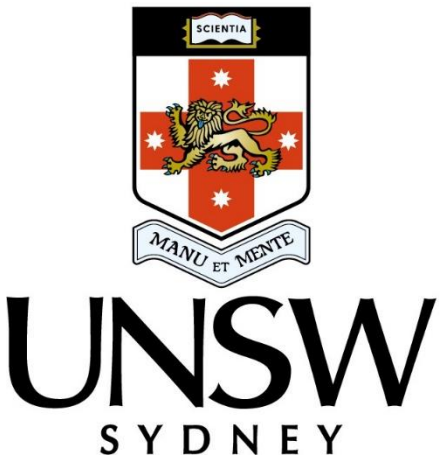


School of Computer Science and Engineering

Faculty of Engineering

The University of New South Wales

TCGA multi-modal data compilation and pan-cancer data analysis.

by

Ricky Nguyen

Thesis submitted as a requirement for the degree of
Bachelor of Bioinformatics Engineering (Honours)

Submitted: August 2024 Student ID: z5309282

Supervisor: A/Prof. Fatemeh Vafaee

Abstract

Survival analysis is crucial in cancer research, providing insights into patient prognosis and potential outcomes. Advances in imaging, sequencing, and profiling technologies have made multiple omics data increasingly available, enhancing cancer diagnosis and treatment. This thesis aims to improve cancer survival predictions through feature selection methods and machine learning algorithms applied to single modalities from The Cancer Genome Atlas (TCGA). We then integrate these features using early and late fusion approaches to assess the impact on survival predictions. The study identifies optimal modality combinations and explores commonalities across cancers, with a pan-cancer component focusing on Gene Set Enrichment Analysis (GSEA), overlapping functions and pathways, disease association networks, and Kaplan-Meier (KM) plots to examine survival signatures across breast (BRCA), ovarian (OV), cervical (CESC), and uterine corpus endometrial carcinoma (UCEC) cancers. Our findings aim to enhance survival model performance with multi-omics integration and identify multi-modal features contributing to survival outcomes, offering valuable insights for cancer research and personalized medicine.

Abbreviations

TCGA	The Cancer Genome Atlas
DM	DNA methylation
ME	MicroRNA expression
GE	Gene expression
SCNA	Somatic copy-number alteration
CNV	Copy number variation
ME	MicroRNA expression
CPH	Cox proportional hazards
CN	Copy number
SM	Somatic mutations
ML	Machine learning
BRCA	Breast
OV	Ovarian
CESC	Cervical
UCEC	Uterine corpus endometrial carcinoma
GSEA	Gene Set Enrichment Analysis
KM	Kaplan-Meier
RSF VI	Random Survival Forest Variable Importance
RSF MD	Random Survival Forest Minimal Depth
RSF VH	Random Survival Forest Variable Hunting

Contents

Content

List of Figures

List of Tables

Chapter 1

Introduction

Cancer is a complex and heterogeneous disease characterized by diverse genetic, epigenetic, and phenotypic variations [Chai et al., 2021]. Traditional single-omics analyses, such as those focusing on GE, DM, CNV, or ME, often provide only partial insights into cancer's biology. Each omics layer offers unique information: gene expression reveals changes in gene activity, DNA methylation provides insights into gene regulation, CNV identifies changes in DNA segment copies, and miRNA expression highlights regulatory roles in gene silencing [Ezgi et al., 2023]. However, analyzing these omics types in isolation fails to capture the full spectrum of interactions driving cancer progression. For a holistic understanding, it is essential to integrate multiple omics data types. This multi-omics approach leverages the strengths of each data type, uncovering complex biological networks, identifying novel biomarkers, and developing more accurate predictive models. Survival analysis is crucial for identifying factors that influence patient outcomes, aiding in the development of targeted therapies and personalized treatment plans. By integrating multi-omics data with survival analysis, researchers can identify multi-modal features significantly contributing to survival outcomes, providing insights beyond single-omics approaches. For instance, combining gene expression with DNA methylation can reveal how epigenetic modifications influence gene activity, while adding CNV and miRNA data shows how gene dosage and regulation impact patient outcomes. This thesis focuses on utilizing multi-omics integration methods on TCGA data to improve survival analysis and predictions across various women-related cancers, specifically breast cancer (BRCA), ovarian cancer (OV), cervical cancer (CESC), and uterine corpus endometrial carcinoma (UCEC). By employing advanced machine learning methodologies, the project aims to identify important features and multi-modal signatures shared across different cancer types. These signatures will enhance our understanding of the factors contributing to survival outcomes in cancer patients, potentially leading to improved diagnostic and therapeutic strategies.

Chapter 2 provides an in-depth overview of cancer biology and the limitations of single-omics data analysis. It discusses the importance of integrating multiple omics data types for a comprehensive understanding of cancer and personalized treatment strategies. Additionally, it

highlights the role of survival analysis in identifying key factors influencing patient outcomes. This chapter will review previous studies that have explored these themes, emphasizing the necessity of multi-omics approaches. Chapter 3 outlines the scope of the project, detailing the requirements and the overall pipeline used in the study. It describes the steps involved in data collection, preprocessing, feature selection, and the integration of multiple omics data types as well as the pan-cancer analysis component. Chapter 4 presents the results of the analysis and evaluation of the implemented pipelines. It discusses the methodologies used for feature selection and multi-omics integration and evaluates their impact on survival predictions. The chapter also investigates the optimal modality combinations and explores results of the pan-cancer analysis performed. Chapter 5 summarizes the findings of the study and discusses their implications for cancer research. It also outlines potential directions for future work, including further exploration of multi-omics integration methods and the application of the findings to other cancer types.

Chapter 2

Background

2.1 Context

Cancer is characterized by its intricate interactions between genetic factors and environmental influences [Chai et al., 2021]. Despite significant advancements in cancer research, substantial variations in outcomes among patients with the same cancer type persist across clinical studies. This variability poses a significant challenge in developing effective therapeutic strategies for cancers. Globally recognized as a major public health concern, cancer arises from a multitude of factors, extending beyond genetic and epigenetic control to include regulatory elements like miRNAs [Zhao et al., 2020]. The diverse regulatory factors contributing to cancer heterogeneity result in a low cure rate and poor prognosis, emphasizing the need for more precise and targeted approaches in cancer treatment. The inherent heterogeneity of cancer is evident in its diverse forms, variables, and multiple subgroups.

In 2020, there were 19.3 million new cancer cases globally, leading to nearly 10 million cancer-related deaths. Commonly diagnosed cancers included breast (11.7%), lung (11.4%), and

colorectal (10%), while prominent causes of cancer-related deaths included lung (18%), colorectal (9.4%), liver (8.3%), stomach (7.7%), and breast (6.9%) cancers. Projections estimate a potential surge to 28.4 million new cancer cases in 2040 if current incidence rates persist [Ezgi et al., 2023].

The heterogeneous nature of cancer biology transcends analysis through a singular omic data type. Cancer is a multifaceted and intricate disease, influenced by diverse factors. Single omics data types, such as genomics or transcriptomics, offer only partial insights into the complex mechanisms underlying the disease [Ezgi et al., 2023]. The inherent heterogeneity of cancer, manifested through various subtypes and molecular profiles, underscores the need for a comprehensive approach. Critical to understanding cancer is recognizing the intricate interactions among genes, proteins, metabolites, and epigenetic changes. Relying solely on individual omics data may lead to the oversight of crucial biomarkers and valuable treatment insights [Spooner et al., 2020]. Therefore, the integration of multiple omics data types becomes imperative, fostering a holistic comprehension of cancer's intricate biology and paving the way for personalized treatment strategies tailored to the unique characteristics of each patient's condition [Tabakhi et al., 2023].

The wealth of omics data, spanning genomes, transcriptomes, proteomes, metabolomics, ionomics, and epigenomes, offers researchers a comprehensive lens to delve into cancer biology [Tabakhi et al., 2023]. The advent of next-generation sequencing techniques has ushered in various genomic data types, with researchers currently identifying such biomarkers through the differential analysis of DM and other omics data, including GE, SCNA, and ME [Zhao et al., 2020]. However, multi-omics data has its various challenges, including the curse of dimensionality, imbalance, missing values, noisiness, and heterogeneity, which collectively present significant hurdles for machine learning models [Tabakhi et al., 2023].

To harness the full potential of this data, integrative methods are imperative. There are a wide range of multi-omics data integration methods, such as deep learning networks, voting strategies, network-based, clustering, feature extraction, transformation, and factorization, serve diverse applications such as disease subtyping, biomarker discovery, pathways analysis, and drug repurposing [Tabakhi et al., 2023]. This integrated approach holds promise for a more nuanced and comprehensive understanding of cancer prognosis, bridging diverse omics dimensions to unravel key insights for improved clinical outcomes.

In the realm of cancer studies, survival analysis serves as a crucial tool, addressing pivotal questions such as the factors influencing patients' survival in the face of this complex disease. It

provides insights into the probability that a patient will endure beyond a specific timeframe, often exemplified by the five-year mark. By delving into these inquiries, survival analysis not only enhances our understanding of the intricate dynamics of cancer but also contributes valuable information for prognosis and treatment planning [Spooner et al., 2020]. Cancer survival analysis encompasses binary classification and risk regression. In binary classification, patients are segregated into short- and long-survival groups based on predefined thresholds, while risk regression studies involve calculating a risk score for each patient, typically employing the CPH model and its extensions [Spooner et al., 2020]. Traditional methods, notably the widely used CPH model, encounter limitations in scaling to high dimensions and large datasets. To address this, machine learning techniques, adept at handling high-dimensional data, have been adapted for censored data analysis, offering more flexible alternatives. The challenge intensifies with multi-omics data, known for high dimensions and limited training samples, leading to overfitting. Thus, the crucial step of learning pertinent features from multi-omics data is pivotal for machine learning model-based survival and recurrence predictions [Tabakhi et al., 2023].

Publicly available multi-omics datasets, exemplified by TCGA, have significantly accelerated cancer research. TCGA has sequenced multiple types of omics data from more than ten thousand samples across 33 cancer types, enabling integrative cancer analyses based on multi-omics data and facilitating precise grading, staging, and survival predictions [Weinstein et al., 2013]. TCGA data can be accessed via the GDC, which serves as a unified repository and knowledge base for the cancer research community, encompassing approximately 68 primary sites and over 88,000 cases.

2.2 Machine Learning Algorithms

In this thesis, we will be examining several machine learning algorithms for survival analysis, including the Cox Proportional Hazards (CPH) model, Penalized Cox Regression, Boosted Cox Regression, and Random Survival Forest.

The standard statistical tool for analyzing censored survival data is the CPH model, which evaluates the effect of multiple features simultaneously on the time to an event of interest, in our case, death [Spooner et al., 2020]. While CPH is a robust model, it does not generalize well to high-dimensional data, which is a common challenge when handling multi-omics data. Despite this limitation, CPH is included in this study as a baseline for comparison with other models that extend the capabilities of the CPH model.

The Cox model is expressed by the hazard function, which represents the risk of an event occurring at time t:

$$h(t \mid X) = h0(t)exp(\beta 1X1 + \beta 2X2 + \cdots + \beta pXp)$$

where $h(t \mid X)$ is the hazard function given covariates X, $h0(t)$ is the baseline hazard function, $\beta 1, \beta 2, \ldots, \beta p$ are coefficients, and $X1, X2, \ldots, Xp$ are the covariates (features).

To overcome the limitations of CPH with high-dimensional data, Penalized Cox Regression adds a constraint to the equation. This constraint reduces coefficient values towards zero, decreasing variance and ensuring that less important features have less impact [Spooner et al., 2020]. This study will use Elastic Net regression, which is particularly useful when the number of features is larger than the number of samples. Boosting is an ensemble technique that trains weak learners sequentially, so each new model added learns from the mistakes of previous models. This makes boosting resistant to overfitting and capable of handling high-dimensional data [Spooner et al., 2020]. This study will focus on gradient boosting, which iteratively refits the residuals of the ensemble model at each step. GLMBoost is a gradient boosting algorithm that uses penalized Cox regression models as its base learners. Random Survival Forests (RSF) are becoming more widely accepted as an alternative to the CPH model due to their ability to model complex, non-linear, and high-dimensional data. RSFs can identify interactions and naturally impute missing data, reducing the tendency to overfit [Spooner et al., 2020].

By comparing these models, we aim to assess their predictive accuracy and ability to handle high-dimensional multi-omics data, ultimately enhancing our understanding of cancer survival outcomes.

2.3 Feature Selection Methods

Feature selection is the process of selecting a subset of relevant features essential for analyzing and dealing with high-dimensional data. Feature selection techniques can be divided into filter, wrapper, and embedded methods [Spooner et al., 2020]. In this study, we will employ filter methods, which rank features according to an external measure and then use a threshold to select the most important ones.

The filter methods applied in this study include univariate and multivariate Cox filters. The univariate Cox filter fits a CPH model to each feature individually, while the multivariate Cox filter fits a CPH model to all features together. The features are then ranked by the performance

of the resulting Cox model, given by the C-index. Another method, random variable importance, works by adding random noise to the features and calculating the difference in prediction error before and after the noise addition. If adding noise significantly decreases performance, the feature is considered highly predictive. The features are ranked by this importance score.

Random forest minimal depth evaluates feature importance based on the shortest distance from the root of the tree to the largest subtree that has that feature as the root. The idea is that the closer a feature's maximum subtree is to the root, the more important it is. Lastly, random forest variable hunting is designed for high-dimensional datasets. This method combines variable importance with minimal depth methodologies to create a joint importance score. Features are ranked by their frequency of occurrence in the model and are added one at a time until there is no increase in the joint importance score.

By using these filter methods, we aim to identify the most relevant features from each omics dataset, thereby reducing dimensionality and improving the performance of our survival analysis models.

2.4 Previous Studies

In the pursuit of enhancing cancer survival predictions through multi-omics integration, several seminal studies have provided crucial insights and laid the groundwork for ongoing research in this field.

Yuan et al. (2014) conducted a pioneering study by integrating genomic data, including SCNA, DM, GE, ME, and protein expression, to forecast patient survival. Their findings demonstrated that the combination of molecular data with clinical variables significantly improved the accuracy of survival predictions across various cancers [Yuan et al., 2014]. This study was instrumental in establishing the value of multi-omics data for cancer prognosis and provided a foundational framework for subsequent investigations.

Building upon this foundation, Zhu et al. (2017) introduced a kernel machine learning method to systematically evaluate the prognostic significance of clinical information, GE, SCNA, DM, and ME across 14 different cancer types [Zhu et al., 2017]. Their research aimed to compare the efficacy of various omics data types in predicting patient survival and found that GE and ME data had the most robust prognostic capabilities. This study highlighted the ongoing debate regarding the optimal selection of omics data for accurate cancer prognosis and contributed

significantly to the discourse on the comparative strengths of different data types.

In the realm of survival prediction, Zhao et al. (2018) conducted a comprehensive exploration of various classification algorithms to forecast 5-year survival in breast cancer [Zhao et al., 2018]. By integrating gene expression data with clinical and pathological factors, Zhao et al. enhanced the predictive accuracy of their models. Their study underscored the importance of dimensional reduction techniques and showcased the discriminative power of various machine learning models in cancer survival prediction. This research provided valuable insights into the application of machine learning techniques for predictive modeling and the stability of these models in breast cancer survival prediction.

These studies collectively advance the field of multi-omics integration for cancer survival analysis. Yuan et al. (2014) established the utility of integrating molecular and clinical data, Zhu et al. (2017) refined our understanding of the comparative prognostic value of different omics data types, and Zhao et al. (2018) demonstrated the effectiveness of machine learning models in survival prediction for breast cancer.

While these studies emphasize the value of integrating multiple data types, they do not fully address the challenges of integrating different modalities into a cohesive predictive model. Further research is needed to develop robust methods for combining various omics data types in a way that maximizes their collective predictive power. The focus of Zhu et al. (2017) on multiple cancer types highlights the potential for generalization, but there is still a need for comprehensive studies that systematically evaluate the performance of integrated multi-omics models across a broader spectrum of cancers.

Chapter 3

Scope and Method

3.1 Project Scope and Outline

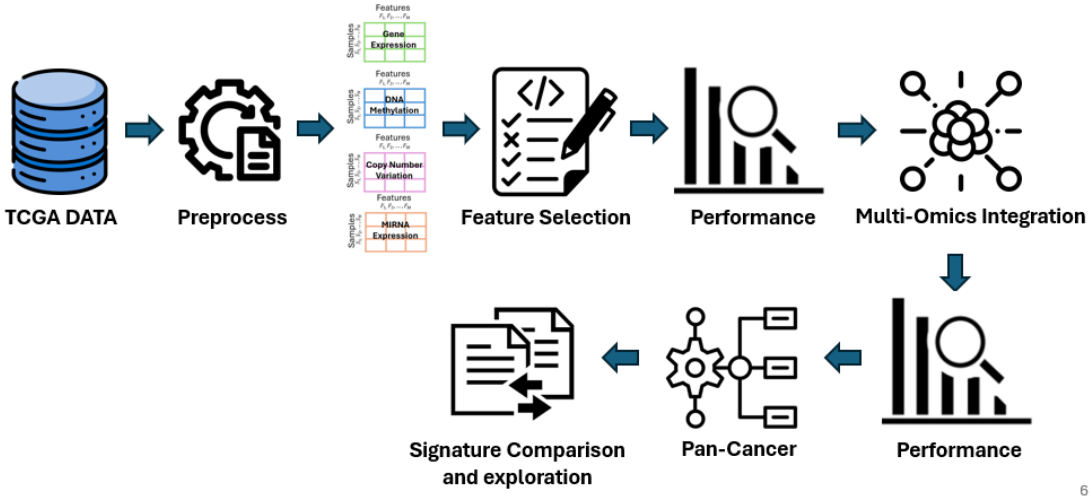


Figure 1: Entire Project pipeline

The project involves a comprehensive analysis of multi-omics data from TCGA, including gene expression, DNA methylation, miRNA expression, and copy number variations. Initially, the focus will be on preprocessing these diverse data types to ensure consistency and quality. Subsequently, feature selection methods will be applied to each single-omics dataset to identify the most relevant features for survival analysis. Machine learning algorithms will then be employed to develop survival prediction models based on these single-omics features. Following this, the project will integrate the features from different omics data types using both early fusion (combining features before modeling) and late fusion (combining model predictions). The performance of these multi-omics integration approaches will be assessed to determine the most effective modality combinations. This process will be repeated for various cancer types, including BRCA, OV, CESC, and UCEC, to identify common features and signatures across

different cancers. The final stage will involve evaluating the performance of multi-omics models compared to single-omics models and exploring the biological significance of the identified signatures. The findings will be documented and discussed, highlighting the strengths, limitations, and potential implications for cancer research and treatment.

3.2 Omic Modalities and Cancers

Table 1: Overview of number of samples for all omics data

Cancer	GE	CNV	DM	ME	Clinical	Common
BRCA: Breast Invasive Carcinoma	1111	1050	1096	1096	1098	739
OV: Ovarian Serous Cystadenocarcinoma	421	557	582	490	608	397
CESC: Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	304	294	307	307	307	291
UCEC: Uterine Corpus Endometrial Carcinoma	553	536	438	545	560	421

Multi-omics data were obtained from The Cancer Genome Atlas (TCGA) using the TCGAbiolink R package. Table 1 provides an overview of the number of samples available for each omics data type across the cancers studied. The focus is on four women-related cancers: Breast Invasive Carcinoma (BRCA), Ovarian Serous Cystadenocarcinoma (OV), Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (CESC), and Uterine Corpus Endometrial Carcinoma (UCEC). For each cancer type, we collected data on gene expression (GE), copy number variations (CNV), DNA methylation (DM), miRNA expression (ME), and clinical variables. Clinical data, including patient vital status and the number of days to death or last follow-up, were integrated for survival analysis. We selected only those samples that had complete information across all five data categories to ensure consistency and comprehensiveness in our analysis.

Table 2: Overview of four omics data modalities

Data Modality	Gene Expression (GE)	Copy Number Variation (CNV)	DNA Methylation (DM)	MiRNA expression (ME)
Measure	Fragments per kilobase of transcript per million mapped reads (FPKM)	Gain/Loss/Neutral	Beta Value	Reads per million mapped reads (RPM)
Type	Continuous	Discrete	Continuous	Continuous
Range	[0, Billions]	[Loss < 0; Neutral = 0; Gain > 0]	[0,1]	[0, Millions]
Features	Ensembl Gene ID	Ensembl Gene ID	cg Probe ID	MiRNA ID

For each omics modality, feature extraction was conducted using bioinformatics pipelines provided by TCGA. The preprocessed multi-omics tabular data were retrieved from the GDC Portal via the TCGAbiolink R package. Table 2 offers a summary of the data types used in this study. The GE data were normalized to FPKM (Fragments Per Kilobase of transcript per Million mapped reads) using the TCGA mRNA pipeline and include over 60,000 features, covering different isoforms for each gene and various non-coding RNA transcripts. For CNV, we utilized the "Gene Level Copy Number Scores," which provides CNV scores reflecting gains and losses at the gene level across all samples. Specifically, “Gain” indicates an increased number of gene copies, while “Loss” signifies a reduced number of copies compared to normal. The DM data consist of beta values derived from Illumina Human Methylation 450K or 27K assays, with values ranging from 0 to 1. A beta value of 0 indicates no methylation at the probe, whereas a value of 1 signifies complete methylation of the CpG site [Tong et al., 2020]. The ME data were obtained from a quantification table generated by TCGA using a modified profiling pipeline developed by the British Columbia Genome Sciences Centre (BCGSC).

3.3 Data pre-processing

Once each data modality was obtained in tabular form, with samples represented by rows and features by columns, we retained samples with a sample type code of "01" as these represent "Primary Solid Tumor." All omics data were then converted into matrices based on TCGA annotation information, aligning GE features with their respective gene names and matching them with CNV data. In cases where a single gene had multiple signals within one sample, the average of these values was calculated as the final signal. Clinical survival data were subsequently integrated into each omics matrix.

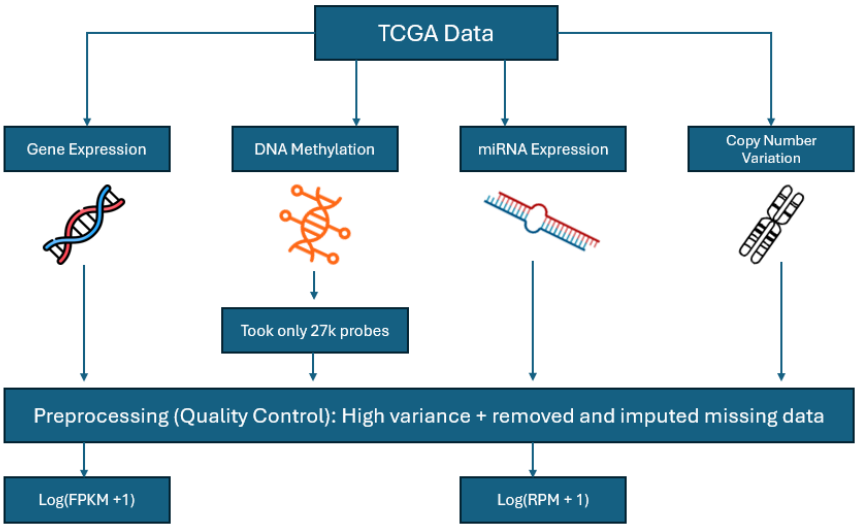


Figure 2: Overview of omics data preprocessing

The individual preprocessing steps for each omics data type are outlined in Figure 2. For GE, CNV, and ME data, features and samples with more than 20% missing values were removed, and the data were filtered for highly variable features. Table 3 illustrates the number of features and samples from the BRCA dataset, showcasing how the preprocessing steps were used to filter out low-quality samples and features, thereby reducing data dimensionality.

Table 3: Overview of BRCA data preprocessing

Data Modality	**Gene Expression (GE)**	**Copy Number Variation (CNV)**	**DNA Methylation (DM)**	**MiRNA expression (ME)**
Features (Original)	60660	60623	485577	1881
Features (Filtered)	3033	2689	1216	737
Samples (Original)	1111	1050	1096	1096
Samples (Filtered)	737	737	746	737

To normalize the data, we applied a log transformation $\log(X + 1)$ to the features, where X is FPKM for GE and RPM for ME. For DM data, to ensure compatibility with the 27k platform and facilitate pan-cancer comparisons, we reduced the CpG sites to 27k, extracting them from the 450k samples. This step was crucial as some cancers, such as OV, only had data for the 27k platform. Samples and features with more than 50% missing values were filtered out, and the

remaining missing values were imputed using the mean. Additionally, for ME data, we retained miRNAs with values greater than 0 in more than 50% of the samples and with values greater than 1 in more than 10% of the samples, following the approach by Zhao et al. (2020).

3.4 Single-Omics pipeline

Figure 3 details the single-omics pipeline. Once we have pre-processed the TCGA data for each omics, we run it through our feature selection pipeline to obtain a final set of features for each omics. These selected features will then be evaluated using our predictive survival models.

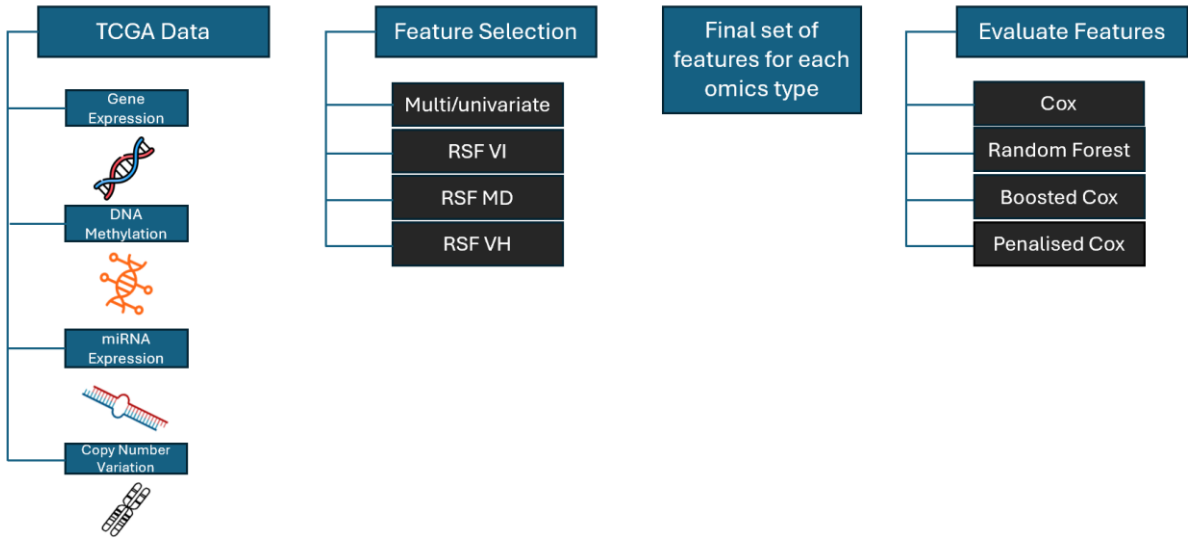


Figure 3: Overview of single-omics pipeline

However, before developing our feature selection pipeline, we first needed to evaluate the performance of each filter method and compare them with having no feature selection as a baseline.

Table 4: Overview of ML and FS Methods

Method	**R package**	**Function**	**Hyper-parameters and values**
Learning Algorithms			
Cox PH model	*survival*	coxph	
Elastic Net	*glmnet*	cv.glmnet	alpha = 0.5, nfolds=5
Gradient boosting with linear models as base learner	*mboost*	glmboost	

Random Survival Forests	*ranger*	*ranger*	splitrule = "maxstat", importance = "permutation", mtry: sqrt(#features) -> 100, min.node.size: 50, num.trees=1000
Feature Selection Methods (filters)			
Univariate Cox filter	*mlr*	various	
RSF variable importance	*randomForestSRC*	rfsrc	ntree = 1000, nsplit = 10 mtry = sqrt(#features), nodesize=3
RSF minimal depth	*randomForestSRC*	var.select	method = "md" ntree=1000, nsplit=20, nodesize=5, splitrule="logrank"
RSF variable hunting	*randomForestSRC*	var.select	method = "vh", ntree=1000, nodesize=5, splitrule="logrank", nsplit=20, nrep=3, K=10, nstep=1

Table 4 provides an overview of the machine learning algorithms and filter methods used in this study, including their respective R packages and the hyperparameters for each one.

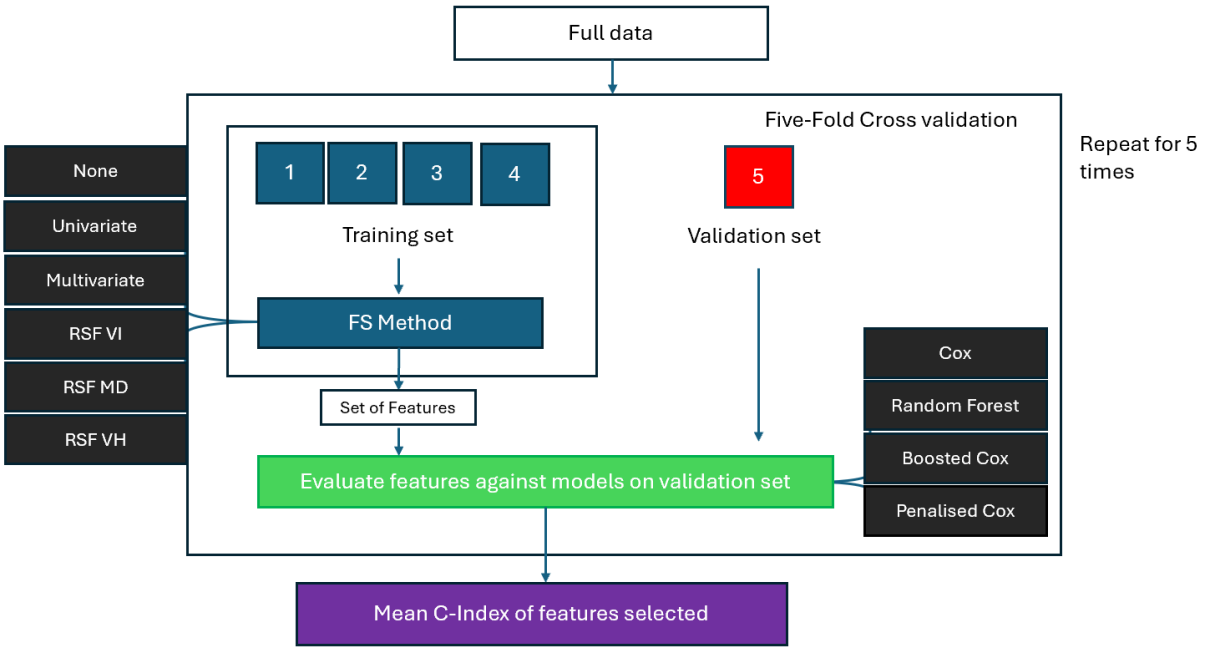


Figure 4: Evaluation of individual FS Methods Pipeline

Figure 4 illustrates the process for evaluating each feature selection method. We employed 5 repeats of 5-fold cross-validation, splitting the data into five folds, with four folds for training and one-fold for validation. Each filter method was individually applied to the training set, yielding a set of features from each method. Features were deemed important by thresholding coefficients and importance scores. For the univariate and multivariate Cox filter methods, we selected features that were both significant (p-value < 0.05) and had non-zero coefficients. For the random survival forest filter methods, we selected features with an importance score greater than 0. The performance of these selected features was then evaluated using all four of our survival machine learning algorithms with the validation set to obtain a concordance index (C-index) for each model. The C-index is a non-parametric measure that quantifies the discriminatory power of a predictive model, ranging from 0.5 to 1. A C-index of 1 represents perfect prediction accuracy, while a C-index of 0.5 indicates random guessing [Zhao et al., 2020].

This process was repeated 5 times with 5-fold cross-validation, resulting in 25 evaluations for each set of features by a predictive model, producing 25 different C-index values for each feature selection method and survival model. The mean of these C-index values was taken for each model to provide the final assessment of the feature selection method's performance.

3.5 Feature Selection Pipelines

For the single-omics pipeline, we developed two feature selection pipeline, one that utilizes

cross-validation and other that uses bootstrapping.

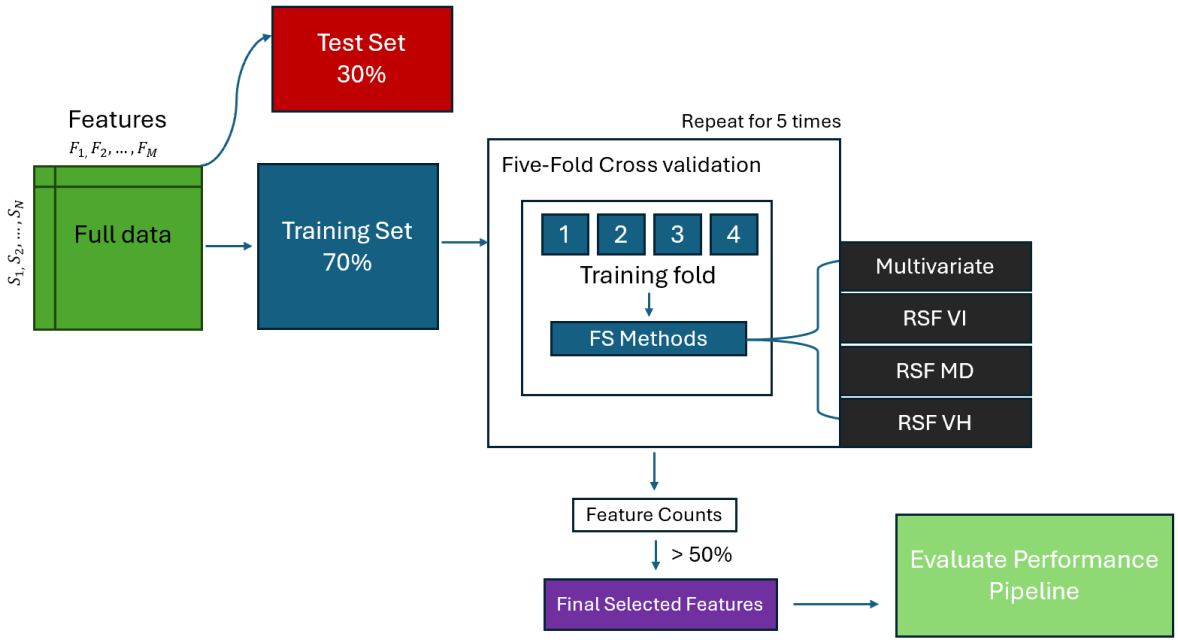


Figure 5: Cross-validation Feature Selection pipeline

Figure 5 illustrates our cross-validation (CV) pipeline. Initially, we perform a 70:30 training-test split, using the training set for feature selection. The training set is further divided into 5 folds, with one-fold being ignored each time. Each filter method is applied to the remaining training folds to obtain a set of important features.

We utilize 5 repeats of 5-fold cross-validation, running each model 25 times. Each feature's occurrence in the selected sets is recorded, allowing us to rank features by their total counts. Since there are four filter methods and each method is applied in 5 repeats of 5-fold CV, the maximum possible occurrence for a feature is 25*4=100. The more frequently a feature is selected, the more predictive it is. A feature appearing 100 times indicates its selection in every set from all filter methods across the CV process.

To ensure stability and robustness, we set a threshold of 50 for feature selection. Features that appear at least 50% of the time (i.e., in at least 50 sets) are included in the final set of selected features. This final set is then evaluated separately using our performance pipeline on the test set, ensuring no information leakage as feature selection is solely performed on the training set.

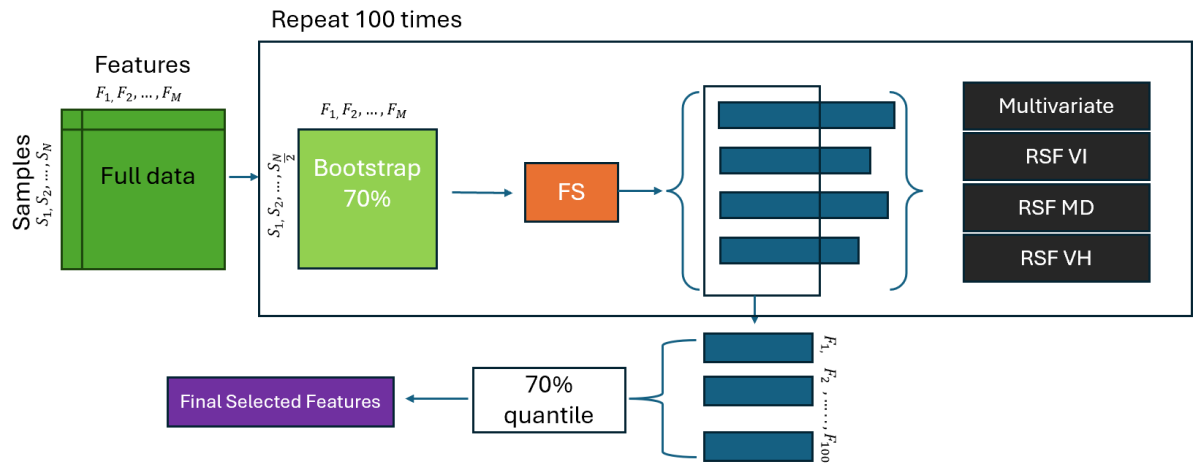


Figure 6: Bootstrapping Feature Selection pipeline

The second feature selection pipeline method developed is the bootstrapping pipeline, as shown in Figure 6. In this method, we start by bootstrapping 70% of the data and then applying each of our filter methods to the bootstrapped data. Each method provides a set of features deemed important. We then identify the common features that appear across all sets, forming the first bootstrapped set of common features. This process is repeated 100 times, generating 100 common feature sets. We track the frequency of each feature's occurrence in these common sets. A frequency of 100 indicates that a feature has appeared in all filter methods across all 100 bootstrapped samples, representing the maximum occurrence. To ensure robust feature selection, we apply a quantile-based threshold. We select features that appear at or above the 70th quantile in terms of frequency, meaning only those features that are consistently deemed important across a substantial number of bootstrapped samples are included in the final set. This approach helps identify highly predictive features while accounting for variability in the data.

3.6 Performance Pipeline

To evaluate the performance of our feature selection pipelines, we developed the evaluation pipeline shown in Figure 7. This pipeline tests the final set of selected features using the test set that we separated earlier in the process. By doing so, we ensure that the selected features are assessed independently of the training data, providing an unbiased evaluation of their predictive performance.

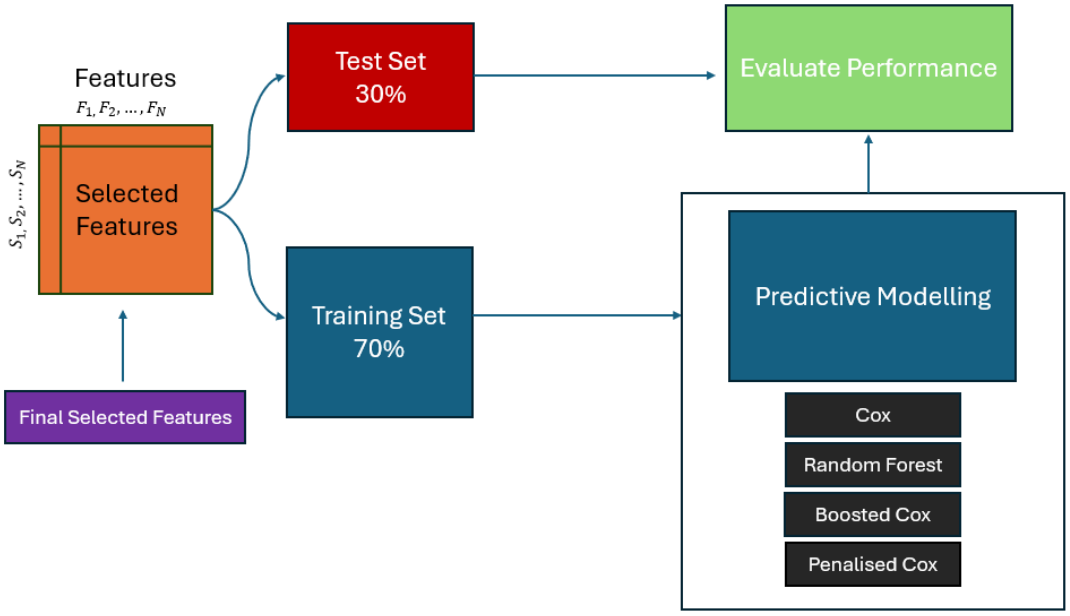


Figure 7: Feature Evaluation Pipeline

Both the test and training data are first filtered to include only the final selected features from the feature selection pipeline. The training set, which was previously used for feature selection, is then used to train each of the predictive models. Subsequently, the performance of these models is evaluated using the test set. This approach ensures that the models are trained on the most relevant features and that their performance is assessed on an independent dataset, providing a robust measure of their predictive capability.

3.7 Multi-Omics pipeline

Once we obtain our final sets of features for each omics modality using our feature selection pipeline, we proceed to perform multi-omics integration using two fusion techniques: early fusion and late fusion.

Early fusion involves combining the selected features from multiple omics modalities into a single dataset before training a predictive model. This approach integrates the features at the data level, allowing the model to learn from the comprehensive set of multi-omics data simultaneously [Tabakhi et al., 2023].

Late fusion, on the other hand, involves training separate models for each omics modality using their respective selected features. The predictions from these individual models are then combined to make the final prediction [Tabakhi et al., 2023].

By utilizing both early and late fusion techniques, we aim to determine which method provides the best performance for multi-omics integration in survival analysis.

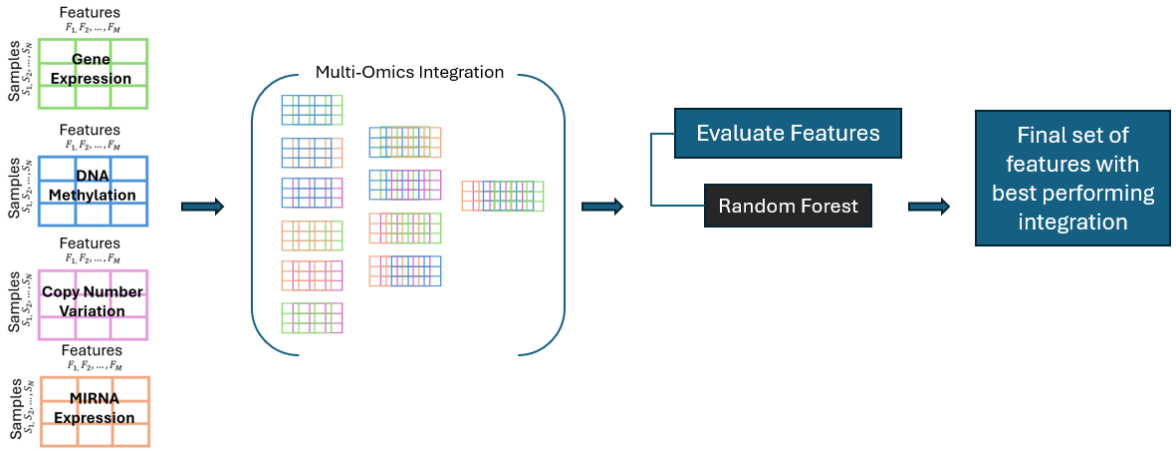


Figure 8: Overview of Multi-Omics Pipeline

As shown in Figure 8, we will apply both fusion techniques to every possible omics combination to determine which modality combination best complements each other in achieving higher survival prediction accuracy. This approach is similar to the study by Tong et al. (2020), where autoencoders were used for integration. By evaluating all potential combinations, we aim to identify the optimal multi-omics integration strategy for enhancing survival prediction.

To evaluate the performance of each integration, we will use Random Forest. The best modality combination, based on the highest predictive accuracy, will be selected as our final set of multi-modal signatures for each particular cancer.

3.8 Early Fusion Integration

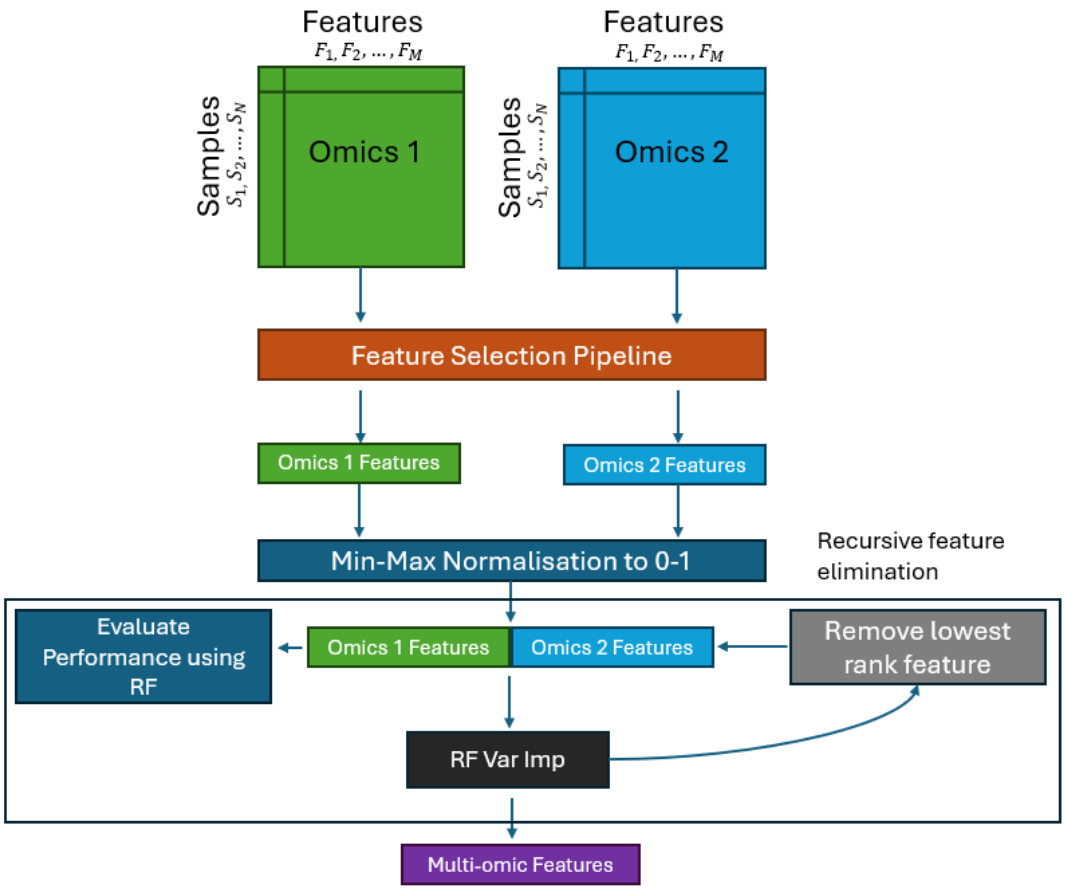


Figure 9: Early Fusion Integration

Figure 9 illustrates the early fusion process with two different modalities for clarity, though this method will be extended to all four omics types in our study. Once each omics dataset has been processed through our feature selection pipeline, we obtain a set of features deemed highly important for each respective modality. Given that these omics datasets are on different scales, we first normalize the data using min-max normalization to standardize values between 0 and 1.

After normalization, we concatenate the data from different omics modalities. We then apply Recursive Feature Elimination (RFE) to identify the optimal set of combined features that achieves the highest predictive performance, measured by the concordance index (c-index). The RFE process begins by evaluating the performance of the current feature set using Random Forest (RF). The importance of each feature is ranked using RF variable importance scores. We then iteratively remove the lowest-ranked feature and reassess the performance of the remaining

features. This process continues recursively until we identify the set of features that yields the highest predictive accuracy. This final set of multi-omic features represents the optimal feature combination for that particular integration of modalities.

3.9 Late Fusion Integration

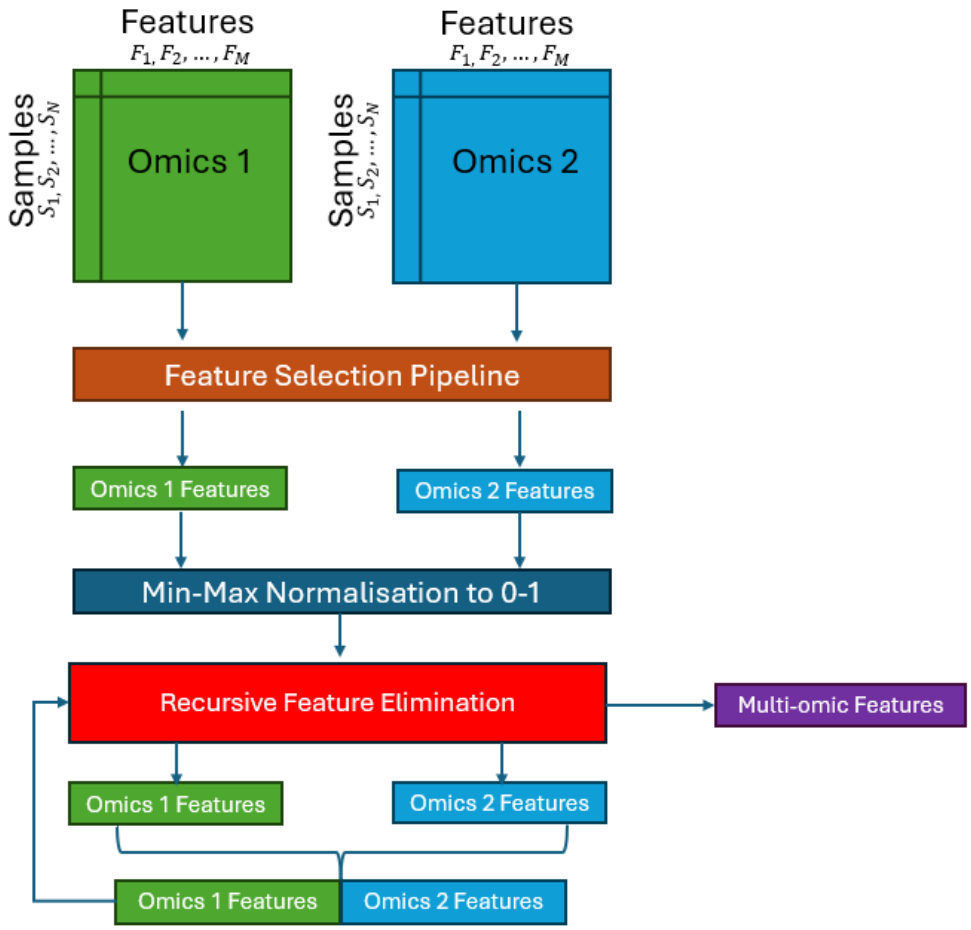


Figure 10: Late Fusion Integration

Our second integration method is Late Fusion, illustrated in Figure 10. This approach starts similarly to Early Fusion: we first obtain our set of features from each omics modality and normalize the data. However, instead of concatenating the omics features at this stage, we apply RFE individually to each omics dataset. By performing RFE separately on each omics modality, we identify the optimal set of features for each type. Once these individual feature sets have been determined, we concatenate the features from all modalities. We then apply RFE once again to the combined dataset to identify the final set of multi-omic features that provides the

highest predictive performance.

This method allows us to refine the feature selection process for each modality separately before integrating them, potentially enhancing the overall performance of the multi-omics model.

3.10 Pan-cancer Analysis

With the best multi-omics combinations identified for each cancer—BRCA, OV, CESC, and UCEC—we will now delve into our pan-cancer analysis. This analysis aims to uncover multi-modal signatures shared across these cancers. Our approach includes several key steps:

- Overlap Analysis: We will directly examine the overlapping features across the four cancers. This involves identifying common features and further exploring potential overlaps by analyzing miRNA and methylation targets associated with these features.

- Disease Association Networks: We will explore disease association networks to understand the relationships between the overlapping targets, providing insights into their interconnected roles in cancer biology.

- Gene Set Enrichment Analysis (GSEA): To identify shared biological themes, we will perform GSEA on the targets from each cancer. This will help us uncover common Gene Ontology (GO) Terms and KEGG Pathways that are significant across cancers.

- Validation: To ensure the robustness and biological relevance of our findings, we will validate some of the overlapping features using Kaplan-Meier (KM) plots and cross-reference with existing literature. While our feature selection was driven by predictive power in survival analysis, this validation step aims to confirm that these features have meaningful biological implications.

By integrating these analyses, we aim to confirm that our results are not only statistically significant but also provide valuable biological insights into the commonalities and differences across the cancers under study.

Chapter 4

Results

4.1 Evaluation of Individual Filter Methods

As discussed previously, before developing our feature selection pipeline, we first needed to examine each of the filter methods that we will be utilizing (Figure 4). Due to computational complexity and resources, we decided to opt for only four filter methods for the purposes of this study.

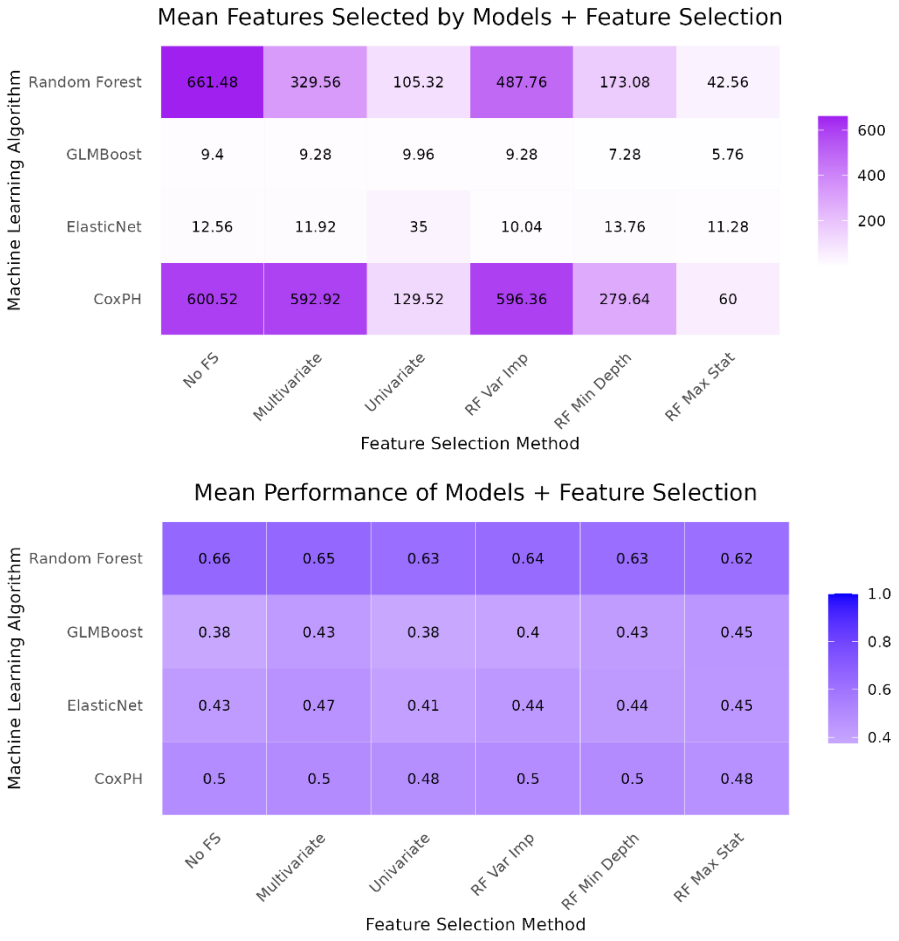


Figure 11: Heatmap of both mean performance and features selected from each model and filter methods, using BRCA DM data.

The results of our experiments are presented in Figure 11 as heatmaps, illustrating both the mean c-index values and the number of features selected across 5 repeats of 5-fold cross-validation for each combination of ML algorithms (rows) and filter methods (columns). The Cox Proportional Hazards (CPH) model, while not an ML algorithm, is included as a benchmark for comparison with the other models. Column 1, which shows results without feature selection, serves as a baseline for comparison with other filter methods.

The example shown uses the BRCA DNA Methylation (DM) dataset. However, the trends observed in this example are consistent with other modalities and cancer types, making it a representative illustration of the overall trends for these ML algorithms and filter methods. We expect the performance of each filter method to be at least comparable to having no feature selection. As seen in the top heatmap, there was a significant reduction in the number of features, while the predictive performance remained stable or slightly decreased, indicating the removal of many redundant features. Across the board, the Random Forest (RF) model had the highest overall mean performance, with the lowest c-index being 0.62.

Surprisingly, both Penalized and Boosted Cox models performed worse than the baseline CPH, with c-index values dropping below 0.5. This is unexpected, considering these models are designed to handle high-dimensional data better. However, it is worth noting that applying the filter methods generally improved the performance of the two extended Cox models. Overall, we opted to remove the univariate filter method due to its poor overall performance and the notion that it evaluates individual features in isolation rather than considering their combined effect, potentially limiting the identification of features with contextual synergy.

4.2 Performance of Feature Selection Pipelines

In this study, we developed two feature selection pipelines: one using cross-validation and the other using bootstrapping, both incorporating our four main filter methods. We then evaluated their performance, as shown in Figures 5, 6, and 7.

Table 5: Overview of Feature Selection pipeline performance using BRCA dataset

	Omics	None	CV	BS
CoxPH (1)	GE	0.463	0.394	0.363
	ME	0.509	0.581	0.392
	DM	0.496	0.461	0.490
	CNV	0.473	0.497	0.421
Ranger (2)	GE	**0.725**	**0.655**	**0.635**
	ME	**0.619**	**0.654**	**0.587**
	DM	**0.658**	**0.640**	**0.634**
	CNV	**0.611**	**0.629**	**0.579**
GLMBoost (3)	GE	0.323	0.406	0.423
	ME	0.405	0.406	0.523
	DM	0.376	0.507	0.509
	CNV	0.422	0.420	0.431
ElasticNet (4)	GE	0.325	0.407	0.340
	ME	0.422	0.4	0.493
	DM	0.430	0.5	0.517
	CNV	0.425	0.426	0.435

Table 5 shows the performance of each of our feature selection pipelines compared to a baseline of having no feature selection. Column 1 lists each of the predictive models, while column 2 presents each of the modalities for which we evaluated performance. Overall, our cross-validation pipeline produced decent results, showing improvements in some areas and minor reductions in others. However, the bootstrapping pipeline did not perform as well as anticipated, requiring significantly higher computational resources and achieving overall poorer performance compared to the cross-validation pipeline.

Just like our filter method performance test, it seems that Random Forest produces the highest results overall. Based on the results of our BRCA dataset, we decided to use cross-validation (CV) as our main feature selection pipeline moving forward.

Table 6: Overview of Feature Selection pipeline performance for all Cancers

	Omics	None	CV
CoxPH (1)	GE	0.515	0.496
	ME	0.495	0.434
	DM	0.483	0.449
	CNV	0.471	0.524
Ranger (2)	GE	**0.530**	**0.530**
	ME	**0.557**	**0.603**
	DM	**0.534**	**0.543**
	CNV	**0.536**	**0.532**
ElasticNet (3)	GE	0.480	0.491
	ME	0.448	0.384
	DM	0.422	0.467
	CNV	0.482	0.479
GLMBoost (4)	GE	0.500	0.485
	ME	0.445	0.395
	DM	0.5	0.460
	CNV	0.484	0.489

OV dataset

	Omics	None	CV
CoxPH (1)	GE	0.550	0.400
	ME	0.677	0.428
	DM	0.540	0.520
	CNV	0.441	0.560
Ranger (2)	GE	**0.626**	**0.525**
	ME	**0.651**	**0.688**
	DM	**0.553**	**0.496**
	CNV	**0.471**	**0.471**
ElasticNet (3)	GE	0.386	0.355
	ME	0.441	0.412
	DM	0.576	0.595
	CNV	0.520	0.573
GLMBoost (4)	GE	0.360	0.372
	ME	0.473	0.427
	DM	0.5	0.583
	CNV	0.497	0.570

CESC dataset

	Omics	None	CV
CoxPH (1)	GE	0.447	0.465
	ME	0.411	0.371
	DM	0.470	0.491
	CNV	0.387	0.382
Ranger (2)	GE	**0.666**	**0.598**
	ME	**0.686**	**0.676**
	DM	**0.636**	**0.632**
	CNV	**0.696**	**0.685**
ElasticNet (3)	GE	0.369	0.373
	ME	0.315	0.341
	DM	0.360	0.361
	CNV	0.383	0.357
GLMBoost (4)	GE	0.359	0.370
	ME	0.317	0.354
	DM	0.359	0.342
	CNV	0.392	0.5

UCEC dataset

Table 6 shows the results for the other cancers investigated using the cross-validation (CV) pipeline. Overall, Random Forest appears to be the highest-performing predictive model across all four cancers. It is important to note that ovarian cancer (OV) has an overall lower c-index compared to the other cancers, with none reaching 0.6.

4.3 Multi-omics Integration RFE

For our multi-omics integration, we applied two techniques: early and late fusion, as seen in Figures 9 and 10. Both techniques utilized recursive feature elimination (RFE).

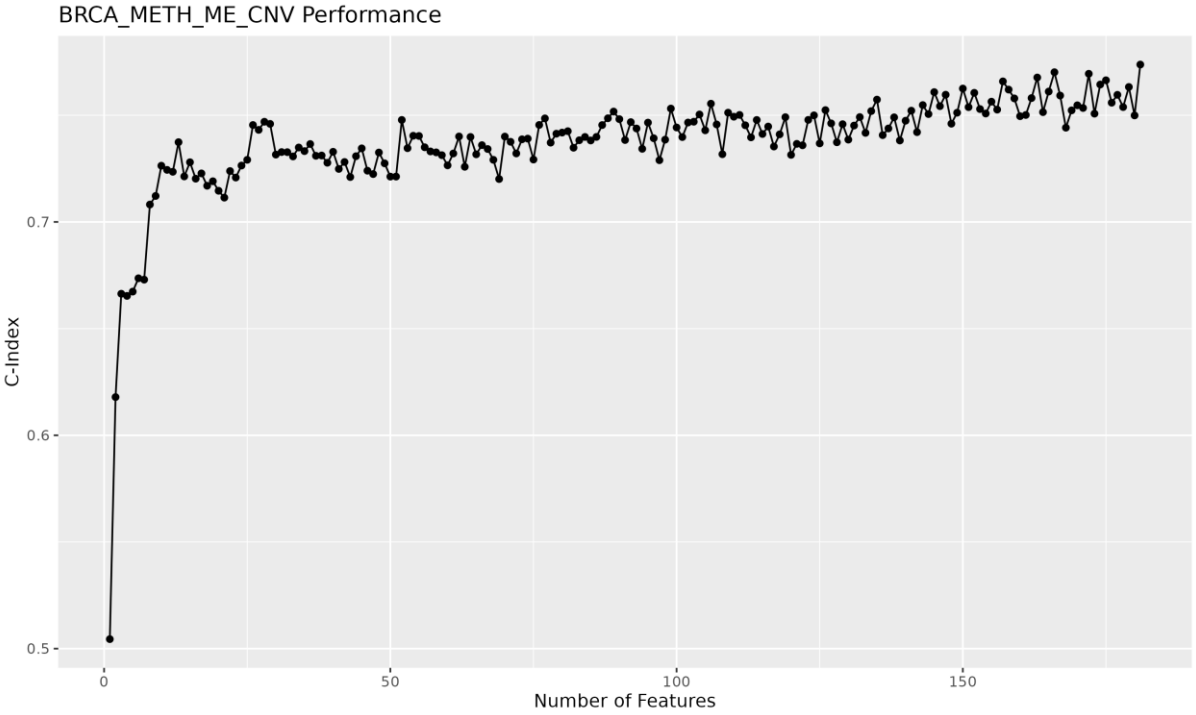


Figure 12: Graph showing performance of feature set at each iteration for BRCA dataset, where the modalities integrated were Methylation, miRNA and copy number variation data.

Figure 12 illustrates an example of how we use RFE to obtain the best-performing set of features for each modality combination. In this example, the modalities integrated were DNA methylation (DM), miRNA expression (ME), and copy number variation (CNV) data. The best performance was observed at the start of the RFE process, where no features were eliminated, indicating that retaining all features produced the highest performance. The majority of the RFE graphs displayed a similar trend, with small fluctuations in performance and a slight decrease as features were removed from the set.

As demonstrated in this example, the fusion of modalities has already produced better predictive results than the single-omics pipeline, achieving a c-index of 0.774.

4.4 Late and Early Fusion

Performance using Early Fusion:

Performance	ME + GE	ME + DM	GE + DM	GE + CNV	ME + CNV	DM + CNV	DM + GE + ME	DM + GE + CNV	DM + ME + CNV	ME + GE + CNV	ME + GE + CNV + DM
C-index	0.728	0.740	0.653	0.608	0.672	0.523	**0.730**	0.713	0.718	0.673	0.653

Performance using Late Fusion:

Performance	ME + GE	ME + DM	GE + DM	GE + CNV	ME + CNV	DM + CNV	DM + GE + ME	DM + GE + CNV	DM + ME + CNV	ME + GE + CNV	ME + GE + CNV + DM
C-index	0.654	0.640	0.699	0.738	0.632	0.718	0.822	0.737	**0.835**	0.693	0.720

Figure 13: Comparison between Late and Early Fusion using CESC dataset

In Figure 13, we illustrate the performance comparison between Late and Early Fusion using the CESC dataset. This trend is consistent across all cancer types tested, not just CESC. Our results show that Late Fusion consistently outperforms Early Fusion. For example, in the CESC dataset, Early Fusion achieved a c-index of 0.730, while Late Fusion significantly improved this to 0.835. This suggests that the optimal modality combination can vary depending on the integration strategy used. In this case, Late Fusion identified the combination of DM, ME, and CNV as superior to the DM, GE, and ME combination preferred by Early Fusion. Overall, Late Fusion demonstrated better performance across all cancer datasets, leading us to select it as our primary integration strategy.

4.5 Multi-omics results

Using Late fusion as our main integration method, we tested every possible modality combination and observed their c-index.

Table 7: Multi-omics results for BRCA

Performance	ME + GE	ME + DM	GE + DM	GE + CNV	ME + CNV	DM + CNV	DM + GE + ME	DM + GE + CNV	DM + ME + CNV	ME + GE + CNV	ME + GE + CNV + DM
C-index	0.768	0.702	0.694	0.714	0.718	0.654	0.692	0.707	**0.774**	0.752	0.655

Omics	Features
DM	cg15520279, cg18087514, cg02776251, cg24262376, cg01671575, cg14717170, cg20786074, cg09656934, cg20308679, cg10634551, cg12626411, cg01664666, cg07837085, cg20261167, cg15147516, cg08205865, cg25583174, cg12958813, cg15761405, cg25167447, cg04574507, cg20676475, cg11902458, cg00807586, cg01637734, cg02836529, cg26583078, cg16232126, cg17214107, cg08422599, cg09952204, cg10978355, cg05619712, cg21137417, cg09563216, cg02983451, cg15077070, cg11657808, cg20311501, cg05275752, cg02431964, cg17453778, cg18335243, cg18182399, cg19616230, cg10777851, cg03679305, cg23300372, cg01612158, cg12815916, cg03614513, cg21591742, cg18006568, cg27403635, cg23282674, cg23557926, cg22176895, cg13320626, cg06933072, cg18110483, cg22359606, cg09220361, cg13847070, cg22855405, cg07038400, cg13986130, cg07354440, cg01114088, cg18794577, cg16557944, cg18482268, cg11428724, cg20613889, cg00995327, cg02774439, cg01216369, cg24512973, cg17701886, cg16516400, cg11819637, cg09478478, cg05200628, cg12105450, cg17525406, cg04098585, cg08474603, cg00884221, cg25462303, cg17105014, cg17542385, cg17020834, cg07548313, cg01868128, cg14925024, cg05768141, cg17108819, cg08826839, cg05909475, cg25057743, cg25631352, cg08047907, cg10861599, cg08658594, cg10691387, cg18277754, cg08578023, cg12880658, cg17281600, cg08690031, cg19228118, cg10031651, cg03702236, cg08057475, cg22605643, cg25437385, cg03852144, cg21948655, cg03329572, cg06168050, cg06274159, cg23363832, cg10370591, cg11724759, cg05248781, cg10774440, cg07664856, cg21972382, cg20535085, cg07694025, cg24928687, cg12864235, cg20579480, cg14419187, cg06043114, cg19884658, cg17430393, cg19526600, cg00498604, cg03199651, cg11976166, cg16139316, cg06940792, cg00567749, cg10946435, cg21991396, cg18908499, cg01367992, cg26117023
ME	hsa.mir.31, hsa.mir.30a, hsa.mir.22, hsa.mir.150, hsa.mir.4742, hsa.mir.221
CNV	ADAM2, AC139365.1, RAB11FIP1, CYP4F44P, AC239800.2, U3, RNU1.124P, HSPA8P13, POMK, AC110275.1, AC139365.2, AC123767.1, RNU6.356P, AC103726.1, AC067817.2, ZNF703, SLC20A2, RPS20P22, AC048387.1, MIR1204, ADAM3A, RN7SL709P, AC118650.1, SNORD65B, SFRP1, MIR548AO, AC091182.1

Table 7 presents the results for the BRCA dataset, where the highest performance was achieved with a c-index of 0.774 using a modality combination of DM, ME, and CNV. This combination outperformed the baseline single-modality Random Forest model, which achieved a maximum c-index of 0.725 with GE data alone. The accompanying table at the bottom details the features from each modality that contributed to this peak performance. Notably, the DM data contributed the most to this result, providing a substantial number of features that were crucial to the model's success.

Table 8: Multi-omics results for OV

Performance	ME + GE	ME + DM	GE + DM	GE + CNV	ME + CNV	DM + CNV	DM + GE + ME	DM + GE + CNV	DM + ME + CNV	ME + GE + CNV	ME + GE + CNV + DM
C-index	0.673	**0.683**	0.646	0.593	0.559	0.616	0.666	0.585	0.630	0.633	0.637

Omics	Features
ME	hsa.mir.1301, hsa.mir.3200, hsa.let.7a.1, hsa.let.7a.3, hsa.mir.135b, hsa.mir.139
DM	cg04533291, cg03874199, cg27342801, cg20676475, cg06101324

Table 8 displays the results for the OV dataset, which exhibited the lowest performance relative to the other cancers, a trend consistent with the single-omics results. Despite this, the best-performing modality combination for OV was ME and DM, achieving a c-index of 0.683. This combination outperforms the single-omics models. However, it's important to note that OV had relatively few features contributing to this predictive power compared to the other cancers.

Table 9: Multi-omics results for CESC

Performance	ME + GE	ME + DM	GE + DM	GE + CNV	ME + CNV	DM + CNV	DM + GE + ME	DM + GE + CNV	DM + ME + CNV	ME + GE + CNV	ME + GE + CNV + DM
C-index	0.654	0.640	0.699	0.738	0.632	0.718	0.822	0.737	**0.835**	0.693	0.720

Omics	Features
DM	cg25514503, cg16607065, cg01612158, cg12958813
ME	hsa.mir.502, hsa.mir.101.2, hsa.mir.140, hsa.mir.150, hsa.mir.500b, hsa.mir.1306, hsa.mir.142, hsa.mir.204, hsa.mir.144, hsa.mir.188, hsa.mir.155, hsa.mir.205, hsa.mir.335, hsa.mir.148a, hsa.mir.196a.1, hsa.mir.151a, hsa.mir.193a, hsa.let.7e
CNV	FAM91A2P, RNU6.488P

Similarly to BRCA, Table 9 reveals that the combination of DM, ME, and CNV yields the highest performance for CESC, achieving a notable c-index of 0.835. ME was the most influential modality contributing to this high predictive power.

Table 10: Multi-omics results for UCEC

Performance	ME + GE	ME + DM	GE + DM	GE + CNV	ME + CNV	DM + CNV	DM + GE + ME	DM + GE + CNV	DM + ME + CNV	ME + GE + CNV	ME + GE + CNV + DM
C-index	0.654	0.723	0.684	0.650	0.650	0.609	0.678	0.606	**0.766**	0.667	0.687

Omics	Features
DM	cg25514503, cg16607065, cg01612158, cg12958813
ME	hsa.mir.502, hsa.mir.101.2, hsa.mir.140, hsa.mir.150, hsa.mir.500b, hsa.mir.1306, hsa.mir.142, hsa.mir.204, hsa.mir.144, hsa.mir.188, hsa.mir.155, hsa.mir.205, hsa.mir.335, hsa.mir.148a, hsa.mir.196a.1, hsa.mir.151a, hsa.mir.193a, hsa.let.7e
CNV	FAM91A2P, RNU6.488P

Table 10 shows that for UCEC, the combination of DM, ME, and CNV achieved the highest performance with a c-index of 0.766, with ME being the major contributor. Interestingly, 3 out of 4 cancer types identified this modality combination as the highest in predictive power. It's also notable that DM and ME consistently appeared in the final integration for all four cancers.

4.6 Pan-cancer Analysis - Overlapping Features

After obtaining the final set of multi-modal features for each cancer using our multi-omics pipeline, we conducted a pan-cancer analysis to identify overlapping features among the cancers.

Table 11: Overlapping Features across all cancers

Cancer Types	Overlapping Features
BRCA, OV	METH_cg20676475
BRCA, CESC	METH_cg01612158, METH_cg12958813, ME_hsa.mir.150
BRCA, UCEC	CNV_MIR1204, CNV_U3, METH_cg07548313, METH_cg09478478, METH_cg11657808, METH_cg12105450, METH_cg14419187, METH_cg17525406, METH_cg18277754, METH_cg22605643, ME_hsa.mir.150, ME_hsa.mir.22, ME_hsa.mir.30a ME_hsa.mir.31
OV, UCEC	ME_hsa.let.7a.1, ME_hsa.let.7a.3, ME_hsa.mir.135b
CESC, UCEC	METH_cg16607065, ME_hsa.mir.140, ME_hsa.mir.142, ME_hsa.mir.144, ME_hsa.mir.148a, ME_hsa.mir.150, ME_hsa.mir.155, ME_hsa.mir.196a.1, ME_hsa.mir.335
BRCA, CESC, UCEC	ME_hsa.mir.150

METH = Methylation, CNV = Copy Number Variation, ME = miRNA

Table 11 presents the results of our direct overlap analysis, highlighting the shared pan-cancer signatures among the cancers. The most overlaps were observed between BRCA and UCEC, with methylation signatures comprising many of the overlapping features, and between CESC and UCEC, where miRNA signatures predominated.

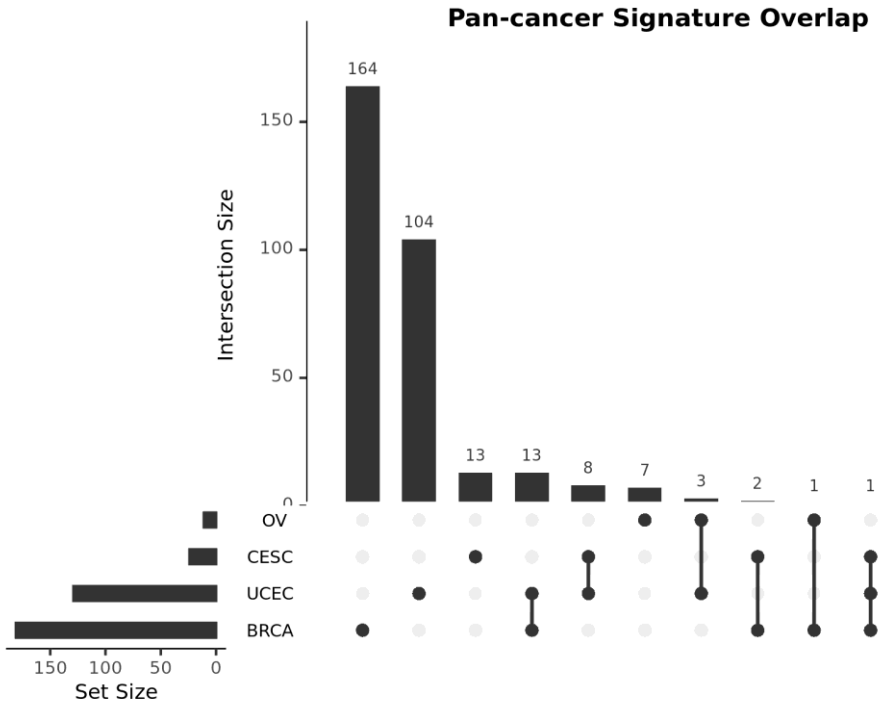


Figure 14: Upset plot showing direct overlapping features between cancers

Figure 14 offers additional insights. Notably, BRCA had the greatest number of features, while OV contributed very few. No feature appeared in all four cancers, with only the miRNA signature mir.150 being common to three cancers: BRCA, CESC, and UCEC.

4.7 MiRNA Targets

The goal of this analysis is to uncover further overlaps among our cancers beyond the direct feature overlap comparison. This involves examining the miRNA targets of our miRNA features from each cancer to identify overlapping gene targets. Using the multiMiR package, we extracted validated miRNA-target interactions and filtered these validated gene targets to include only those expressed in our studied cancers.

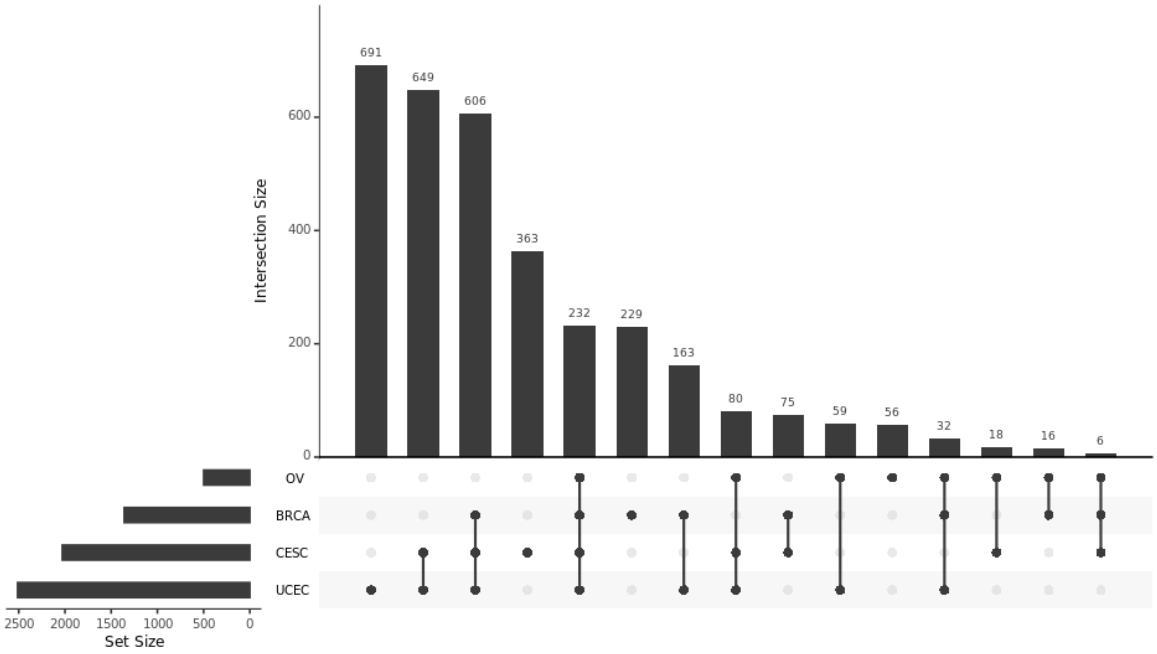


Figure 15: Upset plot showing direct overlapping miRNA Targets between cancers

Figure 15 shows the results of this analysis. We identified 232 gene targets shared across all four cancers. Notably, UCEC contributed the most gene targets, while OV provided the least.

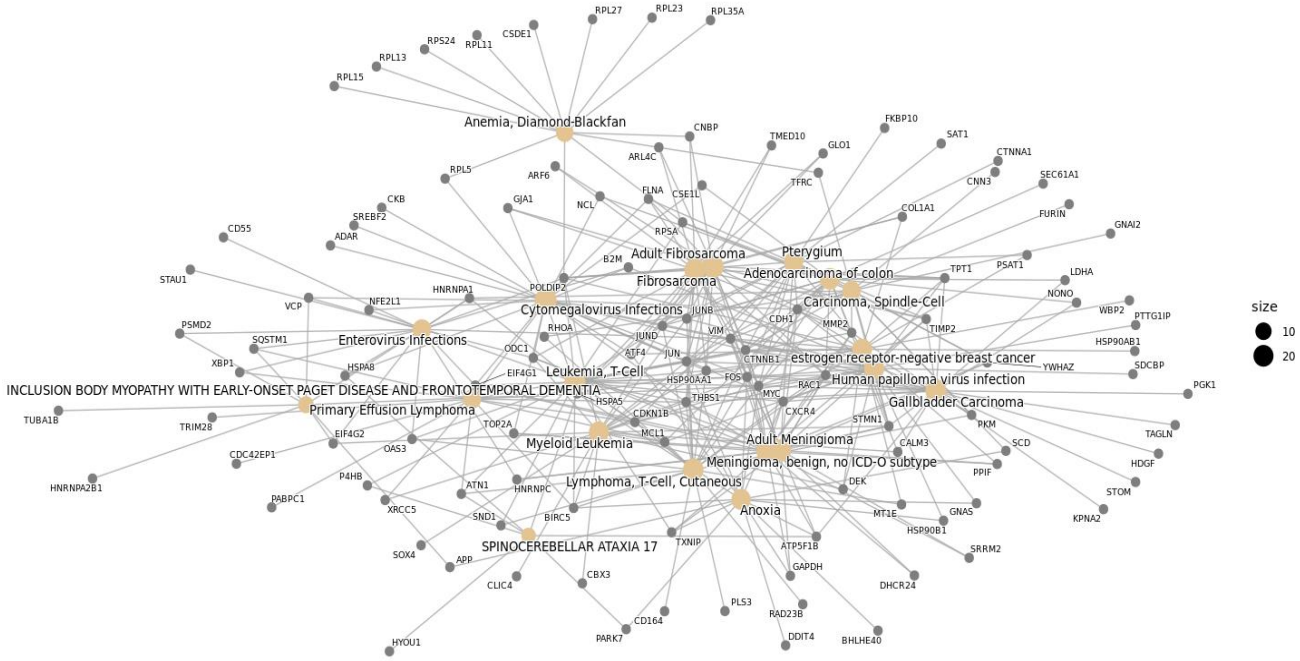


Figure 16: Disease gene network of overlapped gene targets

Using the overlapping 232 gene targets, we performed enrichment analysis with DisGeNET, a database of gene-disease associations, to determine whether these overlapping targets are biologically meaningful and involved in oncogenic pathways. This analysis can help uncover other shared pathways among these targets, potentially providing further therapeutic targets for future research.

As shown in Figure 16, the analysis revealed key oncogenic diseases such as estrogen receptor-negative breast cancer, lymphoma, and various carcinoma types, confirming that these predictive features are associated with oncogenic pathways. We also identified diseases linked to our cancers, such as HPV infection, a well-known risk factor for cervical cancer [Castellsagué et al., 2008]. Interestingly, leukemia-related genes were also present, indicating potential shared pathways between blood cancers and solid tumors. Additionally, Anemia, Diamond-Blackfan, a rare blood disorder that can predispose individuals to cancer [Lipton et al., 2006], was identified.

Most of these gene targets are associated with malignancies, highlighting common oncogenic pathways and suggesting new therapeutic targets across these cancers and related diseases.

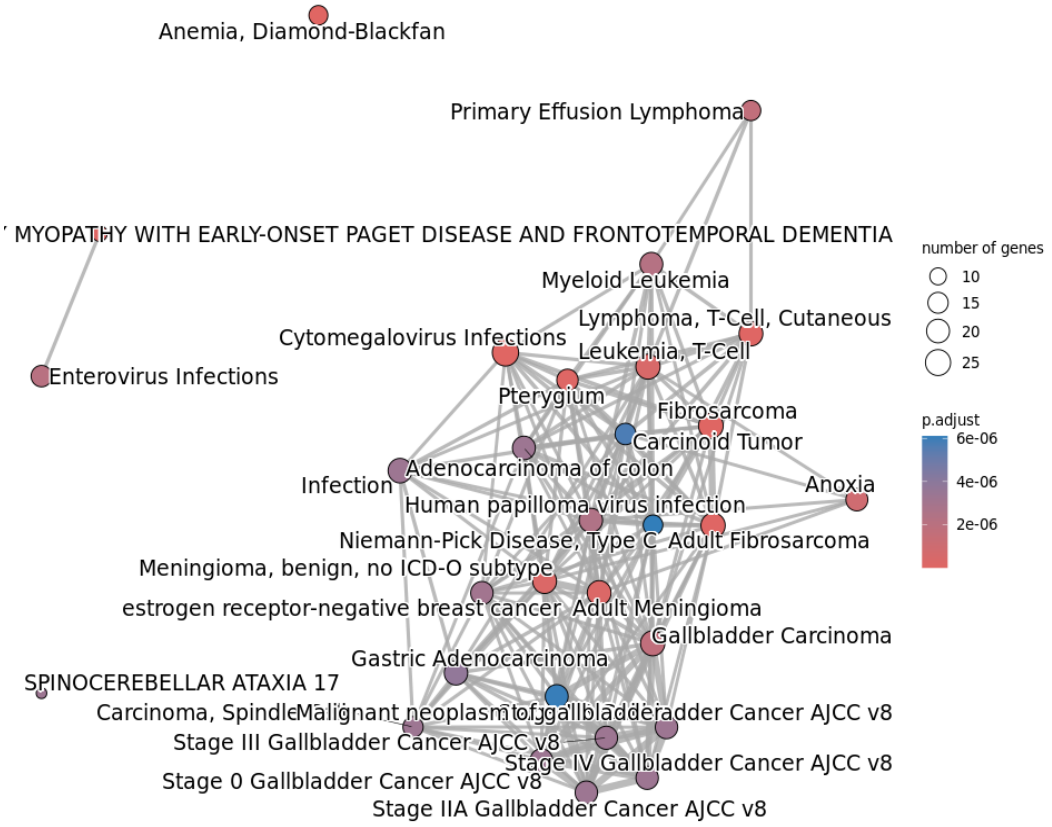


Figure 17: Enrichment Map of gene disease network

Aside from examining the gene-disease network, Figure 17 presents an enrichment map that identifies clusters within these diseases, as mutually overlapping genes tend to form clusters, helping us pinpoint functional modules. The map reveals that cancers such as estrogen receptor-negative breast cancer, adenocarcinoma of the colon, and gallbladder carcinoma cluster together, indicating common pathways among these malignancies.

As expected, hematologic malignancies like myeloid leukemia, T-cell leukemia, and primary effusion lymphoma also form clusters, along with different stages of gallbladder cancer. Interestingly, the map also shows that neurodegenerative and genetic disorders, such as Niemann-Pick disease, appear within the cancer clusters. This finding suggests potential overlapping gene targets involved in both cancer and these conditions.

4.8 Overlapping MiRNA

In this analysis, instead of examining the miRNA gene targets, we focused on the actual overlapping miRNA features from our direct overlap analysis—those miRNA signatures that appeared in two or more cancers. To further validate the biological significance of our predictive features, we explored the miRNAs' disease associations using the mir2disease database.

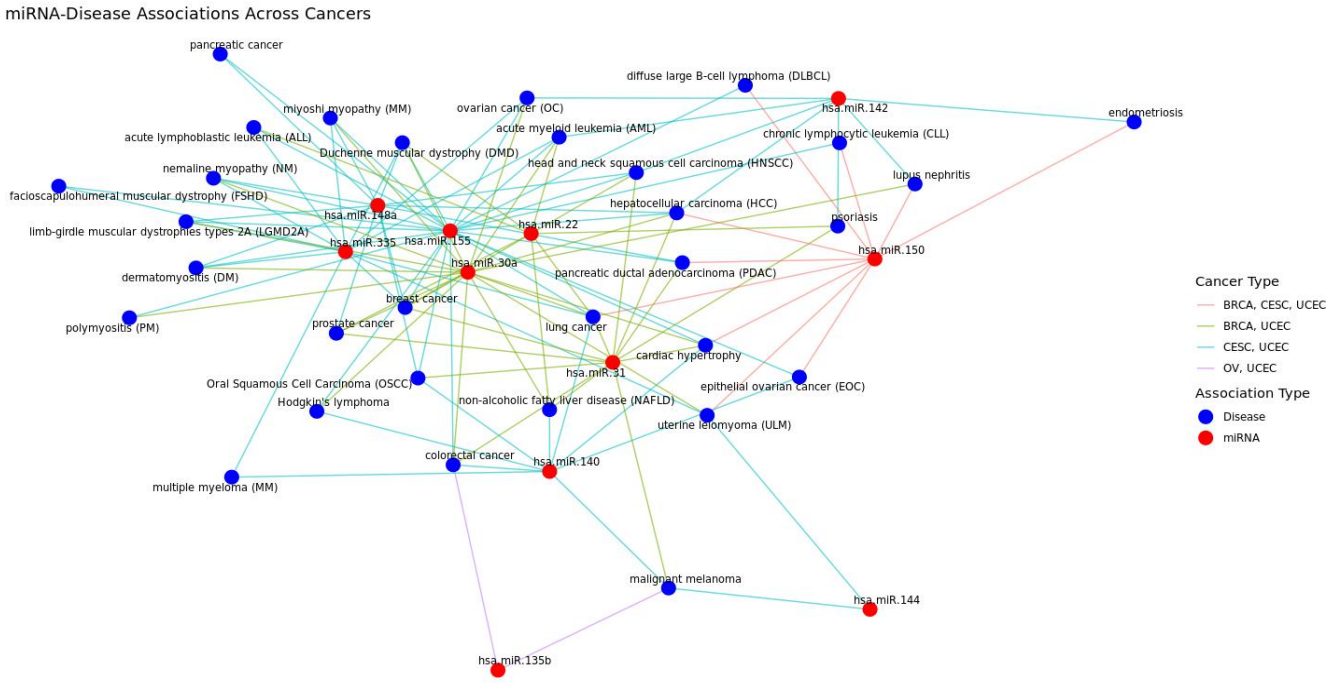


Figure 18: MiRNA Disease associations across cancers

Similar to the disease networks derived from gene targets, the miRNA disease network further validates our findings. As shown in Figure 18, most of the disease terms are related to various forms of cancer, including carcinoma, lymphoma, leukemia, myeloma, and melanoma. This indicates that these miRNAs may be involved in multiple oncogenic processes and molecular pathways. The recurrence of these cancer types in the disease association networks suggests that these specific miRNAs might regulate genes and pathways crucial for tumorigenesis and cancer progression across diverse cancer types. This highlights the possibility of shared oncogenic mechanisms, where the same miRNAs influence multiple forms of cancer, potentially pointing to common therapeutic targets for intervention.

Additionally, this network reveals the presence of muscle-related diseases, such as muscular dystrophy and polymyositis. This suggests that these miRNAs may also play a role in muscle-related pathways, potentially indicating shared biological mechanisms between muscle diseases

and certain cancers.

4.9 GSEA

To further identify shared biological themes and validate our signatures, we performed Gene Set Enrichment Analysis (GSEA) on our miRNA targets to uncover common Gene Ontology (GO) Terms and KEGG Pathways that are significant across cancers.

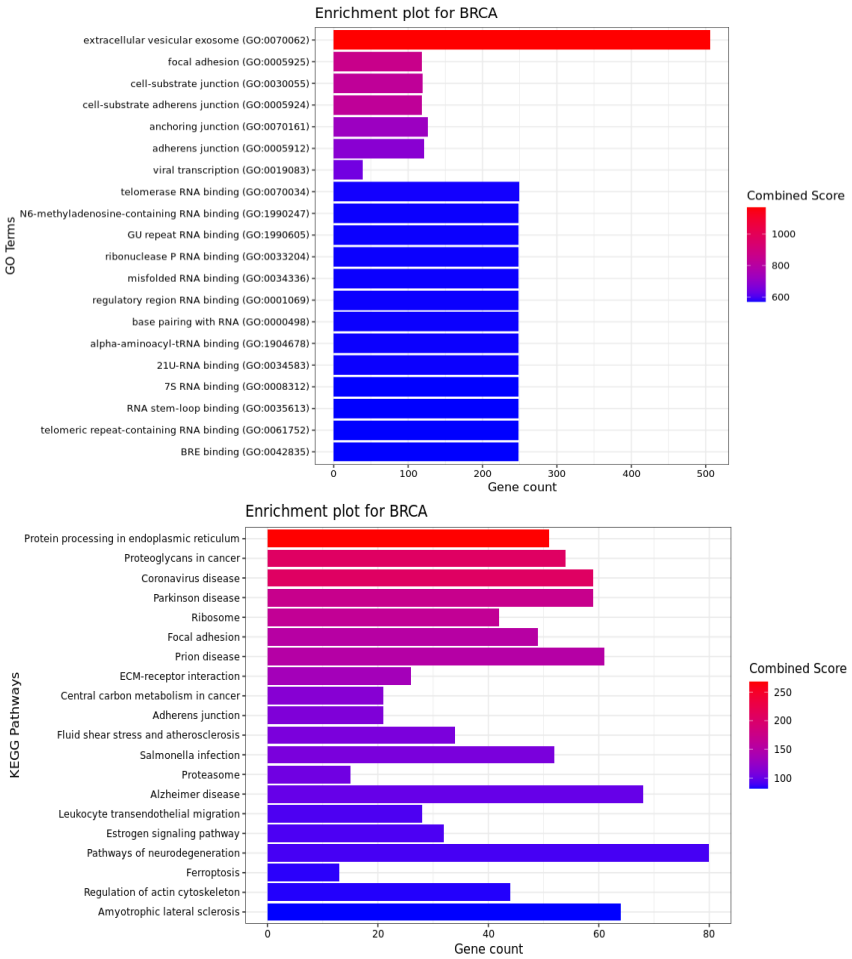
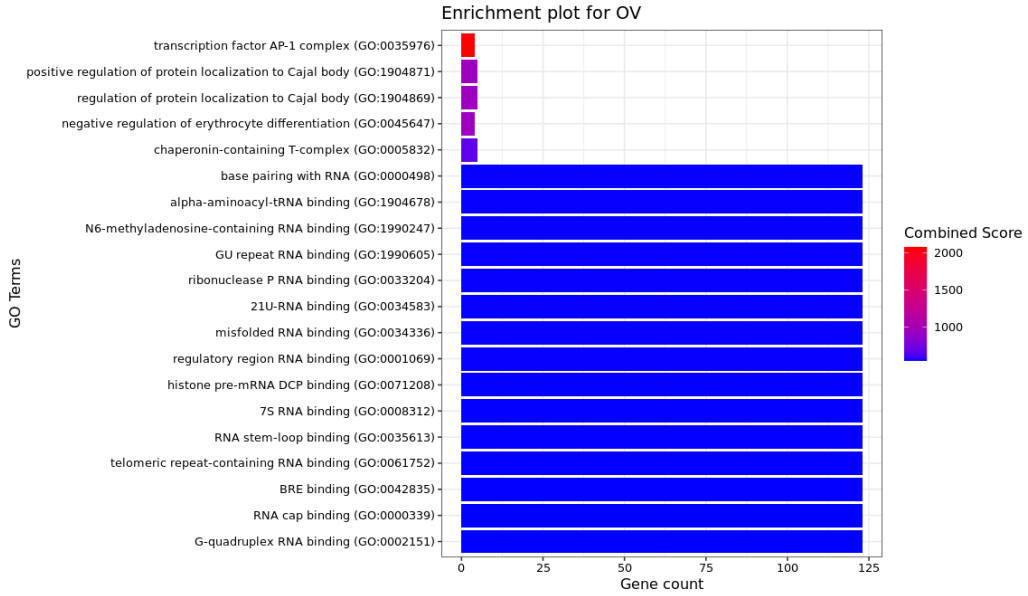


Figure 19: GSEA for BRCA. Top 20 GO Terms (TOP), Top 20 KEGG Pathways (BOTTOM)

Figure 19 shows the results of the enrichment plot for the BRCA miRNA targets. The enrichment plots are ordered by their combined score, which is calculated using the adjusted p-value and z-score, ensuring the robustness of the significance of our gene targets within the associated GO Term or KEGG Pathway. The gene count on the x-axis indicates how many of our gene targets are associated with the GO Term or KEGG Pathway. This GO enrichment plot reveals significant involvement in extracellular vesicular transport, cell adhesion, and RNA binding/processing. These findings suggest that the gene targets may regulate key processes

related to tumor progression. For example, miRNAs that regulate extracellular vesicular transport could influence how cancer cells send and receive signals, aiding in tumor growth and metastasis. Dysregulated cell adhesion is a hallmark of cancer progression [Sousa, Pereira, and Paredes, 2019]. miRNAs involved in RNA binding and processing can impact gene expression at the post-transcriptional level. By modulating RNA dynamics, miRNAs can control the expression of genes involved in tumor suppression or oncogenesis.

The KEGG enrichment plot reveals significant involvement in protein processing and pathways such as proteoglycans in cancer, which play roles in cell adhesion, signaling, and are associated with the tumor microenvironment [Iozzo and Sanderson, 2011]. Their dysregulation can promote cancer cell proliferation, invasion, and metastasis. Interestingly, we also see infectious pathways like Coronavirus, Salmonella infection, and prion disease, indicating potential intersections with immune responses and infectious diseases. This reflects the complex interplay between cancer and the immune system. Similar to our disease association networks, pathways associated with neurodegenerative diseases, which often intersect with cancer pathways [Seo and Park, 2020], are also highlighted here.



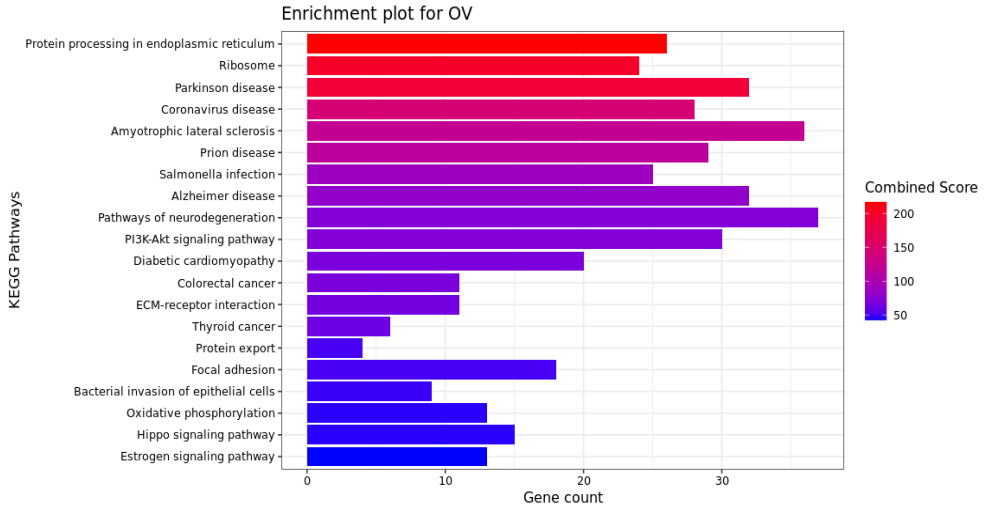
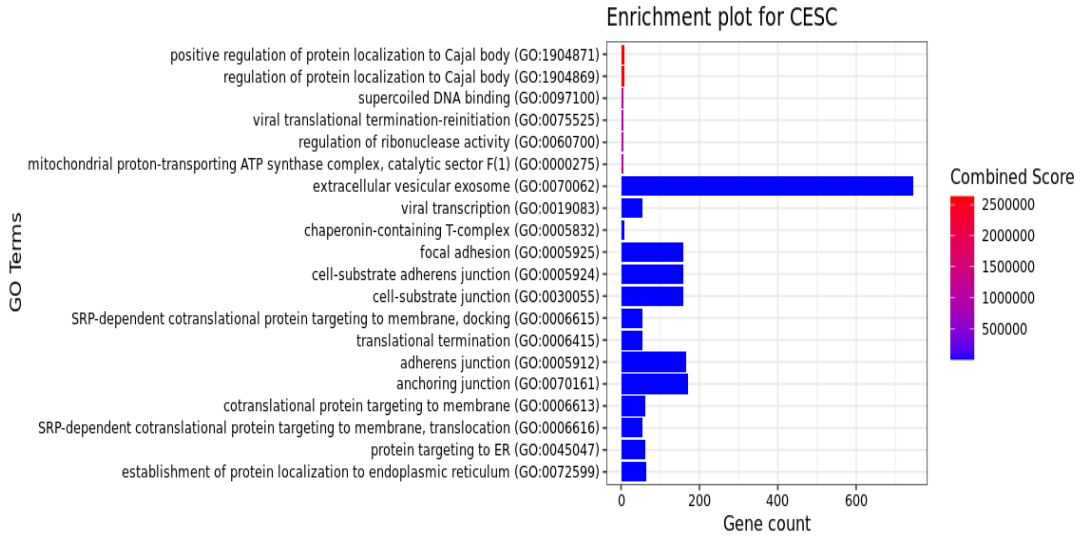


Figure 20: GSEA for OV. Top 20 GO Terms (TOP), Top 20 KEGG Pathways (BOTTOM)

Figure 20 shows the results for the OV targets, revealing that the significant GO Terms involve targets in diverse RNA-related processes. These include binding to specific RNA motifs, regulating protein complexes, and potentially influencing broader biological functions. Dysregulation of these functions can implicate cancer initiation and progression.

For the KEGG plot, similar to BRCA, these targets are involved in protein processing, infectious, and neurodegenerative disease pathways. Additionally, the targets are implicated in several other cancer pathways as well as various signaling pathways.



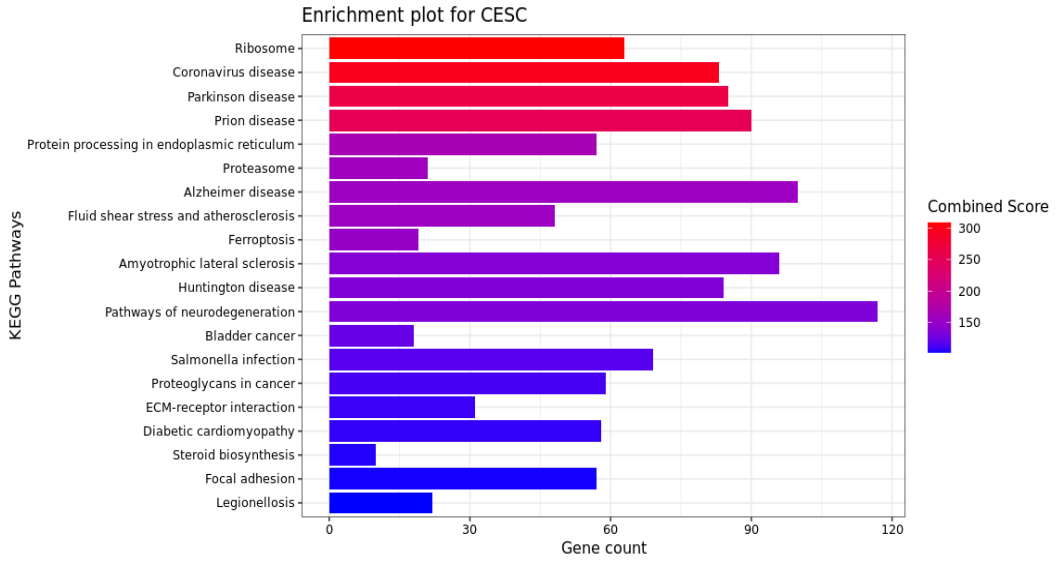
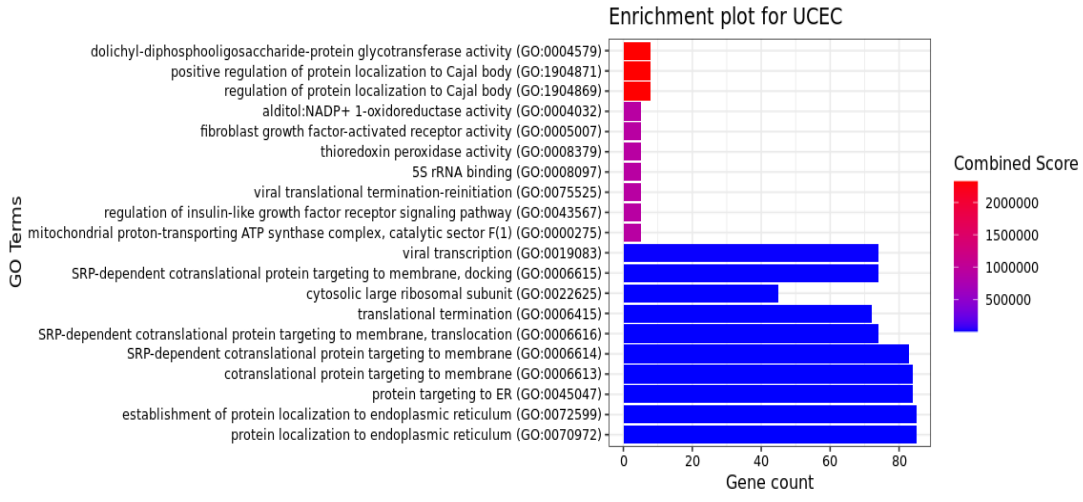


Figure 21: GSEA for CESC. Top 20 GO Terms (TOP), Top 20 KEGG Pathways (BOTTOM)

Figure 21 displays the results for the CESC miRNA targets. The significant GO Terms indicate that these gene targets may play crucial roles in regulating cellular processes such as protein localization, DNA/RNA interactions, cellular junctions, and organelle function. Dysregulation of these processes and abnormal protein localization are associated with tumorigenesis and cancer progression.

For the KEGG Pathways, there is a noticeable trend involving fundamental cellular processes, various cancer-related pathways, and disease-related pathways, including those linked to neurodegenerative disorders and infectious diseases. Notably, pathways such as proteoglycans in cancer and focal adhesion highlight the importance of extracellular matrix interactions and cell signaling in cancer progression.



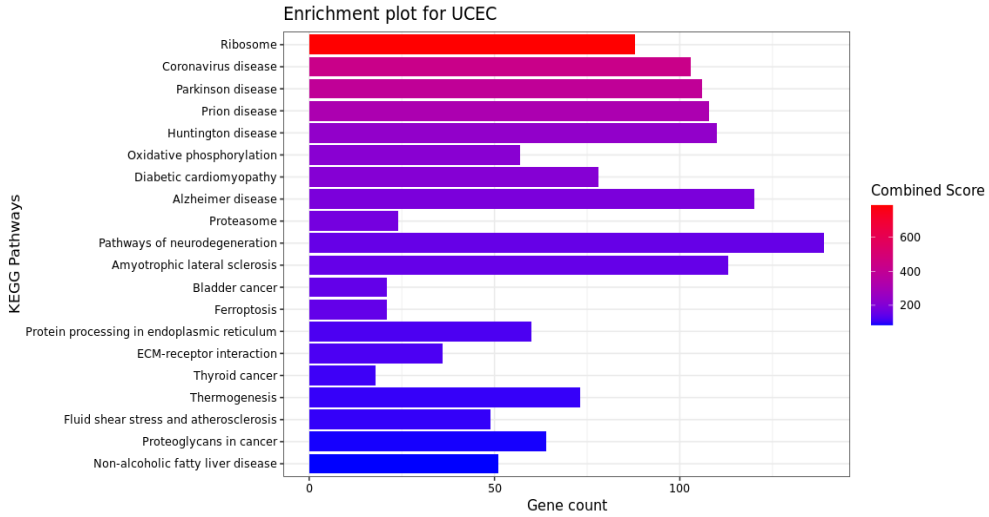


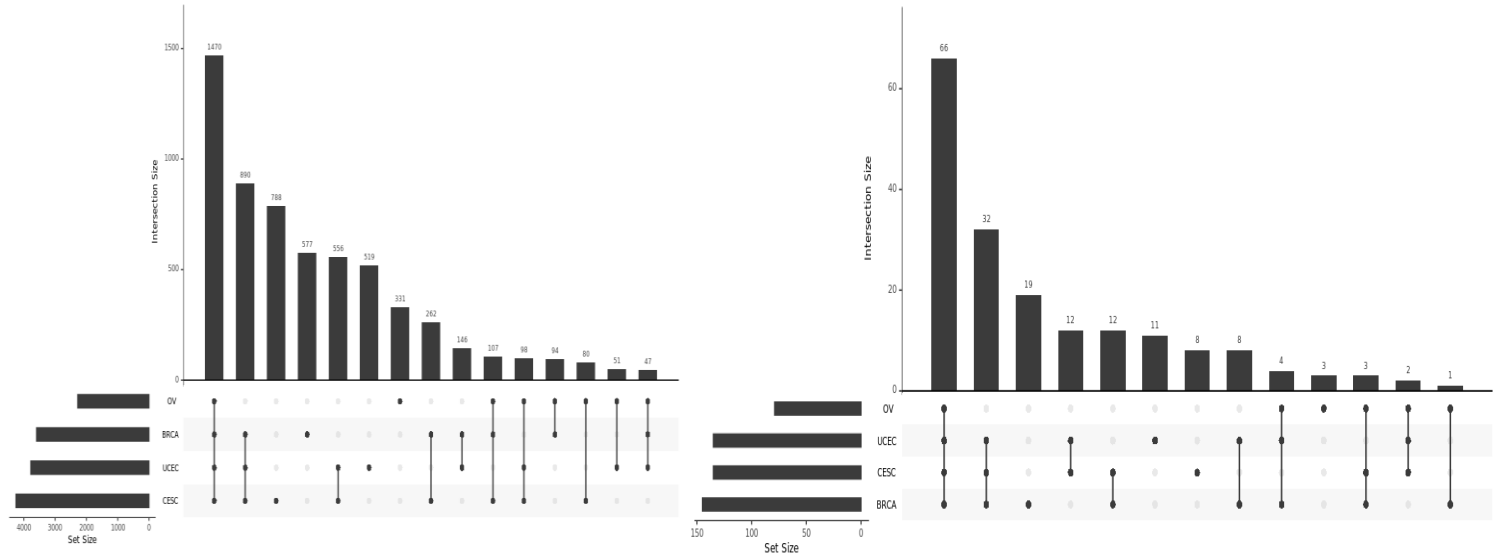
Figure 22: GSEA for UCEC. Top 20 GO Terms (TOP), Top 20 KEGG Pathways (BOTTOM)

Figure 22 displays the results for the UCEC miRNA targets. The significant GO Terms highlight the targets' roles in protein regulation, RNA metabolism, cellular signaling, and membrane dynamics. Dysregulation of these processes is implicated in tumorigenesis and cancer progression.

For the KEGG Pathways, a notable trend across all four cancers is the involvement in fundamental cellular processes, cancer-related pathways, and disease-related pathways, including neurodegenerative disorders and infectious diseases. Additionally, metabolic pathways and those associated with metabolic diseases suggest that dysregulation may play a role in UCEC.

4.10 GSEA Overlaps

Having reviewed the GSEA results for each individual cancer's miRNA targets and observed notable trends, we have validated their biological relevance through associations with oncogenic pathways and other disease networks. Our next step is to analyze the overlapping GO Terms and KEGG Pathways across the different cancers. This will help us identify common biological themes and pathways shared by the miRNAs across the cancers, further elucidating the underlying mechanisms and potential therapeutic targets.



GO Term KEGG Pathway

Figure 23: Upset plot showing GSEA overlaps across all cancers for miRNA targets

Figure 23 presents the upset plot for GO Term and KEGG Pathway overlaps. The plot reveals that the highest intersection size occurs when all four cancers (BRCA, OV, UCEC, and CESC) share common GO terms. This suggests that these cancers may have overlapping biological processes or functions regulated by the miRNA gene targets analyzed. This finding highlights potential common pathways or biological mechanisms influenced by these miRNAs across different cancer types. Identifying these shared regulatory mechanisms is significant as it may reveal crucial targets for further investigation or therapeutic intervention.

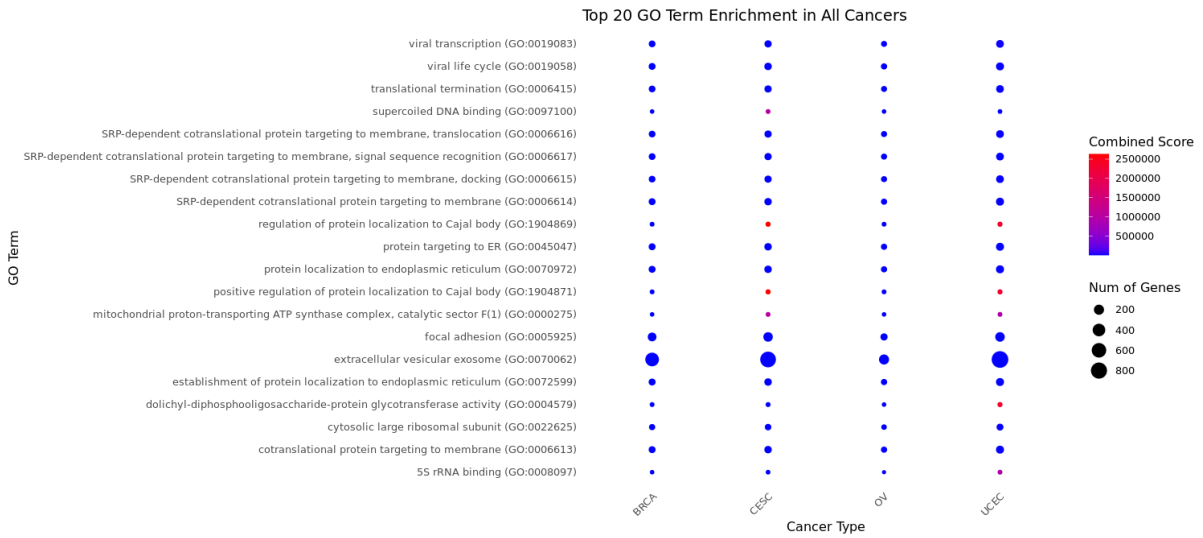


Figure 24: Enrichment dot plot showing TOP 20 GO Terms shared in all four cancers

Figure 24 displays the top 20 common GO Terms shared across all four cancers. The enrichment dot plot highlights several key biological processes critical for cancer progression, including protein localization and targeting, mitochondrial and ribosomal functions, DNA and RNA processes, and viral mechanisms. Terms such as "protein targeting to ER" and "focal adhesion" underscore the importance of accurate protein synthesis, cell adhesion, and migration in tumor development. The involvement of extracellular vesicles points to their role in cell communication within the tumor microenvironment [Lopez et al., 2023]. These findings provide valuable insights into crucial processes for cancer cell survival, growth, and metastasis, potentially guiding the identification of therapeutic targets.

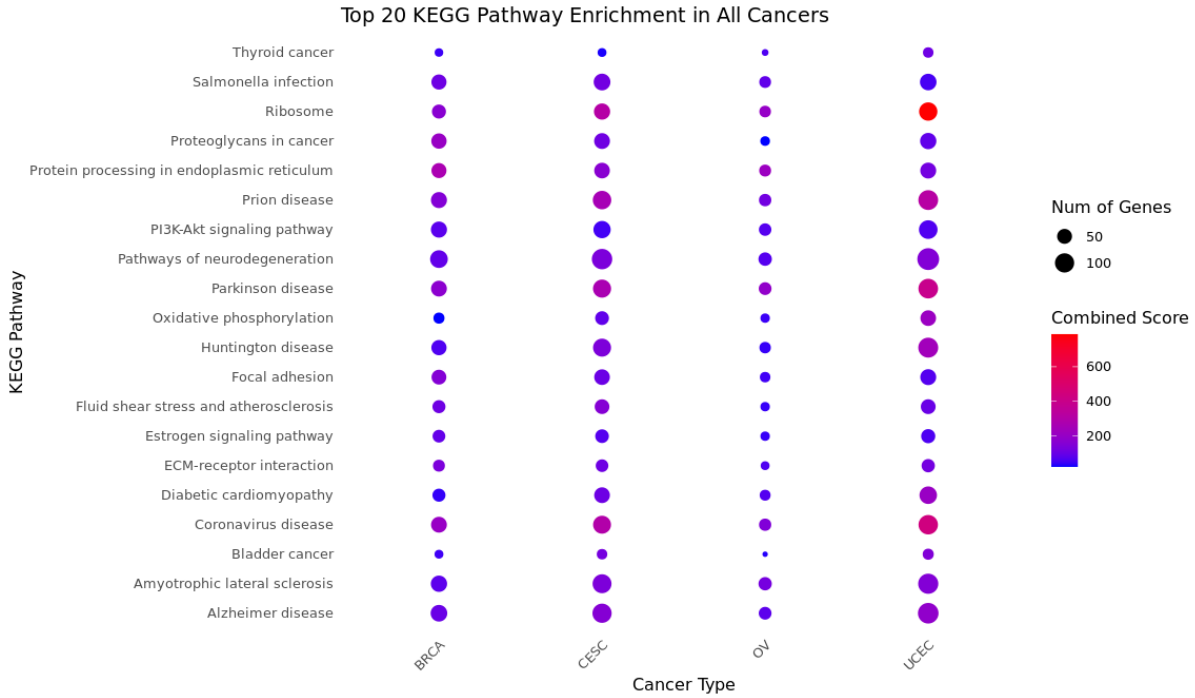


Figure 25: Enrichment dot plot showing TOP 20 KEGG Pathways shared in all four cancers

Figure 25 illustrates the KEGG Pathways for our miRNA targets. Consistent with the trends observed in the individual GSEA results, the targets are involved in fundamental cellular processes. Key pathways such as "Ribosome" and "Protein processing in endoplasmic reticulum" are crucial for protein synthesis and folding. Infectious pathways, including Coronavirus and prion diseases, are also notably shared. Additionally, neurodegenerative diseases like "Parkinson's," "Alzheimer's," and "Huntington's" are represented, highlighting the intersection between neurodegeneration, infectious diseases, and cancer.

These results emphasize not only the biological relevance of the miRNA targets to cancer,

through their association with key cancer-related pathways and GO terms, but also reveal other potentially clinically significant disease pathways. This cross-disease relevance underscores the broader impact of these targets on multiple health conditions.

4.11 Methylation Targets

In this analysis, we focused on methylation targets instead of miRNA targets. Using the Illumina Human Methylation 27k database, we obtained annotated cg probe sites. It's important to note that each cg probe site has a one-to-one gene target association, as opposed to the one-to-many associations observed with miRNA and gene targets.

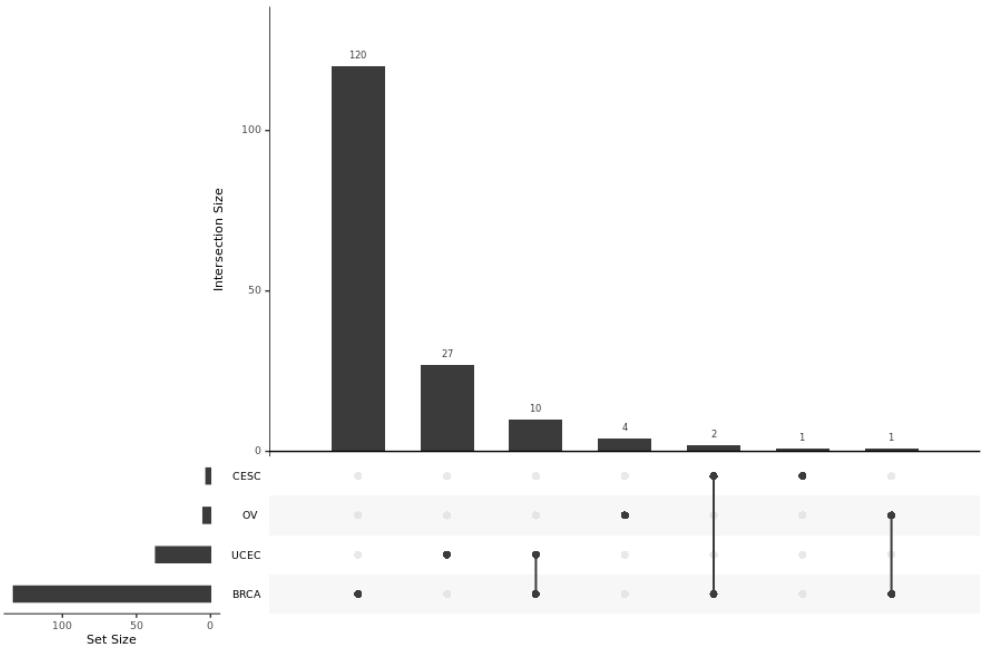


Figure 26: Upset plot showing Methylation target overlaps across all cancers

Due to these one-to-one associations, we obtained fewer gene targets compared to the miRNA analysis (Figure 26). Additionally, BRCA has the most methylation signatures out of all the cancers studied, which skews the results towards BRCA, resulting in more gene targets compared to the other cancers.

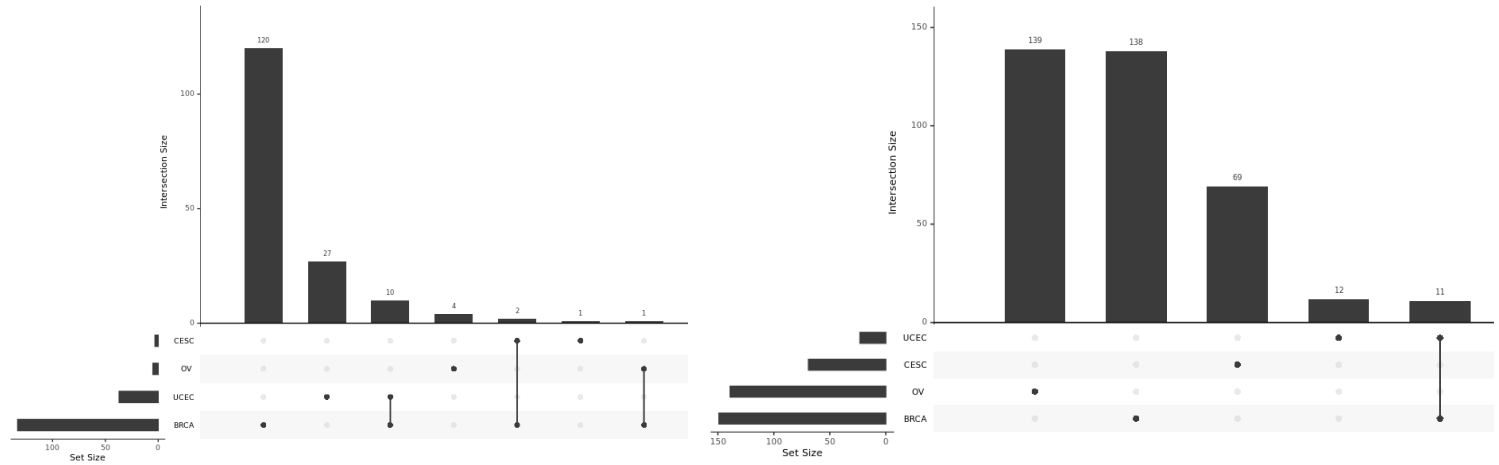


Figure 27: Upset plot showing GSEA overlaps across all cancers for methylation targets

Figure 27 presents the upset plot for GO Terms and KEGG Pathways for the methylation targets. Due to the limited number of gene targets and the unbalanced distribution of methylation signatures, our methylation target analysis did not yield ideal results. There are few overlaps between Pathways and GO Terms across the different cancers. Notably, only UCEC and BRCA share GO Terms.

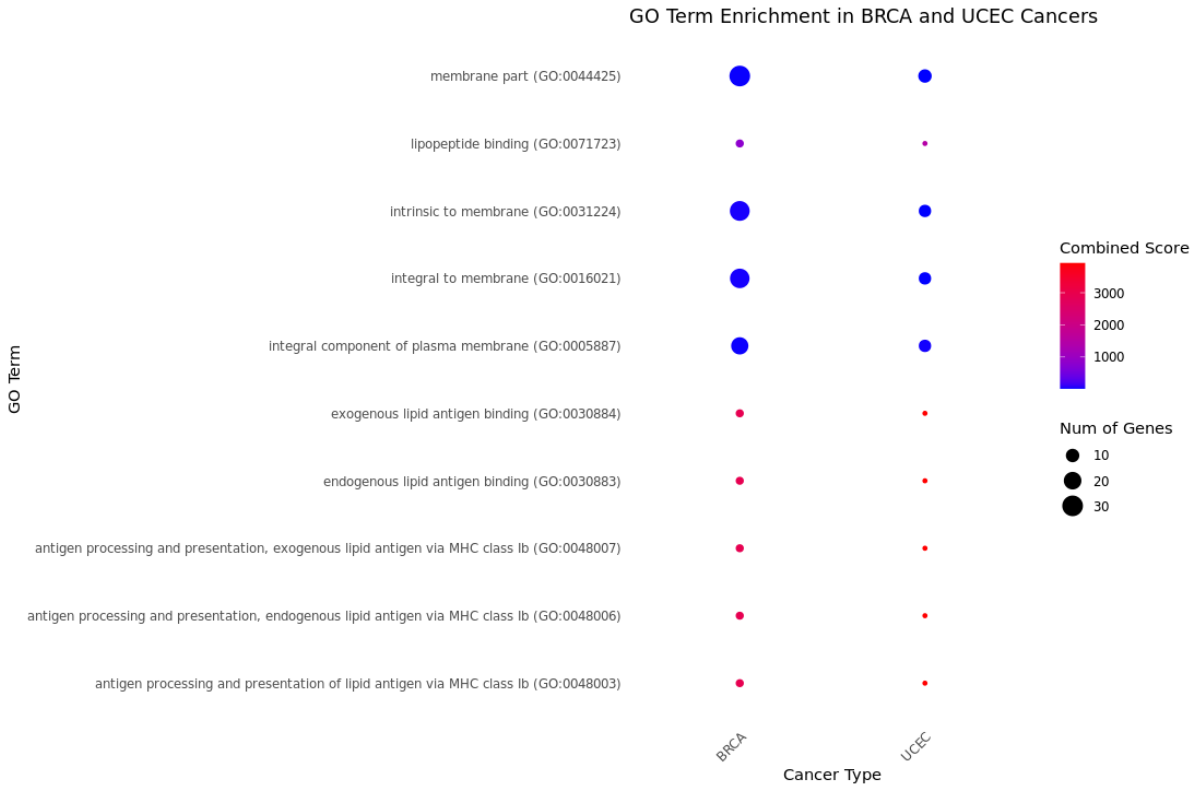


Figure 28: Enrichment dot plot showing TOP 20 GO Terms shared in BRCA and UCEC

Using the overlapping methylation targets obtained from Figure 27, we performed a GSEA and plotted an enrichment dot plot shown in Figure 28. This figure highlights the top 20 GO Terms shared between BRCA and UCEC. The results suggest that these targets are involved in processes related to membrane structure and function, as well as specific immune-related responses linked to lipid antigens. This might indicate immune evasion in the context of cancer, with dysregulation of these processes contributing to cancer progression. However, it is important to note that due to the limited number of overlapping gene targets, there is a lack of diversity in the GO Terms.

4.12 Survival Analysis – KM plots

To further validate the biological relevance of our overlapping multi-modal signatures, we employed KM plots and cross-referenced the survival results with existing literature. This validation helps to confirm the significance of the overlapping features and their potential impact on cancer prognosis.

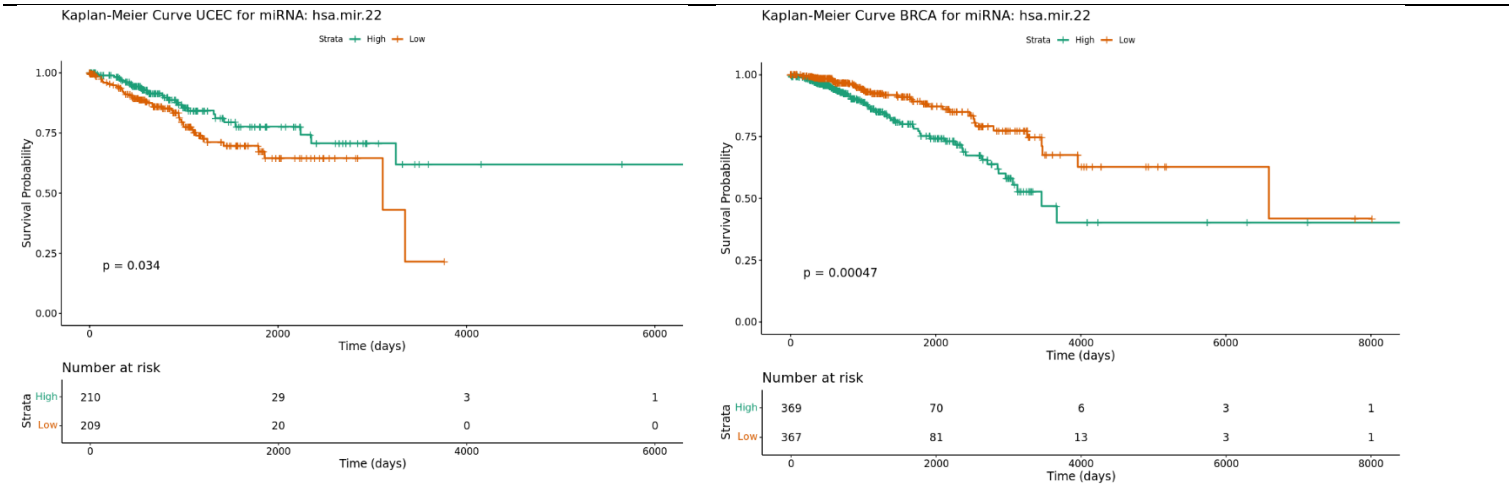


Figure 29: KM Survival plot for miR-22

The first signature we explored is miR-22, which, from our analysis, was present in both UCEC and BRCA (Table 11). We used two studies, Koufaris et al., 2023 and Cui et al., 2023, for literature validation.

Koufaris et al.'s study showed that miR-22, when enriched in BRCA, repressed glycolytic metabolism, which reduced survival outcomes in patients, indicating that miR-22 can serve as an oncomir. Conversely, Cui et al. demonstrated that in liver cancer, miR-22 inhibits cancer

cell EMT via regulation of SPRY2, showcasing its ability to act as a tumor suppressor. These studies highlight that miR-22 can function as either an oncomir or tumor suppressor, depending on the cancer type.

Our KM plot (Figure 29) further supports these findings. In BRCA, high levels of miR-22 are associated with lower survival rates over a 5-year period, aligning with Koufaris et al.'s study. In contrast, for UCEC, low levels of miR-22 are associated with lower survival rates, suggesting that miR-22 acts as a tumor suppressor in this context, similar to its role in liver cancer as described by Cui et al.

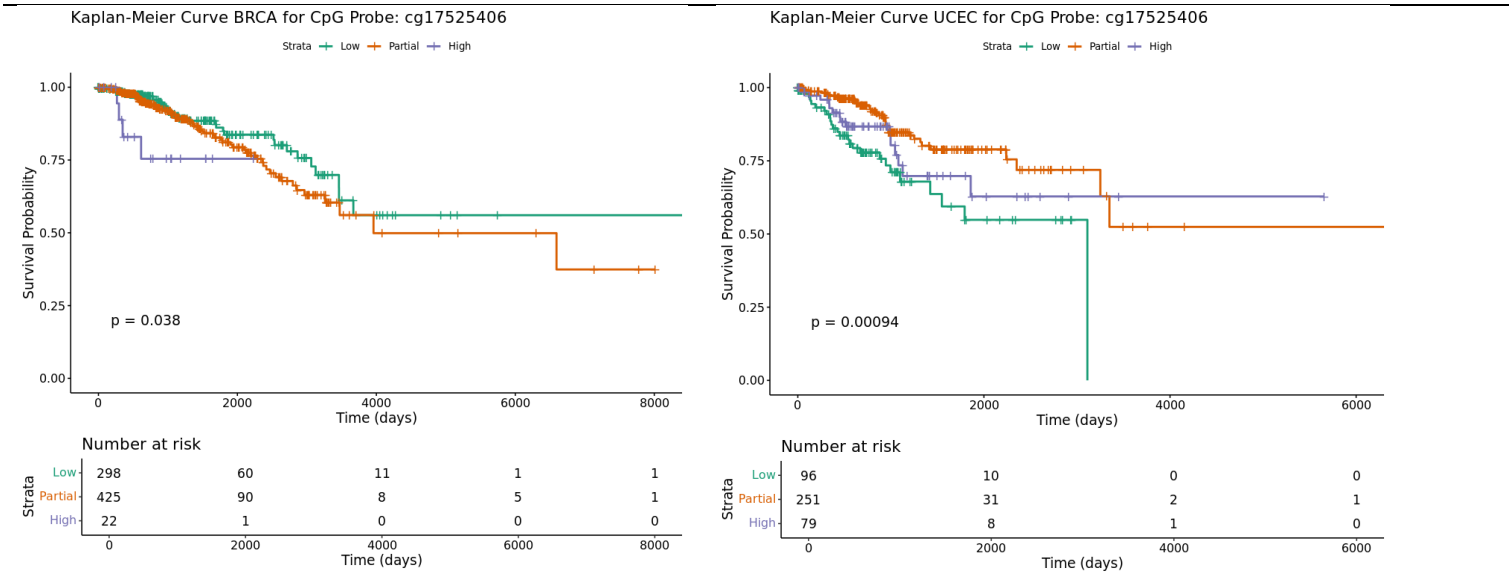


Figure 30: KM Survival plot for cg17525406

The second signature we explored is the cg probe cg17525406, identified in several lung cancer studies as significantly hypermethylated, leading to the silencing of tumor suppressor genes (Shen et al., 2019; Shi, 2021).

In Figure 30, we observe that in BRCA, within a five-year period, a high level of methylation at this probe is associated with lower survival rates, corroborating the findings from lung cancer studies. This suggests that hypermethylation of cg17525406 silences tumor suppressor genes in BRCA, contributing to poorer outcomes. Interestingly, the opposite trend is observed in UCEC. Here, low levels of methylation at cg17525406 are linked to lower survival rates. This could indicate that in UCEC, hypomethylation at this probe might activate a potential oncogene, suggesting a dual role where the cg17525406 probe can act as both a tumor suppressor and an oncogene, depending on the cancer type. This duality highlights the complexity of epigenetic regulation in cancer and underscores the necessity for further studies to understand the specific

roles and mechanisms of cg17525406 in different cancer contexts.

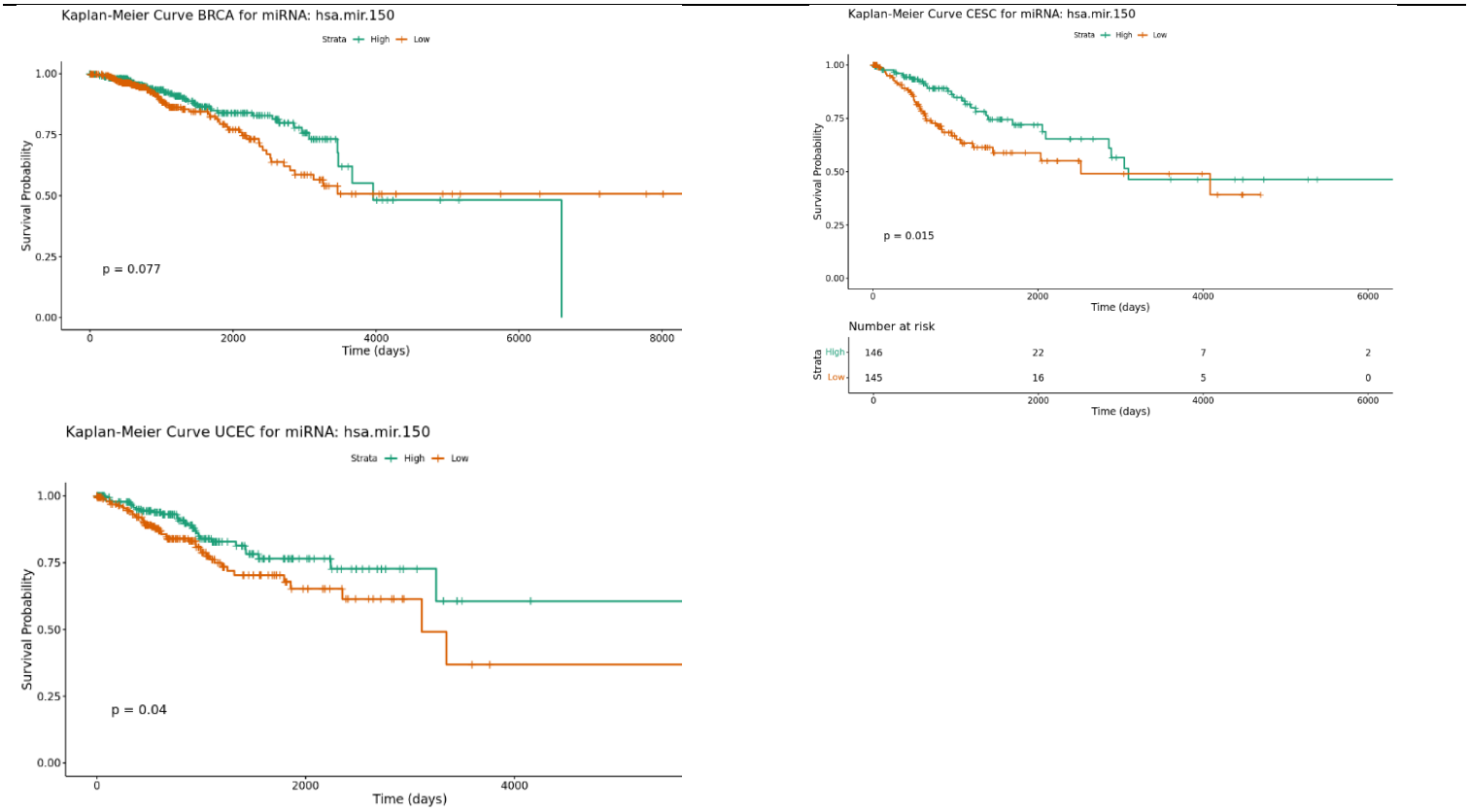


Figure 31: KM Survival plot for miR-150

For our last signature, we explored miR-150, the only multi-modal feature appearing in at least three cancers. To validate its relevance, we refer to three studies: Wang, Ren, and Zhang (2015), Sugita et al. (2022), and Sun et al. (2019), all of which have shown that miR-150 is frequently deregulated in cancers.

In Figure 31, we observe that in UCEC, BRCA, and CESC, lower expression of miR-150 correlates with a lower survival rate over a 5-year period. This trend across multiple cancer types supports the notion that miR-150 could be a significant biomarker for cancer prognosis.

The consistent pattern of lower miR-150 expression leading to reduced survival rates across these cancers aligns with findings from the cited studies, which also highlight miR-150's deregulation in various cancers. This validation strengthens the credibility of our study's features and suggests that miR-150 might play a crucial role in cancer progression and could be a potential target for therapeutic intervention.

Chapter 5

Discussion and Future Work

5.1 Objective

In this thesis, we pursued two primary objectives:

1. To improve the precision of cancer survival predictions by utilizing multi-omics data and leveraging a combination of feature selection methods and survival analysis machine learning algorithms.
2. To identify the pivotal multi-modal features that exert substantial influence on survival outcomes across diverse types of cancer.

To achieve these goals, we drew inspiration from the approaches of prominent researchers. Zhao et al. (2020) conducted research on the identification of pan-cancer prognostic biomarkers through the integration of multi-omics data. Tong et al. (2020) explored deep learning-based feature-level integration of multi-omics data for breast cancer survival analysis. In terms of machine learning models and feature selection methods, we delved into Spooner et al. (2020) comprehensive list of existing methods to inform our exploration and analysis.

5.2 Findings

We have demonstrated that multi-modal integration significantly enhances survival predictions, surpassing baseline predictions derived from single-omics data. Among the survival machine learning algorithms evaluated, Random Forest emerged as one of the most effective, particularly in handling high-dimensional data. The best-performing multi-modal combination involved the late fusion of methylation, miRNA, and copy number variation data across three of the four cancers studied, with methylation and miRNA data being common components in all successful combinations.

These results offer new insights into the ideal machine learning survival model for handling the high-dimensional nature of multi-omics data. Previous studies, including those foundational to this thesis, primarily utilized the baseline CPH model. Our thesis demonstrated that Survival Random Forest provides a superior concordance index (c-index), regardless of the feature selection methods used. This improvement is clearly evident when comparing our study to Tong et al. (2020), who used a different multi-omics integration technique but also utilized the CPH model for survival predictions. They concluded that DM and ME data contain complementary information that improves survival predictions, achieving a c-index of 0.641. In contrast, our BRCA results achieved a c-index of 0.774, and we also discovered CNV data provide complementary information with ME and DM.

Even without considering our highest-performing combination, our integration of DM and ME data alone yielded a higher c-index of 0.702 using a basic fusion multi-omics integration strategy, compared to Tong et al.'s more refined approach. This indicates that selecting the most appropriate survival model is crucial before considering any multi-omics integration methods. These findings underscore the importance of model selection in improving survival predictions, as well as further showcasing that different omics data indeed complement each other in enhancing survival predictions; in our case, it was the combination of ME, DM, and CNV data.

After obtaining our multi-omics results, we moved on to our pan-cancer analysis to explore overlapping features across the cancers studied. This step aimed to validate the biological relevance of these features and uncover potential common oncogenic pathways and therapeutic targets. While no single feature was present across all four cancers, miR-150 was notable for its presence in three cancers. Further investigation into gene targets revealed significant overlaps in Gene Ontology (GO) terms and KEGG pathways, suggesting shared biological processes. The features analyzed were found to be involved in common oncogenic pathways as well as pathways related to infectious and neurodegenerative disorders, as evidenced by both Gene Set Enrichment Analysis (GSEA) and disease-association network analysis.

The Kaplan-Meier (KM) analysis of the multi-modal signatures underscored their significance in survival outcomes and validated their biological relevance through cross-referencing with existing literature. This analysis also revealed that, depending on the cancer type, certain features can function as either oncogenes or tumor suppressors. These findings underscore the potential of multi-modal data integration to uncover critical biomarkers and pathways that are instrumental in cancer progression and patient survival, offering valuable insights for future therapeutic strategies.

5.3 Areas of Improvement

Despite the positive outcomes of our analysis, several areas require improvement before expanding upon this thesis.

Firstly, while the thesis utilized pre-processing techniques similar to those employed by Zhao et al. (2020) and Tong et al. (2020), we found these methods to be somewhat simplistic and generalized. Applying a uniform pre-processing pipeline across all cancers may have been too broad, particularly evident in the OV dataset, where the lower c-index suggests that the pre-processing approach might not have been optimal for this cancer type. More tailored pre-processing strategies specific to each cancer type could potentially yield better results.

Additionally, our use of basic imputation techniques, such as mean imputation, contrasts with more sophisticated methods like Multiple Imputation by Chained Equations (MICE) employed in other studies. Enhancing imputation strategies could improve data quality and, consequently, the performance of our models.

The pre-processing of copy number variation (CNV) data was also quite basic. We used direct Gene Level Scores from TCGA, while other studies, such as those using GISTIC2.0 processing from Firebrowse, might provide more refined CNV data. Incorporating more advanced CNV processing techniques could improve the accuracy and relevance of the data.

Furthermore, our analysis did not account for differentially expressed genes (DEGs) and differentially methylated regions (DMRs), which are known to be significant for data filtration and could have enriched the quality of gene features and cg probes used in the study. Addressing these aspects in future work will be crucial for obtaining higher-quality data and refining the analysis.

5.4 Future Work

Building on the achievements of this thesis, several promising directions for future research can be pursued:

- Expansion of Multi-Omics Integration: Extending the scope of multi-omics studies to

include additional data types, such as proteomics and metabolomics, could offer a more comprehensive view of the molecular landscape of cancers. Integrating these new omics layers with existing DM, ME, GE, and CNV data may enhance our understanding of cancer biology and improve survival predictions.

- Advanced Multi-Omics Integration Techniques: Exploring and comparing a wider range of multi-omics integration techniques beyond basic fusion approaches could provide valuable insights. Tabakhi et al. (2023) offers an extensive review of various multi-omics strategies and available tools. Investigating and benchmarking these strategies in terms of survival prediction performance could lead to more refined and effective integration methods.

- Enhanced Survival Machine Learning Algorithms and Filtering Methods: Further exploration of survival machine learning algorithms and filtering methods is essential. Spooner et al. (2020) provided a benchmark of various survival algorithms and filters, yet due to time constraints and computational limitations, this thesis could not exhaustively explore all available methods. Future work should aim to refine and identify the most effective survival models and filtering techniques tailored to specific cancer types.

- Functional Validation of Pan-Cancer Signatures: For the pan-cancer signatures identified in this thesis, conducting experimental validation is crucial. This includes validating the roles of identified miRNAs and methylation targets in cell lines and animal models to confirm their involvement in cancer progression. Longitudinal studies tracking changes in miRNA and methylation patterns over time in patients could provide insights into the dynamic nature of cancer progression and treatment response.

- Therapeutic Exploration: Investigating the therapeutic potential of targeting identified miRNAs and methylation sites is an important next step. This could involve developing and testing miRNA mimics or inhibitors and exploring epigenetic drugs. Conducting preclinical and clinical trials to evaluate the efficacy and safety of these therapeutic approaches will be vital for translating these findings into potential treatments.

- Interdisciplinary Investigations: Detailed studies into shared pathways between cancer, neurodegenerative diseases, and infectious diseases could uncover novel insights into

the interplay between these conditions. Understanding these shared mechanisms may reveal new therapeutic strategies and improve our comprehension of disease interactions.

By pursuing these avenues, future research can build upon the foundation laid by this thesis, leading to more advanced, integrative, and clinically relevant discoveries in cancer research and beyond.

5.5 Conclusion

This report has thoroughly addressed the background motivations and objectives set forth for this thesis. We successfully achieved our goals by demonstrating that multi-modal integration significantly enhances survival predictions compared to single-omics approaches. Through the application of feature selection methods and survival analysis algorithms, we found that Random Forest is particularly adept at managing high-dimensional data and delivering precise survival predictions.

Our study identified the most effective multi-modal combination for survival prediction as the integration of DM, ME, and CNV data, which showed superior performance in three out of four cancers examined. This finding highlights the complementary nature of these omics data types. Additionally, our pan-cancer analysis uncovered substantial overlaps in biological processes and pathways across different cancers, validating the biological relevance of the multi-modal signatures and identifying potential common oncogenic pathways and therapeutic targets. The KM analysis further supported the survival significance of these signatures, reinforcing their biological relevance through cross-referencing with existing literature.

In summary, this thesis offers a robust framework for integrating multi-omics data to enhance cancer survival predictions, reveals key biological features crucial for cancer prognosis, and suggests important directions for future research to build upon these findings.

References

1. Zhao, N., Guo, M., Wang, K., Zhang, C. and Liu, X. (2020). Identification of Pan-Cancer Prognostic Biomarkers Through Integration of Multi-Omics Data. *Frontiers in Bioengineering and Biotechnology*, 8. doi:https://doi.org/10.3389/fbioe.2020.00268.

2. Chai, H., Zhou, X., Zhang, Z., Rao, J., Zhao, H. and Yang, Y. (2021). Integrating multi-omics data through deep learning for accurate cancer prognosis prediction. *Computers in Biology and Medicine*, [online] 134, p.104481. doi:https://doi.org/10.1016/j.compbiomed.2021.104481.

3. Spooner, A., Chen, E., Sowmya, A., Sachdev, P., Kochan, N.A., Trollor, J. and Brodaty, H. (2020). A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific Reports*, 10(1).

4. Ezgi Demir Karaman and Zerrin Işık (2023). Multi-Omics Data Analysis Identifies Prognostic Biomarkers across Cancers. *Medical Sciences*, 11(3), pp.44–44. doi:https://doi.org/10.3390/medsci11030044.

5. Tabakhi, S., Mohammod N. I. Suvon, Pegah Ahadian and Lu, H. (2023). Multimodal Learning for Multi-omics: A Survey. World Scientific Annual Review of Artificial Intelligence, 01. doi:https://doi.org/10.1142/s2811032322500047.

6. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C. and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics, 45(10), pp.1113–1120. doi:https://doi.org/10.1038/ng.2764.

7. Yuan, Y., Van Allen, E. M., Omberg, L., Wagle, N., Amin-Mansour, A., Sokolov, A., et al. (2014). Assessing the clinical utility of cancer genomic and proteomic data across tumor types. Nat. Biotechnol. 32, 644–652. doi: 10.1038/nbt.2940

8. Zhu, B., Song, N., Shen, R., Arora, A., Machiela, M. J., Song, L., et al. (2017). Integrating clinical and multiple omics data for prognostic assessment across human cancers. Sci. Rep. 7:16954. doi: 10.1038/s41598-017-17031-8

9. Zhao M, Tang Y, Kim H, Hasegawa K. (2018). Machine learning with k-means dimensional reduction for predicting survival outcomes in patients with breast cancer. Cancer Informat. 17:1176935118810215.

10. Tong, L., Mitchel, J., Chatlin, K. and Wang, M.D. (2020). Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis. BMC Medical Informatics and Decision Making, 20(1). doi:https://doi.org/10.1186/s12911-020-01225-8

11. Castellsagué, X. (2008). Natural history and epidemiology of HPV infection and cervical cancer. Gynecologic Oncology, 110(3), pp.S4–S7. doi:https://doi.org/10.1016/j.ygyno.2008.07.045.

12. Lipton, J.M., Atsidaftos, E., Zyskind, I. and Vlachos, A. (2006). Improving clinical care and elucidating the pathophysiology of Diamond Blackfan anemia: An update from the Diamond Blackfan Anemia Registry. Pediatric Blood & Cancer, 46(5), pp.558–564. doi:https://doi.org/10.1002/pbc.20642.

13. Sousa, B., Pereira, J. and Paredes, J. (2019). The Crosstalk Between Cell Adhesion and Cancer Metabolism. International Journal of Molecular Sciences, [online] 20(8). doi:https://doi.org/10.3390/ijms20081933.

14. Iozzo, R.V. and Sanderson, R.D. (2011). Proteoglycans in cancer biology, tumour microenvironment and angiogenesis. Journal of Cellular and Molecular Medicine, 15(5), pp.1013–1031. doi:https://doi.org/10.1111/j.1582-4934.2010.01236.x.

15. Seo, J. and Park, M. (2020). Molecular crosstalk between cancer and neurodegenerative diseases. Cellular and Molecular Life Sciences, [online] 77(14), pp.2659–2680. doi:https://doi.org/10.1007/s00018-019-03428-3.

16. Lopez, K., Tsuen, W., De, E., Dávila, R.G. and Shuck, S.C. (2023). Extracellular vesicles: A dive into their role in the tumor microenvironment and cancer progression. Frontiers in Cell and Developmental Biology, 11. doi:https://doi.org/10.3389/fcell.2023.1154576.

17. Costas Koufaris, Papandreou, M.E., Ellis, J.K., Nicolaidou, V. and Keun, H.C. (2023). miR-22-enriched breast cancer cells display repressed glycolytic metabolism, increased glycogen synthesis, and reduced survival in low glucose conditions. Molecular Biology Reports, [online] 50(6), pp.5185–5193. doi:https://doi.org/10.1007/s11033-023-08458-6.

18. Cui, S., Chen, Y., Guo, Y., Wang, X. and Chen, D. (2023). Hsa-miR-22-3p inhibits liver cancer cell EMT and cell migration/ invasion by indirectly regulating SPRY2. PloS One, [online] 18(2), p.e0281536. doi:https://doi.org/10.1371/journal.pone.0281536.

19. Shi, Y.-X. (2021). Identification of the molecular function of tripartite motif containing 58 in human lung cancer. Oncology Letters, 22(3). doi:https://doi.org/10.3892/ol.2021.12946.

20. Shen, N., Du, J., Zhou, H., Chen, N., Pan, Y., Hoheisel, J.D., Jiang, Z., Xiao, L., Tao, Y. and Mo, X. (2019). A Diagnostic Panel of DNA Methylation Biomarkers for Lung Adenocarcinoma. Frontiers in Oncology, 9. doi:https://doi.org/10.3389/fonc.2019.01281.

21. WANG, F., REN, X. and ZHANG, X. (2015). Role of microRNA-150 in solid tumors. Oncology Letters, 10(1), pp.11–16. doi:https://doi.org/10.3892/ol.2015.3170.

22. Sugita, B.M., Rodriguez, Y., Fonseca, A.S., Emanuelle Nunes Souza, Bhaskar Kallakury, Cavalli, I.J., Ribeiro, F., Aneja, R. and Cavalli, L.R. (2022). MiR-150-5p Overexpression in Triple-Negative Breast Cancer Contributes to the In Vitro Aggressiveness of This Breast Cancer Subtype. Cancers, [online] 14(9), pp.2156–2156. doi:https://doi.org/10.3390/cancers14092156.

23. Sun, X., Zhang, C., Cao, Y. and Liu, E. (2019). miR-150 Suppresses Tumor Growth in Melanoma Through Downregulation of MYB. Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics, 27(3), pp.317–323. doi:https://doi.org/10.3727/096504018x15228863026239.