associations, as shown in Table 6.7. The results indicate that there is no direct relationship between the disease groups within Cluster_0 and the established subtypes of breast cancer. However, by analyzing the expression of the three biomarker genes, we were able to identify specific associations between diseases. This observation aligns with the notion of "semantic similarity," where the connection between

the disease under study and clustered diseases is not genetically rooted but rather determined by behavior or patterns. This analysis highlights the possible existence of shared biological pathways and offers insights into potential comorbidities or shared underlying biological mechanisms between breast cancer subtypes and other diseases.

6.6 Discussion

In this study, we introduced the concept of semantic similarity, which computes the similarity/diversity between disease groups and the target disease group under study. The aim is to discover the cause of the similarity of diseases using machine learning, specifically the G-S-M-P model of GediNETPro employing the Rn and SC components.

The results show that the similarity of those diseases that were scored similarly over all the 100 iterations is likely due to biological relationships and not to the content of the groups (common genes). Although the genetic dissimilarity of these groups suggests inherent differences, the concept of semantic similarity highlights their shared behaviors or characteristics, which correspond to the disease under study. This suggests that disease relationships can extend beyond genetic makeup, emphasizing the importance of considering behavior and other factors when grouping diseases.

The concept of semantic similarity used in this study can be a useful tool in disease research. Semantic similarity measures how closely related two diseases are based on their biological relationships. By utilizing the machine learning algorithm, specifically the G-S-M-P of GediNETPro, and extending the new SC and Rn components to compute semantic similarity between disease groups, this study was able to identify groups that were similar in terms of their underlying biology. This extends beyond conventional content-based similarity approaches, enabling the identification of groups with shared biological behaviors.

CHAPTER SEVEN

CONCLUSIONS AND FUTURE PERSPECTIVES

7.1 Conclusions

The main contribution in the first study in this thesis is the development of a novel approach, GediNET, for discovering disease-disease associations and detecting the genes/biomarkers associated with those diseases. With the widespread use of machine learning techniques, the seek for genes linked to specific diseases has yielded valuable insights. However, the integration of pre-existing biological knowledge into DDA has emerged as a powerful tool, promising the identification of novel biomarkers with practical implications for clinical practice.

GediNET stands as a pioneering approach that not only harnesses the capabilities of machine learning and gene analysis but also strategically incorporates established biological knowledge. GediNET introduces a three-step approach: Grouping, Scoring, and Modeling (G-S-M), which collectively forges a path toward the elucidation of intricate disease-disease associations. This novel approach integrated pre-existing biological knowledge about the disease to enhance group selections rather than perform feature selections within the classification task. In this way, GediNET reveals hidden patterns related to the DDA. The method mainly analyzes gene expression data utilizing DisGeNET v7, an external database of biological knowledge that links genes to diseases. The G component of GediNET groups genes based on disease information obtained from the DisGeNET database and generates a two-class subdataset related to each group/disease. The S component computes a score for each group based on the differential expression between the two classes and ranks the groups accordingly. The M component considers the top-ranked disease groups and merges their genes to form the top-ranked associated genes, on which a Random Forest model is trained and evaluated. The evaluation part was done using Monte Carlo cross-validation (MCCV). This process is iterated 100 times by splitting the data into training and testing. This process extracts hidden patterns and potentially yields novel insights into biological knowledge.

The true innovation of GediNET lies in its capacity to uncover associations that extend beyond the initial disease focus. The application of DDA based machine learning enables the identification of novel unknown associations between seemingly unrelated diseases. This revelation not only enriches our fundamental understanding of disease pathways but also holds immense potential for reshaping medical approaches.

Nonetheless, the GediNET tool failed to leverage the potential insights extractable from the MCCV iterations. Hence, we've introduced a new component, denoted as "P," designed to unearth hidden patterns within MCCV through the new pro version, GediNETPro. In our second study, The new component P detect clusters or patterns of disease groups based on

their rank values assigned by the S component. A new cluster-score is computed to detect the most significant cluster of groups. Traditional approaches mainly use CV or other cross-validation techniques to evaluate performance measurements. However, GediNETPro utilizes the ranks or scores all over the iterations to be used in the P component to detect hidden patterns of the groups' ranks.

In the last study, we introduced the concept of "semantic similarity," a measure that evaluates the similarity or dissimilarity between groups of diseases and the target group of diseases being studied. The purpose was to unravel the reasons behind the similarities observed among diseases using a machine learning approach, specifically the G-S-M-P model within GediNETPro, which incorporates the Rn and SC components.

The outcomes indicated that diseases exhibiting similar scores across all 100 iterations are likely to share biological connections rather than mere group contents (common genes). Although their genetic differences suggest distinctiveness, the notion of semantic similarity brings attention to shared behaviors or traits aligning with the disease under examination. This underscores the idea that disease relationships may stretch beyond genetic makeup, emphasizing the significance of considering behavior and other factors when grouping diseases.

7.2 Limitations and Future Prospects

For each of the three studies carried out as part of this thesis work, we can briefly describe our prospects for the future as follows:

7.2.1 Future prospects and limitations of study 1:

The uniqueness of the GediNET approach is its method of evaluating gene groupings by taking into account the contribution of all its member. One potential limitation of this approach that might be considered, is whether some members (genes) within a group may have a noisy impact and as a result adversely affect the overall classification performance. Other feature selection methods that consider each gene independently, will not have this problem. However, to avoid this, we used a statistical t-test on the training dataset to first detect the top differentially expressed genes. The top 2000 genes with the most significant differences in expression were then used to extract the training datasets that were used as input to the G component. Thus, GediNET will always be dealing with the least noisy genes. One direction of future work is to perform internal gene scoring for each gene group to consider only those genes with the highest scores.

Another potential limitation of our approach is the possibility that the size of the (gene) group could influence the performance. For example, by influencing Scoring component. Groups that contain larger numbers of gene would tend to have higher scores. This issue might be solved by considering a fixed number of representative genes from each group. One aspect of feature selection (scoring) that we have not covered in this study is the potential for two groups of features that are ineffective on their own to become effective when combined. Within GediNET, the scoring component evaluates each group separately. One possible future strategy could involve enhancing the S component to score groups at the same time to tackle this issue.

7.2.2 Future prospects and limitations of study 2:

While GediNETPro's incorporation of ranks or scores from numerous iterations to unveil hidden patterns in the P component is a novel approach, it also presents a limitation worth acknowledging. The reliance on ranks or scores across iterations might introduce a degree of variability, which could potentially impact the robustness and consistency of the hidden patterns detected. This variability could be attributed to factors such as noise in the data or the specific characteristics of the scoring mechanism. To address this limitation and further enhance the effectiveness of GediNETPro, future plans could involve the implementation of techniques to mitigate the impact of variability. One approach could involve introducing more advanced statistical methods to stabilize the ranks or scores, reducing potential fluctuations. Additionally, considering ensemble techniques that aggregate results from multiple iterations could offer a way to improve the reliability of the hidden pattern detection process.

Furthermore, to validate the hypothesis that disease groups sharing the same cluster possess similar biological functions, future endeavors could encompass comprehensive biological validation. This might involve conducting in-depth functional enrichment analyses on the identified clusters to identify common biological pathways, molecular functions, or cellular processes. Collaborative efforts between computational biologists and domain experts could offer a deeper understanding of the fundamental biology within these clusters.

7.2.3 Future prospects and limitations of study 3:

This study has important implications for the future development of GediNETPro. By adding a pre-processing step, we can improve the accuracy and efficiency of the tool. Besides, by removing redundant groups and including diverse groups in the final model, GediNETPro can provide more accurate predictions and insights into the underlying biology of diseases. Besides, by decreasing the computational time required for the S component of GediNETPro, this pre-processing step can help make the tool more efficient.

In addition to improving the accuracy and efficiency of GediNETPro in disease research, this study can also have important implications for disease co-morbidity studies. Disease co-morbidity refers to the presence of many diseases in a single patient, by understanding the relationships between different diseases can be important for drug repurposing and other therapeutic strategies. By revealing both the similarities and diversities among different disease groups, the semantic similarity concept can facilitate a enhanced comprehension of the biochemical connections among diseases.

References

Acharya, S., Saha, S., & Nikhil, N. (2017). Unsupervised gene selection using biological knowledge: Application in sample clustering. *BMC Bioinformatics*, *18*(1), 513. https://doi.org/10.1186/s12859-017-1933-0

Aiello, S., Click, C., Roark, H., Rehak, L., & Lanford, J. (n.d.). *Machine Learning with Python and H2O*.

Anggraeni, A. N., Mustofa, K., & Priyanta, S. (2021). Comparison of Filter and Wrapper Based Feature Selection Methods on Spam Comment Classification. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *15*(3), Article 3. https://doi.org/10.22146/ijccs.66965

Asif, M., Martiniano, H. F. M. C. M., Vicente, A. M., & Couto, F. M. (2018). Identifying disease genes using machine learning and gene functional similarities, assessed through Gene Ontology. *PLoS ONE*, *13*(12), e0208626. https://doi.org/10.1371/journal.pone.0208626

Bach, M., Werner, A., & Palt, M. (2019). The Proposal of Undersampling Method for Learning from Imbalanced Datasets. *Procedia Computer Science*, *159*, 125–134. https://doi.org/10.1016/j.procs.2019.09.167

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., & Soboleva, A. (2013). NCBI GEO: Archive for functional genomics data sets--update. *Nucleic Acids Research*, *41*(Database issue), D991-995. https://doi.org/10.1093/nar/gks1193

Bellazzi, R., & Zupan, B. (2007). Towards knowledge-based gene expression data mining. *Journal of Biomedical Informatics*, *40*(6), 787–802. https://doi.org/10.1016/j.jbi.2007.06.005

Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton University Press.

Ben-dor, A. (2002). Gene-Expression Profiles in Hereditary Breast Cancer. *Advances in Anatomic Pathology*. https://www.academia.edu/17533809/Gene_Expression_Profiles_in_Hereditary_Breast_Cancer

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., & Wiswedel, B. (2008). KNIME: The Konstanz Information Miner. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data Analysis, Machine Learning and Applications* (pp. 319–326). Springer. https://doi.org/10.1007/978-3-540-78246-9_38

Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., Sampas, N., Dougherty, E., Wang, E., Marincola, F., Gooden, G., Lueders, J., Glatfelter, A., Pollock, P., Carpten, J., & Sondak, V. (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, *406*, 536–540. https://doi.org/10.1038/35020115

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification And Regression Trees* (1st ed.). Routledge. https://doi.org/10.1201/9781315139470

Brown, G., Pocock, A., Zhao, M.-J., & Luján, M. (2012). Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection. *Journal of Machine Learning Research*, *13*(2), 27–66. http://jmlr.org/papers/v13/brown12a.html

Brown, M. B., & Forsythe, A. B. (1974). Robust Tests for the Equality of Variances. *Journal of the American Statistical Association*, *69*(346), 364–367. https://doi.org/10.1080/01621459.1974.10482955

Chen, B., Shang, X., Li, M., Wang, J., & Wu, F.-X. (2016). Identifying Individual-Cancer-Related Genes by Rebalancing the Training Samples. *IEEE Transactions on NanoBioscience*, *15*, 1–1. https://doi.org/10.1109/TNB.2016.2553119

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Cheng, L., Zhao, H., Wang, P., Zhou, W., Luo, M., Li, T., Han, J., Liu, S., & Jiang, Q. (2019). Computational Methods for Identifying Similar Diseases. *Molecular Therapy - Nucleic Acids*, *18*, 590–604. https://doi.org/10.1016/j.omtn.2019.09.019

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

DAVID: Functional Annotation Tools. (n.d.). Retrieved April 8, 2022, from https://david.ncifcrf.gov/tools.jsp

Elati, M., & Rouveirol, C. (2010). Unsupervised Learning for Gene Regulation Network Inference from Expression Data: A Review. In *Algorithms in Computational Molecular Biology: Techniques, Approaches and Applications* (pp. 955–978). https://doi.org/10.1002/9780470892107.ch41

El-Hadj Imorou, S. (2020). Socio-Economic and Health Determinants of Rural Households Consent to Prepay for Their Health Care in N'Dali (North of Benin). *Open Journal of Social Sciences*, *08*(05), 348–360. https://doi.org/10.4236/jss.2020.85024

Ersoz, N. S., Bakir-Gungor, B., & Yousef, M. (2023). GeNetOntology: Identifying affected gene ontology terms via grouping, scoring, and modeling of gene expression data utilizing biological knowledge-based machine learning. *Frontiers in Genetics*, *14*. https://www.frontiersin.org/articles/10.3389/fgene.2023.1139082

Fang, O. H., Mustapha, N., & Sulaiman, M. N. (2014). An integrative gene selection with association analysis for microarray data classification. *Intelligent Data Analysis*, *18*(4), 739–758. https://doi.org/10.3233/IDA-140666

Fix, E., & Hodges, J. L. (1989). Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. *International Statistical Review / Revue Internationale de Statistique*, *57*(3), 238. https://doi.org/10.2307/1403797

Franz, M., Lopes, C. T., Huck, G., Dong, Y., Sumer, O., & Bader, G. D. (2016). Cytoscape.js: A graph theory library for visualisation and analysis. *Bioinformatics*, *32*(2), 309–311. https://doi.org/10.1093/bioinformatics/btv557

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (With discussion and a rejoinder by the authors). *The Annals of Statistics*, *28*(2). https://doi.org/10.1214/aos/1016218223

GeneMANIA. (n.d.). Retrieved April 8, 2022, from https://genemania.org/

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., & Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics*, *10*(5), e1004383. https://doi.org/10.1371/journal.pgen.1004383

Goldman, M. J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., Banerjee, A., Luo, Y., Rogers, D., Brooks, A. N., Zhu, J., & Haussler, D. (2020). Visualizing and interpreting cancer genomics data via the Xena platform. *Nature Biotechnology*, *38*(6), 675–678. https://doi.org/10.1038/s41587-020-0546-8

Greene, C. S., Krishnan, A., Wong, A. K., Ricciotti, E., Zelaya, R. A., Himmelstein, D. S., Zhang, R., Hartmann, B. M., Zaslavsky, E., Sealfon, S. C., Chasman, D. I.,

FitzGerald, G. A., Dolinski, K., Grosser, T., & Troyanskaya, O. G. (2015). Understanding multicellular function and disease with human tissue-specific networks. *Nature Genetics*, *47*(6), 569–576. https://doi.org/10.1038/ng.3259

Hadj Taieb, M. A., Ben Aouicha, M., & Ben Hamadou, A. (2013). Computing semantic relatedness using Wikipedia features. *Knowledge-Based Systems*, *50*, 260–278. https://doi.org/10.1016/j.knosys.2013.06.015

Hamzeh, O., & Rueda, L. (2019). *A Gene-disease-based Machine Learning Approach to Identify Prostate Cancer Biomarkers*. 633–638. https://doi.org/10.1145/3307339.3343479

Hanchuan Peng, Fuhui Long, & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1226–1238. https://doi.org/10.1109/TPAMI.2005.159

Hand, D., & Till, R. (2004). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Undefined*. https://www.semanticscholar.org/paper/A-Simple-Generalisation-of-the-Area-Under-the-ROC-Hand-Till/b04db132c033b31010281baa44ce547463367453

Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology*, *18*(1), 83. https://doi.org/10.1186/s13059-017-1215-1

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, *1*, 278–282 vol.1. https://doi.org/10.1109/ICDAR.1995.598994

Inza, I., Larrañaga, P., Blanco, R., & Cerrolaza, A. J. (2004). Filter versus wrapper gene selection approaches in DNA microarray domains. *Artificial Intelligence in Medicine*, *31*(2), 91–103. https://doi.org/10.1016/j.artmed.2004.01.007

Irani, J., Pise, N., & Phatak, M. (2016). Clustering Techniques and the Similarity Measures used in Clustering: A Survey. *International Journal of Computer Applications*, *134*, 9–14. https://doi.org/10.5120/ijca2016907841

Jovic, A., Brkic, K., & Bogunovic, N. (2015). A review of feature selection methods with applications. *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 1200–1205. https://doi.org/10.1109/MIPRO.2015.7160458

Kent, J. T. (1983). Information gain and a general measure of correlation. *Biometrika*, *70*(1), 163–173. https://doi.org/10.1093/biomet/70.1.163

Kolde, R., Laur, S., Adler, P., & Vilo, J. (2012). Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, *28*(4), 573–580. https://doi.org/10.1093/bioinformatics/btr709

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W., & Ma'ayan, A. (2016). Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, *44*(W1), W90–W97. https://doi.org/10.1093/nar/gkw377

Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., de Schaetzen, V., Duque, R., Bersini, H., & Nowé, A. (2012). A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *9*(4), 1106–1119. https://doi.org/10.1109/TCBB.2012.33

Lu, M., Hou, Q., Qin, S., Zhou, L., Hua, D., Wang, X., & Cheng, L. (2023). A Stacking Ensemble Model of Various Machine Learning Models for Daily Runoff Forecasting. *Water*, *15*(7), 1265. https://doi.org/10.3390/w15071265

Luo, P., Tian, L.-P., Chen, B., Xiao, Q., & Wu, F.-X. (2020). Ensemble disease gene prediction by clinical sample-based networks. *BMC Bioinformatics*, *21*(Suppl 2), 79. https://doi.org/10.1186/s12859-020-3346-8

Martens, M., Ammar, A., Riutta, A., Waagmeester, A., Slenter, D. N., Hanspers, K., A. Miller, R., Digles, D., Lopes, E. N., Ehrhart, F., Dupuis, L. J., Winckers, L. A., Coort, S. L., Willighagen, E. L., Evelo, C. T., Pico, A. R., & Kutmon, M. (2021). WikiPathways: Connecting communities. *Nucleic Acids Research*, *49*(D1), D613–D621. https://doi.org/10.1093/nar/gkaa1024

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., & Barabási, A.-L. (2015a). Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science (New York, N.Y.)*, *347*(6224), 1257601. https://doi.org/10.1126/science.1257601

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., & Barabási, A.-L. (2015b). Uncovering disease-disease relationships through the incomplete interactome. *Science*, *347*(6224), 1257601. https://doi.org/10.1126/science.1257601

Meng, L., Huang, R., & Gu, J. (2013). A Review of Semantic Similarity Measures in WordNet. *International Journal of Hybrid Information Technology*, *6*(1).

Missori, G., Serra, F., Prestigiacomo, G., Ricciardolo, A. A., Brugioni, L., & Gelmini, R. (2020). Case Report: Metastatic breast cancer to the gallbladder. *F1000Research*, *9*, 343. https://doi.org/10.12688/f1000research.23469.1

Mungloo-Dilmohamud, Z., Jaufeerally-Fakim, Y., & Peña-Reyes, C. (2020). Exploring the Stability of Feature Selection Methods across a Palette of Gene Expression Datasets. *Proceedings of the 2019 6th International Conference on Biomedical and Bioinformatics Engineering*, 7–12. https://doi.org/10.1145/3375923.3375938

Nacu, Ş., Critchley-Thorne, R., Lee, P., & Holmes, S. (2007). Gene expression network analysis and applications to immunology. *Bioinformatics*, *23*(7), 850–858. https://doi.org/10.1093/bioinformatics/btm019

Papachristoudis, G., Diplaris, S., & Mitkas, P. A. (2010). SoFoCles: Feature filtering for microarray classification based on gene ontology. *Journal of Biomedical Informatics*, *43*(1), 1–14. https://doi.org/10.1016/j.jbi.2009.06.002

Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J. F., Stijleman, I. J., Palazzo, J., Marron, J. S., Nobel, A. B., Mardis, E., Nielsen, T. O., Ellis, M. J., Perou, C. M., & Bernard, P. S. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, *27*(8), 1160–1167. https://doi.org/10.1200/JCO.2008.18.1370

Peng, J., Bai, K., Shang, X., Wang, G., Xue, H., Jin, S., Cheng, L., Wang, Y., & Chen, J. (2017). Predicting disease-related genes using integrated biomedical networks. *BMC Genomics*, *18*(1), 1043. https://doi.org/10.1186/s12864-016-3263-4

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, A. L., Brown, P. O., & Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature*, *406*(6797), 747–752. https://doi.org/10.1038/35021093

Perscheid, C. (2021). Integrative biomarker detection on high-dimensional gene expression data sets: A survey on prior knowledge approaches. *Briefings in Bioinformatics*, *22*(3), bbaa151. https://doi.org/10.1093/bib/bbaa151

Pes, B., Dessì, N., & Angioni, M. (2017). Exploiting the ensemble paradigm for stable feature selection: A case study on high-dimensional genomic data. *Information Fusion*, *35*, 132–147. https://doi.org/10.1016/j.inffus.2016.10.001

Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., & Furlong, L. I. (2017). DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, *45*(D1), D833–D839. https://doi.org/10.1093/nar/gkw943

Qi, J., & Tang, J. (2007). Integrating gene ontology into discriminative powers of genes for feature selection in microarray data. *Proceedings of the 2007 ACM Symposium on Applied Computing*, 430–434. https://doi.org/10.1145/1244002.1244101

Qumsiyeh, E., Showe, L., & Yousef, M. (2022). GediNET for discovering gene associations across diseases using knowledge based machine learning approach. *Scientific Reports*, *12*(1), Article 1. https://doi.org/10.1038/s41598-022-24421-0

Qumsiyeh, E., Yazıcı, M., & Yousef, M. (2023a). GediNETPro: Discovering Patterns of Disease Groups. *Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies - BIOINFORMATICS*, 195–203. https://doi.org/10.5220/0011690800003414

Qumsiyeh, E., Yazıcı, M., & Yousef, M. (2023b). *GediNETPro: Discovering Patterns of Disease Groups*. *2*, 195–203. https://doi.org/10.5220/0011690800003414

Raghu, V. K., Ge, X., Chrysanthis, P. K., & Benos, P. V. (2017). Integrated Theory- and Data-driven Feature Selection in Gene Expression Data Analysis. *Proceedings. International Conference on Data Engineering*, *2017*, 1525–1532. https://doi.org/10.1109/ICDE.2017.223

Real, R., & Vargas, J. (1996). The Probabilistic Basis of Jaccard's Index of Similarity. *Systematic Biology - SYST BIOL*, *45*, 380–385. https://doi.org/10.1093/sysbio/45.3.380

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Suleman, R. M., & Korkontzelos, I. (2021). Extending latent semantic analysis to manage its syntactic blindness. *Expert Systems with Applications*, *165*, 114130. https://doi.org/10.1016/j.eswa.2020.114130

Suratanee, A., & Plaimas, K. (2015). DDA: A Novel Network-Based Scoring Method to Identify Disease-Disease Associations. *Bioinformatics and Biology Insights*, *9*, BBI.S35237. https://doi.org/10.4137/BBI.S35237

Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015a). The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemporary Oncology*, *19*(1A), A68–A77. https://doi.org/10.5114/wo.2014.47136

Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015b). The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemporary Oncology (Poznan, Poland)*, *19*(1A), A68-77. https://doi.org/10.5114/wo.2014.47136

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., & Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, *415*(6871), Article 6871. https://doi.org/10.1038/415530a

Wang, J., Zheng, J., Wang, Z., Li, H., & Deng, M. (2018). Inferring Gene-Disease Association by an Integrative Analysis of eQTL Genome-Wide Association Study and Protein-Protein Interaction Data. *Human Heredity*, *83*(3), 117–129. https://doi.org/10.1159/000489761

Wang, R. (2012). AdaBoost for Feature Selection, Classification and Its Relation with SVM, A Review. *Physics Procedia*, *25*, 800–807. https://doi.org/10.1016/j.phpro.2012.03.160

Wang, X., Gulbahce, N., & Yu, H. (2011). Network-based methods for human disease gene prediction. *Briefings in Functional Genomics*, *10*(5), 280–293. https://doi.org/10.1093/bfgp/elr024

Wang, Z., Gu, Y., Zheng, S., Yang, L., & Li, J. (2023). MGREL: A multi-graph representation learning-based ensemble learning method for gene-disease association prediction. *Computers in Biology and Medicine*, *155*, 106642. https://doi.org/10.1016/j.compbiomed.2023.106642

Wong, T.-T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, *48*(9), 2839–2846. https://doi.org/10.1016/j.patcog.2015.03.009

Xiang, J., Zhang, J., Zhao, Y., Wu, F.-X., & Li, M. (2022). Biomedical data, computational methods and tools for evaluating disease–disease associations. *Briefings in Bioinformatics*, *23*(2), bbac006. https://doi.org/10.1093/bib/bbac006

Xu, Q.-S., & Liang, Y.-Z. (2001). Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, *56*(1), 1–11. https://doi.org/10.1016/S0169-7439(00)00122-2

Yao, X., Ouyang, S., Lian, Y., Peng, Q., Zhou, X., Huang, F., Hu, X., Shi, F., & Xia, J. (2024). PheSeq, a Bayesian deep learning model to enhance and interpret the gene-disease association studies. *Genome Medicine*, *16*(1), 56. https://doi.org/10.1186/s13073-024-01330-7

Ye, X., Shen, H., Ma, X., Bunescu, R., & Liu, C. (2016). *From Word Embeddings To Document Similarities for Improved Information Retrieval in Software Engineering*. https://doi.org/10.1145/2884781.2884862

Yousef, M., Abdallah, L., & Allmer, J. (2019). maTE: Discovering expressed interactions between microRNAs and their targets. *Bioinformatics*, *35*(20), 4020–4028. https://doi.org/10.1093/bioinformatics/btz204

Yousef, M., Bakir-Gungor, B., Jabeer, A., Goy, G., Qureshi, R., & C. Showe, L. (2021). Recursive Cluster Elimination based Rank Function (SVM-RCE-R) implemented in KNIME. *F1000Research*, *9*, 1255. https://doi.org/10.12688/f1000research.26880.2

Yousef, M., Goy, G., & Bakir-Gungor, B. (2022). miRModuleNet: Detecting miRNA-mRNA Regulatory Modules. *Frontiers in Genetics*, *13*, 767455. https://doi.org/10.3389/fgene.2022.767455

Yousef, M., Goy, G., Mitra, R., Eischen, C. M., Jabeer, A., & Bakir-Gungor, B. (2021). miRcorrNet: Machine learning-based integration of miRNA and mRNA expression profiles, combined with feature grouping and ranking. *PeerJ*, *9*, e11458. https://doi.org/10.7717/peerj.11458

Yousef, M., Jabeer, A., & Bakir-Gungor, B. (2021). SVM-RCE-R-OPT: Optimization of Scoring Function for SVM-RCE-R. In G. Kotsis, A. M. Tjoa, I. Khalil, B. Moser, A. Mashkoor, J. Sametinger, A. Fensel, J. Martinez-Gil, L. Fischer, G. Czech, F. Sobieczky, & S. Khan (Eds.), *Database and Expert Systems Applications—DEXA 2021 Workshops* (pp. 215–224). Springer International Publishing. https://doi.org/10.1007/978-3-030-87101-7_21

Yousef, M., Jung, S., Showe, L. C., & Showe, M. K. (2007). Recursive Cluster Elimination (RCE) for classification and feature selection from gene expression data. *BMC Bioinformatics*, *8*(1), 144. https://doi.org/10.1186/1471-2105-8-144

Yousef, M., Ketany, M., Manevitz, L., Showe, L. C., & Showe, M. K. (2009). Classification and biomarker identification using gene network modules and support vector machines. *BMC Bioinformatics*, *10*(1), 337. https://doi.org/10.1186/1471-2105-10-337

Yousef, M., Kumar, A., & Bakir-Gungor, B. (2020). Application of Biological Domain Knowledge Based Feature Selection on Gene Expression Data. *Entropy (Basel, Switzerland)*, *23*(1), E2. https://doi.org/10.3390/e23010002

Yousef, M., Ozdemir, F., Jaaber, A., Allmer, J., & Bakir-Gungor, B. (2022). *PriPath: Identifying Dysregulated Pathways from Differential Gene Expression via Grouping, Scoring and Modeling with an Embedded Machine Learning Approach* [Preprint]. In Review. https://doi.org/10.21203/rs.3.rs-1449467/v1

Yousef, M., Sayıcı, A., & Bakir-Gungor, B. (2021). Integrating Gene Ontology Based Grouping and Ranking into the Machine Learning Algorithm for Gene Expression Data Analysis. In G. Kotsis, A. M. Tjoa, I. Khalil, B. Moser, A. Mashkoor, J. Sametinger, A. Fensel, J. Martinez-Gil, L. Fischer, G. Czech, F. Sobieczky, & S. Khan (Eds.), *Database and Expert Systems Applications—DEXA 2021 Workshops* (pp. 205–214). Springer International Publishing. https://doi.org/10.1007/978-3-030-87101-7_20

Yousef, M., Ülgen, E., & Uğur Sezerman, O. (2021). CogNet: Classification of gene expression data based on ranked active-subnetwork-oriented KEGG pathway enrichment analysis. *PeerJ. Computer Science*, *7*, e336. https://doi.org/10.7717/peerj-cs.336

Zhang, Y., Xuan, J., Clarke, R., & Ressom*, H. W. (2013). Module-based breast cancer classification. *International Journal of Data Mining and Bioinformatics*, *7*(3), 284–302. https://doi.org/10.1504/IJDMB.2013.053309

اكتشاف ارتباطات الجينات عبر الأمراض باستخدام نهج تعلم الآلة المستند إلى المعرفة

اعداد: ايما ممدوح جريس قمصية

اشارف: د. مالك يوسف
مشرف مشارك: د. رشيد جيوسي
مشرف مشارك: د. زيدون صلاح

ملخص

إن الأمراض المعقدة مثل السكري ومرض الزهايمر والسرطان تتأثر بتركيبة من العوامل الوراثية ونمط الحياة والعوامل البيئية التي لا تتبع أنماطًا وراثية مباشرة. وتُعَدُّ الأنظمة البيولوجية معقدة للغاية ومتنوعة ومن أجل حل تلك التعقيدات المُحيطة بهذه الأنظمة يتم إجراء أبحاث واسعة النطاق في المختبرات، مما يُوَفِّر كميات هائلة من البيانات البيولوجية. في هذه الرسالة وفي دراستنا الأولى، تم تطوير نهج جديد يستند إلى تعلم الآلة يسمى GediNET لدمج المعرفة البيولوجية السابقة في مجموعات الجينات المرتبطة بالأمراض. يستخدم GediNET نهج التجميع والتسجيل والنمذجة (G-S-M) لتحديد مجموعات الجينات الأفضل أداءً التي يتم استخدامها بعد ذلك لتدريب نموذج تعلم آلي. وبعد خطوات استكشاف البيانات وتحضيرها تم بناء نماذج تصنيف متنوعة باستخدام تقنية Monte Carlo Cross-Validation-100 fold ومن ثم تقييم أداء هذه النماذج. من خلال تطبيق تعلم الآلة القائم على ارتباط المرض بالمرض(DDA) ، يكتشف GediNET علاقات جديدة بين الأمراض، مما يحسن التشخيص وتوقع الإصابة بالمرض والمقاربات العلاجية.

في الدراسة الثانية قمنا بتطوير GediNETPro ، وهو إصدار متقدم من GediNET .يستخدم هذا الإصدار معلومات التحقق المتقاطع Cross-Validation (CV) وتقنيات التجميع مثل K-means لتحديد أنماط ارتباط مجموعات الأمراض. ويوفر GediNETPro أدوات تصور مثل الخرائط الحرارية وتحليلًا عميقًا لتجمعات مجموعات الأمراض مما يساهم في تطوير تداخلات تشخيصية فعالة.
أما بالنسبة للدراسة الثالثة فقد استفادت من البيانات على المستوى الجزيئي لتطوير أساليب فعالة لتوقع ارتباط الأمراض ببعضها وذلك من خلال تطوير تقنية إحصائية باستخدام نموذج G-S-M-P لـ GediNETPro لحساب مقاييس التشابه الدلالي بين الأمراض. تعمل طريقة التشابه الدلالية على اكتشاف الأمراض الممثلة ضمن التجمعات وإنشاء علاقة دلالية بين المرض المعني فيه بالبحث وبين أمراض أخرى. تسهم الدراسات المقدمة في هذه الرسالة في فهم تعقيدات الأمراض واكتشاف ارتباطاتها ببعضها البعض وتحديد العلامات البيولوجية المحتملة وأهداف العقاقير.