Maham Khokhar ENHANCING GROUPING-SCORING-MODELING (G-S-M) APPROACH THROUGH A STATISTICAL PRE-SCORING COMPONENT: A CASE STUDY FOR HIGH-DIMENSIONAL TRANSCRIPTOMIC DATA ANALYSIS AGU 2024

ENHANCING GROUPING-SCORING-MODELING (G-S-M) APPROACH THROUGH A STATISTICAL PRE-SCORING COMPONENT: A CASE STUDY FOR HIGH-DIMENSIONAL TRANSCRIPTOMIC DATA ANALYSIS

A THESIS

SUBMITTED TO ABDULLAH GÜL UNIVERSITY

SOCIAL SCIENCES INSTITUTE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF MASTER OF SCIENCE

By

Maham Khokhar

July 2024

Kayseri, Türkiye

ENHANCING GROUPING-SCORING-MODELING (G-S-M) APPROACH THROUGH A STATISTICAL PRE-SCORING COMPONENT: A CASE STUDY FOR HIGH-DIMENSIONAL TRANSCRIPTOMIC DATA ANALYSIS

A THESIS

SUBMITTED TO THE DEPARTMENT OF DATA SCIENCE AND THE GRADUATE SCHOOL OF SOCIAL SCIENCES OF ABDULLAH GUL UNIVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

By

Maham KHOKHAR

July 2024

Kayseri, Türkiye

SCIENTIFIC ETHICS COMPLIANCE

I hereby declare that all information in this document has been obtained in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name-Surname: Maham Khokhar

Signature :

REGULATORY COMPLIANCE

M.Sc. thesis titled Enhancing Grouping-Scoring-Modeling (G-S-M) Approach Through A Statistical Pre-Scoring Component: A Case Study For High-Dimensional Transcriptomic Data Analysis. Data has been prepared in accordance with the Thesis Writing Guidelines of the Abdullah Gül University, Graduate School of Social Sciences Institute.

Prepared By	Advisor
Maham Khokhar	Associate Prof. Burcu BAKIR GÜNGÖR
Signature	Signature

Head of the DATA SCIENCE Program

Assoc. Prof. Umut TÜRK

Signature

ACCEPTANCE AND APPROVAL

M.Sc. thesis titled “Enhancing Grouping-Scoring-Modeling (G-S-M) Approach Through A Statistical Pre-Scoring Component: A Case Study For High-Dimensional Transcriptomic Data Analysis “ and prepared by Maham Khokhar has been accepted by the jury in the Data Science Graduate Program at Abdullah Gül University, School of Social Sciences Institute.

…….. /…….. / ……..

JURY:

Advisor : Assocaite. Prof. Burcu BAKIR GÜNGÖR

Member: Associate. Prof. Umut TÜRK

Member: Prof. Malik YOUSEF

APPROVAL:

The acceptance of this M.Sc. thesis has been approved by the decision of the Abdullah Gül University, Graduate School of Social Sciences Institute, Management Board dated ….. /….. / …….. and numbered ……..……. .

…….. /…….. / ……..

(Date)

Associate. Prof. Umut TÜRK

Name Surname : Maham Khokhar
Program : Data Science (Master of Science)
Advisor : Assist. Prof. Burcu BAKIR GÜNGÖR
Thesis Title : Enhancing Grouping-Scoring-Modeling (G-S-M) Approach Through A Statistical Pre-Scoring Component: A Case Study For High-Dimensional Transcriptomic Data Analysis
Date and Place : July, 2024 – Kayseri, Turkey

ABSTRACT

Rapid advancements in transcriptomic technologies have significantly increased the volume of data available for analysis, which presents challenges in terms of efficiency and computational demand. This thesis introduces a Pre-Scoring component to the Grouping-Scoring-Modeling (G-S-M) framework to address inefficiencies caused by the excessive number of gene groups generated by traditional GSM. By selectively prioritizing gene groups based on their statistical significance, this innovation aims to reduce the computational demands associated with scoring these groups using machine learning models, thereby streamlining the analysis process. Assessed across nine diverse Gene Expression datasets, the Pre-Scoring G-S-M framework not only maintained accuracy comparable to the traditional approach but did so with significantly fewer genes. This refinement conserves resources while maintaining the robustness and reliability of the data analysis, crucial for advancing research in personalized medicine and therapeutic strategies. The findings suggest that the modified G-S-M framework serves as a valuable tool in bioinformatics, offering a more efficient approach to handling large-scale genomic datasets. Future work will focus on adapting this enhanced framework to incorporate diverse types of omics knowledge, such as proteomics and metabolomics, further optimizing its performance to broaden its applicability in both clinical and research settings

Keywords: Gene Selections, Machine Learning, Grouping Scoring Modeling, Transcriptomics, Feature Selection

Ad Soyad	: Maham Khokhar
Anabilim Dalı, Program	: Veri Bilimi Yüksek Lisansı
Tez Danışmanı	: Assist. Prof. Burcu BAKIR GÜNGÖR
Tez Başlığı	: İstatistiksel Ön Puanlama Bileşeni İle Gruplama Puanlama Modellemesi (Gsm) Yaklaşimin Geliştirilmesi: Yüksek Boyutlu Transkriptomik Veri Analizi İçin Bir Vaka Çalişmasi
Tarih ve Yer	: Temmuz, 2024 – Kayseri, Türkiye

ÖZET

Transkriptomik teknolojilerdeki hızlı ilerlemeler, analiz için kullanılabilir veri miktarını önemli ölçüde artırmış, bu da verimlilik ve hesaplama talepleri açısından zorluklar oluşturmuştur. Bu tez, geleneksel GSM tarafından üretilen aşırı sayıdaki gen gruplarından kaynaklanan verimsizlikleri ele almak için Gruplandırma-Puanlama-Modelleme (G-S-M) çerçevesine bir Ön-Puanlama bileşeni tanıtmaktadır. İstatistiksel öneme göre seçici bir şekilde gen gruplarını önceliklendirerek, bu yenilik, bu grupların makine öğrenimi modelleri kullanılarak puanlanmasıyla ilişkili hesaplama taleplerini azaltmayı hedeflemekte ve böylece analiz sürecini daha verimli hale getirmektedir. Dokuz çeşitli Gen İfadesi veri seti üzerinde değerlendirildiğinde, Ön Puanlama G-S-M çerçevesi, geleneksel yaklaşımla karşılaştırılabilir doğrulukta performans göstermekle kalmamış, aynı zamanda önemli ölçüde daha az gen ile bunu başarmıştır. Bu iyileştirme, kişiselleştirilmiş tıp ve tedavi stratejilerinde araştırmaları ilerletmek için hayati olan veri analizinin sağlamlığını ve güvenilirliğini korurken kaynakları korur.

Anahtar Kelimeler: Gen Seçimi, Makine Öğrenimi, Gruplandırma Puanlama Modelleme, Transkriptomik, Özellik Seçimi

ACKNOWLEDGEMENTS

I would like to extend my heartfelt thanks to my advisor, Assist. Prof. Burcu Bakır Güngör, for her invaluable efforts and support throughout my thesis work. Her constant helpfulness, ever-present smile, and encouragement have profoundly supported me during this journey.

Moreover, I am incredibly thankful to my parents, Farzan Nazir and Muhammad Ud Din, for their financial support and, above all, their love and understanding, even in matters they did not fully comprehend. I would also like to thank my siblings for their jokes and support that kept my life interesting.

Lastly, I owe a special thank you to my husband, Daniel, for his gentle teaching, nudging me towards the right resources and courses to help me become the best in my field. His unwavering belief in my capabilities and his physical support have been crucial throughout this process.

TABLE OF CONTENT

LIST OF ABBREVIATIONS

ACC	Accuracy
AUC	Area Under the Curve
FN	False Negatives
FP	False Positives
GEO	Gene Expression Omnibus
GFS	Group-based feature selection
G-S-M	Grouping, Scoring, and Modeling
IFS	Individual Feature Selection
MCCV	Monte Carlo cross-validation
MLM	Machine Learning Model
NLP	Natural language processing
RNA-Seq	RNA Sequencing
SEN	Sensitivity
SPE	Specificity
TN	True Negatives
TP	True Positives

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

1 INTRODUCTION

The recent surge in data generation within the field of genomics can be attributed to advancements in high-throughput technologies and their associated cost reductions [1]. This immense increase in data often referred to as 'big data,' encompasses a wealth of information crucial for unraveling the complex interweaving of biological systems [2].

Researchers have shifted their focus from examining diseases from a single 'omics' perspective, such as proteomics or genomics, to a more integrative approach. This shift aims to achieve a comprehensive understanding of the molecular causation of diseases at various levels. To understand diseases and disorders holistically, the sphere of multi-omics has emerged, encompassing epigenetics, metabolome, transcriptome and other 'omics disciplines [3]. By viewing diseases through different lenses, researchers can now discern the complex interconnectivity between molecular levels and integrate these insights to have a clearer picture [4].

The multi-omics approach is critical for comprehending disease initiation and progression [5]. However, the extensive data generated through multi-omics studies pose significant challenges for analysis and interpretation. Several technological approaches have been established to tackle this problem, including machine learning and cloud computing [6][7]. Machine learning is employed to uncover hidden patterns within biological datasets, while cloud computing facilitates the processing and management of large-scale data.

Machine learning models offer innovative methods for integrating and analyzing varied omics information, substantially aiding in discovering novel biomarkers. These

discoveries are pivotal in advancing drug discovery, improving the accuracy of disease prediction and contributing to the field of personalized medicine in the future [8].

Machine learning adeptly and swiftly navigates through extensive omics datasets to assist in detecting potential biomarkers. Supervised machine learning techniques applied to labeled omics data, enable the creation of patient profiles for specific diseases. These detailed profiles are crucial because they provide a more nuanced categorization of people in smaller groups, paving the way for personalized drug therapies [9]. Machine learning aids in a personalized medicine approach over the one-size-fits-all method presently in use [10][8]. This approach is essential for complex diseases such as cancer, where patients' profiles vary a lot, and the 'general drug medicine' approach does not work well. While using unsupervised machine learning, we can discover biomarkers associated with subgroups of the population, which can help us make better diagnostic tools for subgroups [11].

Beyond biomarker discovery, Machine learning is vital for creating predictive models trained on omics data, offering insights into the likely path of disease progression and treatment response [9]. An exciting aspect of using Machine learning with multi-omics data is the discovery of biomarkers that would not be apparent with a single omics dataset [12]. The healthcare sector is being revolutionized with the synergy of omics and machine learning.

However, one of the downsides of any omics data set is the number of features [13]. In data analysis, feature selection is one of the ways to decrease data dimensionality. In the feature selection process, a subset of features is usually selected from the entire set for two main reasons. Firstly, it helps prevent the data from being "overfitted"; secondly, it involves removing several features that might negatively affect the results or are essentially just noise [14]. Feature selection becomes even more critical in the omics dataset since it has a vast number of features. Given this problem in biological datasets, creating sophisticated methods for managing and navigating this information-rich environment is paramount [15]. Otherwise, we risk encountering data processing bottlenecks and unintelligible insights due to too much irrelevant data [16].

Therefore, in a biological setting, the principal target of feature selection is to sift through the immense landscape of features and select only the most pertinent features for the specific aim, which can range from biomarker discovery to disease classification. Feature selection is favored in omics data sets analysis over other data reduction techniques because it maintains the features' core structure and semantic integrity, thus offering scientists transparent data interpretation [17]. Diverse feature selection methods can be applied, including lasso regression, recursive feature elimination, and mutual information-based techniques. Feature Selection (FS) enables the discovery of biomarkers which might otherwise be missed in the immense number of features [18][19]. It is essential to recognize that the features used during model training can also serve as biomarkers [20]. Nevertheless, challenges like coping with noisy or unfinished data, determining the right FS strategy, and confirming the validity of the chosen features remain [21].

With the progression of the field, enhancing feature selection for more efficient handling of high-dimensional omics data will grow in significance. Feature selection can be divided into two categories. One is Individual Feature Selection; in this type, we focus on assessing each feature's value and score independently without considering other features [22]. Conventional feature selection approaches remove features with lower rankings in the dataset, using either the forward selection or the backward elimination technique. This strategy is simple and less computationally intensive. Nevertheless, it is based on the assumption that the features have no interconnectivity or codependency. However, in biological datasets, features are highly codependent, and these dependencies are imperative for grasping the complexities of biological mechanisms. With individual feature selection, we risk losing insights from their synergistic influence.

Meanwhile, group-based feature selection (GFS) considers existing interdependencies between the features; the features are clustered based on preset criteria and then grouped together [23]. This technique is more computationally expensive. However, such an approach commonly leads to a deeper insight into the data. Consequently, for biological data, GFS is often used as it helps with understanding underlying connections in omics datasets [24].

Each approach offers unique benefits depending on the research goal. In cases where the analysis calls for only a preliminary data overview and computational resources are scarce, then Individual Feature Selection (IFS) is recommended. Alternatively, if your purpose is to unearth underlying patterns in the dataset and thoroughly grasp the biological process, employing a grouping-based feature selection method that incorporates prior biological data from diverse databases may be preferable, regardless of the increased computational requirements. While IFS can identify isolated biomarkers, GFS can reveal groups of biomarkers functioning synergistically, potentially leading to a more precise and robust characterization of diseases [25].

The GFS approach can group features based on either domain knowledge, such as biological knowledge, or on statistical grouping. Statistical clustering employs data-oriented algorithms to establish clusters, often revealing unpredicted feature associations among features. Feature clustering is executed based on specific patterns or statistical characteristics in the data without pre-existing knowledge [24]. Conversely, knowledge-based clustering utilizes pre-existing information for grouping features, offering a more innate comprehension of the fundamental processes [25]. Building upon this foundation, the Grouping, Scoring, and Modeling (G-S-M) technique, developed by Malik Yousef, is a systematic approach to incorporating biological data into machine learning models.

1.1 Introducing Traditional G-S-M approach

The Grouping-Scoring-Modeling (G-S-M) framework, represents a sophisticated and comprehensive approach for integrative feature selection in omics data analysis, with broad applications beyond this initial scope. The G-S-M framework fundamentally differs from traditional feature selection techniques by not only estimating the importance of individual features but also considering groups of features as a set that is organized based on prior knowledge, thus acknowledging and utilizing the interdependence among features [26].

The methodology leverages an ensemble of machine learning and domain knowledge to group and score features based on their association with a binary-labeled

target such as control and disease. Prior knowledge can include a variety of compiled information, such as microRNA-target interactions, protein-protein interactions, and pathway associations. This multi-faceted approach is unique in that it concurrently utilizes computational and domain knowledge, with embedded feature selection at the heart of the algorithms [27].

The G-S-M framework has been employed in a myriad of bioinformatic tools, offering a testament to its versatility and robustness in handling complex biological data. Some of the notable implementations include PriPath [28], CogNet [29], maTE [30], SVM-RCE-R [31], GediNET [32], miRcorrNet [33], 3Mint [34], TextNetTopics [35], miRdisNET [36], and miRModuleNet [37]. Each of these applications harnesses the G-S-M methodology to systematically organize data into functionally significant groups and score them, thereby facilitating a nuanced exploration of biological systems and enhancing the analysis of omics data.

G-S-M's application extends to various omics analyses, from genomics and transcriptomics to proteomics and lipidomic. By incorporating prior domain knowledge, such as disease associations and pathway information, the framework enables researchers to make new discoveries and gain insights into the underlying mechanisms of diseases.

The G-S-M tool, is freely available for download from the GitHub repository, it aims to make the feature selection approach accessible to a broader audience, with potential applications in medical practice. It is noteworthy that the approach is designed to accommodate data from two classes, e.g., control and disease, which are essential for the grouping phase of the analysis. Moreover, the approach allows for the inclusion of multi-omics data, potentially improving patient stratification and the understanding of complex relationships among genes, diseases, and drug interactions.

In summary, the G-S-M framework presents a rich, integrative, and holistic approach to the analysis of complex omics data. It stands out for its rigorous combination of machine learning and prior knowledge to extract the most discriminative groups of features, offering a powerful tool for researchers in advancing the field of personalized medicine and targeted therapeutic strategies. Building upon

this robust foundation, my thesis introduces a pivotal enhancement: the Pre-Scoring component. This component is designed to boost the G-S-M approach's efficiency and precision, a critical improvement in the face of the burgeoning volume and intricacy of omics data.

1.2 Research question

This thesis draws on the significant contributions of the G-S-M framework and the latest advances in omics data analysis to propose a novel enhancement in the form of the Pre-Scoring component. While the G-S-M framework is effective, it generates a considerable large number of groups requiring extensive computational resources for scoring with machine learning models like Random Forest Classifiers—a process that can be both time-consuming and inefficient.

Addressing this challenge, the research introduces the Pre-Scoring component to refine the G-S-M approach by selecting biological groups based on their statistical significance for prioritized analysis. This innovation is intended to optimize computational efficiency in a landscape increasingly dominated by data-intensive research. Through strategic statistical filtering, the Pre-Scoring component performs initial group scoring and ranking, thus reducing computational demand and enabling a focus on the most pertinent groups for further study.

Designed to be versatile, the Pre-Scoring component can be applied to various transcriptomic datasets, reflecting the dynamic and innovative essence of this research. The anticipated advancement in feature selection precision holds promise for impactful contributions to personalized medicine, fostering targeted and efficient patient care strategies.

The driving force behind the development of the Pre-Scoring component is the need to effectively manage the vast output of groups generated by the G-S-M framework. The objective of this paper is to introduce a strategic pre-scoring component designed for integration into the original G-S-M approach, aiming to enhance the existing methodology significantly. The pre-scoring component will do initial ranking and prioritize gene groups based on statistical filtering, enabling faster

data processing with reduced computational demand. Additionally, it will allow researchers to focus on the most relevant groups based on specific criteria, avoiding the exhaustive processing of the entire dataset. The methodology section will compare the original and enhanced G-S-M approaches, detailing the anticipated outcomes and the rationale for this significant extension to the existing framework.

1.3 Research philosophy

This experiment is designed to leverage quantitative data obtained from content analysis, experimental procedures, and secondary data sources. Adopting a deductive approach, our primary goal is to investigate the impact of introducing a new statistical filter for pre-scoring groups on both the precision and efficiency of the results obtained. The methodological framework guiding this study will be rooted in positivism, contrasting with interpretivism, given that the outcomes will be quantitatively measured and include a scoring component. Specifically, groups will be evaluated using a scoring range from 0 to +1, based on different statistic [38]. This approach underscores our commitment to empirical rigor and the objective assessment of our hypothesis.

CHAPTER 2

2 MATERIALS AND METHODS

2.1 Overview of Grouping, Scoring, and Modeling (G-S-M) Approach

2.1.1 *Introduction to G-S-M*

The G-S-M approach is a comprehensive framework used in bioinformatics to integrate machine learning with prior biological knowledge for feature selection in omics data analysis. This method is distinct from traditional feature selection techniques as it evaluates sets of features (or groups) rather than individual ones, considering their interdependencies.

2.1.2 *Purpose and General Process*

The primary purpose of the G-S-M approach is to enhance the analysis and interpretation of complex biological data sets by leveraging existing knowledge about the relationships between features. It involves grouping features based on this knowledge, scoring the groups according to their relevance to the research question, and modeling to predict or classify biological outcomes based on these scores.

2.1.3 *Grouping Component*

The Grouping component, designated as "G" and illustrated in the referenced Figure 2.1, initiates the process in the G-S-M Tool. This component strategically organizes features into smaller, distinct groups based on pre-existing biological knowledge. It utilizes a user-provided grouping file to categorize features according to their established relationships with specific diseases or biological pathways, drawing

from resources like miRTarBase, KEGG pathway, etc. Examples of such groupings, including the names and associated features of each group, are detailed in the provided tables. This step is crucial as it channels the subsequent analysis towards specific predictors that are considered relevant to the groups of interest. By focusing on biologically coherent groups of features, this component enhances the contextual accuracy and relevance of the analysis, ensuring that the investigation aligns with known biological interactions and dependencies.

2.1.4 *Scoring Component*

Following the grouping, the "G" component produces multiple groups that the "S" component then evaluates. This evaluation involves partitioning the training data into a 90% training set and a 10% testing set. A classifier is trained with the training set and used to predict outcomes on the testing set, thereby generating various performance metrics like accuracy, specificity, and precision. This procedure is repeated five times for each group, and the average of these metrics from the iterations is used to score each group. The groups are subsequently ranked according to their average scores, with higher-scoring groups prioritized. This ranking is used to inform the next component and ensures that the most statistically significant and reliable groups advance in the analysis process. Table 2.1 shows an example output of the Scoring component.

Table 2.1 Example output of the Scoring(S) Component

Disease	Associated Genes	Score	Rank
LYMPHOMA, DIFFUSE	SPRY2(10), CASP10(10), CSF3(9),...	0.97	1
SKULL BASE CHORDOMA	PIK3CD(7), VEGFA(10), PIK3CB(10).....	0.94	2
MYOFIBROBLASTOMA	ALK(10), WT1(10), VIM(10), CD34(10).....	0.9	3
LEUKEMIA, MYELOID, ACCELERATED PHASE	DNMT1(10), BCR(4), ABL1(10), GADD45A(10).....	0.89	4
CARCINOMA IN SITU OF PROSTATE	ESR2(5), ESR1(8), IGF1(10), RXRA(10).....	0.88	5

In the table 2.1 above, the first column represents the disease name, also known as the group name. The second column contains all the genes associated with the disease group. The third column represents the score assigned to the disease group as computed by the scoring component. The final column provides information about the Rank of the disease group.

2.1.5 *Modeling Component*

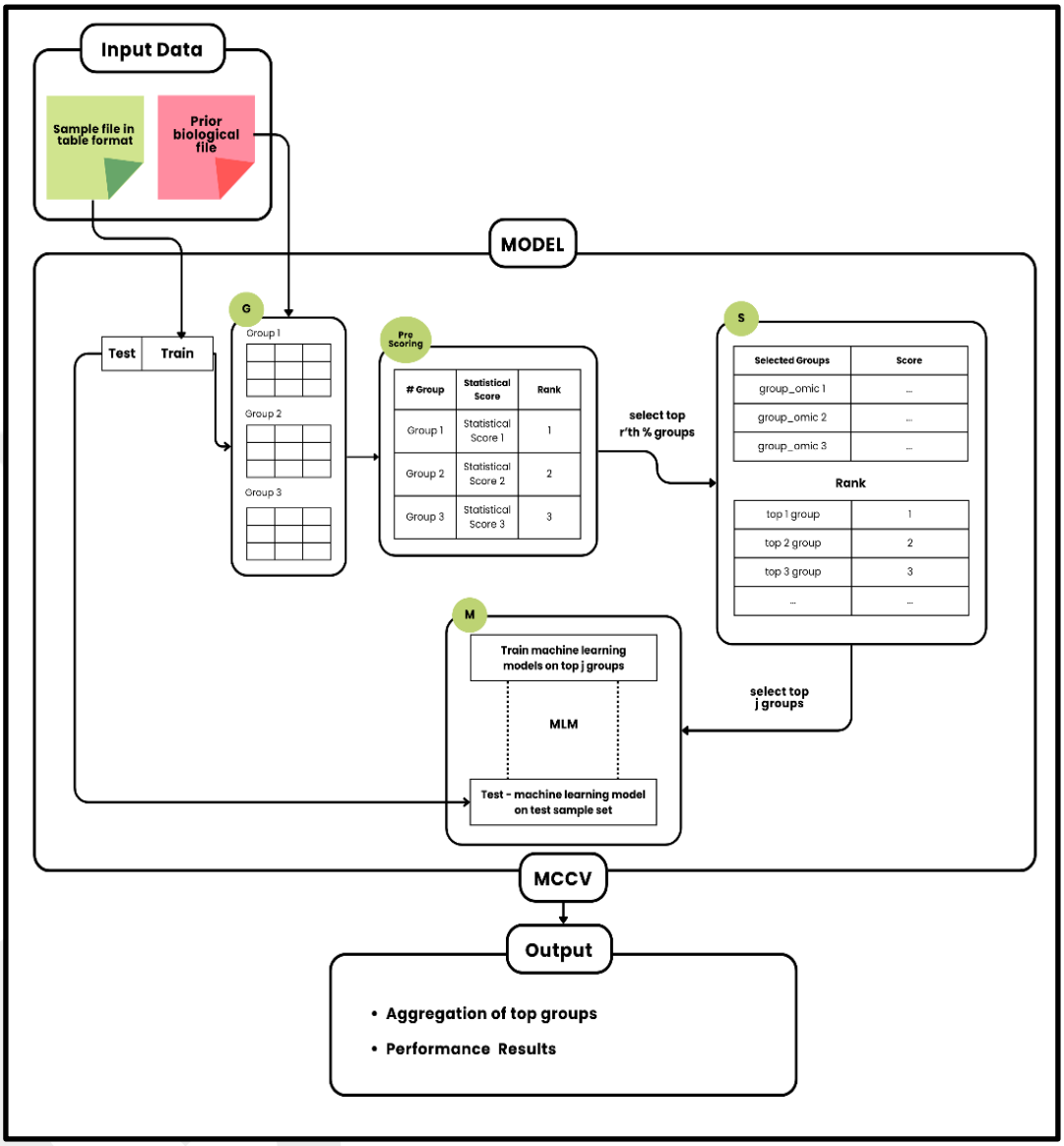
In the final phase of the G-S-M approach, the Modeling component utilizes the outcomes of the scoring phase to develop robust predictive models. This process involves training models using features from the top-ranked groups in a cumulative manner. Initially, the model is trained using only the features from the top group. It is then incrementally enriched by adding features from the next highest-ranked groups—one group at a time—until the top ten groups are included. This sequential integration allows each model to be evaluated for its effectiveness in utilizing the combined features, thereby assessing the incremental value added by each subsequent group. A specific machine learning algorithm, chosen at the outset—such as a Decision Tree, Support Vector Machine, or Random Forest—is consistently used throughout this process to ensure that the evaluation of each setup is consistent and reliable. The performance of each model setup is meticulously analyzed using accuracy, specificity, and other critical metrics, helping to pinpoint the most effective combination of groups for classification or prediction tasks.

2.1.6 *Overall Operational Workflow of the G-S-M Approach*

In the operational workflow of the G-S-M approach, Monte Carlo cross-validation (MCCV) plays a pivotal role in ensuring the robustness and reliability of the entire analysis, not just isolated components. MCCV is utilized to rigorously test and validate the models developed from the grouped and scored data. By repeatedly partitioning the data into training and testing sets at random, MCCV allows for multiple iterations of model training and evaluation, ensuring that the predictive performance is consistent across different subsets of data. This method effectively addresses potential overfitting and biases, providing a more accurate assessment of the

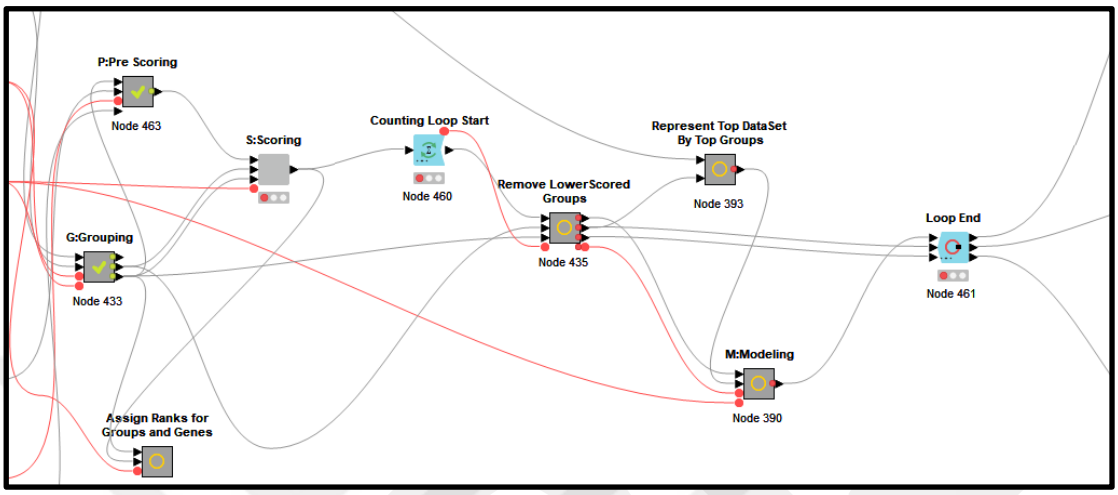
model's predictive power and generalizability. Alongside MCCV, under sampling is employed to manage data imbalance, particularly in datasets where some classes are overrepresented. This technique adjusts the composition of the training data to ensure a fair representation of all classes, facilitating more equitable and effective model training. Together, MCCV and under sampling enhance the G-S-M framework's capacity to deliver reliable and applicable insights from complex biological data. Figure 2.1 illustrates the basic flow of the Pre scoring G-S-M approach

Figure 2.1 Displays the operation of the Pre-Scoring G-S-M Tool. The primary function of Pre-Scoring G-S-M combines existing biological data to categorize genes according to their association with a grouping factor, such as diseases. This information is supplied by the user.



Users interact with the system at a higher level, setting input files and parameters without needing to delve into the workflow's complexities. With the Pre-Scoring component integrated, users can further optimize this setup by pre-selecting the most promising groups based on statistical significance before they enter the scoring phase, enhancing the framework's efficiency. Figure 2.2 illustrates the four components used in the workflow.

Figure 2.2 Show the central workflow of the G-S-M framework, emphasizing the Grouping, Pre-Scoring, Scoring, and Modeling (G-S-M) stages. The processing is managed within meta-nodes, which users can expand to explore detailed operations.



2.2 Data Collection and Preparation

In my thesis, I have introduced a new component called the 'pre-scoring node' to the standard Grouping, Scoring, and Modeling (G-S-M) framework to enhance its applicability and efficiency. While the principles of the G-S-M approach allow it to be adapted for various types of omics data, for the purpose of validating my modifications, I specifically employed a disease-gene association database alongside gene expression datasets from GEO. This choice was driven by the need to test the pre-scoring node with datasets that are rich in biological connections, focusing particularly on gene-disease pairs. However, it's important to note that the pre-scoring node is designed to be versatile and can be integrated into the G-S-M workflow with any omics dataset, maintaining the flexibility to address a wide range of biological questions and datasets in future applications

2.2.1 *DisGeNET Data Source Description*

The data that will be used to train and test the machine learning model will be coming from two sources namely DisGeNet and GEO. “DisGeNET is a discovery

platform containing one of the largest publicly available collections of genes and variants associated with human diseases [39].” It is a publicly open database frequently used for validation or experimental purposes. This dataset is generated by independent scientists and then cataloged by those intending to publish their results. The scientist generated this data in labs through in vivo testing and collected and then validated their result before publishing their data. DisGeNET retrieves data from multiple public sources, including scientific literature, public databases, and expert-curated resources. These sources include PubMed, OMIM, Orphanet, GWAS Catalog, UniProt, and others. Moreover, this platform employs Natural language processing (NLP) techniques which extract relevant information from scientific articles. Text mining algorithms search for gene-disease associations mentioned in the literature and extract details such as gene symbols, disease names, and the strength of the association [40].

2.2.2 *Criteria for Filtering DisGeNET Data*

The dataset we used, sourced from DisGeNET version 7.011, comprises genes linked with their related diseases, featuring 30,170 diseases and 21,666 genes, creating a total of 3,241,576 associations between genes and diseases. Due to the extensive nature of this dataset, we applied two specific filters to manage the data more feasibly and lessen the computational load. These filters targeted the 'diseaseType' and 'diseaseSemanticType' columns within the DisGeNET dataset. For 'diseaseType,' we narrowed our focus to entries classified under 'disease,' excluding those labeled as 'phenotype' or 'group' to align with our research objectives. Similarly, within 'diseaseSemanticType,' we selected entries tagged either as 'Neoplastic Process' or 'Disease' to enhance our study's relevance and clarity in interpreting the results. Post-filtering, the refined dataset included 15,991 genes and 3,929 diseases, amounting to 329,936 associations.

2.2.3 *Preparation of GEO Datasets*

The second data is coming from the GEO database “The Gene Expression Omnibus (GEO) database is an international public repository that archives and freely

distributes high-throughput gene expression and other functional genomics data sets [41].”

Researchers design and conduct experiments to investigate gene expression patterns in various biological contexts. These experiments can involve different conditions, treatments, tissues, cell types, and organisms. The experimental design determines the factors being studied and the samples that will be analyzed. Then researchers collect this data in a specific format and officially submit it to the GEO database where the platform runs further standardization and normalization before integrating the data into the mainframe public server [42].

We sourced nine datasets from the GEO database related to human gene expression for different complex diseases as shown in Figure 2.3. Each dataset was cataloged with the disease it pertains to and the total number of samples it includes, distinguishing between positive and negative samples. Moreover, Table 2.2 provides an example of one of these GDS 2545 Dataset.

Figure 2.3 Overview of the nine datasets used.

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 suspect lung cancer	Lung cancer	192	Negative = 90 Positive = 102
GDS3257	Cigarette smoking 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
GDS3268	Colon epithelial biopsies of ulcerative colitis patients	Colitis	202	Negative = 73 Positive = 129

Table 2.2 Displays a dataset from GEO GDS 2545

class	MAPK3	TIE1	CYP2C19	CXCR5....	MMP10
neg	535.40	121.40	11.80	62.50	32.90
neg	459.90	126.10	10.70	16.50	7.30
neg	404.10	119.60	7.00	49.80	4.10
neg	360.40	63.80	6.20	16.20	2.60
neg	527.50	109.90	10.80	22.80	6.90
neg	345.40	88.90	36.60	71.00	34.10
neg	561.70	136.10	32.10	81.50	27.60
neg	718.90	187.90	20.50	98.70	15.00
neg	389.80	56.60	4.00	12.80	1.90
neg	446.10	60.20	6.60	54.00	11.00
neg	429.80	65.60	2.30	13.60	23.40
pos	319.90	43.30	8.90	27.20	26.10
pos	685.60	93.50	5.60	38.40	6.20
pos	709.80	112.90	13.30	30.60	7.30
pos	460.60	83.80	20.10	19.40	6.40
pos	558.50	72.80	10.60	23.70	9.70
pos	698.20	47.00	9.20	50.00	8.10

The first column in Table 2.2 indicates class label information, labeled 'pos' for positive patients and 'neg' for control patients. Each row corresponds to a sample, while the columns represent genes. Each column displays the gene expression level of the patients.

2.3 Statistical Methods and Tools

This section discusses the statistical methodologies deployed in enhancing the G-S-M framework.

2.3.1 *Implementation of Levene's Test*

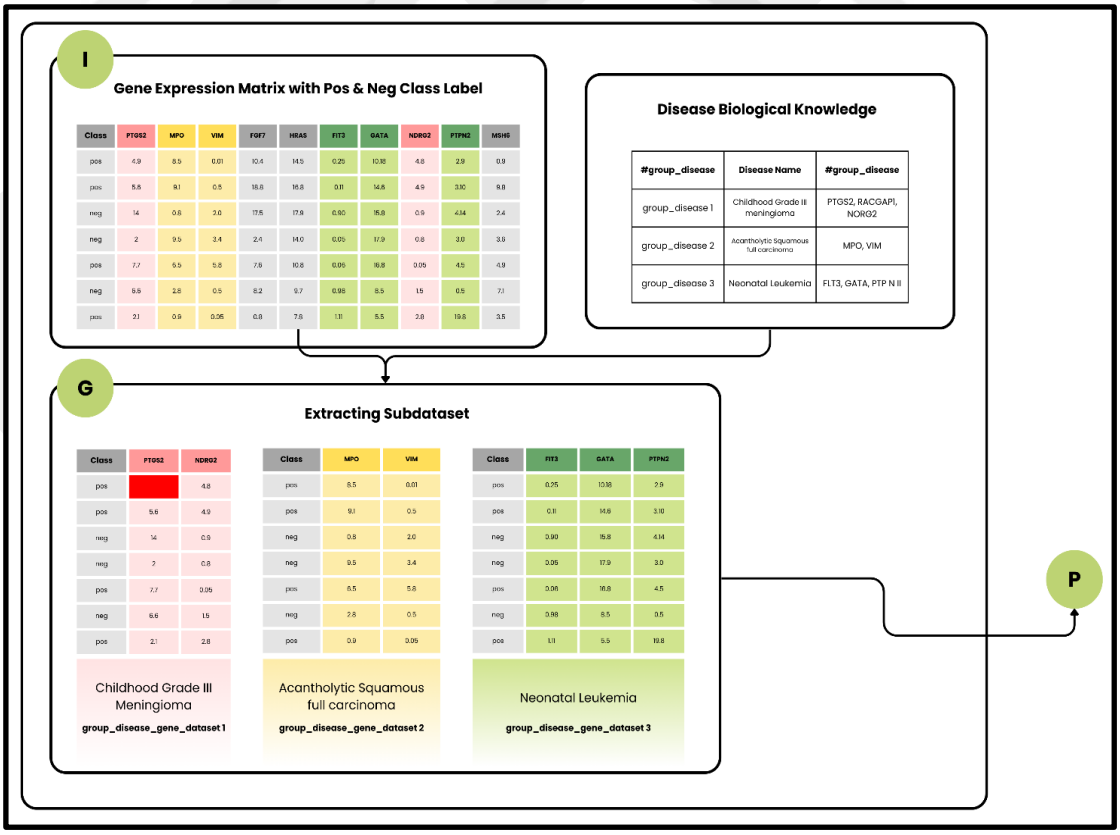
Levene's Test is utilized in our methodology both the standard G-SM and Pre scoring G-S-M assess the equality of variances across different samples for gene expression data [43]. This statistical test is essential for identifying genes that do not exhibit significant variance among conditions, allowing us to focus on genes that demonstrate meaningful differences in expression. By applying Levene's Test, we effectively remove genes with low differentiation, enhancing our dataset's quality and ensuring that our subsequent analyses are based on genes more likely to be relevant to the biological phenomena under investigation.

This preprocessing step utilizes the ANOVA test node in KNIME with a focus on Levene's T-test and p-values. An important feature of this initial filtering process was the flexibility in selecting the number of genes to proceed to the differential expression analysis. This selection is controlled by a flow variable within the KNIME analytics platform, allowing the researcher to adjust the number of genes based on their specific research objectives, desired sensitivity, and specificity thresholds. Such customization enhances the study's adaptability to different research needs, facilitating a more targeted and efficient analysis.

2.3.2 *Processing our datasets into Groups for Grouping component*

To evaluate each group effectively, two-class sub datasets is construct specific to each group (disease). This involves isolating the genes associated with each disease from the main dataset, D. For every group or disease, the relevant gene columns and their corresponding original class labels are extracted to form these sub-datasets. Suppose there are m groups; then, m two-class sub datasets are generated for input into the Pre-Scoring (P) component. An illustration of this process is shown in Figure 2.4, where the input panel (I) displays two matrices: the left matrix represents the gene expression matrix, including a 'Class' column indicating each sample's class label. The right matrix details the pre-existing biological knowledge, enumerating diseases (or group names) alongside their associated genes. For instance, in the matrix, group_disease1 is associated with 'Childhood Grade III Meningioma' and includes genes such as PTGS2, and NORG2.

Figure 2.4 This describes the creation of two-class sub datasets based on disease-group names, which are then processed by the Pre-Scoring component for Statistical scoring.



In the Grouping (G) component, the extraction process for these two-class sub datasets is conducted. As depicted, three such sub-datasets are prepared by extracting gene columns relevant to each disease group from the dataset, maintaining the original class labels where 'pos' denotes the positive class and 'neg' the negative class. These sub-datasets are being prepared to then be forwarded to the Pre-Scoring component to be evaluated and ranked based on their expression profiles relative to the disease condition.

2.3.3 *Pre-Scoring Groups using Leven T test P values*

The following formula was used for Levene test as shown in Figure 2.5

Figure 2.5 Presents a detailed view of the Levene's test equation as utilized in our analysis. It highlights the mathematical formulation used to verify the homogeneity of variances across various groups.

$$W = \frac{(N-k)}{(k-1)} \cdot \frac{\sum_{i=1}^{k} n_i (Y_{i.} - Y_{..})^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - Y_{i.})^2}$$

Where:

- W is the test statistic, which follows an F-distribution under the null hypothesis that the variances across the groups are equal.
- N is the total number of observations across all groups.
- k is the number of groups (in your case, 2: positive and negative patients).
- n_i is the number of observations in group i (number of patients in either positive or negative group).
- Y_{ij} is the absolute deviation of the j-th observation (gene expression level) in the i-th group from its group mean.
- $Y_{i.}$ is the mean of the absolute deviations in group i.
- $Y_{..}$ is the overall mean of the absc ↓ deviations across all groups.

In the initial setup for the pre-scoring node, Levene's test was selected to provide a statistical score and rank to groups based on their F-values and p-values,

reflecting the variance comparisons among groups. I collected all necessary data tables, which included detailed gene information, along with their associated F-values, p-values, and details on each disease-gene pair.

These tables were initially merged to streamline the preprocessing step. After merging the data, a groupby node was used to organize diseases with their corresponding genes. This organization was crucial for efficiently grouping the genes based on their disease associations, setting the stage for the subsequent scoring process.

The arranged data then enabled the scoring of each group by calculating the average p-value and F-value from Levene's test, with each gene within a group given equal weight. The scoring method not only quantified but also ranked the groups based on their p-values and F-values. This ranking prioritized groups with lower p-values and higher F-values, which are statistically more significant and, thereby more likely to provide biologically relevant insights.

This average p-value was employed to score the groups, as depicted in the formula in Figure 2.6 Groups were then ranked in ascending order, prioritizing those with lower p-values, indicative of higher statistical significance and potential for yielding biologically relevant insights.

Figure 2.6 Illustrates the equation used to calculate the average p-value, which is then employed to score the groups based on the p-value scores of their constituents.

$$P_g = \frac{1}{N_g} \sum_{i=1}^{N_g} p_i$$

Where:

- P_g is the average p-value for the group.
- N_g is the number of genes in the group.
- p_i is the p-value for the i-th gene in the group.

$$F_g = \frac{1}{N_g} \sum_{i=1}^{N_g} F_i$$

Where:

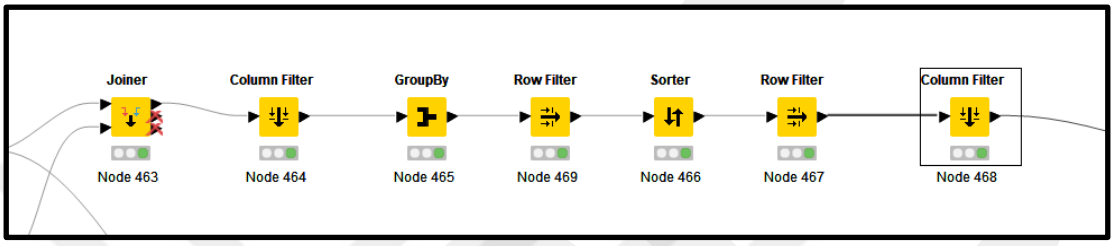
- F_g is the average F-value for the group.
- N_g is the number of genes in the group.
- F_i is the F-value for the i-th gene, indicating the test statistic associated with the variance analysis of that gene between groups (like positive vs. negative patients).

However, a notable challenge was that most top-ranked groups consisted of only one gene, which could potentially skew downstream analysis. To mitigate this, groups containing only a single gene were excluded. The workflow was designed to allow specification of the number of top-ranked groups to advance to the scoring component—initially set at 100 for testing purposes. This parameter is transformed into a flow variable, dynamically selecting the top 'k' groups for further processing.

Despite the utility of Levene's test in scoring, it was often found inadequate as it tended to select very small groups. This did not accurately reflect the variations in group sizes and adversely affected downstream analysis. Initially, the response was to remove single-gene groups; however, it was soon recognized that this might lead to the loss of significant genes. Consequently, more robust statistical methods were explored. The Limma package was integrated to provide a more sophisticated approach

to handling these statistics, ensuring decisions about group exclusion or inclusion were based on comprehensive statistical evaluation rather than merely on group size. Figure 2.7 illustrates the sequential data processing steps executed within the initial pre-scoring node.

Figure 2.7 Breaks down in detail the nodes used to transform and manipulate the data to provide statistical scores to the groups based on Levene's p-values.



The workflow encompasses several important stages:

1. **Integration**: The gene cluster table was joined with the statistics table, and unnecessary columns, such as d1 and d2, were removed to streamline the analysis.
2. **Grouping**: The data was grouped based on clusters, ensuring that each disease association was clearly demarcated for subsequent aggregation steps.
3. **Aggregation**: This crucial step involved computing counts of unique genes, concatenating gene identifiers for uniqueness, and calculating the mean of t-statistics and p-values for each group.
4. **Row** Filtering: At this stage, single-gene groups were filtered out. Despite their high individual scores, these groups were considered outliers that could skew the analysis in downstream processes.
5. **Sorting**: The remaining groups were sorted in descending order based on their computed scores, preparing the dataset for the selection of top-performing groups.
6. **Row Filtering for Selection:** A further row filter was applied to trim the dataset to the desired number of groups for deeper analysis. For instance, from an initial set of 2140 groups, only the top 100 were retained for the scoring component.

Each node in the diagram corresponds to one of these steps, illustrating a clear path from data integration to the final selection of groups for the scoring phase.

2.3.4 *Selection of the Limma Package*

This section elucidates the underlying considerations that led to the selection of the Limma package for the analysis of gene expression datasets within this research. The datasets in question are characterized by small sample sizes, presenting a set of analytical challenges that Limma is well-equipped to address.

Data Type and Analytical Precision: Limma offers a robust framework capable of analyzing both microarray and RNA-seq data. The package's empirical Bayes methods are particularly valuable for contexts with limited sample sizes, as they provide more stable inferences by borrowing strength across genes, thus enhancing the reliability of the results [44]

Flexibility in Experimental Design: Given the multifaceted experimental designs typical in genomic studies, a versatile analysis tool is imperative. Limma's ability to manage complex designs, integrating multiple factors or covariates, is instrumental. Its functionality enables the construction of various contrasts and supports nuanced comparisons across different conditions [44], aligning perfectly with the multifarious design of the current study.

Statistical Integrity: The use of linear models in conjunction with empirical Bayes statistics positions Limma as a robust and reliable option for data analysis. Its methodological robustness is key to achieving an optimal balance between the detection of genuine biological signals (True positive) and the minimization of false discovery rates, which is especially pertinent for datasets of small to moderate sizes [45].

Provenance and Support Network: The extensive history of Limma's application within the bioinformatics field, coupled with its comprehensive documentation and the breadth of community support, underscores its reliability and

trustworthiness as a research tool where it is used in experiments involving microarrays [46][47] and RNA sequencing (RNA-seq) data [48].

Alignment with Project Goals: Considering the specific objectives of my project, which involve conducting gene expression analysis, categorizing genes into disease groups, and further analysis through machine learning techniques, Limma's proficiency in identifying differentially expressed genes with high statistical confidence is invaluable. The output it generates, including log-fold changes and adjusted p-values, furnishes critical insights for downstream analysis and informed decision-making in the subsequent stages of the research.

Furthermore, the utility of the Limma package extends beyond the analysis of gene expression datasets. It's equally adept at processing mRNA sequencing data and other omics datasets, given that these data types also require the identification and statistical assessment of differentially expressed features. Limma accommodates the unique characteristics of these datasets by applying suitable normalization and variance modeling techniques, ensuring that the findings are robust and reliable. This flexibility reinforces Limma's selection for the present study, as it could cater to future extensions involving varied molecular data types while maintaining the statistical integrity central to this research

2.3.5 *Key Statistical Metrics Provided by Limma*

The Limma package offers a suite of statistical metrics essential for the robust examination of gene expression data. While several metrics, such as Log Fold Change and t-statistic, contribute to our analytical arsenal, the primary focus remains on the adjusted p-value due to its relevance in addressing the complexities of our experimental design.

Log Fold Change (logFC): This metric quantitatively reflects the change in gene expression between two conditions, indicating both the magnitude and direction of this change. Significant deviations from zero (whether positive or negative) signal substantial shifts in gene expression levels.

Adjusted P-Value (adj.P.Val): Given the vast number of genes typically analyzed simultaneously in gene expression studies, the adjusted p-value is indispensable. It provides a corrected probability measure for the observed changes in gene expression, taking into account the multiple-testing scenario. A lower adjusted p-value denotes a higher likelihood that the detected gene expression difference is statistically significant.

Average Expression (AveExpr): This metric represents the average expression level of a gene across all samples, offering a contextual backdrop for the logFC values. It helps in evaluating the biological significance of gene expression changes and understanding the gene's overall activity within the experimental framework.

t-statistic: This statistic encapsulates the statistical support for a gene's differential expression by considering both the change's extent and the expression's variability. It is a crucial indicator of the strength of evidence for differential expression.

B-statistic: The B-statistic provides an odds ratio that a particular gene is differentially expressed based on the combined evidence of the change magnitude and its consistency. High B-statistic values indicate strong evidence for differential expression.

While Limma provides a broad spectrum of statistical tools, our research primarily utilizes the adjusted p-value. This choice is strategic and aimed at enhancing the reliability of our findings by effectively addressing the multiple comparisons issue, a common challenge in gene expression studies.

2.4 Pre-Scoring Node with Limma Integration

2.4.1 *Data Preparation for Limma Analysis*

The gene expression dataset, structured with genes as columns and samples as rows as shown in Figure 2.8, progresses through the workflow to the pre-scoring component after undergoing under sampling and variance stabilization via Levene's test. This stage is crucial as it prepares the dataset for integration with the Limma package. The data manipulation and transformation required for Limma compatibility are efficiently handled by an R Snippet node within KNIME. This step ensures that the dataset is correctly formatted and optimized for the sophisticated statistical analyses that Limma will perform.

Figure 2.8 Shows the gene expression dataset used as input for the Knime workflow.

Table "default" - Rows: 146 | Spec - Columns: 12581 | Properties | Flow Variables

Row ID	S class	D MIR4640	D PAX8	D EPHB3	D MFAP3	D BRF1	D PTPRC	D PTPN11	D NMNAT2	D STMN1
Row89	pos	0.384	0.096	0.735	0.4	0.034	0.676	0.152	0.025	0.326
Row110	pos	0.265	0.175	0.334	0.527	0.09	0.353	0.154	0.228	0.428
Row120	pos	0.482	0.151	0.506	0.246	0.175	0.209	0.317	0.064	0.214
Row101	pos	0.274	0.103	0.364	0.484	0.043	0.169	0.052	0.051	0.206
Row159	pos	0.232	0.12	0.413	0.239	0.4	0.007	0.04	0.219	0.252
Row149	pos	0.559	1	0.142	0.231	0.32	0.456	0.243	0.162	0.145
Row94	pos	0.215	0.131	0.062	0.608	0.061	0.313	0.467	0.134	0.253
Row85	pos	0.253	0.073	0.152	0.077	0.311	0.254	0.127	0.034	0.402
Row124	pos	0.181	0.098	0.45	0.639	0.434	0.196	0.22	0.04	0.313
Row132	pos	0.311	0.218	0.486	0.244	0.097	0.234	0.424	0.119	0.516
Row83	pos	0.218	0.145	0.476	0.478	0.071	0.153	0.273	0.016	0.275
Row139	pos	0.449	0.128	0.37	0.324	0.056	0.243	0.102	0.077	0.348
Row143	pos	0.347	0.179	0.498	0.157	0.165	0.183	0.074	0.073	0.808
Row112	pos	0.212	0.12	0.275	0.587	0.433	0.391	0.006	0.04	0.309
Row103	pos	0.358	0.115	0.443	0.665	0.086	0.087	0.174	0.101	0.546
Row91	pos	0.309	0.113	0.213	0.046	0.124	0.01	0.047	0.036	0.696
Row117	pos	0.221	0.124	0.44	0.472	0.074	0.39	0.459	0.046	0.362
Row148	pos	0.362	0.147	0.775	0.102	0.104	0.233	0.428	0.193	0.454
Row87	pos	0.439	0.172	0.478	0.663	0.972	0.383	0.143	0.102	0.446
Row164	pos	0.022	0.099	0.069	0.059	0.497	0.153	0.049	0.583	0.437
Row155	pos	0.401	0.112	0.354	0	0.508	0.363	0.007	0.325	0.219

2.4.2 R Code Implementation:

A segment of the R code used within this node is presented here to illustrate the analytical process:

```
library(limma)
# Assuming 'data' is the input data frame from KNIME
data <- kIn
```

```
# Extract and store RowID (assuming it's the first column)
original_rowIDs <- data[, 1]
# Ensure that all the data except 'class' column is numeric
data[, -c(1,2)] <- sapply(data[, -c(1,2)], as.numeric)
# Filter out rows with NAs in 'class' column or in the expression data
valid_rows <- complete.cases(data)
data <- data[valid_rows, ]
original_rowIDs <- original_rowIDs[valid_rows]
# Transpose the data so that genes are rows and samples are columns
exprsData <- t(data[, -c(1,2)])
# Ensure that the class column is a factor and matches the number of samples
group <- factor(data$class)
# Create the design matrix
design <- model.matrix(~ group)
# Fit the linear model
fit <- lmFit(exprsData, design)
# Apply empirical Bayes statistics
fit2 <- eBayes(fit)
# Obtain top-table results including desired statistics
results <- topTable(fit2, adjust="BH", sort.by="B", number=Inf)
# Since we have the complete cases, we can be confident that the rowIDs match the order of the results
results$GeneName <- rownames(results)
# Finally, we can reorder the data frame to have RowID as the first column
results <- results[, c(ncol(results), 1:(ncol(results)-1))]
# Pass the results to KNIME
rOut <- results
```

The results from this script are detailed in Figure 2.9**,** illustrating the statistical values obtained.

Figure 2.9 Displays the output from one of the R snippets we created in knime in which we implement the Limma package.

GeneName	logFC	AveExpr	t	P.Value	adj.P.Val	B
PAFAH1B1	-0.208	0.44	-7.259	0	0	15.104
TTLL12	0.167	0.339	6.587	0	0	11.842
SIM2	0.181	0.19	6.448	0	0	11.187
MEIS2	-0.191	0.413	-6.078	0	0	9.475
RPL29	0.12	0.268	5.872	0	0	8.546
32878_f_at	0.11	0.366	5.862	0	0	8.501
LHFPL2	-0.19	0.522	-5.811	0	0	8.274
MEN1	0.143	0.263	5.723	0	0	7.881
PYCR1	0.173	0.28	5.714	0	0	7.845
BCAM	0.15	0.259	5.691	0	0	7.74
GJA1(1)	-0.145	0.25	-5.634	0	0	7.491
GSTP1	-0.19	0.45	-5.619	0	0	7.427
TGFB3(2)	-0.151	0.253	-5.603	0	0	7.36
TRAF4	0.169	0.357	5.592	0	0	7.311
MSN	-0.162	0.339	-5.582	0	0	7.265
LDHB	-0.152	0.379	-5.576	0	0	7.239
FAM168B(1)	-0.14	0.461	-5.551	0	0	7.133
ANKHD1-EIF4EBP3	0.146	0.33	5.544	0	0	7.103
PRKAR1A(2)	-0.167	0.536	-5.484	0	0	6.845
PKP3	0.115	0.285	5.483	0	0	6.838
DPYSL3	-0.19	0.343	-5.466	0	0	6.768
RGS10(1)	0.147	0.21	5.449	0	0	6.695
RRAGA	-0.16	0.602	-5.446	0	0	6.679

2.4.3 Statistical Values and Gene Group Scoring:

Subsequent to the Limma output, the ‘disease- gene’ pair table is merged with Limma's output. Utilizing a 'groupby' node, diseases are transformed into groups with associated genes compiled in a list. Aggregation is then performed, whereby average adjusted p-values are computed for each group, ensuring equal weighting is assigned to each gene. This computation culminates in the scoring of groups based on the average adjusted p-values. Similarly, scores based on other statistical measures are also calculated.

2.4.4 Exploration of Statistical Scoring Methods:

To enhance the robustness of the scoring mechanism in identifying biologically relevant gene groups, an initial approach using, based on a square root adjustment for group size, was initially introduced. The intention behind this adjustment was to

provide a counterbalance for the influence of group size, thereby enabling a more equitable comparison among groups of varying sizes. The adjustment factor, was calculated using the formula as visualized in the provided Figure 2.10.

Figure 2.10 Showcases the formula for the penalty method, which is based on the group size, meaning it is based on the number of genes in each group.

$$\text{Adjustment Factor} = \frac{1}{\sqrt{\text{Group Size}}}$$

By incorporating this adjustment factor, the aim was to attenuate the impact of group size so that larger groups wouldn't unduly overshadow smaller ones. In theory, this penalization would leverage the adjusted p-value against the square root of the group size, offering an advantage to larger groups that might otherwise be overly penalized for their size while not discounting smaller groups with potentially significant genes.

This adjustment was intended to offset the influence of group size on the p-value. The penalty score for each gene group was then obtained by multiplying this adjustment factor with the group's mean adjusted p-value, yielding a score that reflects both the statistical significance and the group size as shown:

Penalty Score=Mean Adjusted P-Value × Adjustment Factor

Despite this approach's theoretical appeal, it presented practical challenges when applied. The adjustment factor skewed the results rather than normalizing the influence of group size across the dataset. It inadvertently favored groups of extreme sizes—very large or very small—leading to an imbalanced representation of gene significance. Particularly, some smaller yet potentially significant groups were marginalized, resulting in their exclusion from further analysis.

In response to these findings, an alternative method was adopted, which centered on the use of mean-adjusted p-values without any group size penalty as shown in Figure 2.11.

Figure 2.11 Presents the equation used to assign a statistical score to each group, derived from the individual scores of the genes within those groups.

$$\text{Mean Adjusted P-Value} = \frac{1}{N} \sum_{i=1}^{N} \text{Adjusted P-Value}_i$$

Where N is the number of genes in the group.

By doing so, each group's score was determined by the average significance of its constituent genes without the confounding influence of group size, allowing for a fair and unbiased evaluation of all gene groups. This methodological refinement proved more effective in discerning biologically meaningful patterns, promoting a more equitable and precise feature selection for downstream analysis in our experiments. Additionally, the pre-scoring component was augmented with key enhancements to further refine the feature selection process.

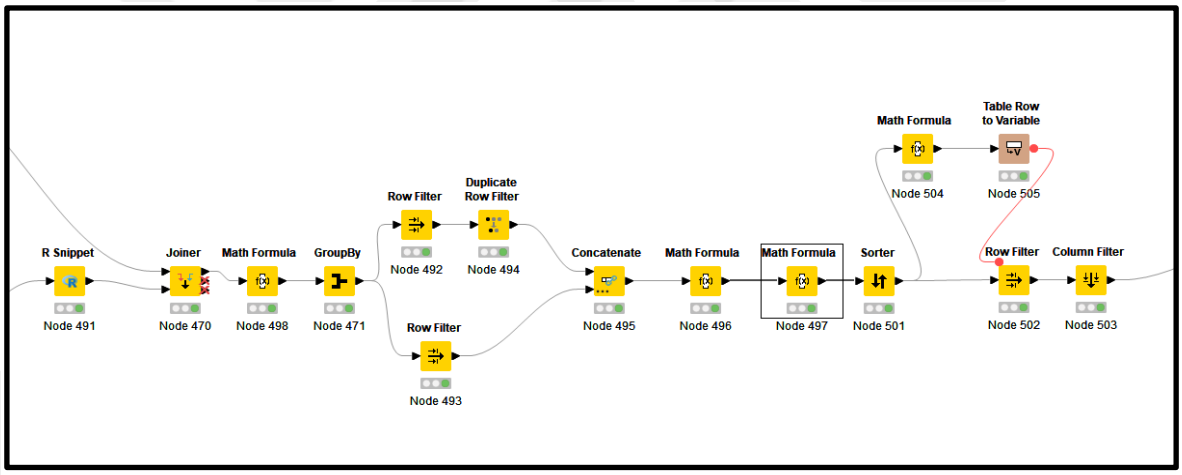
Percentage-Based Selection Mechanism: The percentage-based selection mechanism is a pivotal addition to the pre-scoring component. This innovative feature allows for the selection of a specific proportion of top-ranking groups for further analysis rather than a fixed number. By setting a percentage, the mechanism adapts to the volume of data, enabling a flexible approach that is responsive to the variability inherent in different datasets. This dynamic scaling ensures that selection criteria remain consistent, regardless of the dataset size, promoting a focused analysis of the most statistically relevant groups.

Streamlining Single-Gene Associations: In refining the pre-scoring process, particular attention was given to groups consisting of a single gene linked to multiple diseases. It was observed that certain genes were repeatedly appearing across different groups, each time associated with a distinct disease. To address this redundancy and enhance the clarity of the dataset, a curation step was instituted. For each gene found

in multiple one-gene groups, duplicates were removed, leaving only a single instance of that gene in the dataset. This approach ensures that the presence of a gene in the analysis is singular and not over-represented by its association with multiple disease terms, which might otherwise inflate its importance and skew the data set's overall pattern of disease associations.

The subsequent Figure 2.12 below **il**lustrates the Pre-Scoring Component with Limma Integration Workflow. This schematic provides a visual representation of the sequential steps and analytical nodes that compose the refined pre-scoring component.

Figure 2.12 Displays the comprehensive details of the Pre-Score component integrating the Limma package.



Detailed below are the key stages of the workflow, each underpinning the advanced processing of gene expression data with the integration of Limma's statistical analysis.

1. **R Snippet Execution:** The workflow begins with an R Snippet node, where the Limma package is applied. Here, gene expression data is processed, and linear models are fitted to the dataset using Limma's empirical Bayes methods to compute statistics like t-scores and adjusted p-values.

2. **Data Joining:** A Joiner node merges the output of the Limma analysis with additional data, ensuring that all relevant information for each gene is consolidated into a single dataset for comprehensive scoring.

3. **Mathematical Transformation:** Following the join, a Math Formula node is used to apply the square root adjustment factor to the group sizes, as per the earlier described methodology, preparing for penalty score calculations.

4. **Data Grouping by Disease:** The GroupBy node categorizes the genes into disease-specific clusters, which is critical for understanding gene-disease relationships and for the subsequent scoring based on the computed statistics.

5. **Filtering Redundancies:** Duplicate Row Filter nodes eliminate redundant gene-disease associations, particularly where a single gene is linked to multiple diseases, ensuring a unique representation of each gene.

6. **Data Concatenation:** Concatenate nodes are used to compile the data, making sure that every unique gene-disease association is represented once, providing a clean and accurate dataset for further analysis.

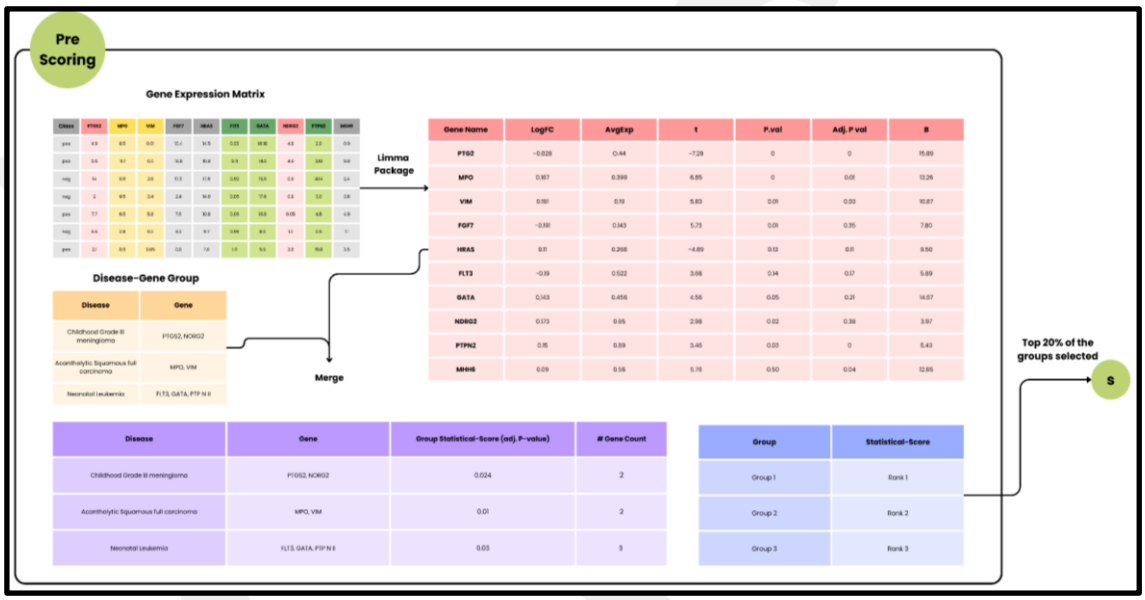
7. **Pre-Score Computation and Sorting:** Additional Math Formula nodes calculate the final penalty scores by multiplying the mean adjusted p-values with the square root adjustment factors. Based on these scores, a Sorter node then organizes the gene groups in descending order.

8. **Top Group Selection**: A Row Filter is utilized to select a specified percentage of top groups based on their penalty scores, aligning with the percentage-based selection mechanism described earlier.

9. **Final Data Preparation:** Lastly, a Table Row to Variable node followed by a Row Filter and a Column Filter ensures that only the most relevant and statistically significant groups are retained for the final analysis. The dataset is formatted to precisely fit the requirements of the scoring component.

Figure 2. 13 illustrates the Pre-Scoring process used in the Grouping-Scoring-Modeling (G-S-M) framework for enhanced feature selection in transcriptomic data.

Figure 2.13 Illustrates the Pre-Scoring process used in the Grouping-Scoring-Modeling (G-S-M) framework for enhanced feature selection in transcriptomic data.



The process for pre-scoring components begins with a Gene Expression Matrix, where class labels and gene expression levels are listed. The data then undergoes analysis via the Limma package, resulting in calculated values such as LogFC, AvgExp, t-statistics, p-values, adjusted p-values, and B statistics for each gene. These results are merged with the disease-gene group information to assess each group's statistical score based on adjusted p-values. The table in the lower section displays these scores alongside the gene count for each group. Ultimately, the top 20% of groups, based on their statistical scores, are selected for further analysis.

As depicted in Figure 2.14, following the selection of the top 20% of gene groups through our Pre-Scoring component, these groups proceed to the Scoring component. Here, they are subjected to a rigorous evaluation using a Random Forest classifier

within a structured cross-validation framework. This phase is crucial for verifying the potential of each group to contribute to accurate disease prediction.

Figure 2.14 Scoring Component of the Pre-Scoring G-S-M Framework

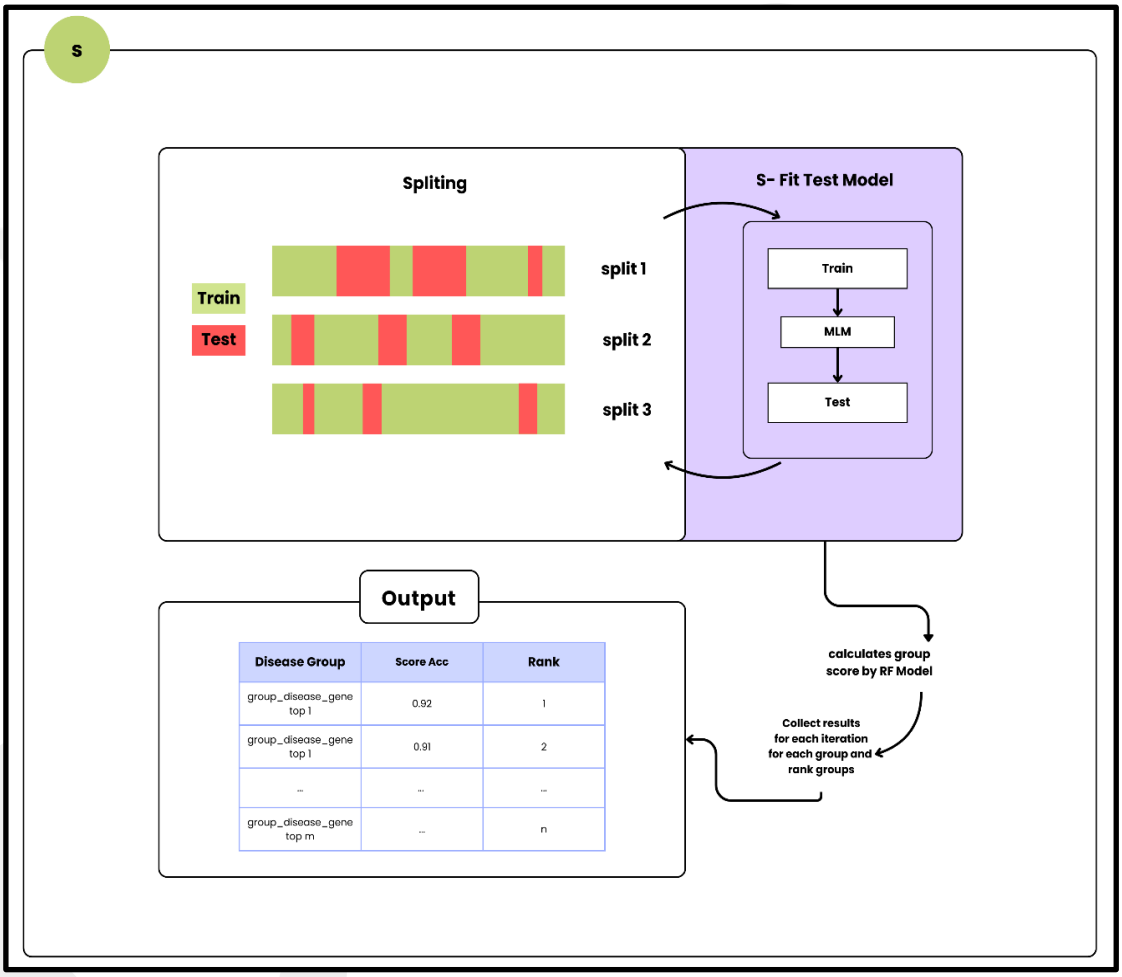


Figure 2.14 demonstrates the Scoring component of the Pre-Scoring Grouping-Scoring-Modeling (G-S-M) framework. It starts with the dataset splitting into three parts for cross-validation, which is then processed through the S-Fit Test Model. This model trains and tests the data using a Machine Learning Model (MLM), specifically a Random Forest classifier, to assess the efficacy of the gene groups selected during the Pre-Scoring phase. The outputs, including accuracy scores and ranks for each

disease-gene group, are then collated and displayed in the output table, showing the ranking based on the calculated scores.

CHAPTER 3

3 RESULTS

3.1 Evaluation of the G-S-M Framework with Pre-Scoring

In the evaluation phase of the G-S-M framework enhanced with a Pre-Scoring component, the methodologies employed are consistent with those used in established G-S-M practices. The Random Forest Classifier is utilized with a split of 90% training data and 10% testing data, which aligns with the typical setup for handling omics datasets in this analytical framework. An under-sampling method is applied to manage the imbalanced nature of the datasets, aiming for equitable representation of 2:1 during model training.

Monte Carlo cross-validation (MCCV) is utilized to rigorously assess the model performance, with performance measures computed as the average of 10 iterations. This technique supports the objective evaluation of the model by reducing variance and enhancing the reliability of results.

The performance metrics calculated include Accuracy, Sensitivity, and Specificity as shown in Figure 3.1 calculated from true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Additionally, the Area Under the Curve (AUC) is used to evaluate the classifier’s effectiveness in distinguishing between classes.

Figure 3.1 Depicts the formulas for accuracy, specificity, and sensitivity, which are a few of the performance metrics provided by the Pre-Scoring GSM.

$$\text{Sensitivity (SEN)} = \text{TP}/\left(\text{TP} + \text{FN}\right),$$

$$\text{Specificity (SPE)} = \text{TN}/\left(\text{TN} + \text{FP}\right),$$

$$\text{Accuracy (ACC)} = \left(\text{TP} + \text{TN}\right)/\left(\text{TP} + \text{TN} + \text{FP} + \text{FN}\right),$$

Rank aggregation methods are employed to manage the variability in lists of disease groups generated across iterations. This technique uses the RobustRankAggreg R package, which assigns a p-value to each element in the aggregated list, providing a statistically rigorous means to prioritize the most significant groups.

3.2 Performance Evaluation of Pre-Scoring GSM

Table 3.1 presents an example of the average 10-fold MCCV performance table for Pre-Scoring GSM for aggregated top-ranked 10 groups for the GDS1962 dataset. The first row presents the performance of the top-ranked group (#Groups = 1). The AUC obtained is 95% using 4.30 genes on average. The row of #Groups = 2 presents the performance metrics obtained for the top 2 groups, where the genes of the first top-ranked group and the second-highest scoring group are aggregated together. That is to say that Pre-Scoring GSM reports the performance results for the top 10 groups cumulatively.

Table 3.1An example showing the cumulative averages from a performance table of 10 MCCV, featuring the top-ranked 10 groups from the Pre-Scoring GSM for the GDS1962 dataset.

#Groups	#Genes	Accuracy	Area Under Curve	Precision	Specifity	F-measure	Cohen's kapp	Sensitivity
1.00	4.30	0.91	0.95	0.98	0.95	0.93	0.81	0.90
2.00	10.20	0.93	0.96	0.98	0.95	0.95	0.84	0.92
3.00	17.70	0.94	0.98	0.98	0.95	0.96	0.87	0.94
4.00	20.90	0.94	0.97	0.98	0.95	0.96	0.87	0.94
5.00	23.60	0.94	0.96	0.98	0.95	0.96	0.87	0.94
6.00	27.20	0.94	0.96	0.98	0.95	0.96	0.87	0.94
7.00	30.20	0.94	0.96	0.98	0.95	0.96	0.87	0.94
8.00	34.40	0.94	0.97	0.98	0.95	0.96	0.87	0.94
9.00	37.20	0.94	0.96	0.98	0.95	0.96	0.87	0.94
10.00	41.20	0.96	0.97	1.00	1.00	0.97	0.91	0.94

Table 3.2 displays the performance of Pre-Scoring GSM across 9 datasets for the top-performing gene groups. All values represent the results of an average of 10-MCCV iterations, with the AUC utilized to showcase performance. This table lists the GEO accession in the first column, and columns labeled #Genes, ACC (accuracy), SEN (sensitivity), SPE (specificity), and AUC (area under the curve) show detailed metrics.

Table 3.2 Performance results of Pre-Scoring G-S-M over the top-ranked groups.

GEO Accession	#Genes	Accuracy	Sensitivity	Specifity	Area Under Curve	#Groups
GDS1962	17.70	0.94	0.94	0.95	0.98	3.00
GDS2545	46.70	0.74	0.74	0.74	0.81	4.00
GDS2771	57.00	0.67	0.71	0.63	0.69	7.00
GDS3257	16.00	0.98	0.98	0.98	1.00	4.00
GDS3837	41.60	0.94	0.92	0.97	0.97	6.00
GSD4206	54.90	0.68	0.25	0.87	0.65	2.00
GDS4516_4718	3.00	0.99	0.99	1.00	1.00	3.00
GDS3268	79.10	0.71	0.69	0.73	0.73	5.00
GDS5499	39.10	0.92	0.96	0.83	0.98	2.00

An output of the Pre-Scoring GSM, similar to standard GSM tools, is a list of ranked disease groups that are assigned a P-value by the robust rank aggregation package. Table 3.3 illustrates this feature for the GDS1962 dataset, showcasing the use of this methodology within the enhanced framework.

Table 3.3 Shows one of the example output of the RobustRankAggreg tool for the dataset GDS1962

Disease name (Group)	p-value	List of genes	Gene (Count)
ADENOMA, VILLOUS	4.38E-07	CKAP4(10), TP53(10), MYC(10), UVRAG(10)...	13.00
HER-2 POSITIVE BREAST CANCER	4.43E-07	VEGFA(10), ALDH1A3(10), MALAT1(10), EGFR(10)...	14.00
ADULT SUBEPENDYMAL GIANT CELL ASTROCYTOMA	1.24E-05	DPYSL3(10), TP53(10), MYC(10), CDK1(10)...	12.00
ECCRINE POROMA	1.32E-05	MAML2(10), EGFR(10), PCNA(10), TP53(10)...	8.00
STAGE III PROSTATE CARCINOMA	1.37E-05	MYC(10), LPL(10)	2.00
MEDULLOEPITHELIOMA	3.59E-05	PTCH1(9), NOTCH2(10), SMO(10), NES(10)...	8.00
GERM CELL NEOPLASIA	4.13E-05	HSD17B3(7), HMMR(10), CKAP4(10), TP53(10)....	15.00
SURFACE EPITHELIAL-STROMAL TUMOR	7.43E-05	CYLD(10), TP53(10), MYC(10)	3.00
POROMA	8.92E-05	MAML2(10), EGFR(10), PCNA(10), TP53(10)...	8.00
DIFFUSE LARGE B-CELL LYMPHOMA OF CENTRAL NERVOUS SYSTEM	0.000122915	MYC(10), EZH2(10), MYD88(10)	3.00

This feature, commonly utilized in the standard GSM framework, can be used to analyze relationships between diseases. For instance, it can prompt biological inquiries about the association between top-ranked diseases such as ADENOMA VILLOUS, HER-2 POSITIVE BREAST CANCER and the target disease of the dataset under study, such as GDS1962 with Glioma as the target disease. The Pre-Scoring GSM framework maintains this approach by compiling a list of significant genes aggregated by the Robust Rank Aggregation tool. As disease groups are scored, the associated genes receive corresponding scores, culminating in a compiled list of significant genes. Table 3.4 presents an example of this list. This list can facilitate

functional and enrichment analyses similar to those conducted in established tools like PriPath [28], using platforms such as David, EnrichR, and GeneMANIA.

Table 3.4 Top 10 Significant Genes Aggregated by the RobustRankAggreg Tool for the GDS2771 Dataset

Name	Score
GRK2	0.000431779
NET1	0.000444852
PLN	0.000885504
INSR	0.000906996
LGR5	0.000947613
TNS4	0.000951393
TSPAN1	0.00123379
HLA-DQB2	0.001311507
NCR2	0.001324875
ATG5	0.002205237

3.3 Comprehensive Evaluation Across Diverse Datasets

Our research employed the enhanced Pre-Scoring G-S-M model on nine diverse human gene expression datasets obtained from the GEO database, covering a range of complex diseases. This extensive application has provided a robust platform for testing the model's efficiency and precision. The datasets included conditions varying widely in genetic expression profiles, which allowed us to assess the model's adaptability and effectiveness across different biological contexts. Preliminary findings indicate that

our modified G-S-M framework not only handles the intrinsic complexities associated with varied gene expression datasets but also showcases promising initial results in terms of identifying biologically significant gene groups. These early observations suggest that the Pre-Scoring component significantly contributes to the model's ability to discern key biological insights efficiently, setting a strong foundation for further detailed analyses to confirm these findings.

3.4 Enhanced Computational Efficiency

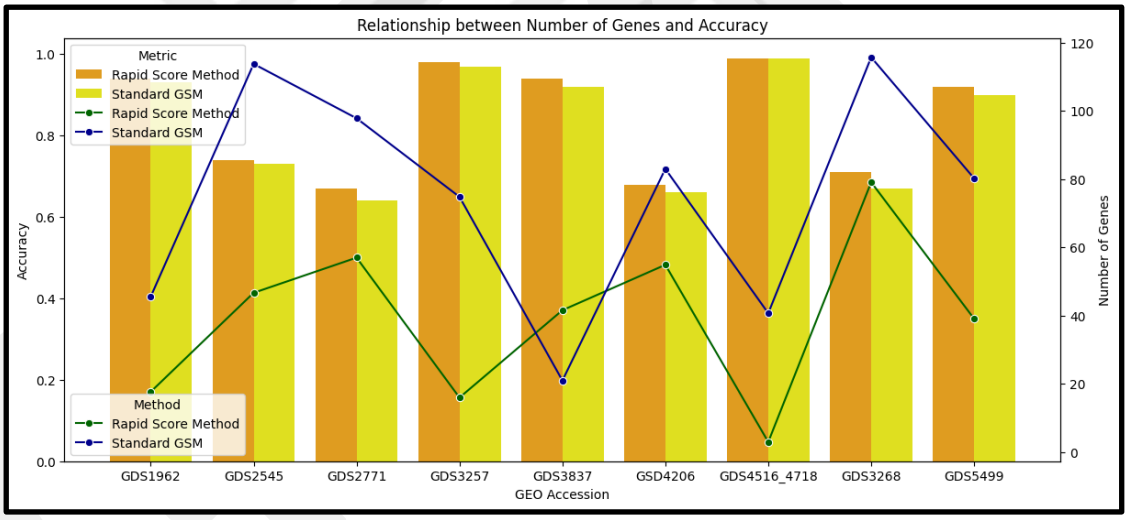
The integration of the Pre-Scoring component into our G-S-M framework has marked a significant improvement in computational efficiency. Early tests indicate a potential reduction in processing time compared to traditional GSM methods. This efficiency gain primarily stems from the component's ability to pre-filter and prioritize gene groups based on their statistical relevance before they undergo the computationally intensive scoring phase with machine learning models. By reducing the number of groups that require deep computational resources, our model not only speeds up the analysis process but also minimizes resource consumption, which is particularly advantageous in settings where computational power is a limiting factor. This preliminary data promises substantial improvements in processing large omics datasets, suggesting that our enhancements could lead to more scalable and expedient research workflows in bioinformatics.

3.5 Analytical Precision and Model Effectiveness

The utilization of the limma package in the Pre-Scoring phase has shown promising improvements in the analytical precision of our G-S-M framework. By applying sophisticated statistical filters early in the analysis process, the model is adept at identifying groups with significant statistical indicators. This enhanced precision not only ensures that the scoring phase focuses on the most promising groups but also

maintains, if not improves, the quality of biological insights derived from the data. Preliminary comparisons suggest that the precision of our enhanced model matches that of traditional G-S-M which is particularly notable given the reduced data volume processed. Figure 3.2 showcases the relationship between Number of Genes and Accuracy. Comparison of gene quantity and model accuracy between the Pre Scoring G-S-M model and the traditional G-S-M framework across nine various GEO datasets. This indicates that our methodology not only streamlines the workflow but does so without compromising on the depth and reliability of the analysis, a crucial aspect for advancing research in personalized medicine and complex disease mechanisms.

Figure 3.2 Compares the accuracy of the Standard GSM, shown in yellow, versus the Pre-Scoring GSM, depicted in orange.



The accuracy of both models is very similar. Additionally, the green line represents the number of features selected by the Pre-Scoring Model, while the blue line indicates the number of features selected by the Standard GSM. As illustrated in the chart, the Pre-Scoring GSM selects a significantly lower number of genes.

3.6 Enhanced Model Performance Through Refined Data

Initial assessments of the machine learning models, which leveraged refined groups from our enhanced G-S-M framework, demonstrate promising results. The performance metrics indicate that the Pre score G-S-M model matches the accuracy of traditional G-S-M implementations. This improved performance is particularly noteworthy as it has been achieved using fewer features, suggesting that the Pre-Scoring component effectively distills the dataset to its most informative elements. These preliminary findings highlight the potential for a more efficient and precise approach to genomic data analysis. Further detailed analyses are planned to substantiate these early results and to better understand the full capabilities of the modified framework.

3.7 Reduction in Data Redundancy

Incorporating systematic approaches to reduce redundancy, our enhanced Pre-Scoring component addresses a common challenge in genomic data analysis—the proliferation of duplicate gene-disease associations. By selectively filtering out repeated instances where a single gene appears across multiple disease categories under various labels, the model enhances the integrity and specificity of the data. This step is crucial for maintaining the semantic clarity of the gene-disease relationships. Preliminary results indicate that this approach effectively minimizes noise and unnecessary complexity in the dataset, which is anticipated to improve the overall performance of the model by focusing on genuinely informative biological signals. Further analysis will verify the impact of this reduction in redundancy on the model's efficacy and its ability to provide clearer, more actionable insights in subsequent analytical phases.

CHAPTER 4

4 DISCUSSION AND FUTURE PROSPECTS

4.1 Relevance of the Pre-Scoring Component in Existing G-S-M Tools

The introduction of the Pre-Scoring component in my thesis addresses a crucial bottleneck in the Grouping-Scoring-Modeling (G-S-M) framework: the computational intensity involved in scoring vast numbers of gene groups. For instance, in the GDS2545 dataset, which features a relatively smaller set of gene features compared to other datasets, the standard grouping process initially generates 2809 groups. However, with the implementation of the Pre-Scoring component, only 563 of these groups are forwarded for detailed scoring. This selective advancement is crucial, as each group undergoes scoring by the classifier five times in each Monte Carlo cross-validation (MCCV) iteration, cumulatively demanding substantial computational resources.

Integrating the Pre-Scoring component eliminates the necessity to score every group generated in the initial phase, focusing only on those with higher statistical significance. This approach not only streamlines the overall scoring process but also ensures that computational efforts are concentrated on groups most likely to yield pertinent biological insights. Such an adjustment markedly enhances the efficiency of the G-S-M framework, potentially reducing the computational load and time required for processing. This example underscores the value of the Pre-Scoring component in optimizing the analysis workflow, making it a pivotal enhancement for bioinformatics tools that handle large-scale genomic data.

This improvement is particularly relevant in the context of tools like GediNET [32], CogNet [29], and PriPath [28], which employ similar methodologies but may benefit from the enhanced efficiency and reduced feature set that this approach offers. When extended across 10 or 100 MCCV iterations, this computational requirement scales dramatically, emphasizing the critical role of the Pre-Scoring component in reducing the total computational burden.

4.2 Future Integration with Established Tools

Existing tools that utilize the G-S-M framework, such as GediNET [32], CogNet [29], and PriPath [28], primarily focus on integrating biological knowledge for grouping genes and scoring these groups using machine learning models like Random Forest. These tools often handle extensive gene sets, which can lead to increased computational demands. By incorporating the Pre-Scoring component, it's plausible that these tools could maintain or even improve their performance while operating more efficiently. To substantiate this potential, future studies are encouraged to systematically explore the integration of the Pre-Scoring component with these platforms. Such investigations could provide critical insights into the consistent benefits of this approach across different implementations and help optimize computational resources in bioinformatics.

4.3 Impact on Feature Selection

As noted in studies like GediNET [32], the number of genes used is high, which may not always be computationally efficient or necessary. For example, in a direct comparison using the GDS1962 dataset, the standard G-S-M approach as implemented in GediNET [32] resulted in an average of 65.5 genes per model with an accuracy of 0.92. Conversely, the implementation of my Pre-Scoring component produced significantly fewer genes, averaging only 4.5, while maintaining a comparable accuracy of 0.914 under the same experimental conditions (10 iterations of MCCV). This reduction not only streamlines the computational process but also enhances the model’s efficiency by concentrating on more relevant biological data. This approach could be particularly beneficial in reducing redundancy and focusing analysis on gene groups with higher potential impact, thus improving the overall precision of disease association studies. By potentially maintaining accuracy levels while using fewer features, the Pre-Scoring component offers a promising avenue for tools like GediNET [32] to achieve computational savings and enhance feature selection efficiency.

4.4 Comparison with Standard G-S-M

In evaluating the enhancements brought by the Pre-Scoring component integrated into the traditional Grouping-Scoring-Modeling (G-S-M) framework, a systematic comparison was conducted across nine diverse gene expression datasets. The comparison-maintained consistency in all parameters across 10 Monte Carlo cross-validation (MCCV) iterations for both the standard G-S-M and the modified Pre-Scoring G-S-M.

Our comparative analysis demonstrates that the Pre-Scoring G-S-M attains similar accuracy and performance metrics to those of the traditional G-S-M method. The distinct advantage of the Pre-Scoring G-S-M lies in its strategic selection process, where only the top 20% of groups—identified through statistical pre-scoring—are advanced for further scoring. This contrasts with the standard G-S-M approach, which proceeds to score every group generated. Additionally, the Pre-Scoring G-S-M introduces a novel feature that allows researchers to dynamically select the percentage of groups they wish to advance to the scoring phase. This flexibility enables researchers to tailor the analysis process to different dataset sizes and characteristics, selecting an optimal 'r' percent of groups generated for detailed scoring. By prioritizing only the most promising groups for detailed analysis and allowing customizable selection thresholds, the Pre-Scoring G-S-M substantially reduces the computational burden without sacrificing the accuracy or the thoroughness of the gene group evaluations, effectively streamlining the analysis while maintaining high standards of data integrity.

Additionally, this approach's efficiency in handling large datasets without an extensive computational burden highlights its potential for scalability and application in more extensive genomic studies. The reduced number of features processed also suggests a decrease in overfitting risks, potentially increasing the model's generalizability across different datasets.

However, a limitation to consider in the Pre-Scoring G-S-M is the risk of excluding potentially significant groups that do not meet the pre-scoring threshold but might still carry meaningful biological signals. While the selection criteria are

designed to prioritize statistical significance, this could inadvertently omit groups with subtler yet relevant biological implications.

This comparison underscores the balance between efficiency and comprehensiveness in genomic analysis, suggesting that while the Pre-Scoring G-S-M offers significant improvements in computational efficiency, it must be continually refined to ensure that it does not overlook biologically significant data. Further research could explore adaptive threshold settings or hybrid scoring mechanisms to optimize both performance and coverage.

Integrating the Pre-Scoring component with existing bioinformatics tools that utilize the Grouping-Scoring-Modeling (G-S-M) framework, such as GediNET [32], PriPath [28], and CogNet [29], could markedly enhance their efficiency and effectiveness. These tools currently leverage extensive datasets to identify disease-gene associations and classify gene expression data, which involves processing a significant number of gene groups. By implementing the Pre-Scoring component, these tools can prioritize gene groups based on their statistical significance before undergoing the computationally intensive scoring process. This would not only streamline their workflows by reducing the number of groups needing detailed analysis but could also maintain or even improve their performance metrics. For example, GediNET [32], which integrates biological knowledge for grouping genes, could see improved computational efficiency and reduced processing times, making it more effective in handling large-scale datasets without sacrificing accuracy.

4.5 Future Directions for the G-S-M Framework

The introduction of the Pre-Scoring component into the Grouping-Scoring-Modeling (G-S-M) framework has opened several pathways for future research and development. An immediate extension could involve adapting the Pre-Scoring component to accommodate a variety of omics datasets, such as proteomics, metabolomics, and epigenomics. This would enhance the versatility of the Pre-scoring G-S-M framework, enabling it to handle diverse biological data types and thus broadening its applicability in systems biology.

Furthermore, there is potential to refine the Pre-Scoring process by experimenting with different statistical metrics provided by tools like Limma. Researchers could be given the flexibility to select or weight specific statistical tests according to their specific research needs or dataset characteristics. This customization could lead to more tailored and precise group selection, aligning closely with the unique variables and demands of each study.

Additionally, implementing a threshold-based selection mechanism within the Pre-Scoring component presents another intriguing direction. Instead of selecting a fixed percentage of top groups, this method would allow groups to be advanced based on exceeding certain statistical thresholds, such as adjusted p-values or log-fold changes. This approach could ensure that only the most statistically significant groups are considered, reducing noise and focusing computational resources more effectively.

These advancements could significantly impact personalized medicine by enabling more nuanced and precise analysis of patient data. By enhancing the capability to discern critical biological markers and their interactions, the Pre-Scoring G-S-M framework could lead to better tailored therapeutic strategies, ultimately improving patient outcomes and the efficacy of treatments. Integrating these innovative approaches would not only advance genomic research but also move us closer to the realization of truly personalized medical interventions.

4.6 Impact on Personalized Medicine

The introduction of the Pre-Scoring component to the Grouping-Scoring-Modeling (G-S-M) framework has demonstrated preliminary efficacy in streamlining the feature selection process by effectively reducing the number of gene groups subjected to intensive analysis. This efficiency potentially accelerates the research process, allowing for quicker transitions from genomic data processing to actionable insights. The refinement in processing large volumes of data could, over time, contribute to more targeted and efficient approaches in biomedical research and therapy development.

Additionally, the adaptability of the Pre-Scoring component suggests its potential applicability across various types of omics data. This could facilitate more precise investigations into complex disease mechanisms, potentially influencing future studies aimed at discovering new therapeutic targets and strategies. By allowing researchers to focus on the most significant groups, there is a scope for enhancing the depth and quality of biomedical insights, which could ultimately lead to innovations in personalized medicine and treatment strategies.

CHAPTER 5

5 CONCLUSION AND LIMITATIONS

5.1 Conclusion

This thesis enhances the Grouping-Scoring-Modeling (G-S-M) framework by integrating a Pre-Scoring component, refining the process of feature selection in transcriptomic data analysis. By prioritizing gene groups based on their statistical relevance before the more resource-intensive scoring phase, this addition addresses the challenges of managing large omics datasets, thus streamlining computational efforts.

The application of the Pre-Scoring component has shown promise in improving efficiency without sacrificing analytical precision, making it a valuable tool for bioinformatics. It offers a practical solution for reducing computational load while maintaining high data integrity, which is crucial for advancing personalized medicine.

Looking ahead, this refined G-S-M framework could significantly impact genomic research and clinical practices by facilitating faster, more accurate analyses. Continued exploration and integration into diverse omics studies will be vital for validating its effectiveness and broadening its application, ultimately contributing to more targeted therapeutic strategies and diagnostics.

5.2 Limitations

One notable limitation of the employed methods is the selection criteria for training the model. After scoring, only the top ten groups are utilized for model training. However, it's common for multiple groups, such as the top 20, to have identical scores. Therefore, the limitation arises in choosing from the top-scored groups with the same ranking if only ten of them are to be used. Moreover, each group individually might contain weak genes, which can end up reducing the accuracy of the model. These questions require a deeper look into the working of the scoring and ranking component. It may necessitate additional measures or refined criteria to

enhance the effectiveness of the scoring system. Further, the variance in group sizes might inadvertently bias the statistical relevance, favoring either larger or smaller groups disproportionately during the statistical scoring process. These issues suggest the need for a more nuanced approach to ranking and selecting groups, potentially incorporating a mechanism to handle ties in scoring and to assess the impact of individual genes within groups more critically.

Furthermore, while the Pre-Scoring component is designed to be adaptable, its performance across vastly different omics datasets has not been fully validated. The effectiveness of the Pre-Scoring filters, primarily optimized for gene expression data, may vary when applied to other types of omics data such as proteomics or metabolomics. This variation could affect the generalizability of the model, necessitating further empirical studies to refine its application across various biological datasets.

6 REFRENCES

[1] Wong, C. (2019). Big data challenges in genome informatics. Biophysical Reviews, 11(1), 51-54. https://doi.org/10.1007/s12551-018-0493-5

[2] Ward, R. M., Schmieder, R., Highnam, G., & Mittelman, D. (2013). Systems Biomedicine, 1(1), 29–34. https://doi.org/10.4161/sysb.24470

[3] Subramanian, I., Verma, S., Kumar, S., Jere, A., & Anamika, K. (2020). Multi-omics Data Integration, Interpretation, and Its Application. Bioinformatics and Biology Insights, 14. https://doi.org/10.1177/1177932219899051

[4] Leng, D., Zheng, L., Wen, Y., Zhang, Y., Wu, L., Wang, J., Wang, M., Zhang, Z., He, S., & Bo, X. (2022). A benchmark study of deep learning-based multi-omics data fusion methods for cancer. Genome Biology, 23, Article 171. https://doi.org/10.1186/s13059-022-02739-2

[5] Wekesa, J. S., & Kimwele, M. (2023). A review of multi-omics data integration through deep learning approaches for disease diagnosis, prognosis, and treatment. Frontiers in Genetics, 14. https://doi.org/10.3389/fgene.2023.1199087

[6] Oh, M., Park, S., Kim, S., & Chae, H. (2021). Machine learning-based analysis of multi-omics data on the cloud for investigating gene regulations. Briefings in Bioinformatics, 22(1), 66-76. https://pubmed.ncbi.nlm.nih.gov/322270

[7] Diogo M. Camacho, Katherine M. Collins, Rani K. Powers, James C. Costello, James J. Collins, Next-Generation Machine Learning for Biological Networks, Volume 173, Issue 7, 2018. https://doi.org/10.1016/j.cell.2018.05.015.

[8] Kang, M., Ko, E., & Mersha, T. B. (2021). A roadmap for multi-omics data integration using deep learning. Briefings in Bioinformatics, 23(1). https://doi.org/10.1093/bib/bbab454

[9] Park, M., Lim, J., Jeong, J., Jang, Y., Lee, J., Lee, J., Kim, H., Koh, E., Hwang, S., Kim, H., & Kim, K. (2022). Deep-Learning Algorithm and Concomitant Biomarker Identification for NSCLC Prediction Using Multi-Omics Data Integration. Biomolecules, 12(12), 1839. https://doi.org/10.3390/biom12121839

[10] Reel, P. S., Reel, S., Pearson, E., Trucco, E., & Jefferson, E. (2021). Using machine learning approaches for multi-omics data analysis: A review. Biotechnology Advances, 49, 107739. https://doi.org/10.1016/j.biotechadv.2021.107739

[**11]** Vahabi, N., & Michailidis, G. (2022). Unsupervised Multi-Omics Data Integration Methods: A Comprehensive Review. Frontiers in Genetics, 13, 854752. https://doi.org/10.3389/fgene.2022.854752

[12] Picard, M., Scott-Boyer, P., Bodein, A., Périn, O., & Droit, A. (2021). Integration strategies of multi-omics data for machine learning analysis. Computational and Structural Biotechnology Journal, 19, 3735-3746. https://doi.org/10.1016/j.csbj.2021.06.030

[13] Li, Y., Mansmann, U., Du, S., & Hornung, R. (2022). Benchmark study of feature selection strategies for multi-omics data. BMC Bioinformatics 23, 412. https://doi.org/10.1186/s12859-022-04962-x

[14] Bhadra, T., Mallik, S., Hasan, N., & Zhao, Z. (2022). Comparison of five supervised feature selection algorithms leading to top features and gene signatures from multi-omics data in cancer. BMC Bioinformatics 23 (Suppl 3), 153. https://doi.org/10.1186/s12859-022-04678-y

[15] Karczewski, K. J., & Snyder, M. P. (2018). Integrative omics for health and disease. Nature Reviews Genetics, 19(5), 299-310. https://doi.org/10.1038/nrg.2018.4

[16] Xu, C., Jackson, S.A. Machine learning and complex biological data. Genome Biol 20, 76 (2019). https://doi.org/10.1186/s13059-019-1689-0

[17] A. Jović, K. Brkić and N. Bogunović. (2015) . A review of feature selection methods with applications, 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, pp. 1200-1205 10.1109/MIPRO.2015.7160458

[18] He, Z., & Yu, W. (2010). Stable feature selection for biomarker discovery. Computational Biology and Chemistry, 34(4), 215-225. https://doi.org/10.1016/j.compbiolchem.2010.07.002

[19] P. Zhang, A. Cox, A. Cripps and N. West. (2017). Integrated biomedical data analysis utilizing various types of data for biomarkers identification. 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1469-1475, 10.1109/BIBM.2017.8217879

[20] Remeseiro, B., & Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications. Computers in Biology and Medicine, 112, 103375. https://doi.org/10.1016/j.compbiomed.2019.103375

[21] Pudjihartono, N., Fadason, T., W., A., & M., J. (2022). A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. Frontiers in Bioinformatics, 2, 927312. https://doi.org/10.3389/fbinf.2022.927312

[22] Zheng, L., Chao, F., Parthaláin, N. M., Zhang, D., & Shen, Q. (2021). Feature grouping and selection: A graph-based approach. Information Sciences, 546, 1256-1272. https://doi.org/10.1016/j.ins.2020.09.022

[23] Perera, K., Chan, J., & Karunasekera, S. (2020). Advances in Knowledge Discovery and Data Mining, 805–817 https://doi.org/10.1007/978-3-030-47426-3_62

[24] Sun, X., Lu, Q., Mukherjee, S., Crane, P. K., Elston, R., & Ritchie, M. D. (2014). Analysis pipeline for the epistasis search – statistical versus biological filtering. Frontiers in Genetics, 5, 81088. https://doi.org/10.3389/fgene.2014.00106

[25] Kuzudisli, C., Bakir-Gungor, B., Bulut, N., Qaqish, B., & Yousef, M. (2023). Review of feature selection approaches based on grouping of features. PeerJ, 11. https://doi.org/10.7717/peerj.15666

[26] Yousef, M., Kumar, A., & Bakir-Gungor, B. (2020). Application of Biological Domain Knowledge Based Feature Selection on Gene Expression Data. Entropy, 23(1), 2. https://doi.org/10.3390/e23010002

[27] Yousef, M., Allmer, J., İnal, Y., & Bakir Gungor, B. (2024). G-S-M: A Comprehensive Framework for Integrative Feature Selection in Omics Data Analysis and Beyond. bioRxiv. https://doi.org/10.1101/2024.03.30.585514

[28] Yousef, M.; Ozdemir, F.; Jaber, A.; Allmer, J.; Bakir-Gungor, B. PriPath: Identifying Dysregulated Pathways from Differential Gene Expression via Grouping, Scoring, and Modeling with an Embedded Feature Selection Approach. BMC Bioinformatics 2023, 24 (1), 60. doi:10.1186/s12859-023-05187-2

[29] Yousef, M.; Ülgen, E.; Uğur Sezerman, O. CogNet: Classification of Gene Expression Data Based on Ranked Active-Subnetwork-Oriented KEGG Pathway Enrichment Analysis. PeerJ Comput. Sci. 2021, 7, e336. doi:10.7717/peerj-cs.336

[30] Yousef, M.; Abdallah, L.; Allmer, J. maTE: Discovering Expressed Interactions between microRNAs and Their Targets. Bioinformatics 2019, 35 (20), 4020–4028. doi:10.1093/bioinformatics/btz204

[31] Kotsis, G., Tjoa, A. M., Khalil, I., Moser, B., Mashkoor, A., Sametinger, J., Fensel, A., Martinez-Gil, J., Fischer, L., Czech, G., Sobieczky, F., Khan, S., EdsYousef, M.; Jabeer, A.; Bakir-Gungor, B. SVM-RCE-R-OPT: Optimization of Scoring Function for SVM-RCE-R. In Database and Expert Systems Applications - DEXA 2021 Workshops; Kotsis, G., Tjoa, A. M., Khalil, I., Moser, B., Mashkoor, A., Sametinger, J., Fensel, A., Martinez-Gil, J., Fischer, L., Czech, G., Sobieczky, F., Khan, S., Eds.; Communications in Computer and Information Science; Springer International Publishing: Cham, 2021; Vol. 1479, pp 215–224. doi:10.1007/978-3-030-87101-7_21

[32] Qumsiyeh, E.; Showe, L.; Yousef, M. GediNET for Discovering Gene Associations across Diseases Using Knowledge Based Machine Learning Approach. Sci. Rep. 2022, 12 (1), 19955. doi:10.1038/s41598-022-24421-0.

[33] Yousef, M.; Goy, G.; Mitra, R.; Eischen, C. M.; Jabeer, A.; Bakir-Gungor, B. miRcorrNet: Machine Learning-Based Integration of miRNA and mRNA Expression Profiles, Combined with Feature Grouping and Ranking. PeerJ 2021, 9, e11458. doi:10.7717/peerj.11458.

[34] Unlu Yazici, M.; Marron, J. S.; Bakir-Gungor, B.; Zou, F.; Yousef, M. Invention of 3Mint for Feature Grouping and Scoring in Multi-Omics. Front. Genet. 2023, 14, 1093326. doi:10.3389/fgene.2023.1093326

[35] Yousef, M.; Voskergian, D. TextNetTopics: Text Classification Based Word Grouping as Topics and Topics' Scoring. Front. Genet. 2022, 13.

[36] Jabeer, A.; Temiz, M.; Bakir-Gungor, B.; Yousef, M. miRdisNET: Discovering microRNA Biomarkers That Are Associated with Diseases Utilizing Biological Knowledge-Based Machine Learning. Front. Genet. 2023, 13.

[37] Yousef, M.; Goy, G.; Bakir-Gungor, B. miRModuleNet: Detecting miRNA-mRNA Regulatory Modules. Front. Genet. 2022, 13, 767455. doi:10.3389/fgene.2022.767455.

[38] Hamarashid H, Utilizing Statistical Tests for Comparing Machine Learning Algorithms Kurdistan Journal of Applied Research July 2021, DOI: 10.24017/science.2021.1.8

[39] Piñero, J., Queralt-Rosinach, N., Bravo, À., Deu-Pons, J., Bauer-Mehren, A., Baron, M., Sanz, F., & Furlong, L. I. (2015). DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes. Database: The Journal of Biological Databases and Curation, 2015.https://doi.org/10.1093/database/bav028

[40] Piñero, J., Saüch, J., Sanz, F., & Furlong, L. I. (2021). The DisGeNET cytoscape app: Exploring and visualizing disease genomics data. Computational and Structural Biotechnology Journal, 19, 2960-2967. https://doi.org/10.1016/j.csbj.2021.05.015

[41] Clough, E., & Barrett, T. (2016). The Gene Expression Omnibus database. Methods in molecular biology (Clifton, N.J.), 1418, 93. https://doi.org/10.1007/978-1-4939-3578-9_5

[42] Wang, Z., Lachmann, A. & Ma'ayan, A. Mining data and metadata from the gene expression omnibus. Biophys Rev 11, 103–110 (2019). https://doi.org/10.1007/s12551-018-0490-8

[43] Levene H. In: Olkin I, Ghurye SG, Hoeffding W, Madow WG, Mann HB, editors. Robust Tests for Equality of Variances. Stanford, Calif: Stanford University Press; 1960. p. 278–292.

[44] Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4402510/#B7

[45] Smyth, G. (2004). Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. Statistical Applications in Genetics and Molecular Biology, 3(1). https://doi.org/10.2202/1544-6115.1027

[46] Hubert, F., Kinkel, S. A., Crewther, P. E., Cannon, P. Z. F., Webster, K. E., Link, M., Uibo, R., O'Bryan, M. K., Meager, A., Forehan, S. P., Smyth, G. K., Mittaz, L., Antonarakis, S. E., Peterson, P., Heath, W. R. and Scott, H. S. (2009). Deletion Mutation Present with Only a Mild Autoimmune Phenotype1. Retrieved from https://journals.aai.org/jimmunol/article/182/6/3902/103679/Aire-Deficient-C57BL-6-Mice-Mimicking-the-Common

[47] Peart, M. J., Smyth, G. K., van Laar, R. K., Bowtell, D. D., Richon, V. M., Marks, P. A., Holloway, A. J. and Johnstone, R. W. (2005). Identification and functional significance of genes regulated by structurally different histone deacetylase inhibitors. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/article

[48] A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. (2014). Retrieved from https://www.nature.com/articles/nbt.2957