Using Ensemble Modeling with Multi-Omics Data to Enhance Cancer Subtype Classification

Signe Hoel[1,2*], Jun Ding[1,2] and Jingtao Wang[1,2]

[1*]Meakins-Christie Laboratories, Research Institute of the McGill University Health Centre, Montreal, QC, Canada.
[2]Department of Medicine, McGill University Health Centre, Montreal, QC, Canada.

*Corresponding author(s). E-mail(s): signe.hoel@gmail.com;
Contributing authors: jun.ding@mcgill.ca; jingtao.wang@mail.mcgill.ca;

Abstract

In cancer research, integrating multi-omics data has become crucial for understanding the complexities of tumor biology and improving clinical outcomes. This study introduces MOIE (Multi-Omics Integration Ensemble), a novel two-tier ensemble model designed to integrate genomic, transcriptomic, and proteomic data for cancer subtype classification and biomarker identification. MOIE combines individual ensemble models for each omics dataset, leveraging a meta-learner to consolidate predictions and enhance the accuracy of cancer classification tasks. Feature importance and survival analyses demonstrate that the biomarkers identified by MOIE are strongly associated with patient survival outcomes, validating its predictive power. Enrichment analysis further confirms the biological relevance of the selected genes, highlighting their involvement in known cancer pathways. This study underscores the efficacy of ensemble learning techniques for multi-omics data integration in cancer research, offering a potent tool for improving diagnostic accuracy and therapeutic interventions. MOIE's enhanced performance and robustness hold promise for advancing personalized medicine and shaping future cancer treatments.

1 Introduction

Recent advancements in cell dynamics research have opened new possibilities for understanding the underlying mechanisms of cellular state changes. The availability

of high-throughput technologies has allowed the generation of large volumes of molecular data - however, due to limitations in sequencing biotechnologies, cellular states are typically based on a single type of biomolecule, such as the quantification of RNA molecules with RNA-seq [1]. Although these cellular state measurements have been proven to lead to significant advancements in the study of cellular dynamics, the analysis of these data sets in isolation cannot properly represent cellular state. For this reason, the integration of multiple omics data sets, such as genomic, transcriptomic, and proteomic, are essential in providing more comprehensive views of the underlying mechanisms for cellular state changes in many biological processes [2].

Machine learning techniques, such as ensemble modeling, have proven indispensable in analysing multi-omics data and identifying predictive biomarkers. These methods can be applied to address critically important problems within healthcare. The classification of tumor subtypes, which plays a leading role in the treatment and prognosis of cancers, is one such example. However, these multi-class classification tasks have always presented a challenge in the field of integrating multi-omics using machine learning. As such, there is an urgent need for a multi-classification model to handle cancer subtype classification and biomarker identification.

In this paper, we introduce a two-tier ensemble model that integrates multi-omics data to 1) accurately classify cancer tumors into predefined subtypes/stages and 2) identify biomarkers predictive of each subtype. Our model comprises individual ensemble models for each single-omic data, and a meta-learner that consolidates predictions from multiple single-omic ensemble learners. This study aims to showcase the effectiveness of ensemble learning approaches when analyzing multi-omics datasets, in contrast to simpler machine learning methods. By contributing to the ongoing efforts in cell dynamics research, we aspire to enhance our comprehension of cellular mechanisms and their practical applications in the clinical setting.

2 Background

Computational models that integrate multi-omics datasets have become essential for discovering new diagnostics and therapies. Ensemble modeling is a popular machine learning technique that can be used to integrate multi-omics data by combining the predictions of several base learners.

Other studies have explored ensemble approaches for multiple data integration and omics feature selection. For example, mixOmics is an R package that can statistically integrate several datasets for multivariate analysis, data exploration, dimension reduction, and visualization [3]. The authors propose frameworks for integrative analysis of multiple independent omics studies, aiming to identify robust molecular signatures that suggest new biological hypotheses. DeepProg is another semi-supervised flexible hybrid machine learning framework that combines multiple omics data matrices and survival information to predict the output using a boosting ensemble model [4]. Other studies use network-based approaches, such as MoGCN, which is a multi-omics integration model designed for cancer subtype analysis based on graph convolutional networks [5].

While these studies have made significant contributions to the field, they also have limitations. mixOmics mainly focuses on identifying small subsets of molecules (molecular signatures) to explain or predict biological conditions, but mainly for a single type of omics data. DeepProg requires additional survival information, which may not be available for all datasets, and its use is limited to predicting the output for a single type of biological outcome [4].

Therefore, this study aims to investigate ensemble approaches that not only combine single omics models but also enhance them individually. By leveraging the strengths of previous studies and addressing their limitations, this study can contribute to the development of more effective computational models for multi-omics data integration in the future.

2.1 Ensemble Model Strategies

Ensemble learning is a machine learning approach that seeks to improve predictive performance by combining predictions from multiple models. The three ensemble learning classes dominate the field are bagging, boosting, and stacking.

Bagging (Bootstrap Aggregating) is an ensemble learning method that uses a diverse group of ensemble members by varying the training data. Typically, this involves using a single machine learning algorithm – usually an unpruned decision tree – and training each model on a different sample of the same training dataset [6]. The resulting predictions made by the ensemble members can then be combined using voting or averaging.

Boosting often involves homogenous weak learners, learning them sequentially in an adaptive way and then combining them using a deterministic strategy. Each model in the sequence is fitted giving more importance to observations that were badly handled by previous models in the sequence [7].

Stacking is an ensemble method that differs from bagging and boosting in that is often considers heterogenous (different) weak learners, rather than homogeneous. Stacking combines the predictions returned by several base models using a meta-model, rather than using deterministic algorithms [7].

3 Materials and Methods

3.1 Datasets

This study utilizes multi-omics data of two different TCGA (The Cancer Genome Atlas) cancer types; breast invasive carcinoma (BRCA) and the pan-kidney cohort (KIPAN). Preprocessed multi-omics data as well as the corresponding clinical information was available through the LinkedOmics portal, a web application which aggregates data from from TCGA and CPTAC cancer cohorts [8]. These datasets were chosen based on their large cohort sizes and their abundant use in other machine learning papers [9].

3.1.1 Pan-kidney (TCGA-KIPAN)

From the TCGA-KIPAN cohort, we create two datasets with separate clinical target variables. The first of integrates RNASeq, Copy Number Variation (CNV) and Reverse Phase Protein Array (RPPA) data with cancer subtype as the target variable. RNASeq was measured as the normalized expression signal of individual gene transcripts, copy number change (CNV) as the normalized copy number (SNPs) and copy number alternations for aggregated/segmented regions, and RPPA was measured as the normalized protein expression for each gene. The KIPAN dataset is comprised of three main subtypes - Kidney Chromophobe (KICH), Kidney Clear Cell Carcinoma (KIRC) and Kidney Papillary Cell Carcinoma (KIRP) - which act as the categories of this classification problem. Out of the 941 possible samples in the KIPAN cohort, there were 736 samples that had all three omics data available (Table 1).

Table 1 KIPAN Subtypes

Histological Type	Number of Samples
Kidney Chromophobe (KICH)	63
Kidney Clear Cell Carcinoma (KIRC)	467
Kidney Papillary Cell Carcinoma (KIRP)	206
Total	736

The second dataset integrates RNASeq, Methylation, and Reverse Phase Protein Array (RPPA) with pathological stage as the target variable. Methylation data was measured in beta values mapped to the genome, per sample. In the clinical data, there are four pathological stages provided - from stage I to IV. As the sample size for each stage is severely unbalanced, patients with clinical tumor stage I and II were labeled as "early" stage and patients with tumor stage III and IV were labeled as "late" stage [10]. This created a binary classification problem, with 350 total samples labeled as early and 208 as late (Table 2).

Table 2 KIPAN Stages

Pathological Stage	Number of Samples
I/II (early)	350
III/IV (late)	208
Total	558

3.1.2 Breast invasive carcinoma (TCGA-BRCA)

Our third dataset comprises of RNAseq, miRNA and methylation data from TCGA-BRCA cohort. RNAseq and miRNA gene-level expression data were normalized and log-transformed, and methylation gene-level beta values have were mapped to the

genome. The four intrinsic subtypes defined by the mRNA based PAM50 signature as phenotypic groups were used as the target variable for this dataset [11].

The TCGA-BRCA cohort has a total of 1098 tumor samples, 430 of which has all three omics data available. Each sample was assigned to one of four cancer subtypes by TCGA: Basal-like, HER2E, Luminal A and Luminal B subtypes (Table 3).

Table 3 BRCA Subtypes

PAM50	Number of Samples
Basal	71
Her2	30
Luminal A	229
Luminal B	100
Total	430

3.2 Preprocessing

To prepare each dataset for downstream analysis, samples were first selected based on whether they were present in each modality, such that the number of samples is equal between each omics dataset. To reduce dimensionality, features with null values were removed and only the top 1000 most variable genes were selected from each modality using the `scanpy` preprocessing tools.

3.3 Base estimators

In this study, we employed a combination of linear and tree-based models as the base estimators for our ensemble machine learning model. All the models were implemented using `scikit-learn` [12], except for the `BalancedRandomForestClassifier` which was implemented from the `imblearn` library [13].

We selected these base estimators based on their ability to perform well on similar classification problems, as well as their versatility and compatibility with our multi-omics dataset.

3.3.1 Logistic Regression

We used the `LogisticRegression` class from scikit-learn's `linear_model` module to implement this estimator . We performed regularized logistic regression using either L2 regularization as the penalty.

3.3.2 Balanced Random Forest

We used the `BalancedRandomForestClassifier` class from imbalanced-learn's `ensemble` module to implement this estimator. Similar to the classical, this meta-estimator fits several decision tree classifiers on sub-samples of the data and uses averaging to control over-fitting and improve predictive accuracy. However, to account for imbalanced data, a balanced random forest differs from the classical version by

drawing a bootstrap sample from the minority class and sampling with replacement the same number of samples from the majority class.

3.3.3 Multi-layer Perceptron Classifier (MLP)

We used the `MLPClassifier` class from scikit-learn's `neural_network` module to implement this estimator . It optimizes the log-loss function using LBFGS or stochastic gradient descent.

3.3.4 Support Vector Classification (SVC)

We implemented this estimator using the `LinearSVC` class from scikit-learn's `svm` module . This linear estimator uses support vector machines to separate data points into different classes.

3.4 Ensemble Modeling

3.4.1 Single-omics Modality Stacking Ensemble

In this study, we are interested in determining if combining several different base learning models into an ensemble model will increase model performance for each omics dataset. Thus, we will be using stacking as the proposed ensemble method.

To create our custom modality ensemble model, we utilized scikit-learn's `BaseEstimator` and `TransformerMixin` classes as a basis for our class. For the stacking ensemble functionality, we used the `StackingClassifier` from the ensemble module in scikit-learn, with `LogisticRegression` as the default final estimator [12]. Each base estimator was fitted on the entire input dataset, and the final estimator was trained using cross-validated predictions from the base estimators, obtained through the `cross_val_predict` function of scikit-learn. This approach helps to improve the model's predictive performance by combining the strengths of individual estimators while minimizing their weaknesses.

3.4.2 Multi-omics Integration Stacking Ensemble

To integrate the strengths of our modality ensembles, we created a multi-omics integration classifier using another custom stacking model. The single-omics modality ensembles act as the estimators for the `StackingClassifier`, with another `LogisticRegression` classifier as the final estimator.

3.5 Performance

To evaluate the performance of our models, we used 5-fold Stratified KFold cross-validation with 3 repeats, implemented with scikit-learn. The f1 score and balanced accuracy score were used as scoring metrics, which were implemented using the `sklearn.metrics` and the `imblearn.metrics` modules, respectively.

The f1 score is the harmonic mean of precision and recall, computed as

$$F_1 = \frac{2 * true\ positive}{2 * true\ positive + false\ positive + false\ negative} \tag{1}$$

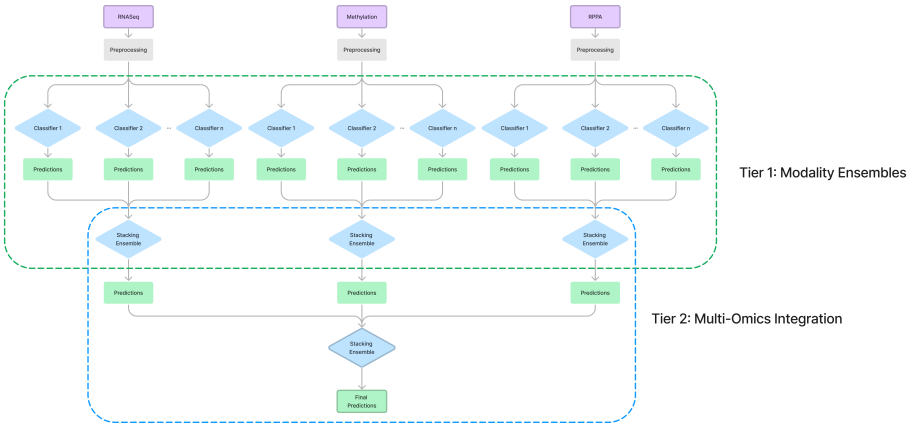

Fig. 1 Overview of our two-tiered ensemble pipeline and integration of multi-omics datasets.

Balanced accuracy is defined as the arithmetic mean of sensitivity and specificity, and is useful in the case where there is imbalanced data. It is computed as

$$BalancedAccuracy = \frac{sensitivity + specificity}{2} \tag{2}$$

We generated performance plots using `matplotlib.pyplot` [14] and `seaborn` libraries [15].

3.6 Statistical Significance

To measure the significance of our results, we tested the 5-fold cross-validation with 3 repeats results, which provided us with 15 data points to compare using a one-sided Wilcoxen statistical test. This was implemented using the `wilcoxen` method from the `scipy.stats` module, with "greater" as the alternative hypothesis [16].

3.7 Feature Importance

To find genes that were most 'important' to our integration ensemble model, we utilized the `permutation_importance` method from scikit-learn [12]. Permutation importance, in this context, measures the average reduction in balanced accuracy caused by permuting (i.e. randomly shuffling) each feature during the calculation of out-of-bag error. The greater the reduction in balanced accuracy, the higher the importance score of a gene.

Each gene was permuted 10 times to get a proper estimate of its impact on the response variable. After obtaining the permutation importances, we normalized them by their standard deviations. These normalized importance scores were then ranked in descending order. Genes with importance scores greater than 0 were selected for subsequent survival and enrichment analyses.

3.8 Survival and Enrichment Analysis

To validate the features selected based on permutation importance, we developed a script to use a Cox proportional hazards model to analyze the relationship between these genes and survival rates. The selected gene sets with permutation importance greater than 0 were separated by modality within their respective omics datasets for comparison. After calculating the Cox model coefficients, we generated Kaplan-Meier plots to visualize the survival outcomes associated with these gene sets. We then used a log-rank statistical test to compare the survival curves to determine if there is a statistically significant difference in survival between them.

Further validation of the selected genes was performed through enrichment analysis using `gseapy`, a python implementation of GSEA and wrapper for Enrichr [17]. The gene sets from each TCGA dataset (i.e. the RNASeq, Methylation, RPPA) were combined into a single set and compared against the Reactome database to identify significantly overlapping biological pathways. We then focused on pathways with the most significant overlaps, utilizing gseapy's functionality to visualize these findings.

3.9 Benchmarking

To validate effectiveness of our method (MOIE), we compared its performance with two other state-of-the-art multi-omics integration methods: MixOmics [3] and MoGCN [5]. Each method was evaluated using 5-fold cross-validation with 3 repeats to ensure robustness in estimating model performance. We conducted these comparisons on each of our three distinct cancer datasets.

MixOmics was employed with its default settings, which involve tuning to find the maximum number of principal components to keep. For the sake of consistency and computational efficiency, we opted to not perform the optional feature selection method and evaluate the resulting model using their `perf` method with `MFold` as the cross-validation method. MixOmics provides the balanced classification error rate, from which we calculated the balanced accuracy (1-BER).

MoGCN, proposed by Xiao Li et. al., is a recent multi-omics integration method that leverages graph convolutional networks for cancer subtype classification [5]. We utilized the implementation publically available at https://github.com/Lifoof/MoGCN. Originally designed to report accuracy and f1 scores through 10-fold cross-validation, we modified their performance validation code to also report balanced accuracy using our 5-fold cross-validation with 3 repeats setup.

4 Results

4.1 Base Learners vs. Modality Ensembles

When comparing the performances of base learners and our modality ensembles for the KIPAN subtyping dataset, the highest scoring models for RNA, CNV and RPPA modalities were all the ensemble models (Table 4), although the difference with the next highest scoring model were only significant for RNA (p=0.019) and CNV (p=0.0002). For KIPAN staging, the modality ensemble is the highest performing in RNA and RPPA modalities, although not significantly different than the highest performing base learners. With BRCA, the miRNA modality ensemble is significantly higher than the next highest base learner (p=0.037), whereas we don't see a significant improvement in RNA or methylation modalities.

Table 4 Mean F1 Scores of Base Learners vs. Modality Ensembles

	KIPAN Subtyping			KIPAN Staging		
Model	RNA	CNV	RPPA	RNA	Methylation	RPPA
Logistic Regression	0.957	0.861	0.872	0.744	0.732	0.687
Balanced Random Forest	0.952	0.857	0.963	0.734	0.738	0.716
Deep NN	0.955	0.833	0.967	0.732	**0.765**	0.717
SVC	0.958	0.708	0.855	0.740	0.712	0.637
ModalityEnsemble	**0.961**	**0.889**	**0.968**	**0.750**	0.757	**0.723**

	BRCA Subtyping		
Model	RNA	miRNA	Methylation
Logistic Regression	0.871	0.785	**0.805**
Balanced Random Forest	0.850	0.779	0.785
Deep NN	0.864	0.782	0.797
SVC	0.867	0.800	0.803
ModalityEnsemble	**0.878**	**0.811**	0.803

4.2 Modality Ensembles vs. Multi-omics Integration Ensemble

Upon combining the modality ensembles to form our second tier multi-omic integration ensembles, the results showcased increases in performance when the modalities were combined rather than individual (Figure 2).

For the KIPAN subtyping target, almost all of the tested combinations performed better than any individual modality. We found that combining rna + rppa modalities received an f1 score 2.3% higher than the highest scoring individual modality, rna (p=0.00003)(Table 5). Several other multi-omics integration results were found to be significantly higher than individual modality ensembles.

When combining all three modality ensembles from the KIPAN staging dataset (rna + meth + rppa), we found that there was a 1.3 % increase in f1 score compared to the highest scoring individual modality, which was methylation (p = 0.012).

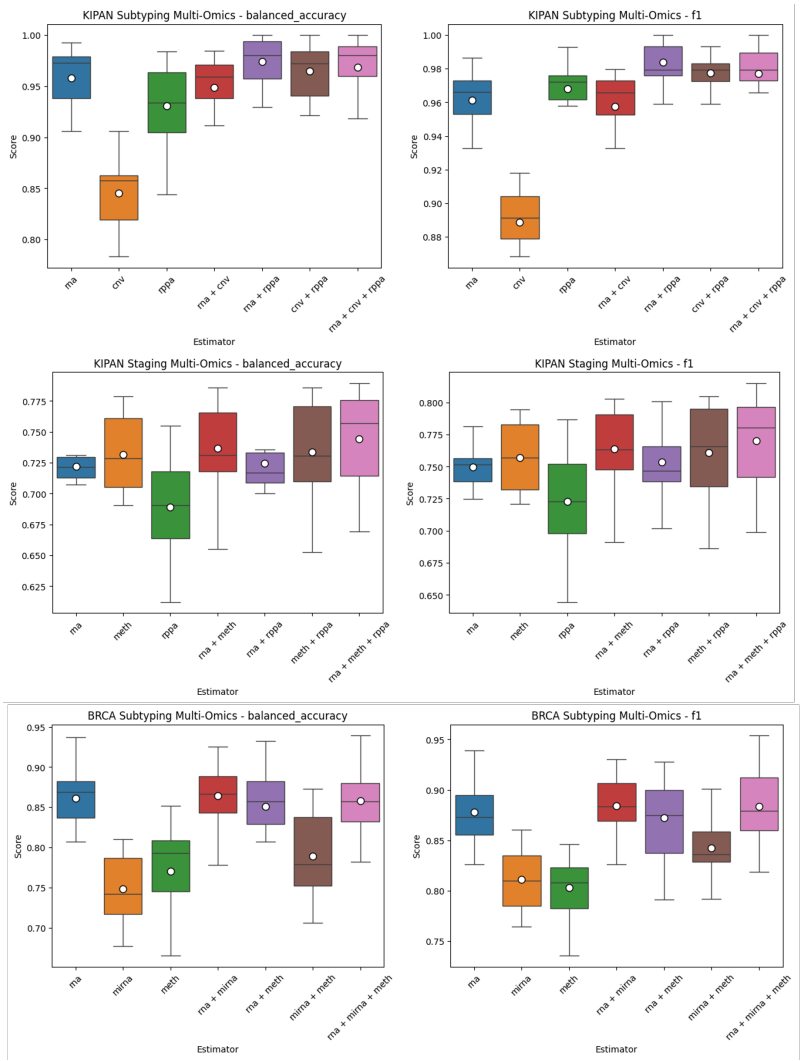

Fig. 2 F1 and Balanced Accuracy Scores for Modality Ensembles and Multi-Omics Integration Ensembles with different combinations of Modality Ensemble estimators with target (from top to bottom) A) KIPAN subtyping B) KIPAN staging C) BRCA subtyping.

In the BRCA dataset, we see that combining the two lesser-performing modalities, mirna + methylation results in a 3.9% increase in f1 score as compared to methylation alone ($p = 0.011$)(Table 5).

4.3 Permutation Importance and Survival Analysis

After calculating the permutation importances of the genes from each modality, 0 RNA, 58 CNV, and 62 RPPA genes were selected as 'important' (permutation importance > 0) from the KIPAN subtyping dataset, the highest scoring being

Table 5 Mean Balanced Accuracy and F1 Scores of Modality Ensembles vs. Multi-Omics Integration Ensembles

KIPAN Subtyping		
Model	Balanced Accuracy	F1 Score
rna	0.958 +/- 0.029	0.961 +/- 0.977
cnv	0.846 +/- 0.034	0.889 +/- 0.022
rppa	0.931 +/- 0.039	0.968 +/- 0.016
rna + cnv	0.949 +/- 0.033	0.958 +/- 0.022
rna + rppa	**0.974 +/- 0.025**	**0.984 +/- 0.013**
cnv + rppa	0.965 +/- 0.024	0.943 +/- 0.066
rna + cnv + rppa	0.969 +/- 0.027	0.977 +/- 0.017
KIPAN Staging		
Model	Balanced Accuracy	F1 Score
rna	0.722 +/- 0.026	0.750 +/- 0.025
meth	0.732 +/- 0.029	0.757 +/- 0.027
rppa	0.689 +/- 0.044	0.723 +/- 0.043
rna + meth	0.737 +/- 0.035	0.764 +/- 0.030
rna + rppa	0.724 +/- 0.031	0.753 +/- 0.029
meth + rppa	0.734 +/- 0.039	0.761 +/- 0.037
rna + meth + rppa	**0.744 +/- 0.038**	**0.770 +/- 0.035**
BRCA Subtyping		
Model	Balanced Accuracy	F1 Score
rna	0.862 +/- 0.034	0.878 +/- 0.033
mirna	0.748 +/- 0.041	0.811 +/- 0.030
meth	0.770 +/- 0.060	0.803 +/- 0.041
rna + mirna	**0.865 +/- 0.038**	**0.884 +/- 0.031**
rna + meth	0.851 +/- 0.055	0.872 +/- 0.040
mirna + meth	0.789 +/- 0.050	0.842 +/- 0.031
rna + mirna + meth	0.858 +/- 0.047	**0.884 +/- 0.039**

CDH2|N-Cadherin. From the KIPAN staging dataset, there were 40 RNA, 15 Methylation and 16 RPPA genes selected, the highest scoring being RAF1|C-Raf_pS338. For the BRCA dataset, there were only RNA genes selected, with 707 of them having permutation importances above 0 (Figure 3).

From the generated KMPlots, we can see for each of these datasets that there is a significant difference between the survival of patients with low and highly expression of these selected genes, with the selected methylation biomarkers from the KIPAN Staging dataset as the exception (Figure 4).

4.4 Enrichment Analysis

The GSEA enrichment analysis results validate the gene features selected by our integration models, demonstrating their biological relevance and effectiveness.

For KIPAN subtyping, the top enriched pathways, including Signal Transduction, PI3K/AKT Signaling, and MTOR Signaling, play critical roles in kidney cancer biology (Figure 5a). These pathways are frequently altered in renal cell carcinoma, promoting cell growth, survival, and poor clinical outcomes [18]. Additionally, pathways

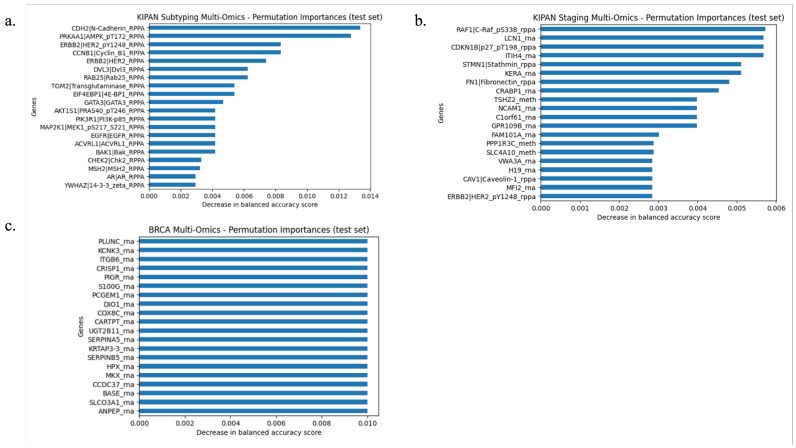

Fig. 3 Top 20 most important genes to the Multi-Omics Integration Ensemble models for a) KIPAN subtyping b) KIPAN staging and c) BRCA subtyping targets, after calculating permutation differences for 10 repeats.

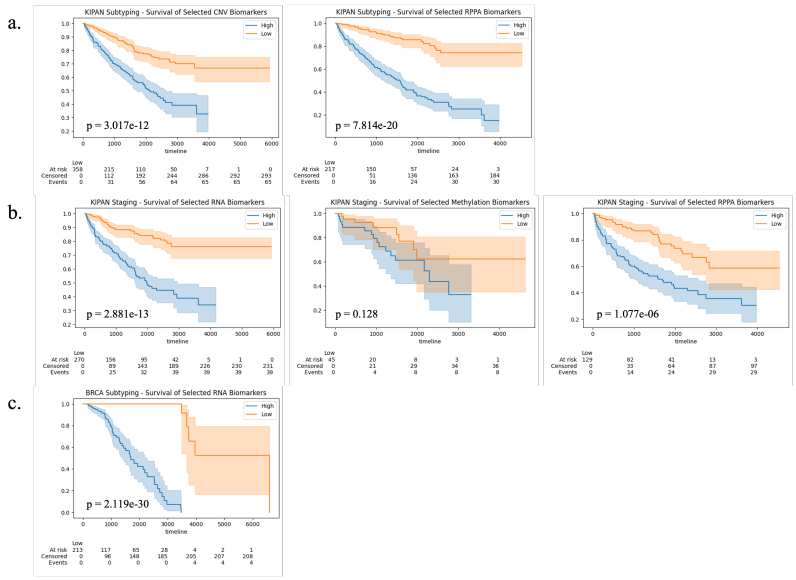

Fig. 4 KMPlots for important genes from a) KIPAN subtyping b) KIPAN staging and c) BRCA subtyping datasets.

related to transcription and gene expression, such as the Generic Transcription Pathway and RNA Polymerase II Transcription, highlight the importance of transcriptional dysregulation in tumor development and progression [19] [20].

For KIPAN staging, pathways such as Signaling by Non-Receptor Kinases and the RAF/MAP Kinase Cascade are essential in regulating cellular processes like growth and apoptosis, distinguishing early from late-stage tumors (Figure 5b) [19] [21]. The

enrichment of pathways like 'ERBB2 Activates PTK6 Signaling' and 'Platelet Degranulation' underscores their role in tumor progression and metastasis, key factors in cancer staging [22] [23]. Furthermore, MAPK Family Signaling Cascades and specific MAPK1/MAPK3 Signaling pathways support the biological relevance of our selected genes, capturing mechanisms involved in tumor invasion and metastasis [19].

For BRCA subtyping, the identified pathways are crucial to breast cancer differentiation and progression. The Metal Sequestration by Antimicrobial Proteins pathway influences tumor growth and immune response, while Keratinization reflects variations in keratin expression among subtypes (Figure 5c) [24] [25]. Glycosylation-related pathways, including those caused by defects in GALNT12, GALNT3, and C1GALT1C1, highlight the significance of glycosylation in cell-cell interactions and signaling [26]. Additionally, pathways such as Diseases of Glycosylation and Diseases of Metabolism underscore the role of post-translational modifications and metabolic reprogramming in distinguishing breast cancer subtypes [27] [28]. Collagen Degradation and Cell Surface Interactions at the Vascular Wall pathways emphasize extracellular matrix remodeling and cell adhesion, critical for invasive and metastatic subtypes [29].

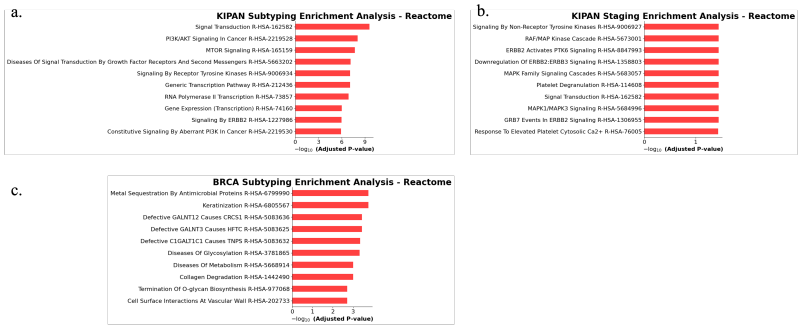

Fig. 5 GSEApy enrichment analysis results from a) KIPAN subtyping b) KIPAN staging and c) BRCA subtyping datasets.

4.5 Benchmarking

To evaluate the performance of our model, we compared it against other popular multi-omics integration methods, specifically MixOmics and MoGCN. Our model, MOIE, demonstrated superior performance across all three classification tasks (Table 6). We see a 2.2% increase in balanced accuracy over the next highest performing model for KIPAN subtyping, a 1.7% increase for KIPAN staging, and a 4.4% increase for the BRCA subtyping dataset.

5 Discussion

Cancer is widely regarded as a highly heterogeneous disease, and the early diagnosis and prognostic of a cancer type has become a focus of cancer research. The integration of multi-omics can help us achieve a better understanding of systems biology

Table 6 Mean Balanced Accuracy and F1 Scores of MultiOmicsIntegrationEnsemble (MOIE) vs. other integration methods

KIPAN Subtyping		
Model	Balanced Accuracy	F1 Score
MultiOmicsIntegrationEnsemble (MOIE)	**0.974 +/- 0.025**	**0.984 +/- 0.013**
MixOmics	0.868 +/- 0.007	0.909 +/- 0.002
MoGCN	0.952 +/- 0.022	0.963 +/- 0.011
KIPAN Staging		
Model	Balanced Accuracy	F1 Score
MultiOmicsIntegrationEnsemble (MOIE)	**0.744 +/- 0.038**	**0.770 +/- 0.035**
MixOmics	0.698 +/- 0.001	0.710 +/- 0.001
MoGCN	0.727 +/- 0.042	0.768 +/- 0.039
BRCA Subtyping		
Model	Balanced Accuracy	F1 Score
MultiOmicsIntegrationEnsemble (MOIE)	**0.865 +/- 0.038**	**0.884 +/- 0.031**
MixOmics	0.821 +/- 0.002	0.806 +/- 0.002
MoGCN	0.811 +/- 0.062	0.859 +/- 0.030

and doing this in an efficient/effective way is an important challenge for research in bioinformatics.

We developed a multi-omics integration pipeline for cancer subtype classification using a two-tier ensemble model approach. The results of this study demonstrate the benefits of using an ensemble model; the `StackingClassifier` model performed provided benefits when used to combine classifiers with each of the datasets, whether it be for combining base learners or modalities.

For KIPAN datasets, the combination of base learners to form an ensemble for each modality showed significant performance improvements, while we didn't see this demonstrated with BRCA. For KIPAN subtyping, KIPAN staging, and BRCA subtyping we saw significant improvements when combining modality ensembles as opposed to using a single modality. For BRCA, we note an improvement when combining two lesser performing modalities such as mirna and methylation together, showcasing the use of ensemble models for boosting performance when you don't have preferred datasets available (such as RNASeq).

To investigate the use of ensemble modelling for biomarker detection, we permuted the features from our three datasets and testing the effects on the performances of our ensemble models. Through this method, we were able to obtain 3 sets of genes that we then performed survival analysis on to validate our ensemble method. For each of our datasets, we saw that there was a significant correlation between the gene sets selected and the survival of patients in that cohort, validating our model's learning process.

Furthermore, the enrichment analysis confirms that our models effectively identify biologically relevant genes for KIPAN subtyping, KIPAN staging, and BRCA subtyping tasks. The alignment of our results with known cancer pathways supports the utility of our models in selecting key genes, enhancing our understanding of the molecular underpinnings of cancer progression and subtyping.

Upon comparing MOIE with other popular multi-omics integration methods, we saw significant improvements in performance. The improvements in balanced accuracy demonstrate MOIE's robustness in handling complex multi-omics data and its effectiveness in capturing the underlying biological signals across different cancer types and classification tasks. The superior performance of MOIE can be attributed to its ability to integrate diverse data types more effectively, leveraging advanced techniques to enhance feature selection and model generalization.

5.1 Limitations and Future work

The current study has demonstrated the effectiveness of an ensemble learning approach in improving the prediction performance for cancer classification tasks using multi-omics data. However, there are several limitations to this study that should be addressed in future research. One limitation is the lack of parameter tuning for the base learners and the ensemble model. Although not the primary focus of this study, parameter tuning can further improve the overall performance of the model. Future research should focus on optimizing the hyperparameters of the base learners and the ensemble model to achieve better performance.

Moreover, as permutation importance methods are very susceptible to correlation bias, we would like to look solutions or alternative methods for feature importance calculation that account for this. As many of the gene expression levels are likely highly correlated with one another, the permutation importance values may be misleading and lead us to believe these genes are less important to the model than they really are. Additional analysis should also be conducted to further validate the selected biomarkers, such as enrichment analysis.

Despite these limitations, this study provides valuable insights into the utility of ensemble learning approaches in feature selection and cancer diagnosis using multi-omics data. Further research can build upon these findings to develop more accurate and efficient models for cancer subtype classification.

6 Conclusions

The results of this study highlight the effectiveness of ensemble machine learning approaches for cancer classification using different 'omics datasets. Combining 'omics datasets into single ensemble models were shown to outperform single dataset ensemble models, in each of these three classification tasks.

Our feature importance and survival analysis methods revealed insightful biomarkers that correlated with the survival times of patients with these two cancer types. Additional enrichment analysis studies would help further validate our ensemble models and showcase it's efficacy at selecting biomarkers.

Overall, this study provides valuable insights into the potential of ensemble machine learning approaches for cancer classification, and suggests that further exploration of these methods can lead to improved diagnostic and treatment strategies for cancer patients.

References

[1] Hughes, T.K., Wadsworth, M.H., Gierahn, T.M., Do, T., Weiss, D., Andrade, P.R., Ma, F., Andrade Silva, B.J., Shao, S., Tsoi, L.C., et al.: Highly efficient, massively-parallel single-cell rna-seq reveals cellular states and molecular features of human skin pathology. BioRxiv, 689273 (2019)

[2] Subramanian, I., Verma, S., Kumar, S., Jere, A., Anamika, K.: Multi-omics data inte- gration, interpretation, and its application. Bioinformatics and biology insights **14**, 1177932219899051 (2020)

[3] Rohart, F., Gautier, B., Singh, A., Lê Cao, K.A.: mixomics: An r package for 'omics feature selection and multiple data integration. PLoS Comput Biol **13**, 11 (2017) https://doi.org/10.1371/journal.pcbi.1005752

[4] Poirion, O.B., Jing, Z., Chaudhary, K., *et al.*: Deepprog: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. Genome Med **13**, 112 (2021) https://doi.org/10.1186/s13073-021-00930-x

[5] Li, X., Ma, J., Leng, L., Han, M., Li, M., He, F., Zhu, Y.: MoGCN: A Multi-Omics Integration Method Based on Graph Convolutional Network for Cancer Subtype Analysis. Frontiers in Genetics **13**, 806842 (2022) https://doi.org/10.3389/fgene.2022.806842

[6] Brownlee, J.: A Gentle Introduction to Ensemble Learning Algorithms. https://machinelearningmastery.com/tour-of-ensemble-learning-algorithms/ Accessed 2023-04-10

[7] Rocca, J.: Ensemble Methods: Bagging, Boosting and Stacking. https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205 Accessed 2023-04-10

[8] Vasaikar, S.V., Straub, P., Wang, J., Zhang, B.: LinkedOmics: analyzing multi-omics data within and across 32 cancer types. Nucleic Acids Research **46**(D1), 956–963 (2018) https://doi.org/10.1093/nar/gkx1090

[9] Liñares-Blanco, J., Pazos, A., Fernandez-Lozano, C.: Machine learning analysis of TCGA cancer data. PeerJ Comput Sci **7**, 584 (2021) https://doi.org/10.7717/peerj-cs.584

[10] Jagga, Z., Gupta, D.: Classification models for clear cell renal carcinoma stage progression, based on tumor RNAseq expression trained supervised machine learning algorithms. BMC Proc **8**(Suppl 6 Proceedings of the Great Lakes Bioinformatics Confer), 2 (2014) https://doi.org/10.1186/1753-6561-8-S6-S2

[11] Parker, J.S., Mullins, M., Cheang, M.C., Leung, S., Voduc, D., Vickery, T., Davies,

S., Fauron, C., He, X., Hu, Z., Quackenbush, J.F., Stijleman, I.J., Palazzo, J., Marron, J.S., Nobel, A.B., Mardis, E., Nielsen, T.O., Ellis, M.J., Perou, C.M., Bernard, P.S.: Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of clinical oncology : official journal of the American Society of Clinical Oncology **27**, 1160–1167 (2019) https://doi.org/10.1200/JCO.2008.18.1370

[12] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)

[13] Lemaître, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research **18**(17), 1–5 (2017)

[14] Hunter, J.D.: Matplotlib: A 2d graphics environment. Computing in Science & Engineering **9**(3), 90–95 (2007) https://doi.org/10.1109/MCSE.2007.55

[15] Waskom, M.L.: seaborn: statistical data visualization. Journal of Open Source Software **6**(60), 3021 (2021) https://doi.org/10.21105/joss.03021

[16] Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods **17**, 261–272 (2020) https://doi.org/10.1038/s41592-019-0686-2

[17] Fang, Z., Liu, X., Peltz, G.: GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. Bioinformatics (2022) https://doi.org/10.1093/bioinformatics/btac757

[18] Glaviano, A., Foo, A.S.C., Lam, H.Y., *et al.*: Pi3k/akt/mtor signaling transduction pathway and targeted therapies in cancer. Molecular Cancer **22**(1), 138 (2023) https://doi.org/10.1186/s12943-023-01827-6

[19] Wang, P., Laster, K., Jia, X., *et al.*: Targeting craf kinase in anti-cancer therapy: progress and opportunities. Molecular Cancer **22**(1), 208 (2023) https://doi.org/10.1186/s12943-023-01903-x

[20] Siebenthall, K.T., Miller, C.P., Vierstra, J.D., Mathieu, J., Tretiakova, M., Reynolds, A., Sandstrom, R., Rynes, E., Haugen, E., Johnson, A., Nelson, J., Bates, D., Diegel, M., Dunn, D., Frerker, M., Buckley, M., Kaul, R., Zheng, Y.,

Himmelfarb, J., Ruohola-Baker, H., Akilesh, S.: Integrated epigenomic profiling reveals endogenous retrovirus reactivation in renal cell carcinoma. EBioMedicine **41**, 427–442 (2019) https://doi.org/10.1016/j.ebiom.2019.01.063

[21] Steelman, L.S., Chappell, W.H., Abrams, S.L., Kempf, C.R., Long, J., Laidler, P., Mijatovic, S., Maksimovic-Ivanic, D., Stivala, F., Mazzarino, M.C., Donia, M., Fagone, P., Malaponte, G., Nicoletti, F., Libra, M., Milella, M., Tafuri, A., Bonati, A., Bäsecke, J., Cocco, L., Evangelisti, C., Martelli, A.M., Montalto, G., Cervello, M., McCubrey, J.A.: Roles of the raf/mek/erk and pi3k/pten/akt/mtor pathways in controlling growth and sensitivity to therapy-implications for cancer and aging. Aging **3**(3), 192–222 https://doi.org/10.18632/aging.100296

[22] Lin, L., Gong, S., Deng, C., Zhang, G., Wu, J.: Ptk6: An emerging biomarker for prognosis and immunotherapeutic response in clear cell renal carcinoma (kirc). Heliyon **10**(7), 29001 (2024) https://doi.org/10.1016/j.heliyon.2024.e29001

[23] Xie, X., Wang, N., Xiang, J., He, H., Wang, X., Wang, Y.: Renal cell carcinoma associated with idiopathic thrombocytopenic purpura. International Journal of Immunopathology and Pharmacology **34**, 2058738420931619 (2020) https://doi.org/10.1177/2058738420931619

[24] Li, S., Jiang, M.: Elevated insulin-like growth factor 2 mrna binding protein 1 levels predict a poor prognosis in patients with breast carcinoma using an integrated multi-omics data analysis. Frontiers in Genetics **13**, 994003 (2022) https://doi.org/10.3389/fgene.2022.994003

[25] Shen, F., Fang, Y., Wu, Y., Zhou, M., Shen, J., Fan, X.: Metal ions and nanometallic materials in antitumor immunity: Function, application, and perspective. Journal of Nanobiotechnology **21**(1), 20 (2023) https://doi.org/10.1186/s12951-023-01771-z

[26] Reily, C., Stewart, T.J., Renfrow, M.B., Novak, J.: Glycosylation in health and disease. Nature Reviews Nephrology **15**(6), 346–366 (2019) https://doi.org/10.1038/s41581-019-0129-4

[27] Serrano-Carbajal, E.A., Espinal-Enríquez, J., Hernández-Lemus, E.: Targeting metabolic deregulation landscapes in breast cancer subtypes. Frontiers in Oncology **10** (2020) https://doi.org/10.3389/fonc.2020.00097

[28] Lei, P., Wang, W., Sheldon, M., Sun, Y., Yao, F., Ma, L.: Role of glucose metabolic reprogramming in breast cancer progression and drug resistance. Cancers **15**, 3390 (2023) https://doi.org/10.3390/cancers15133390

[29] Mensah, S.A., Harding, I.C., Zhang, M., Jaeggli, M.P., Torchilin, V.P., Niedre, M.J., Ebong, E.E.: Metastatic cancer cell attachment to endothelium is promoted by endothelial glycocalyx sialic acid degradation. AIChE Journal **65**(8), 16634 (2019) https://doi.org/10.1002/aic.16634