Multi-Omics Prognostic Marker Discovery and Survival Modeling: A Case Study on Pan-Cancer Survival Analyses in Women's Cancers

Ricky Nguyen[1] and Fatemeh Vafaee[1,2,3]*

[1] School of Biotechnology and Biomolecular Sciences, University of New South Wales (UNSW Sydney), Australia

[2] UNSW Data Science Hub (uDASH), University of New South Wales (UNSW Sydney), Australia

[3] UNSW AI Institute, University of New South Wales (UNSW Sydney), Australia

* Corresponding Author

Correspondence to:

A/Professor Fatemeh Vafaee

School of Biotechnology and Biomolecular Sciences

UNSW SYDNEY NSW 2052 AUSTRALIA

T: +61 (2) 9065 2699

E: f.vafaee@unsw.edu.au

Highlights

- Introducing PRISM, a robust framework that enhances survival predictions and identifies key prognostic markers through multi-omics integration.
- Utilizing a voting-based pipeline, PRISM selects optimal single-omics features and integrates them via feature-level fusion to identify the best modality combinations.
- Highlighting Survival Random Forest as an effective model for high-dimensional data, outperforming other survival algorithms.
- Demonstrating superior survival predictions with integrated DNA methylation, miRNA, and copy number variation data in Breast Invasive Carcinoma (C-index 0.774), Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (CESC, 0.835), and Uterine Corpus Endometrial Carcinoma (UCEC, 0.766).
- Revealing common oncogenic pathways and therapeutic targets through pan-cancer analysis validated by Kaplan-Meier analysis of multi-modal signatures.

Abstract

Survival analysis plays a critical role in cancer research, offering insights into patient outcomes and informing personalized treatments. Advances in sequencing and profiling technologies have unlocked the potential of multi-omics data to improve prognostic models. However, the high dimensionality of multi-omics data often complicates analysis and limits clinical application. Additionally, most studies rely on the traditional Cox proportional hazards model, with limited exploration of alternative survival algorithms or feature selection methods. Few frameworks integrate optimal features across multiple omics modalities, leaving a gap in harnessing the full potential of multi-omics data. To address these challenges, we developed **PRISM** (PRognostic marker Identification and Survival Modelling through Multi-omics Integration), a comprehensive framework for improving survival predictions and identifying key prognostic markers. PRISM systematically evaluates feature selection methods and survival models, employing a voting-based pipeline to identify robust features from single-omics data. These features are integrated using feature-level fusion to determine the best modality combinations for survival prediction. Applied to The Cancer Genome Atlas (TCGA) data for Breast Invasive Carcinoma (BRCA), Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (CESC), and Uterine Corpus Endometrial Carcinoma (UCEC), PRISM demonstrated that integrating DNA methylation, miRNA, and copy number variation data significantly outperforms single-omics approaches (BRCA: C-index 0.774; CESC: 0.835; UCEC: 0.766). Pan-cancer analysis further demonstrated the biological relevance of these multi-modal signatures, revealing shared oncogenic pathways and potential therapeutic targets. By reducing data dimensionality while preserving predictive power, PRISM offers a scalable and general-purpose solution for integrating multi-omics data, advancing cancer research and enabling precision medicine.

Keywords

Multi-omics; TCGA; Pan-cancer; Prognostic marker discovery; Survival Analysis; Machine Learning; Women's Cancer

Introduction

Cancer exhibits complex pathophysiology influenced by diverse genetic, environmental, and molecular factors. Despite significant advances in cancer research, patient outcomes vary

considerably even within the same cancer types, complicating treatment strategies [1,2]. High-throughput molecular profiling technologies, such as next-generation sequencing and mass spectrometry, enable comprehensive analysis of molecules like DNA, RNA, and proteins, facilitating unbiased biomarker discovery. These omics technologies have become a cornerstone of precision medicine. However, relying solely on individual omics data provides only partial insights into the intricate mechanisms underlying the disease, potentially overlooking critical biomarkers and valuable therapeutic opportunities [3,4].

The inherent heterogeneity of cancer, reflected in its diverse subtypes and molecular profiles, necessitates a more integrated approach. Combining multiple omics data types is essential to achieving a holistic understanding of cancer biology and paving the way for personalized treatment strategies tailored to each patient's unique condition [5]. The wealth of omics data—including genomics, transcriptomics, proteomics, metabolomics, and epigenomics—offers a comprehensive view of cancer biology, providing unprecedented opportunities to uncover novel biomarkers and improve clinical outcomes [6,7].

However, multi-omics data presents significant challenges, including high dimensionality, data imbalance, noise, missing values, and heterogeneity, all of which complicate biomarker discovery and hinder robust analysis [5]. To overcome these obstacles, integrative machine learning approaches such as multimodal learning, ensemble strategies, and network-based methods have emerged as essential tools to advance biomarker discovery and the development of clinical outcome models, supporting applications ranging from disease diagnosis to prognosis prediction and treatment response monitoring [8-12].

In this context, survival analysis plays a pivotal role, helping researchers and clinicians assess the factors influencing patient outcomes over time. Predicting survival beyond critical milestones, (e.g., five-year survival) informs treatment planning and also deepens our understanding of the biological mechanisms underlying cancer progression [3]. The importance of multi-omics survival modelling has been widely recognized, leading to multiple studies over the past decade.

Yuan et al. [13] conducted a pioneering study by integrating multiple omics data—including somatic copy-number alterations (SCNA), DNA methylation, mRNA, miRNA, and protein expression—to predict patient survival across four cancer types from The Cancer Genome Atlas (TCGA) project. This work established the value of multi-omics integration for survival analysis. Building on this, Zhu et al. [14] systematically evaluated the prognostic value of various omics types —clinical data, mRNA, SCNA, DNA methylation, miRNA—across 14 cancers using a kernel machine learning approach. They found gene and miR1NA expression to have the strongest predictive power. Liu et al. [15] focused on multi-omics variable selection combined with Cox-regression model for cancer prognosis prediction. Subsequently, Chai et al. [2] integrated multi-omics data to derive representative features, which were then utilized in a Cox proportional hazards (Cox-PH) model for enhanced survival prediction.

Later, Chaudhary et al. [16] employed an autoencoder to integrate DNA methylation, miRNA expression, and RNA-Seq data, selecting survival-associated features using univariate Cox-PH analysis and identifying robust subgroups in hepatocellular carcinoma (HCC). Similarly, Cheerla et al. [17] utilized a deep learning approach with multi-modal representations to enhance cancer prognosis and Cox-PH for outcome prediction. Hao et al. [18] extended deep learning methodologies by proposing a gene- and pathway-based deep neural network (DNN) where the Cox-PH model served as the output layer, embedding multi-omics features for survival prediction. Tong et al. [19] further advanced the field by introducing multi-view learning with autoencoders, integrating multi-omics data through complementary and consensus

principles. Their approach demonstrated how DNA methylation and miRNA expression offer both unique and shared prognostic insights, with the integrated model significantly outperforming single-omics methods.

While these studies underscore the value of integrating multi-omics data in cancer survival modelling, a significant gap remains in systematically evaluating survival algorithms beyond the Cox proportional hazards model. Even fewer studies have examined feature selection methods in conjunction with diverse survival models, and there is a notable lack of a framework that consolidates the best-performing features from individual omics layers into a unified predictive approach. To address these gaps, we developed **PRISM** (PRognostic marker Identification and Survival Modelling through Multi-omics Integration), a comprehensive and versatile framework designed to enhance survival predictions and identify key prognostic signatures. PRISM performs an in-depth analysis of multi-omics data from The Cancer Genome Atlas (TCGA), incorporating gene expression, DNA methylation, miRNA expression, and copy number variations.

The PRISM pipeline begins with data retrieval from TCGA, followed by rigorous preprocessing to ensure quality and consistency across diverse data types. It then applies multiple feature selection methods to each single-omics dataset, identifying the most relevant features for survival analysis. An ensemble voting strategy is used to compile a robust signature panel. Subsequently, machine learning algorithms are employed to develop survival prediction models based on the selected features. To further refine its predictions, PRISM integrates features from different omics layers using two distinct feature-level fusion techniques. The performance of these multi-omics integration strategies is systematically evaluated to identify the most effective combinations of modalities for accurate survival prediction.

To demonstrate its utility, PRISM is applied to women-related cancers from TCGA, including breast (BRCA), ovarian (OV), cervical (CESC), and endometrial (UCEC) cancers. Together, these cancers significantly contribute to cancer incidence and mortality among women worldwide [20,21], emphasizing the clinical relevance of such analyses. By repeating this process across these cancer types, PRISM identified multi-omics signatures, which are subsequently explored through pan-cancer pathway and survival analyses. Despite their distinct molecular profiles and clinical presentations, cancers share overlapping pathways and biological mechanisms that influence progression and therapeutic response [22,23]. Pan-cancer analyses of women's cancers provide insights into shared oncogenic drivers, molecular subtypes, and therapeutic vulnerabilities. This understanding is important for developing broad-spectrum interventions and identifying biomarkers that transcend individual cancer types, ultimately improving patient outcomes across multiple malignancies [24]. While women-related cancers serve as a case study to showcase its capabilities, PRISM is a comprehensive and general-purpose framework that can be applied to any cancer type, making it a valuable tool for advancing precision oncology.

Methods

Omics Modalities and Cancers

Multi-omics data were obtained from The Cancer Genome Atlas (TCGA) (https://portal.gdc.cancer.gov/) 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 1. *Overview of the TCGA sample counts for each omics data type across various cancer types. The "Clinical" column represents samples with available survival data, while "Common" indicates patient samples with complete data across all five omics types*

TCGA Cancer Types	Gene Expression (GE)	Copy Number Variation (CNV)	DNA Methylation (DM)	MiRNA expression (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

For each omics modality, feature extraction was conducted using bioinformatics pipelines provided by TCGA. The pre-processed multi-omics tabular data were retrieved from the GDC Portal via the TCGAbiolink R package. 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 [19]. 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).

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. 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. 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) [1].

Feature Selection Filter Methods

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. 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. Appendix B 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.

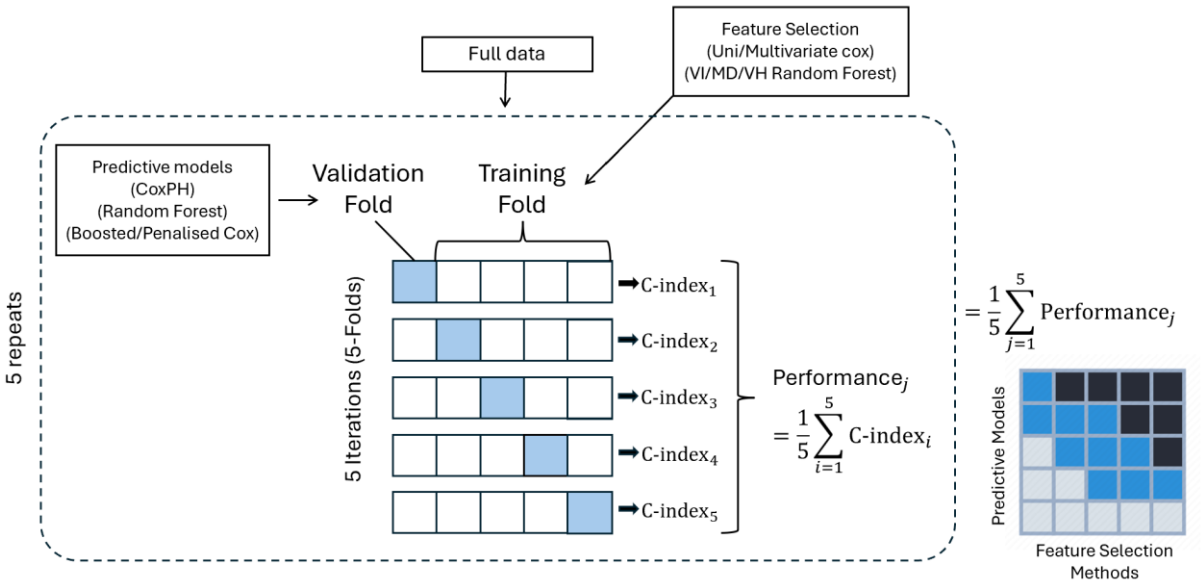


Figure 1. *Evaluation of Individual Filter Methods Pipeline. Workflow for evaluating each filter method independently. Cross-validation is employed to assess the performance of each method, and the results are summarized in a heatmap displaying the mean performance of each filter method when paired with various predictive models*

Figure 1 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 [1]. 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.

Feature Selection Pipelines

We developed two feature selection pipelines, one that utilizes cross-validation and another that uses bootstrapping.

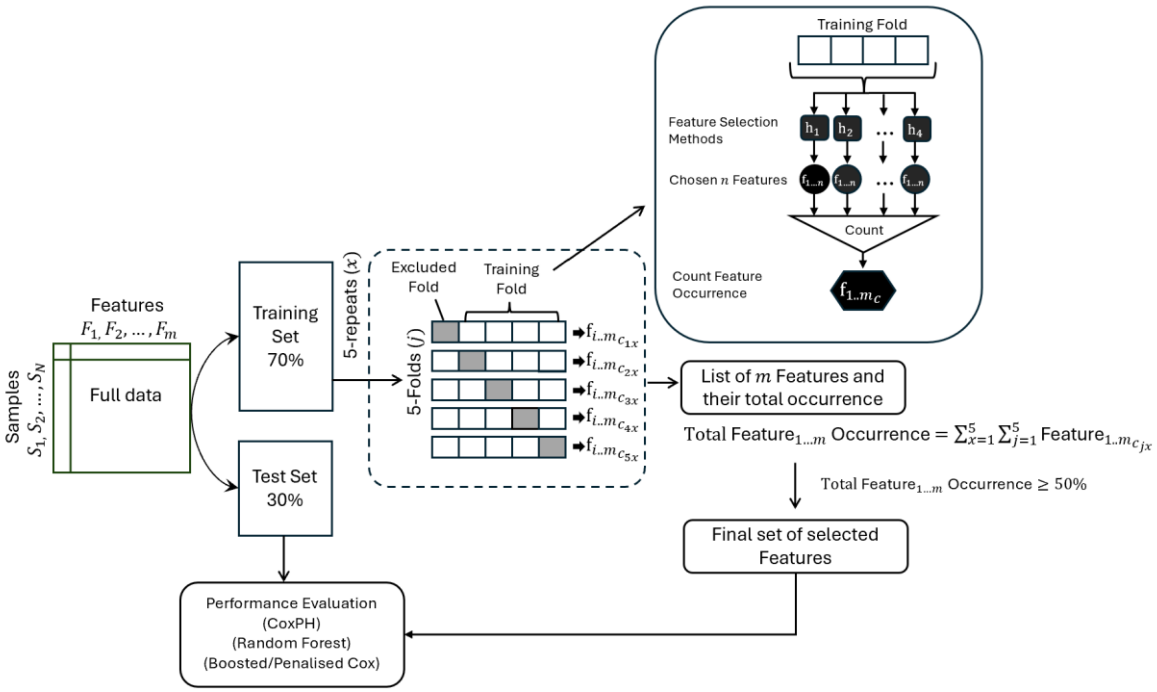


Figure 2. *CV Feature selection pipeline. Applies a cross-validation process combined with a voting technique to identify optimal features. Selected features are determined based on occurrence across the chosen filter methods, enhancing the robustness of feature selection*

Figure 2 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.

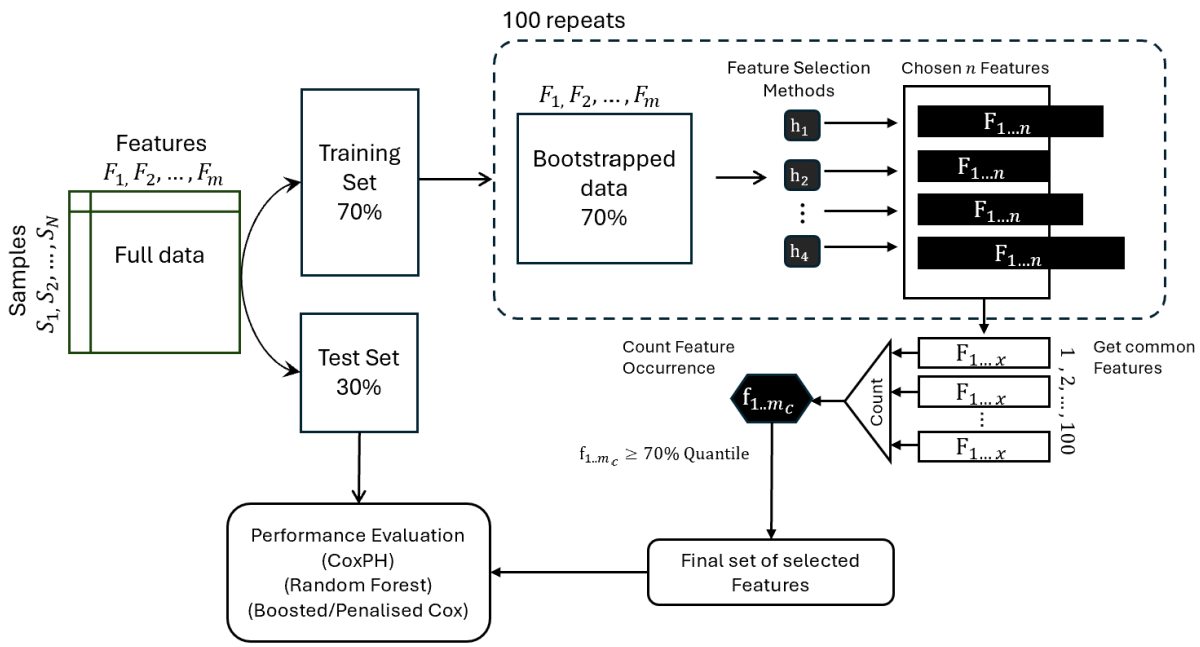


Figure 3. *Bootstrapping Feature selection pipeline. Employs bootstrapping to improve the reliability of selected features by resampling the data across numerous iterations and applying the same voting technique*

The second feature selection pipeline method developed is the bootstrapping pipeline, as shown in Figure 3. 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.

To evaluate the performance of our feature selection pipelines, we tested 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. 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.

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 mid 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. Mid fusion, on the other hand, involves training separate models for each omics modality using their respective selected features. The filtered features from these individual models are then combined and trained together for a final output [5,6]. By utilizing both early and mid-fusion techniques, we aim to determine which method provides the best performance for multi-omics integration in survival analysis.

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) [19], 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.

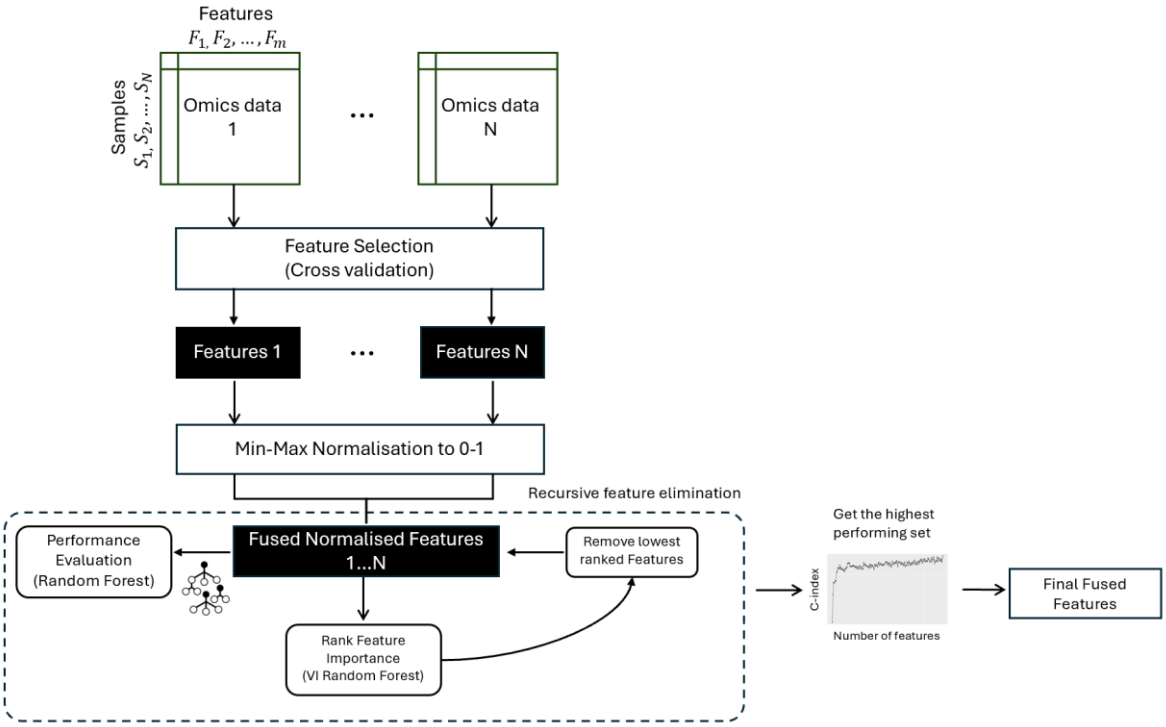


Figure 4. *Early fusion integration pipeline, where two modalities are combined at the feature level. Recursive Feature Elimination (RFE) is then applied to reduce dimensionality, further refining the feature set to enhance model performance*

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.

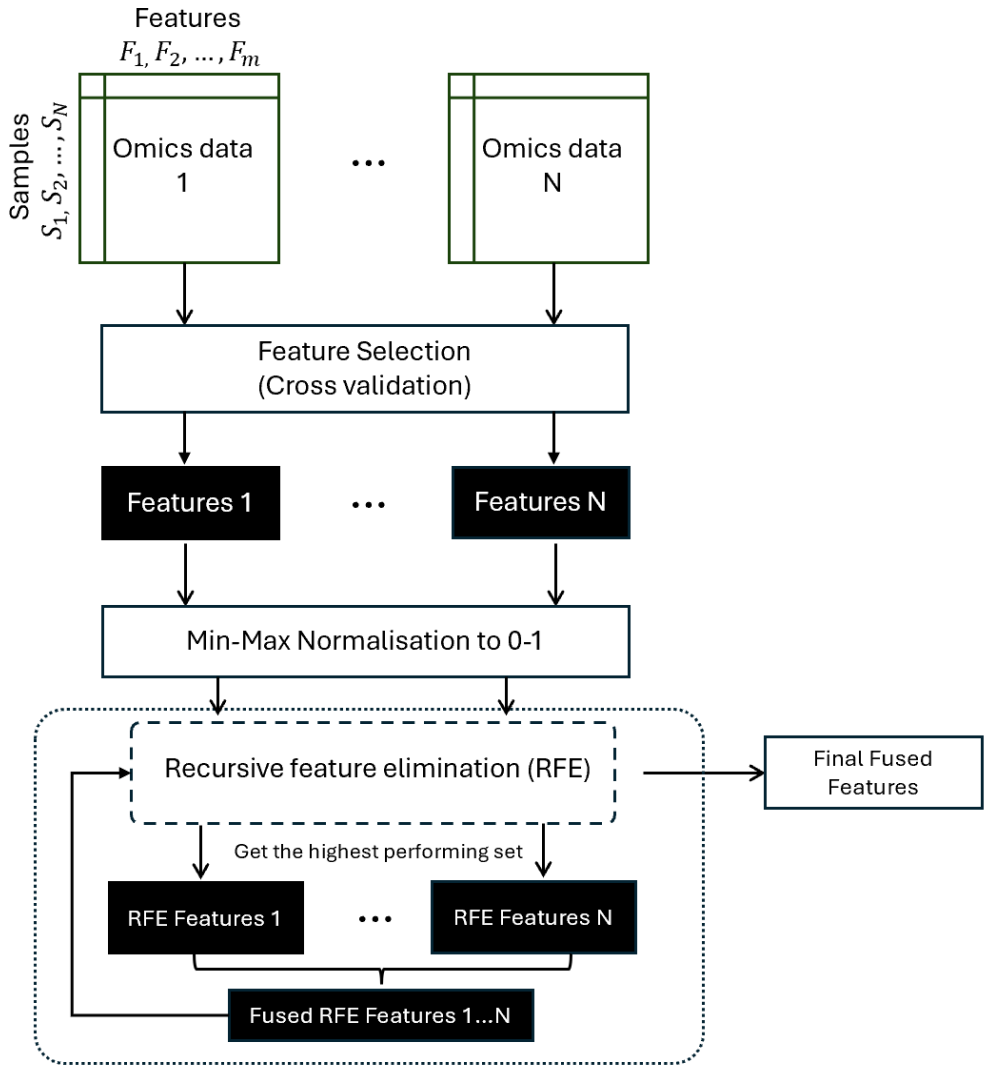


Figure 5. *Mid fusion integration pipeline, where individual modalities first undergo Recursive Feature Elimination (RFE) to reduce dimensionality. The reduced feature sets are then fused together, followed by an additional round of RFE on the fused data to obtain the final set of multi-omics features*

Our second integration method is Mid Fusion, illustrated in Figure 5. 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.

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 analysing miRNA 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 important but also provide valuable biological insights into the commonalities and differences across the cancers under study.

Results

Evaluation of Individual Filter Methods

Before developing our feature selection pipeline, we examined each of the filter methods that we will be utilizing (Figure 1). Due to computational complexity and resources, we decided to opt for only four filter methods for the purposes of this study. The results of our experiments are presented in 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.

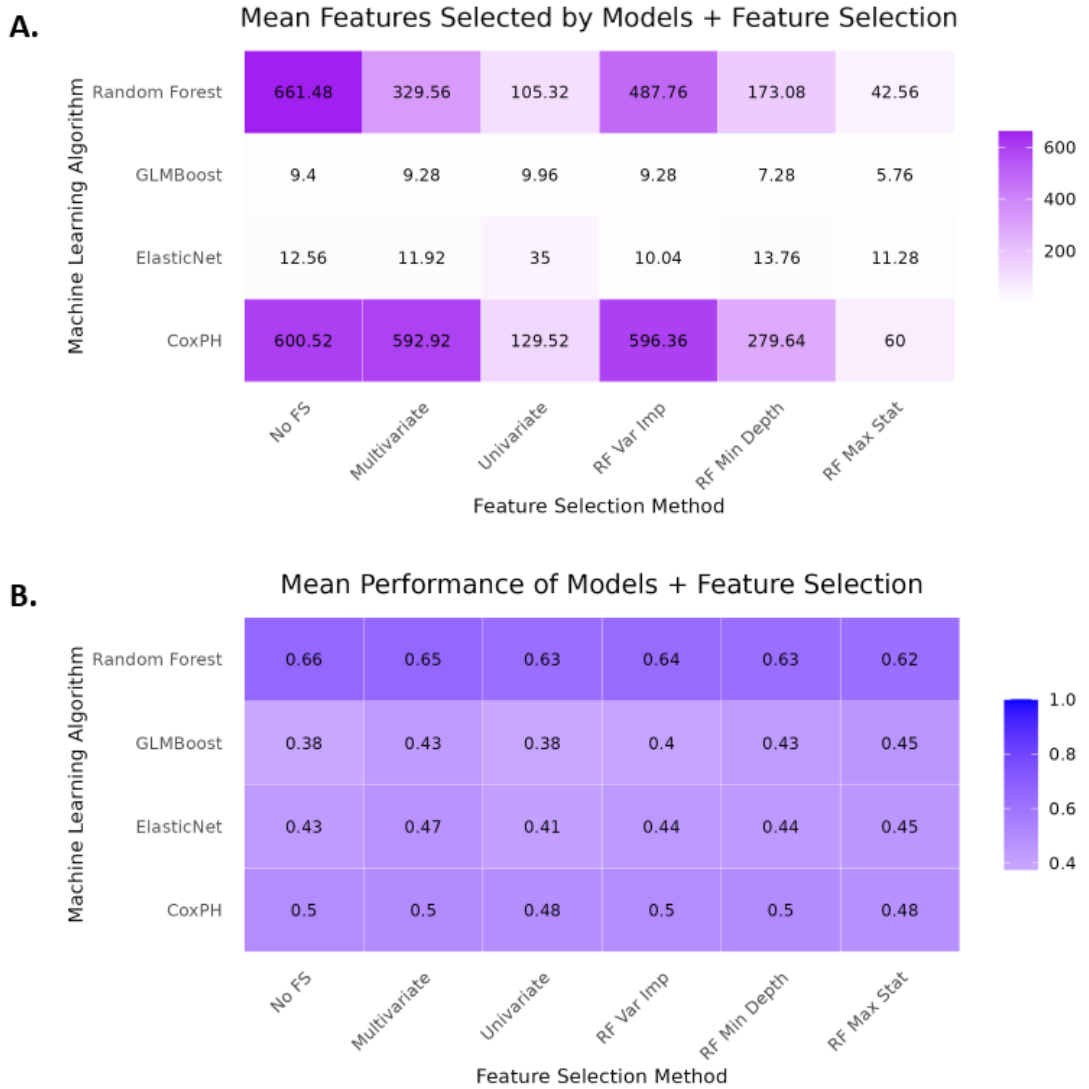


Figure 6. *Heatmap of both features selected from each model and filter methods, using BRCA DM data. Mean number of features selected* ***(a),*** *mean c-index values* ***(b)***

Figure 6 illustrates the results using the BRCA DNA Methylation (DM) dataset. However, the trends observed in this example are consistent with other modalities and cancer types (Appendix C), 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 **(a)**, 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.

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. Table 2 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 2. *Performance of the feature selection pipeline applied to the BRCA dataset. Results include comparisons to a baseline model (denoted as "None"), where no feature selection is applied*

BRCA	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	0.400	0.493
	DM	0.430	0.500	0.517
	CNV	0.425	0.426	0.435

Appendix D 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.

Early and Mid Fusion

For our multi-omics integration, we applied two techniques: early and mid-fusion, as seen in Figures 4 and 5.

Table 3. *Comparison between Mid and Early Fusion using CESC dataset*

Modality combinations	**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 of Early Fusion	0.728	0.740	0.653	0.608	0.672	0.523	**0.730**	0.713	0.718	0.673	0.653
C-index of Mid Fusion	0.654	0.640	0.699	0.738	0.632	0.718	0.822	0.737	**0.835**	0.693	0.720

In Table 3, we illustrate the performance comparison between Mid and Early Fusion using the CESC dataset. This trend is consistent across all cancer types tested, not just CESC (Appendix E). Our results show that Mid Fusion consistently outperforms Early Fusion. For example, in the CESC dataset, Early Fusion achieved a c-index of 0.730, while Mid 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, Mid Fusion identified the combination of DM, ME, and CNV as superior to the DM, GE, and ME combination preferred by Early Fusion. Overall, Mid Fusion demonstrated better performance across all cancer datasets, leading us to select it as our primary integration strategy.

Multi-omics results

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

Table 4. *Multi-omics results for BRCA*

Modality combinations	**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 of Mid Fusion	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 4 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 5. *Multi-omics results for OV*

Modality combinations	**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 of Mid Fusion	0.673	**0.683**	0.646	0.593	0.559	0.616	0.666	0.585	0.630	0.633	0.637
Omics	**Features**										
DM	cg04533291, cg03874199, cg27342801, cg20676475, cg06101324										
ME	hsa.mir.1301, hsa.mir.3200, hsa.let.7a.1, hsa.let.7a.3, hsa.mir.135b, hsa.mir.139										

Table 5 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 6. *Multi-omics results for CESC*

Modality combinations	**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 of Mid Fusion	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 6 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 7. *Multi-omics results for UCEC*

Modality combinations	**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 of Mid Fusion	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	cg11009736, cg12177677, cg17525406, cg22751696, cg06275635, cg13968061, cg19884600, cg10849854, cg16607065, cg06096336, cg14419187, cg22605643, cg04391111, cg12417466, cg26353877, cg05368341, cg14458615, cg04660410, cg23477967, cg15175266, cg08763351, cg08804892, cg13901526, cg14991487, cg09478478, cg06490988, cg10885338, cg11657808, cg21376883, cg02192965, cg25596297, cg18277754, cg25115460, cg10523019, cg25856383, cg06856528, cg12105450, cg01426743, cg14696870, cg22338307, cg07548313
ME	hsa.mir.149, hsa.mir.222, hsa.mir.598, hsa.mir.335, hsa.mir.150, hsa.mir.1224, hsa.let.7a.3, hsa.mir.2355, hsa.mir.31, hsa.mir.146a, hsa.let.7a.1, hsa.mir.193b, hsa.mir.142, hsa.mir.140, hsa.mir.187, hsa.mir.155, hsa.mir.3131, hsa.mir.3614, hsa.mir.1976, hsa.mir.1468, hsa.mir.152, hsa.mir.1249, hsa.mir.143, hsa.let.7a.2, hsa.mir.181b.1, hsa.mir.374a, hsa.mir.1266, hsa.mir.153.2, hsa.let.7f.2, hsa.mir.127, hsa.mir.144, hsa.mir.29a, hsa.mir.194.2, hsa.mir.30e, hsa.mir.29b.2, hsa.let.7f.1, hsa.mir.10b, hsa.mir.181b.2, hsa.mir.22, hsa.mir.16.1, hsa.mir.203b, hsa.mir.191, hsa.mir.1287, hsa.let.7g, hsa.mir.192, hsa.mir.1226, hsa.mir.28, hsa.mir.30c.1, hsa.mir.25, hsa.mir.10a, hsa.mir.301b, hsa.mir.135b, hsa.mir.27b, hsa.mir.133a.1, hsa.mir.133b, hsa.let.7d, hsa.mir.1343, hsa.mir.214, hsa.mir.199a.2, hsa.mir.125b.1, hsa.mir.29b.1, hsa.mir.194.1, hsa.mir.181d, hsa.mir.129.1, hsa.mir.130a, hsa.mir.30a, hsa.mir.20b, hsa.mir.199a.1, hsa.mir.133a.2, hsa.mir.196a.1, hsa.mir.129.2, hsa.mir.223, hsa.mir.148a, hsa.mir.125b.2
CNV	MIR1204, Y_RNA, AL499627.1, AC104653.2, AC017074.1, TERC, RSL24D1P6, AC022634.1, MIR8083, U3, AC104640.1, RNVU1.4, AC022296.2, Metazoa_SRP

Table 7 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.

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 8. *Directly 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 8 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. No feature appeared in all four cancers, with only the miRNA signature mir.150 being common to three cancers: BRCA, CESC, and UCEC.

MiRNA Features and 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.

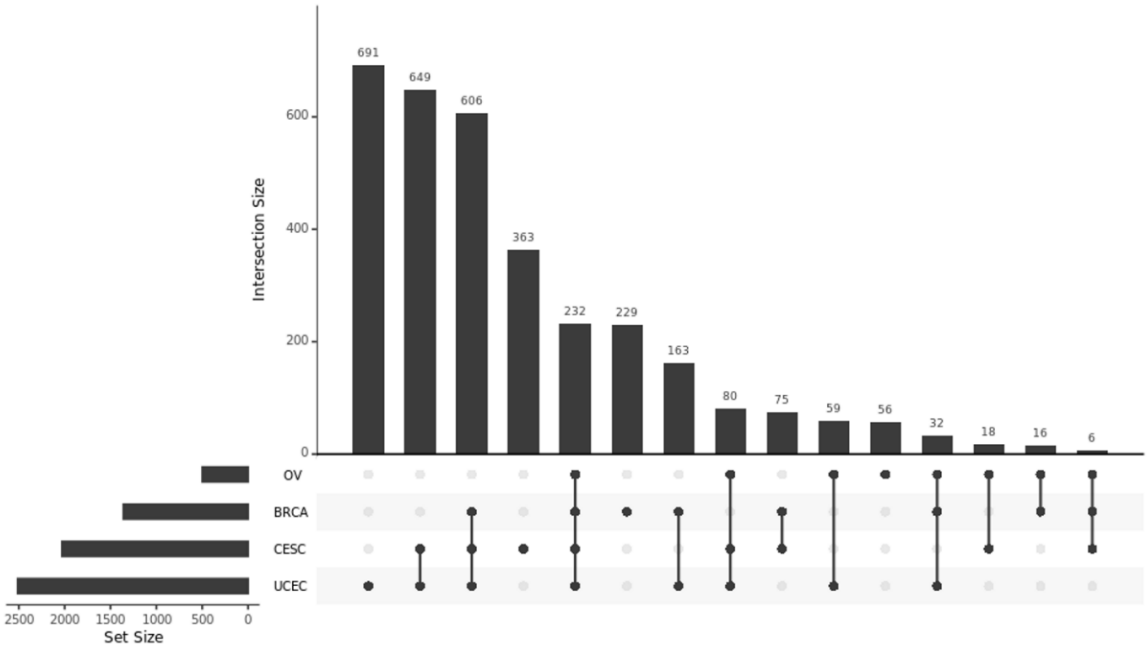


Figure 7. *Upset plot showing direct overlapping miRNA Targets between cancers*

We identified 232 gene targets shared across all four cancers. Notably, UCEC contributed the most gene targets, while OV provided the least.

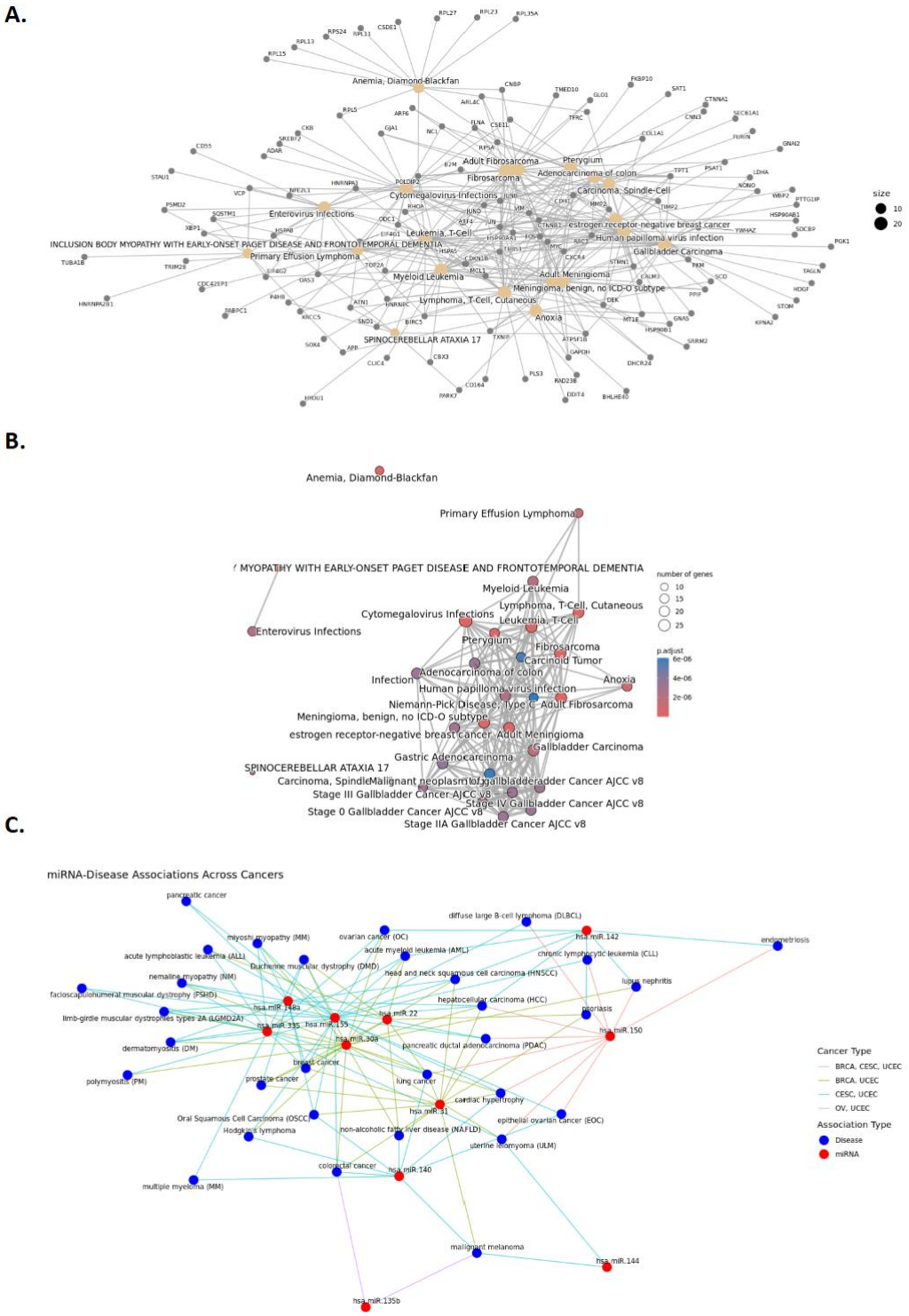


Figure 8. *Disease network for gene targets and miRNA features.* ***(a)*** *Disease gene network of overlapped gene targets. Using the overlapping 232 gene targets, we performed enrichment analysis with DisGeNET, a database of gene-disease associations.* ***(b)*** *Enrichment Map of gene disease network.* ***(c)*** *MiRNA Disease associations across cancers*

As shown in Figure 8(a), 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 [25].

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 [26], 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.

Aside from examining the gene-disease network, Figure 8(b) 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.

In this analysis, Figure 8(c), 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. Majority 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

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.

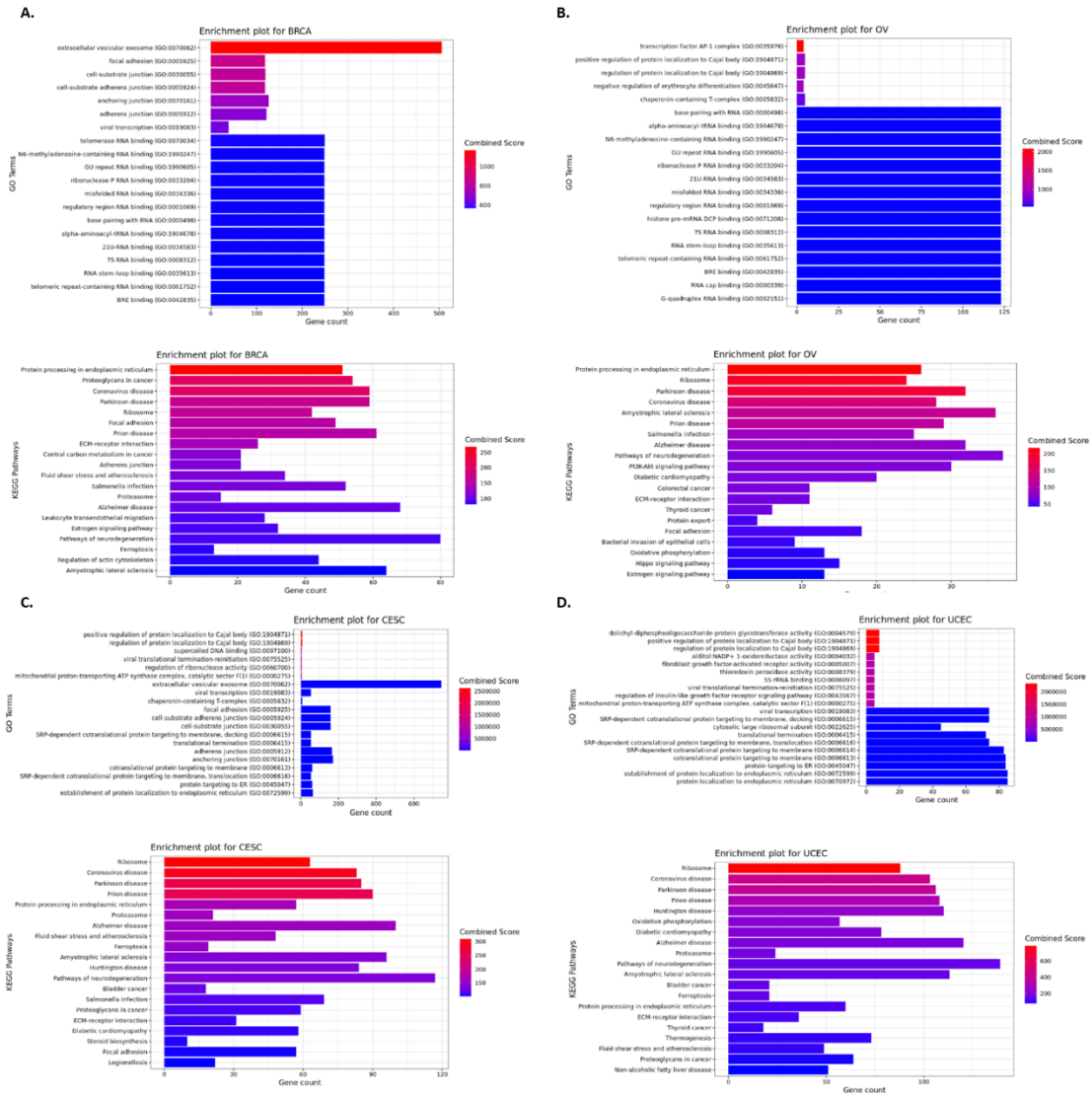


Figure 9. *GSEA enrichment plots of Top 20 GO Terms and KEGG Pathways.* ***(a)*** *BRCA,* ***(b)*** *OV,* ***(c)*** *CSEC and* ***(d)*** *UCEC*

Figure 9(a) depicts the enrichment plot for BRCA miRNA targets. The enrichment plots are ranked by their combined score, calculated using the adjusted p-value and z-score to ensure robust significance of the gene targets within their associated GO Terms or KEGG Pathways. The x-axis represents the number of gene targets associated with each GO Term or KEGG Pathway. Notable GO Terms include extracellular vesicular transport, cell adhesion, and RNA binding/processing, which are key processes implicated in tumor progression. For instance, miRNAs regulating extracellular vesicular transport may influence cancer cell communication, promoting tumor growth and metastasis. Similarly, dysregulated cell adhesion—a hallmark of cancer—can enhance invasiveness [27]. miRNAs involved in RNA binding and processing impact post-transcriptional gene regulation, affecting the expression of tumor suppressor or oncogenic genes. The KEGG enrichment plot highlights significant pathways such as protein processing in the endoplasmic reticulum and proteoglycans in cancer. Proteoglycans play critical roles in cell adhesion, signaling, and interactions within the tumor microenvironment, with dysregulation contributing to cancer cell proliferation, invasion, and metastasis [28]. Intriguingly, pathways related to infectious diseases (e.g., Coronavirus, Salmonella infection, prion diseases) also emerge, suggesting potential links between cancer, immune responses, and infectious diseases. Additionally, neurodegenerative disease pathways, known to

intersect with cancer biology [29], further underscore the complexity of these interactions. Figure 9(b) presents enrichment results for OV miRNA targets. GO Terms emphasize RNA-related processes, including RNA motif binding, regulation of protein complexes, and broader biological functions critical to cancer initiation and progression [30]. The KEGG plot, consistent with BRCA, identifies involvement in protein processing pathways, infectious disease pathways, and neurodegenerative disease pathways. Additionally, the OV targets are implicated in multiple cancer-related and signaling pathways, highlighting their potential roles in tumor progression. Figure 9(c) displays the enrichment results for CESC miRNA targets. Significant GO Terms suggest key roles in cellular processes such as protein localization, DNA/RNA interactions, cellular junctions, and organelle function. Dysregulation in these processes, including abnormal protein localization, is strongly linked to tumorigenesis and cancer progression [31]. KEGG Pathways similarly reveal trends in fundamental cellular processes, cancer-associated pathways, and disease-related pathways. Noteworthy pathways, such as proteoglycans in cancer and focal adhesion, underscore the significance of extracellular matrix interactions and cell signaling in tumor growth and metastasis. Figure 9(d) presents the results for UCEC miRNA targets. GO Terms highlight roles in protein regulation, RNA metabolism, cellular signaling, and membrane dynamics, all of which are critical to tumorigenesis and cancer progression [30,31]. KEGG Pathways show a consistent trend across all cancers, involving fundamental cellular processes, cancer-related pathways, and disease-related pathways such as neurodegenerative disorders and infectious diseases. Unique to UCEC, metabolic pathways and those linked to metabolic diseases are prominently enriched, suggesting a potential role for metabolic dysregulation in this cancer type.

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 analyse 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.

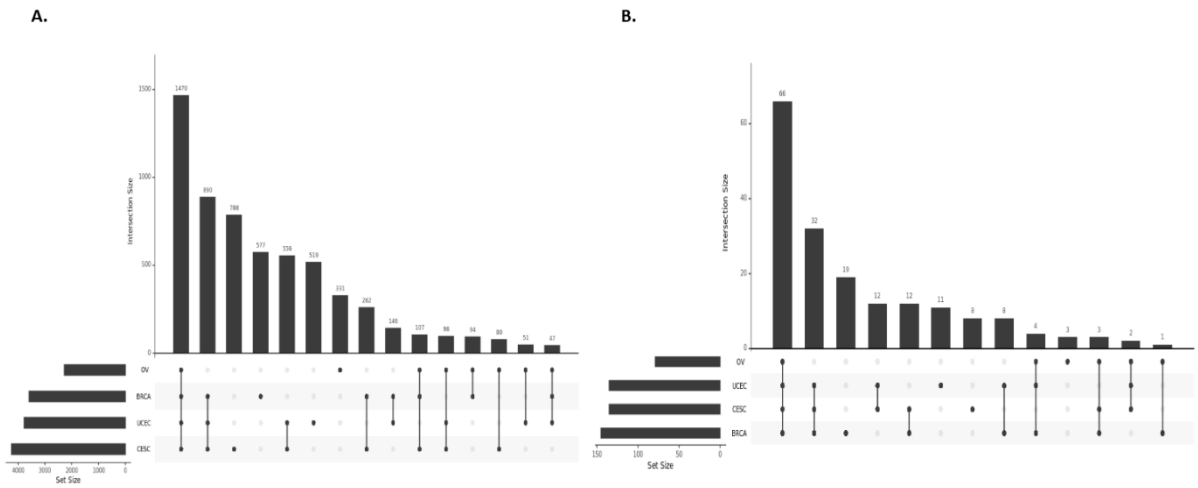


Figure 10. *Upset plot showing GSEA overlaps across all cancers for miRNA targets. GO Terms* ***(a)****, KEGG Pathways* ***(b)***

Figure 10 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 analysed. 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.

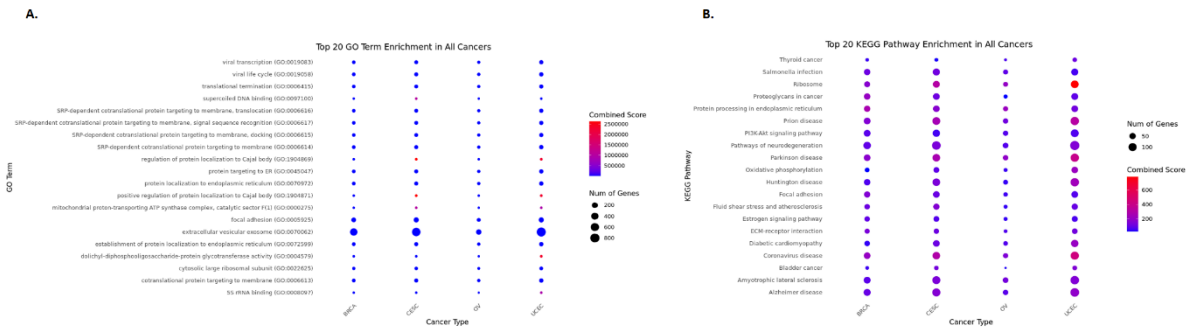


Figure 11. *Enrichment dot plot showing TOP 20 GO Terms* ***(a)*** *and KEGG Pathways* ***(b)*** *shared in all four cancers*

Figure 11(a) 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 [30]. These findings provide valuable insights into crucial processes for cancer cell survival, growth, and metastasis, potentially guiding the identification of therapeutic targets.

Figure 11(b) 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 [32]. 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.

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 analysis helps confirming the biological significance of the overlapping features and their potential impact on cancer prognosis.

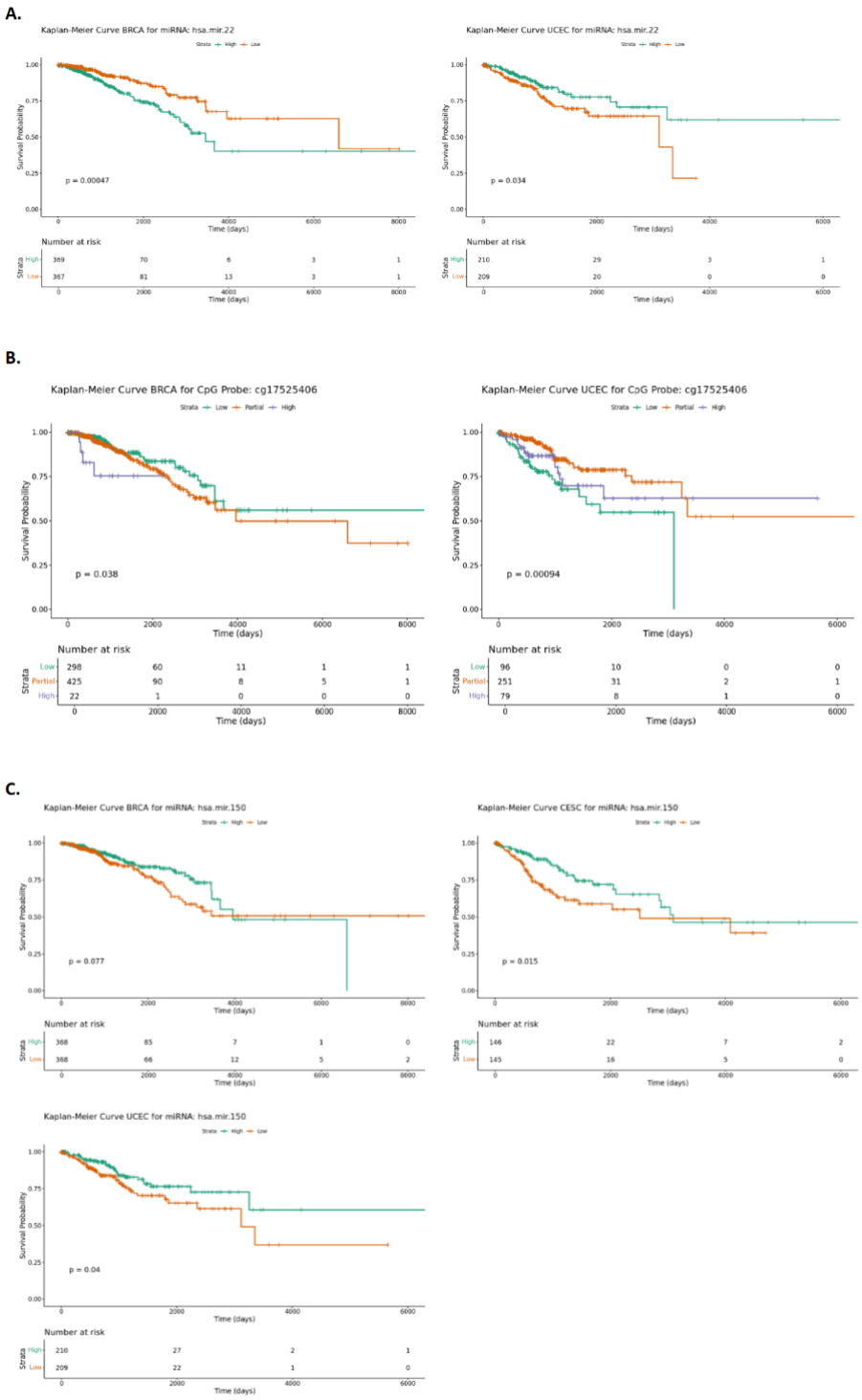


Figure 12. *KM Survival plots* ***(a)*** *miR-22,* ***(b)*** *cg17525406,* ***(c)*** *miR-150*

We first explored miR-22, which from our analysis, was present in both UCEC and BRCA (Table 8). Koufaris et al. [33] 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. [34] 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.

Figure 12(a) 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.

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 [35], [36]. In Figure 12(b), 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.

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 [37], Sugita et al. [38], and Sun et al. [39], all of which have shown that miR-150 is frequently deregulated in cancers. In Figure 12(c), 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.

Discussion

In this study, we developed PRISM to pursue 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, and 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. [1] conducted research on the identification of pan-cancer prognostic biomarkers through the integration of multi-omics data. Tong et al. [19] 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. [3] comprehensive list of existing methods to inform our exploration and analysis.

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 mid 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 study, primarily utilized the baseline CPH model. Our work demonstrated that Survival Random Forest provides a superior concordance index, regardless of the feature selection methods used. This improvement is clear when comparing our study to Tong et al. (2020) [19], 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 miRNA expression and DNA methylation.

Even without considering our highest-performing combination, our integration of DNA methylation and miRNA expression 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 may indicate 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 miRNA expression, DNA methylation, 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 analysed 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. Although our case study focused on women-related cancers, PRISM is a general-purpose framework that can be applied to other cancer types. Exploring its performance and modality combinations across additional cancers, along with their respective pan-cancer analyses, would provide valuable insights into its broader applicability and the identification of shared oncogenic pathways.

Despite the positive outcomes of our analysis, several areas require improvement before expanding upon this study. Firstly, while the study utilized pre-processing techniques similar to those employed by Zhao et al. [1] and Tong et al. [19], we found that 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.

Conclusion

We successfully achieved our goals through PRISM 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 DNA methylation, miRNA expression, 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, PRISM offers a robust general-purpose 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.

Appendix A. Feature and sample reduction after preprocessing

The results of the preprocessing steps are presented in Table A.9 to Table A.12.

Table A.9. Comparison of the BRCA dataset, showing the number of features and samples across all modalities before and after the preprocessing steps.

BRCA	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	737	737

Table A.10. Comparison of preprocessing steps for OV dataset

OV	Gene Expression (GE)	Copy Number Variation (CNV)	DNA Methylation (DM)	MiRNA expression (ME)
Features (Original)	60660	60623	485577	1881
Features (Filtered)	3033	2719	1235	569
Samples (Original)	421	557	582	490
Samples (Filtered)	395	395	395	395

Table A.11. Comparison of preprocessing steps for CESC dataset

CESC	Gene Expression (GE)	Copy Number Variation (CNV)	DNA Methylation (DM)	MiRNA expression (ME)
Features (Original)	60660	60623	485577	1881
Features (Filtered)	3033	2719	1217	417
Samples (Original)	304	294	307	307
Samples (Filtered)	291	291	291	291

Table A.12. Comparison of preprocessing steps for UCEC dataset

UCEC	Gene Expression (GE)	Copy Number Variation (CNV)	DNA Methylation (DM)	MiRNA expression (ME)
Features (Original)	60660	60623	485577	1881
Features (Filtered)	3033	2703	1217	457
Samples (Original)	553	536	438	545
Samples (Filtered)	420	420	420	420

Appendix B. Overview of Methods, Hyperparameters, and R Packages Used in the Study

Table B.13. provides a comprehensive summary of the methods applied in our study, including each method's key hyperparameters and the corresponding R package utilized for implementation.

Method	**R package**	**Function**	**Hyper-parameters & 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)			
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

Appendix C. Mean Feature and c-index for BRCA dataset

The results of the other modalities heatmaps are presented in Figure C.13 to Figure C.15.

Figure C.13. Heatmap of both features selected from each model and filter methods, using BRCA GE data. Mean number of features selected **(a),** Mean c-index values **(b)**

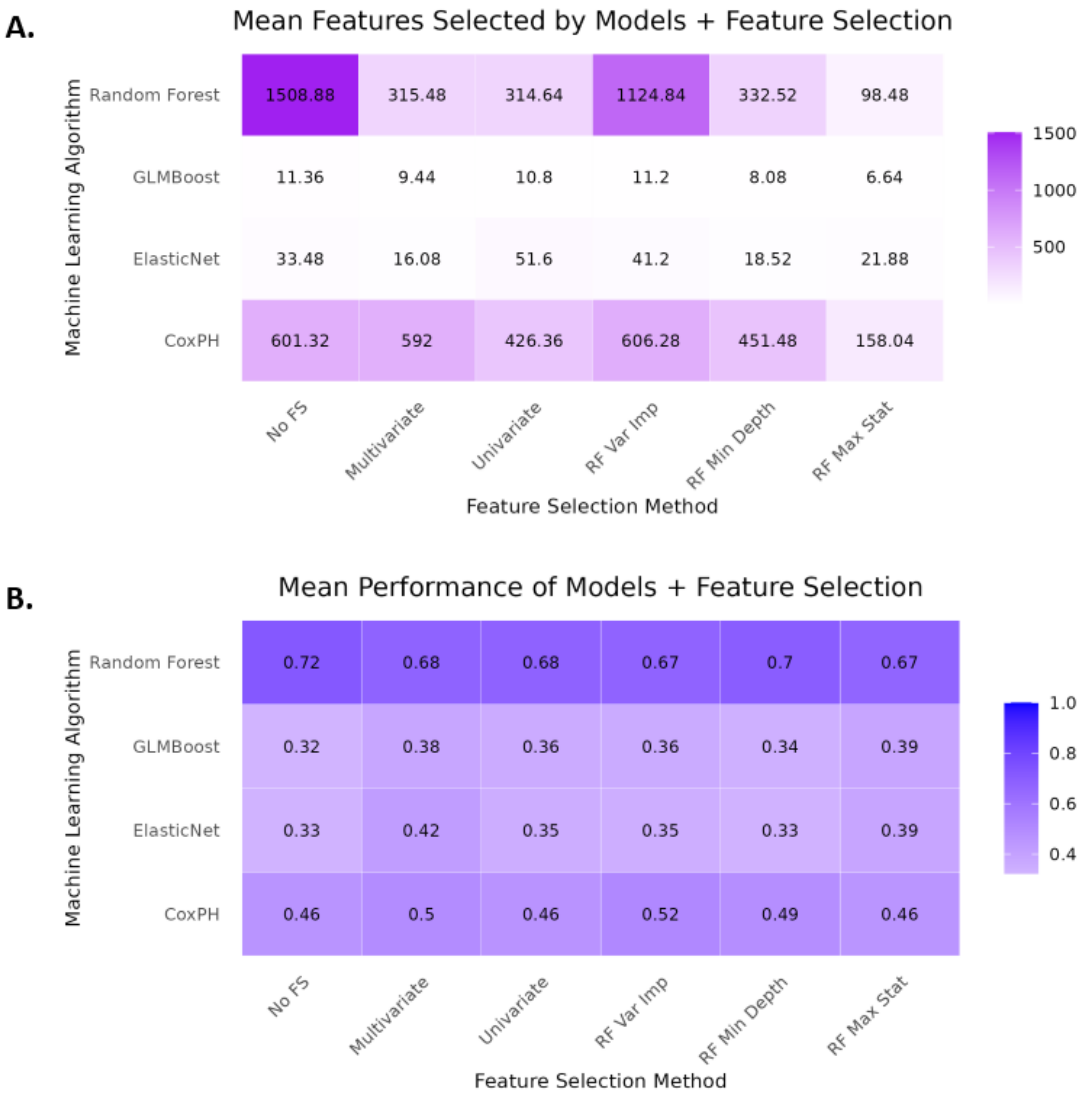


Figure C.14. Heatmap of both features selected from each model and filter methods, using BRCA ME data. Mean number of features selected **(a),** Mean c-index values **(b)**

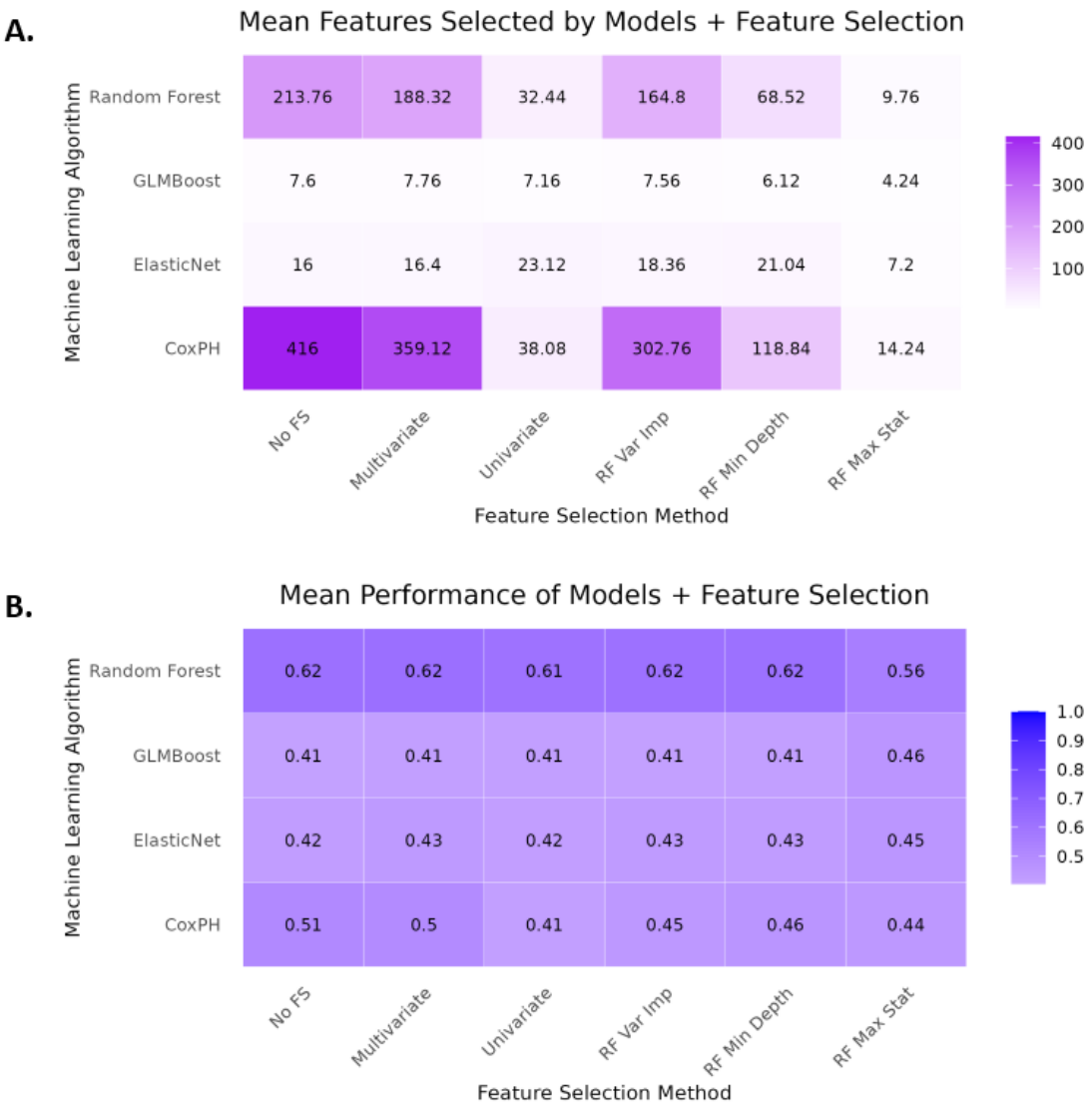
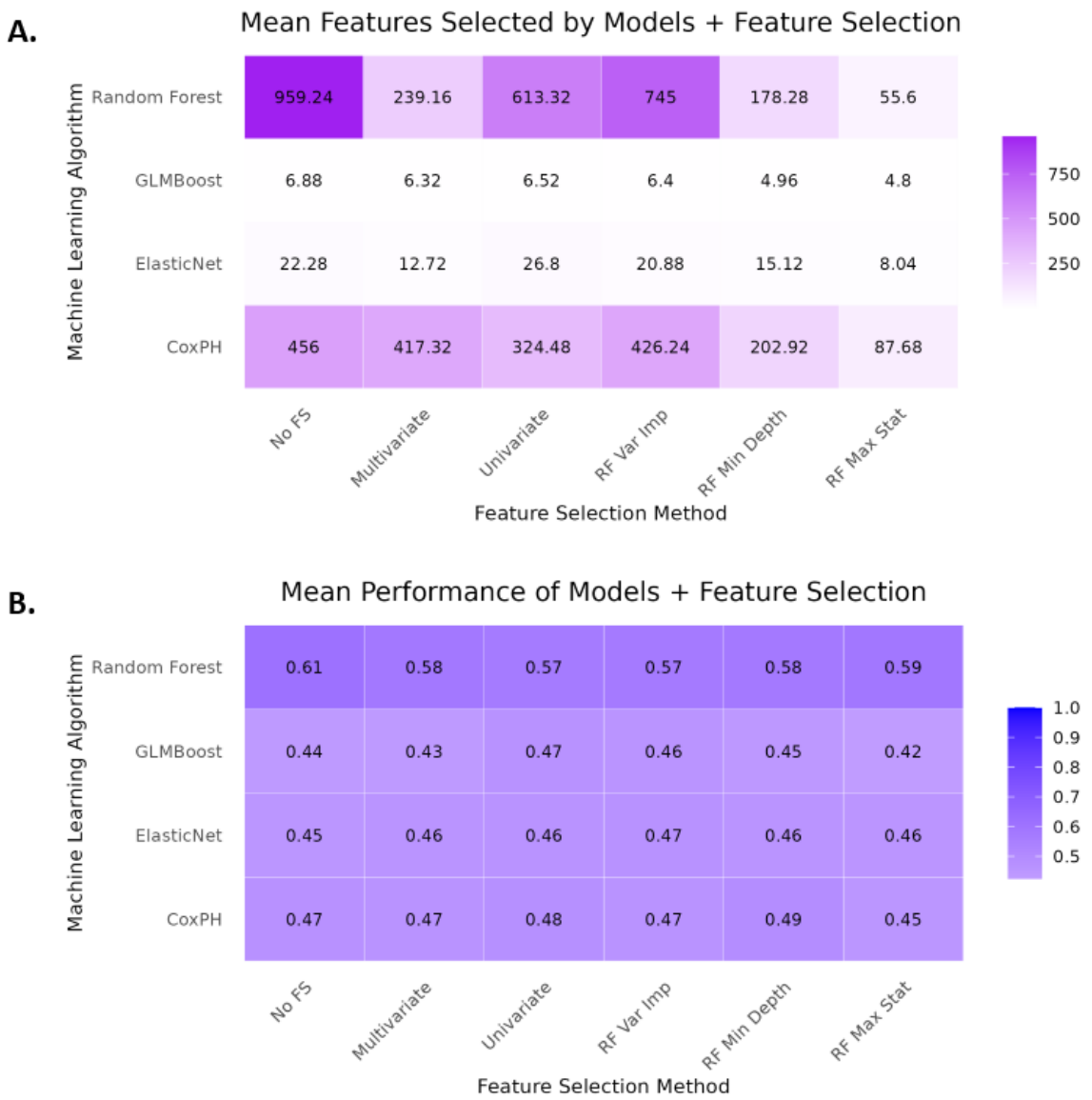


Figure C.15 Heatmap of both features selected from each model and filter methods, using BRCA CNV data. Mean number of features selected **(a),** Mean c-index values **(b)**



Appendix D. Performance of feature selection pipeline across all modalities and survival models.

The results of the comparisons are presented in Table D.14 to Table D.16.

Table D.14. Overview of the performance of the feature selection pipeline applied to the OV dataset.

OV	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	0.500	0.460
	CNV	0.484	0.489

Table D.15. Overview of the performance of the feature selection pipeline applied to the CESC dataset.

CESC	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	0.500	0.583
	CNV	0.497	0.570

Table D.16. Overview of the performance of the feature selection pipeline applied to the UCEC dataset.

UCEC	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	0.500

Appendix E. Early and Mid-fusion performance comparison.

The results of the comparisons are presented in Table E.17 to Table E.18.

Table E.17. Comparison between Mid and Early Fusion using BRCA dataset

Modality combinations	**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 of Early Fusion	0.702	0.699	0.685	0.720	0.662	0.613	**0.762**	0.745	0.715	0.678	0.667
C-index of Mid Fusion	0.768	0.702	0.694	0.714	0.718	0.654	0.692	0.707	**0.774**	0.752	0.655

Table E.18. Comparison between Mid and Early Fusion using OV dataset

Modality combinations	**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 of Early Fusion	0.575	0.583	0.580	0.553	0.618	0.527	0.612	0.568	**0.647**	0.595	0.586
C-index of Mid Fusion	0.673	**0.683**	0.646	0.593	0.559	0.616	0.666	0.585	0.630	0.633	0.637

Acknowledgements

The results published here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Declaration

During the preparation of this work, the authors utilized GPT-4 to enhance the clarity and readability of the content in some sections. All content was originally generated by the authors. Following the use of this tool/service, the authors thoroughly reviewed and edited the material as necessary and take full responsibility for the final content of the publication.

We also confirm that all procedures were performed in compliance with relevant laws and institutional guidelines and have been approved by the appropriate institutional committee(s), when applicable.

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. et al. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. Sci Rep 10, 20410 (2020). https://doi.org/10.1038/s41598-020-77220-w

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. Cai, Z., Poulos, R.C., Liu, J. and Zhong, Q. (2022). Machine learning for multi-omics data integration in cancer. *iScience*, 25(2), p.103798. doi:https://doi.org/10.1016/j.isci.2022.103798.

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

8. Rashid, M.M. and Kumar Selvarajoo (2024). Advancing drug-response prediction using multi-modal and -omics machine learning integration (MOMLIN): a case study on breast cancer clinical data. Briefings in Bioinformatics, [online] 25(4). doi:https://doi.org/10.1093/bib/bbae300.

9. Günther, O.P., Chen, V., Freue, G.C., Balshaw, R.F., Tebbutt, S.J., Hollander, Z., Takhar, M., McMaster, W.R., McManus, B.M., Keown, P.A. and Ng, R.T. (2012). A computational pipeline for the development of multi-marker bio-signature panels and ensemble classifiers. BMC Bioinformatics, 13(1), p.326. doi:https://doi.org/10.1186/1471-2105-13-326.

10. Wang, J., Liao, N., Du, X., Chen, Q. and Wei, B. (2024). A semi-supervised approach for the integration of multi-omics data based on transformer multi-head self-attention mechanism and graph convolutional networks. BMC genomics, 25(1). doi:https://doi.org/10.1186/s12864-024-09985-7.

11. Ahmed, K.T., Sun, J., Cheng, S., Yong, J. and Zhang, W. (2021). Multi-omics data integration by generative adversarial network. Bioinformatics. doi:https://doi.org/10.1093/bioinformatics/btab608.

12. Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. (2021). *Computational and Structural Biotechnology Journal*, [online] 19, pp.949–960. doi:https://doi.org/10.1016/j.csbj.2021.01.009.

13. 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.

14. 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.

15. Liu, C., Wang, X., Genchev, G.Z. and Lu, H. (2017). Multi-omics facilitated variable selection in Cox-regression model for cancer prognosis prediction. Methods, 124, pp.100–107. doi:https://doi.org/10.1016/j.ymeth.2017.06.010.

16. Chaudhary, K., Poirion, O.B., Lu, L. and Garmire, L.X. (2018). Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. Clinical Cancer Research, 24(6), pp.1248–1259. doi:https://doi.org/10.1158/1078-0432.ccr-17-0853.

17. Cheerla, A. and Gevaert, O. (2019). Deep learning with multimodal representation for pancancer prognosis prediction. Bioinformatics, 35(14), pp.i446–i454. doi:https://doi.org/10.1093/bioinformatics/btz342.

18. Hao, J., Masum, M., Oh, J.H. and Kang, M. (2019). Gene- and Pathway-Based Deep Neural Network for Multi-omics Data Integration to Predict Cancer Survival Outcomes. Lecture notes in computer science, pp.113–124. doi:https://doi.org/10.1007/978-3-030-20242-2_10.

19. 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.

20. Siegel, R.L., Miller, K.D. and Jemal, A. (2020). Cancer statistics, 2020. CA: a cancer journal for clinicians, 70(1), pp.7–30.

21. Research Trends in the Early Diagnosis of Ovarian Cancer during 2001–2020: A Bibliometric Analysis. (2022). European Journal of Gynaecological Oncology, 43(2), p.321. doi:https://doi.org/10.31083/j.ejgo4302038.

22. Zhang, J., Yu, S., Peng, Q., Wang, P. and Fang, L. (2024). Emerging mechanisms and implications of cGAS-STING signaling in cancer immunotherapy strategies. Cancer Biology & Medicine, [online] pp.1–20. doi:https://doi.org/10.20892/j.issn.2095-3941.2023.0440.

23. Das, S., Mohan Kumar Dey, Ram Devireddy and Manas Ranjan Gartia (2023). Biomarkers in Cancer Detection, Diagnosis, and Prognosis. Sensors, 24(1), pp.37–37.

24. Zhu, L., Miao, Y., Xi, F., Jiang, P., Xiao, L., Jin, X. and Fang, M. (2022). Identification of Potential Biomarkers for Pan-Cancer Diagnosis and Prognosis Through the Integration of Large-Scale Transcriptomic Data. Frontiers in Pharmacology, 13. doi:https://doi.org/10.3389/fphar.2022.870660.

25. 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.

26. 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.
27. 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.

28. 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.

29. 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.

30. 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.

31. Ma, J., Sun, L., Gao, W., Yang, L. and Dong, D. (2023). RNA binding protein: coordinated expression between the nuclear and mitochondrial genomes in tumors. Journal of Translational Medicine, 21(1). doi:https://doi.org/10.1186/s12967-023-04373-3.

32. Sulima, S.O., Hofman, I.J.F., De Keersmaecker, K. and Dinman, J.D. (2017). How Ribosomes Translate Cancer. Cancer Discovery, 7(10), pp.1069–1087. doi:https://doi.org/10.1158/2159-8290.cd-17-0550.

33. 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.

34. 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.

35. 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.

36. 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.

37. 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.

38. 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.

39. 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.