A denoised multi-omics integration framework for cancer subtype classification and survival prediction

Jiali Pang, Bilin Liang, et al.
Briefings in Bioinformatics, 2023, 24(5),1–12

Presenter: Zhe LIU
Bioinformatics and Biostatistics Lab
December 14, 2023

Outline

1 Introduction

2 Methods

3 Results

4 Conclusion

Introduction

Introduction: Previous multi-omics integration methods

- Similarity-based methods
  - Spectral Clustering
  - Similarity Network Fusion(SNF)
- Dimension reduction-based methods
  - Principle Component Analysis(PCA)
  - Canonical Correlation Analysis(CCA)
  - Non-negative Matrix Factorization(NMF)
- AI-based methods
  - Autoencoder-based network
  - Graph Convolutional Network(GCNs)

Introduction: Main idea of this proposed methods

- Limitations of existing multi-omics integration framework
  - **Model overfitting:** The number of features is much greater than the number of samples.
  - **False positive:** Noisy features are easily to be selected to contribute to models, leading to false positives.
  - **Generalization ability:** Poor Generalizability on different datasets

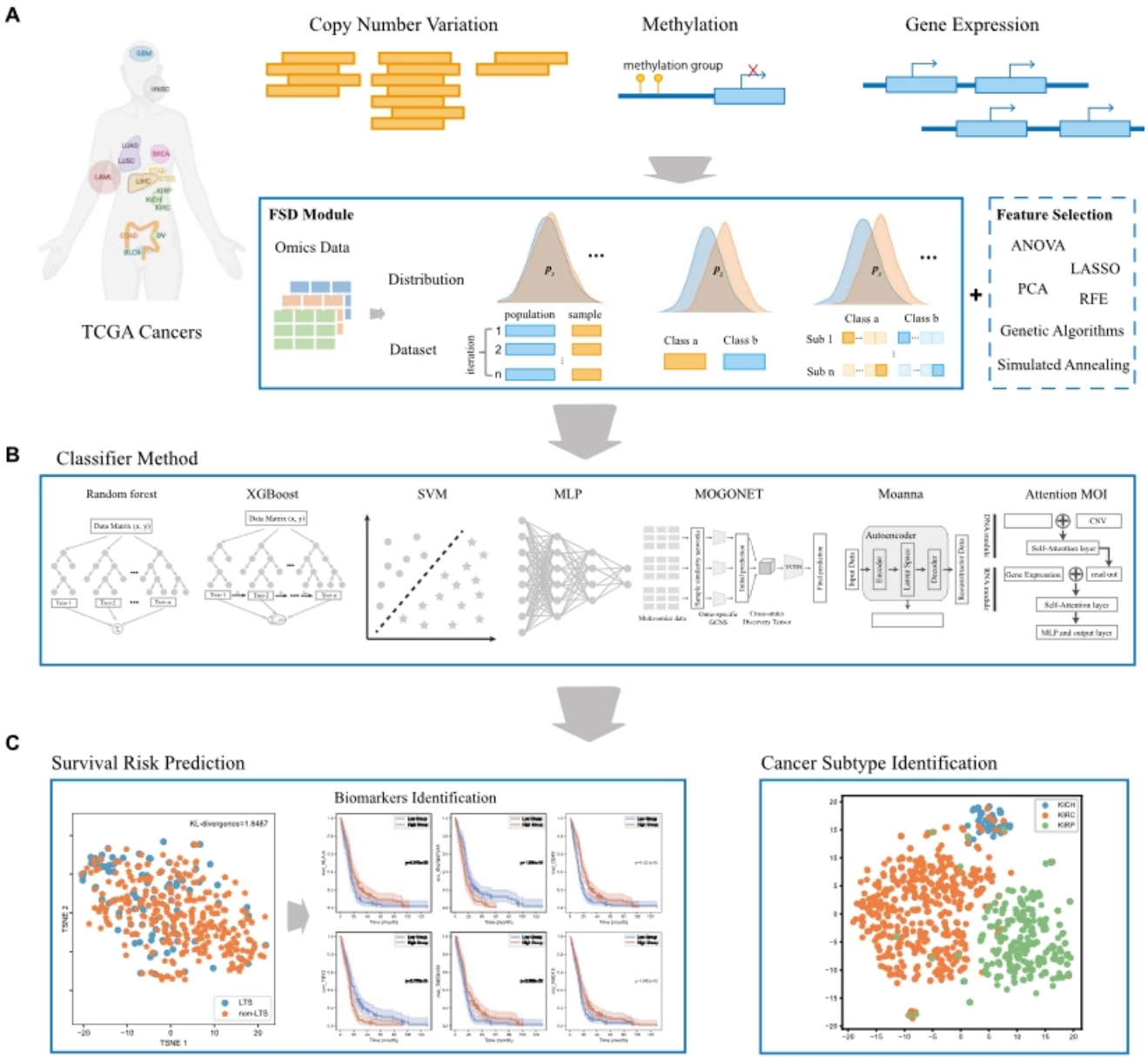
- Novelty of this proposed framework
  - **Feature Selection with Distribution(FSD) module:** Reduce the noise of multi-omics data in the data-preprocessing procedure
  - **Attention Multi-omics Intergratation (AttentionMOI):** Provide a biologically informed multi-omics integration framework

Methods

Methods: Data acquisition

- For task 1: Kidney Cancer Subtype Identification
  - KIPAN dataset (concluding three subtypes: KICH, KIRC, KIRP)
- For task 2: Cancer Survival Time Prediction
  - 15 types of cancer datasets from TCGA project
  - A Glioblastoma(GBM) dataset and a Head and Neck squamous cell carcinoma(HNSC) dataset from the CPTAC project (validation)
- Only patients with all three omics data were selected for prediction
  - Copy Number Variation (CNV)
  - Methylation (Met)
  - RNA Transcriptome (RNA)

Methods: Overview of the framework



Methods: Baseline Feature selection methods[1]

Select a subset of features which can efficiently describe the input data while reducing effects from noise or irrelevant variables and still provide good prediction results.

- **Wrapper methods**
- **Filter methods**
- **Embedded methods**

[1] Girish Chandrashekar and Ferat Sahin. "A survey on feature selection methods". In: *Computers & Electrical Engineering* 40.1 (2014), pp. 16–28.

Methods: Baseline Feature Selection Methods

- **Wrapper methods** use feature subsets to train the model and selects or excludes features based on the model's performance
  - Heuristic Search Algorithms(evaluate different subsets to optimize the objective function)
    - ★ Simulated Annealing (SA)
    - ★ Genetic Algorithm (GA)

- **Filter methods** use variable ranking techniques as the principle criteria for variable selection
  - ANOVA

- **Embedded methods** include variable selection as part of the training process without splitting the data
  - Recursive Feature Elimination(RFE)
  - LASSO

Methods: FSD module

- A subset X_{sub} was randomly selected from the training dataset X. Then, three distribution tests were performed as follows:

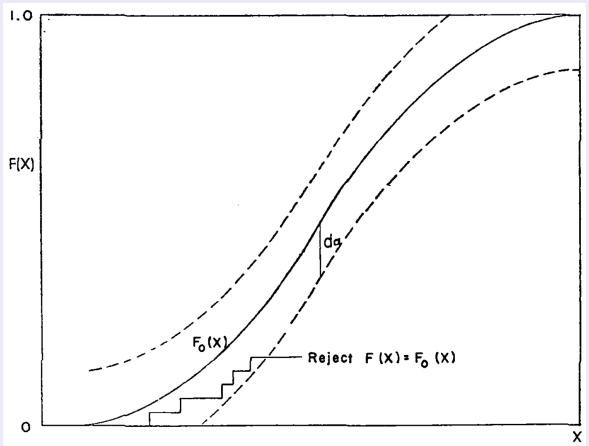
$$
\begin{aligned}
p_1 &\leftarrow \text{KS}(X_{\text{sub}}, X), \\
p_2 &\leftarrow \text{KS}(X_a, X_b, \ldots, X_n), \\
p_3 &\leftarrow \text{KS}(X_{\text{sub},a}, X_{\text{sub},b}, \ldots, X_{\text{sub},n})
\end{aligned}
$$

  - KS indicates the Kolmogorov–Smirnov test
  - p_1, p_2, and p_3 represent P-values of the statistical tests, respectively
  - $X_a, X_b, \ldots, X_n$ are clinical classification $(a, b, \ldots, n)$ data in X
  - $X_{\text{sub},a}, \ldots, X_{\text{sub},n}$ are clinical classification $(a, b, \ldots, n)$ data in X_{sub}
- A feature was considered to be low-noise and high-informative if $p_1 > k$, $p_2 < k$, and $p_3 < k$, where k is 0.05 by default.

Supplement[2]

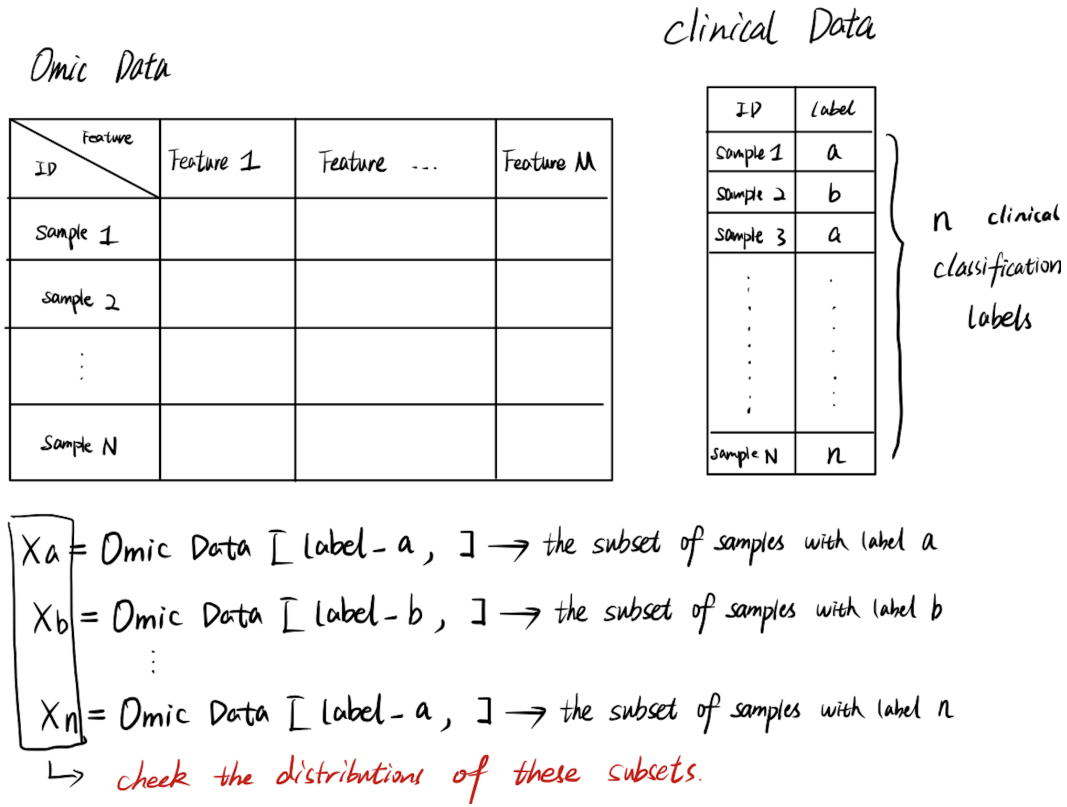
Kolmogorov–Smirnov test

- A non-parametric statistical test used to assess whether a sample comes from a specific distribution.
- H_0: the sample data follows the specified theoretical distribution.
- Test statistic: maximum vertical deviation between the sample CDF and the theoretical CDF.



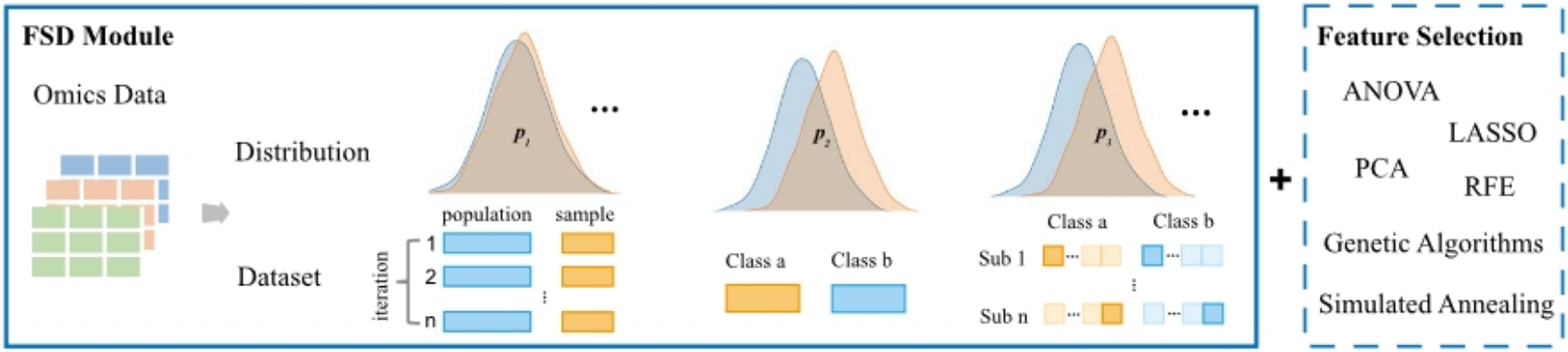
[2] Frank J Massey Jr. "The Kolmogorov-Smirnov test for goodness of fit". In: *Journal of the American statistical Association* 46.253 (1951), pp. 68–78.

Methods: FSD module



- To obtain a more stable feature selection result, they repeated the above process m times
- n represents the number of times that a feature met the above conditions
- If $n/m > j$, the feature proceeds to the subsequent analysis, where j denotes the threshold value

Methods:FSD module + Baseline FS methods



Combinations of FSD and each traditional method are performed to explore whether it is helpful for feature sekection.

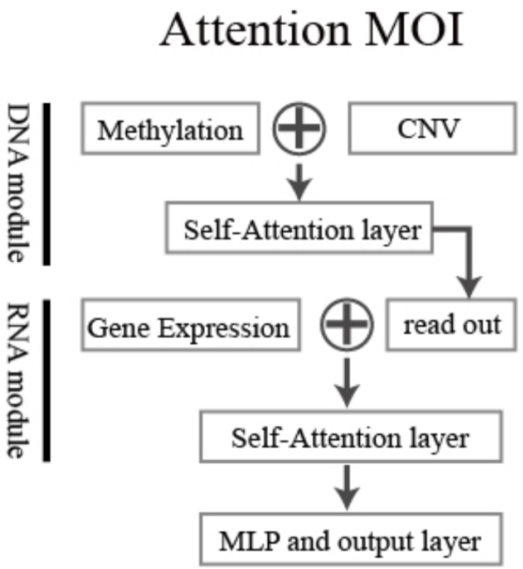
- ANOVA
- LASSO
- PCA
- RFE
- Genetic Algorithms(GA)
- Simulated Annealing (SA)

Methods: AttentionMOI model

Algorithm 1. Attention Multi-omics Integration

Input: Omics data matrix at DNA and RNA levels. M_i^D represents the i-th omics matrix at DNA levels, such as methylation; M_i^R represents the i-th omics matrix at RNA levels, such as gene expression.
Output: Classification (y) of each patient, such as risk stratification or disease subtype.

```
        # Step1: fusing DNA-omics features
1       M^D ← Self_Attention(M_i^D)
2       Readout^D ← MLP(M^D)
        # Step2: fusing RNA-omics features
3       M^R ← Self_Attention(M_i^R)
4       Readout^R ← MLP(M^R)
        # Step3: fusing DNA and RNA features
5       x ← Self_Attention(Readout^D, Readout^R)
        # Step4: classification
6       y ← MLP(x)
```



- 1-4: Put DNA and RNA level features into two attention layers
- 5: Concatenate through an attention layer to weight different features
- 6: Classification task was realized by the fully connected layer

Methods: Biological interpretation

To calculate the contribution of features to the model output, two explaining methods were applied.

- **SHapley Additive exPlanations (SHAP)**
  - explain models building by RF, SVM and XGBoost
- **Integrated gradient (IG)**
  - interpret the MLP and AttentionMOI model

Supplement[3]

SHapley Additive exPlanations (SHAP)

- A method used to interpret model predictions, particularly widely employed in black-box models and ensemble learning
- Calculate the average contribution of each feature to the model output based on shapley value:

$$\phi_i(f) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)]$$

  - N represents the set of features, and $|N|$ is the number of features.
  - S is any subset of N that does not include i.
  - $f(S)$ represents the model's output for a given subset S.
  - $\phi_i(f)$ is the Shapley value for feature i, representing the average marginal contribution across all possible subsets.

[3] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).

Supplyment[4]

Integrated Gradient (IG)

- Evaluate the contribution of each input feature on the deep learning model's output by integrating over the input features.

$$\mathrm{IG}_i(f) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial f(\mathbf{z} + \alpha \times (\mathbf{x} - \mathbf{x}'))}{\partial z_i} d\alpha$$

Where:

- f is the model's output.
- x_i is the i-th feature value of the input.
- x_i' is the i-th feature value at some baseline state.
- $\mathbf{z}$ is a state along the path, where $z_i = x_i' + \alpha \times (x_i - x_i')$.
- $\frac{\partial f(\mathbf{z})}{\partial z_i}$ is the partial derivative of the model output with respect to z_i.

[4] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic Attribution for Deep Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, June 2017, pp. 3319–3328.

Results

Results: 1.Denoising of transcriptome data using FSD

- Applied the FSD for GBM survival group prediction using the transcriptome data from TCGA and CPTAC
  - Long Time Survial(LTS)
  - Non-Long Time Survival (non-LTS)

- Compared the performance based on four machine learning models including MLP, RF, XGBoost and SVM with features selected by FSD, ANOVA, RFE, LASSO, PCA.

Results: 1.Denoising of transcriptome data using FSD

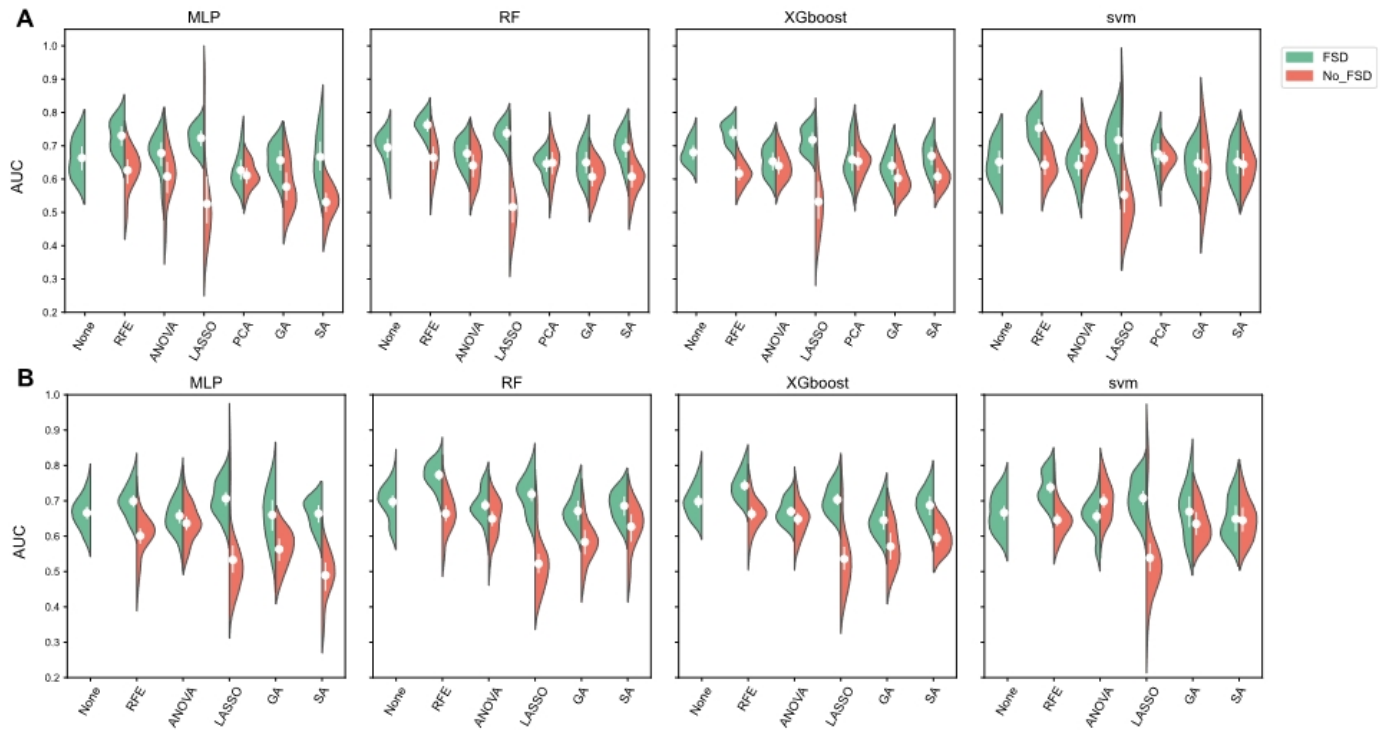
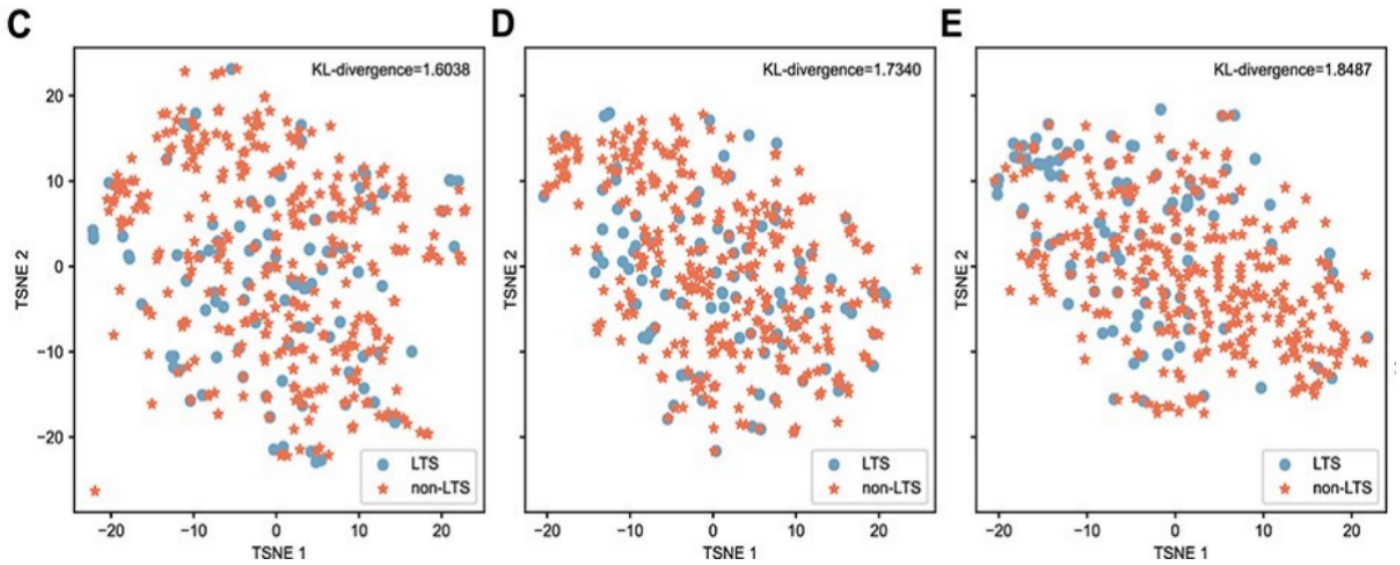


Figure A: TCGA-GBM

Figure B: CPTAC-GBM

- When the FSD module is introduced, all models obtain better performance regardless of the feature selection method used
- LASSO-based methods obtained the worst stability, while **RFE-based models** obtained better performance and stability

Results: 1.Denoising of transcriptome data using FSD



- Compared the **t-SNE visualized GBM gene expressions** of two survival groups with different gene expression inputs. Specifically, genes without selection, RFE selected genes and FSD+RFE selected genes.

- Genes selected from the FSD+RFE method achieved the highest **KL-divergence score** after t-SNE decomposition, indicating that FSD+RFE could better differentiate GBM survival groups

Results: 2.Performance of FSD under different omics data

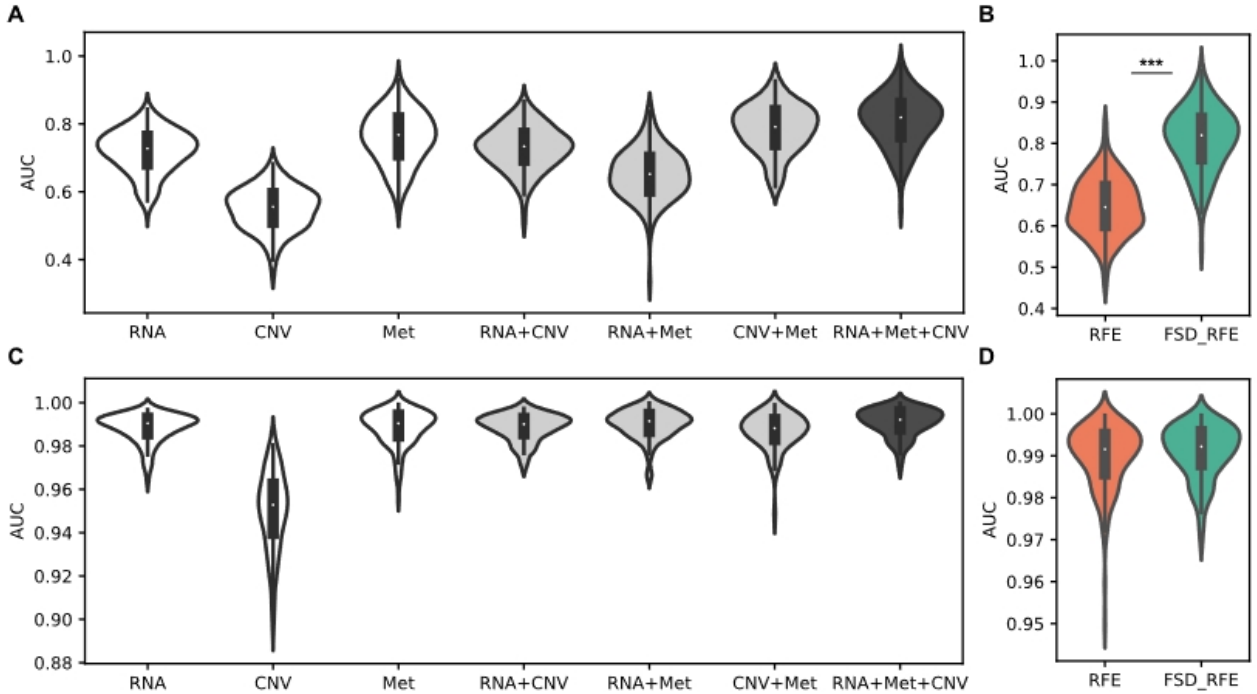
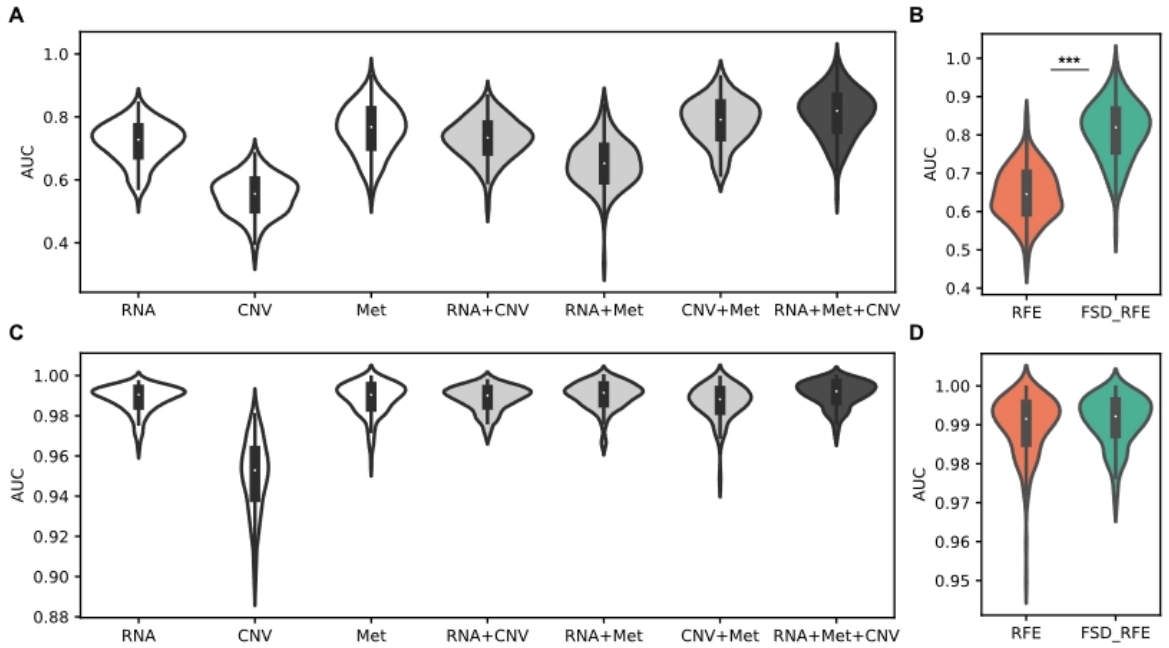


Figure: (A/C) Comparison of AUC under RF model using different combinations of omics in two tasks.

Figure: (B/D) Comparison of AUC under RF model between RFE and FSD + RFE selected features in two tasks based on RNA+Met+CNV data.

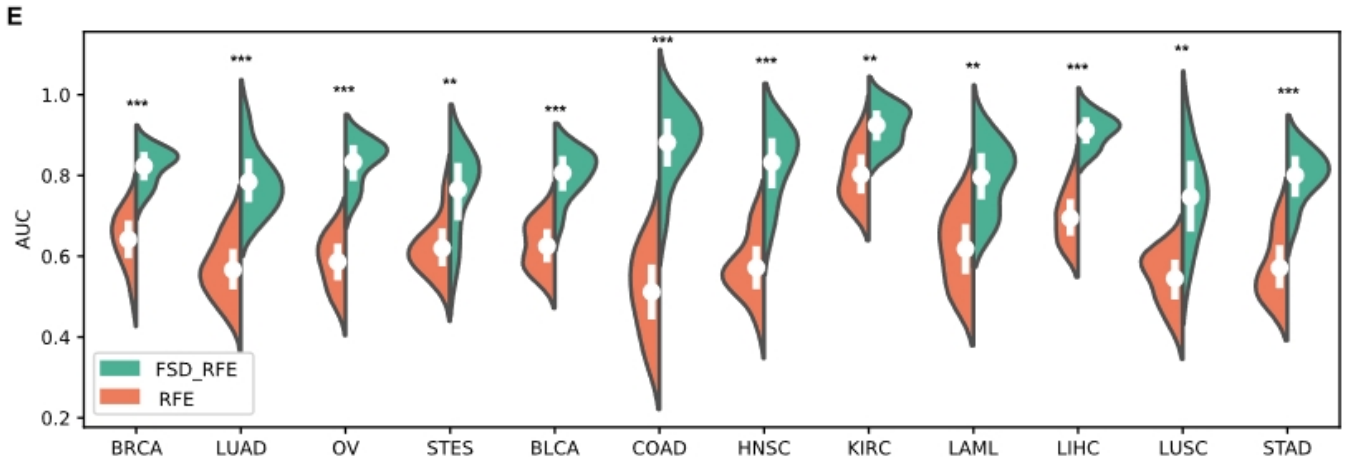
Results: 2.Performance of FSD under different omics data



- Combined prediction of multi-omics achieved better average AUC performance than single omic

FSD module further improved the prediction performance under multi-omics data

Results: 3.The generalizability of FSD



- Comparison of AUC using RFE model and FSD + RFE selected features in prediction of survival among different TCGA cancer types. P-value calculated through Mann–Whitney-U test.
  - p-value < 0.05: $*$
  - p-value < 0.01: $**$
  - p-value < 0.001: $***$
- Among 15 cancer types, 12 cancers using FSD + RFE to select features improved performance significantly.

High generalization ability

Results: 4.FSD selected feature as potential markers I

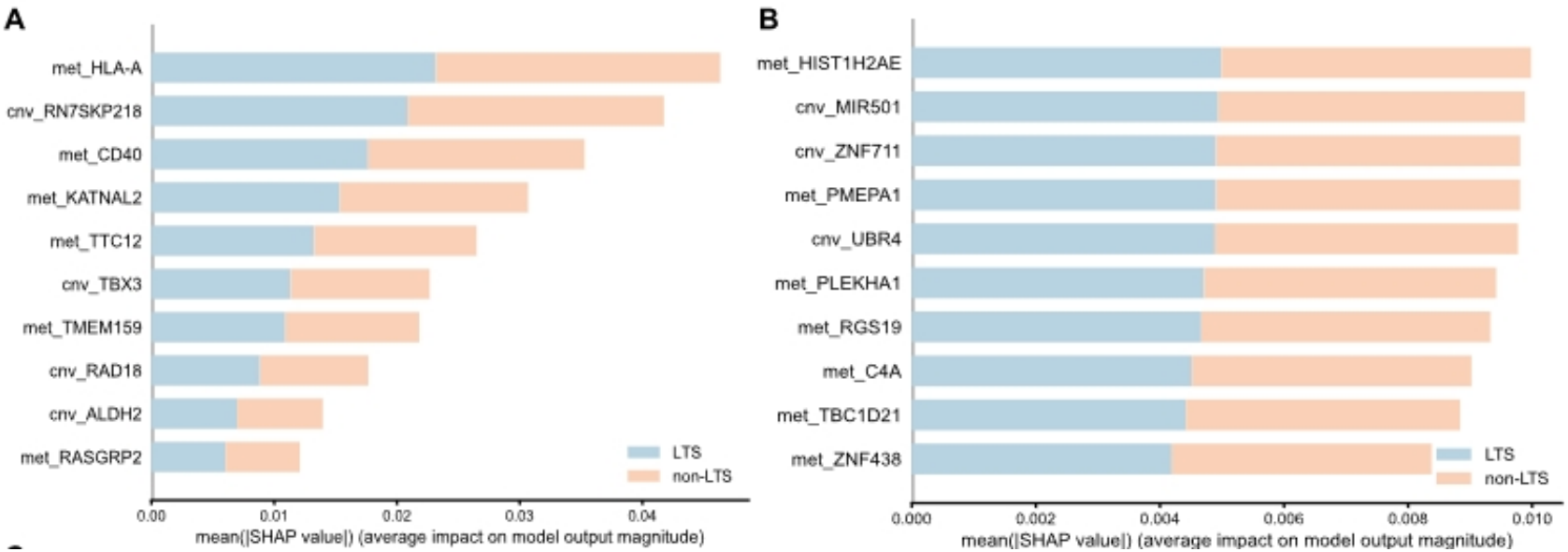


Figure: Feature importance of FSD+RFE / RFE selected features

1. Conducted **SHAP analysis** to evaluate the contributions of features identified by FSD +RFE and RFE under the RF model
   - The estimated feature importances of the top 10 features selected by RFE differed slightly and they did not contribute much

Results: 4.FSD selected feature as potential markers II

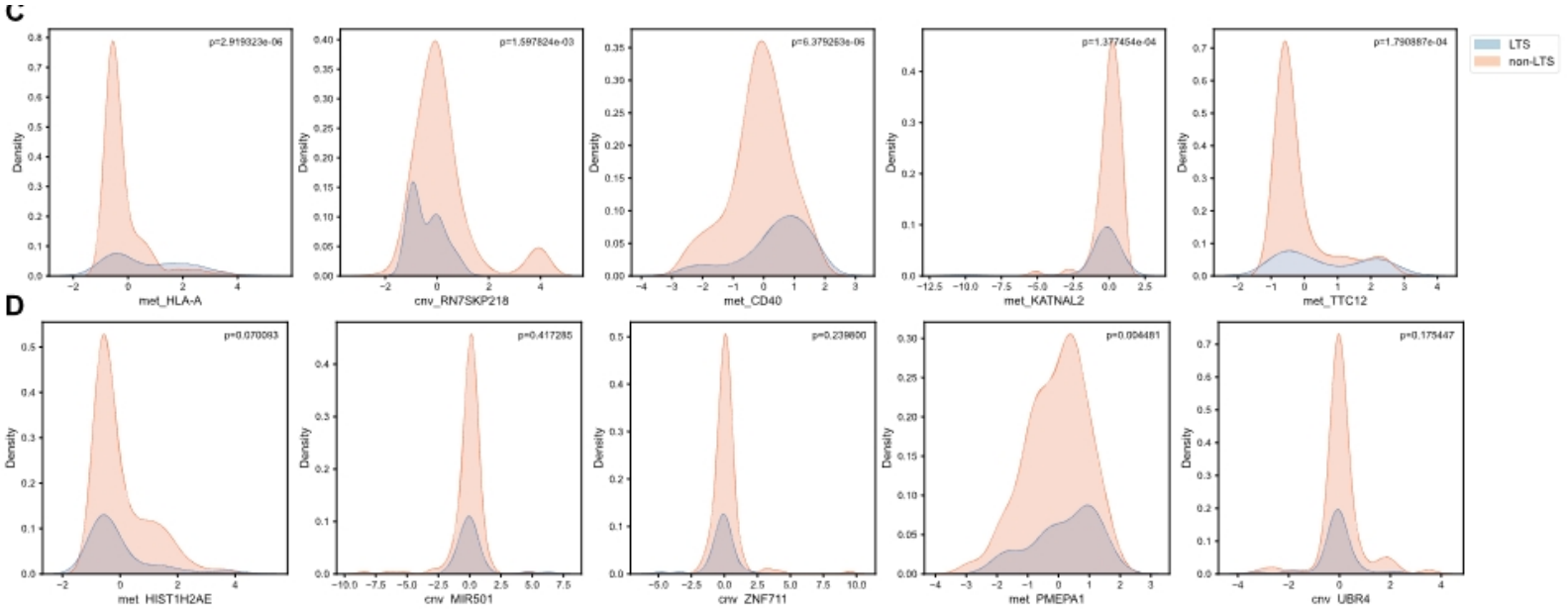
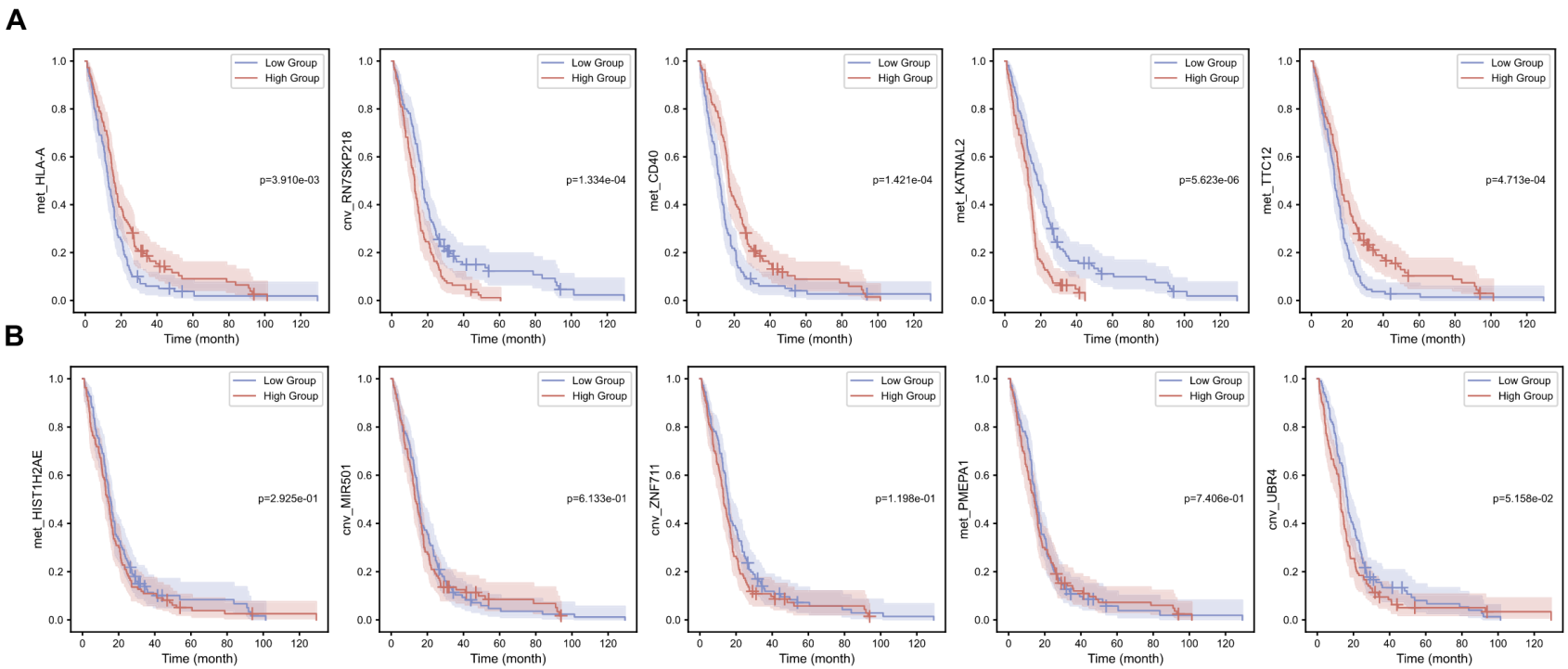


Figure C: FSD+RFE Figure D: RFE

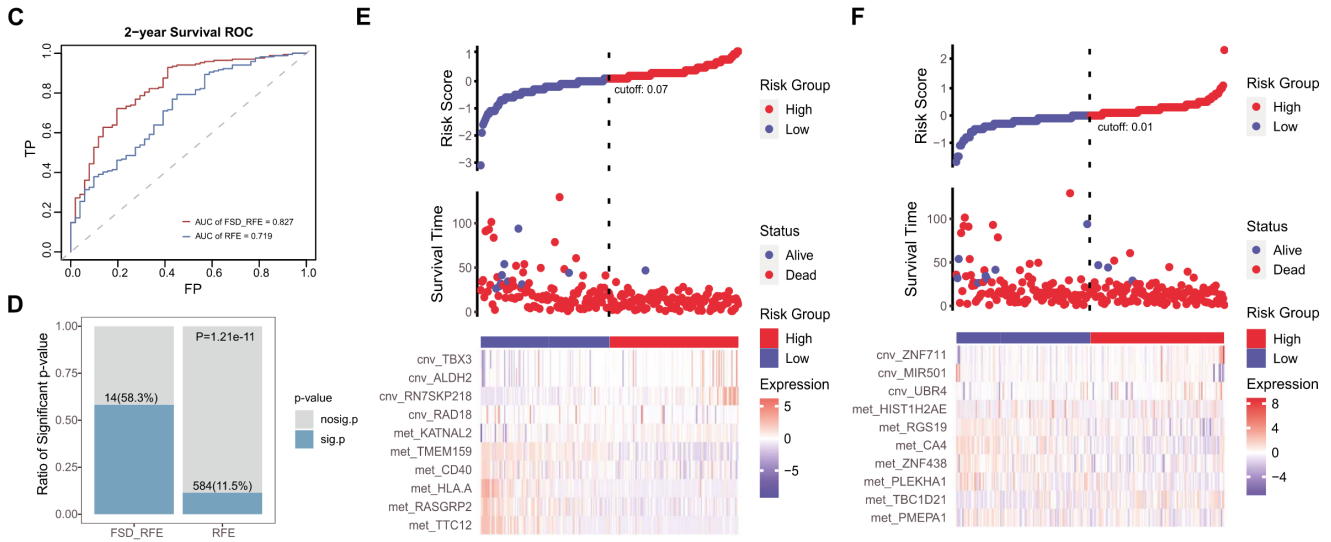
2. Displayed **distributions** of those features to explore how those top features varied between classes.
   - *p*-value calculated through Mann–Whitney-U test.
   - All top 5 FSD + RFE selected features significantly differentiate between classes.
   - Among the top 5 RFE selected features, only one feature significantly different between classes.

Results: 4.FSD selected feature as potential markers III



3. **Kaplan-Miere curves and Log-Rank tests** were conducted.
   - All top 5 features selected by FSD + RFE(Figure A) were significantly influential to patients' survival while none of the features selected by RFE(Figure B) were influential to patients' survival

Results: 4.FSD selected feature as potential markers IV



4. 2-year **survival ROC** estimated by top10 features(C): FSD + RFE selected features achieved higher AUC
5. **Univariate Cox regression analysis**(D): 58.3% features selected by RFE + FSD were significantly associated with patient hazard, while 11.5% for RFE
6. There were clearer patterns of omics data values for REF + FSD selected features when estimated **risk scores** of patients increased(E/F)

The above results all indicated that FSD selected features could be potential prognostic markers

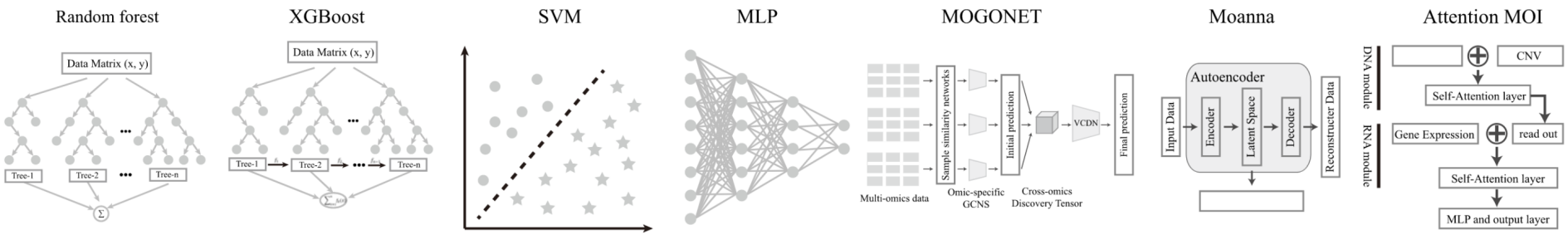
Results: 5.AttentionMOI improved model performance I

Table 1. Performance of AttentionMOI under multi-omics data types

Method	Threshold = 0.2	Threshold = 0.6
MLP	0.7129(0.6751–0.7506)	0.7711(0.7449–0.7972)
AttentionMOI	0.7934(0.7602–0.8265)	0.7200(0.6886–0.7513)
RF	0.7468(0.6844–0.8092)	0.7698(0.7384–0.8012)
XGBoost	0.7338(0.6802–0.7875)	0.7955(0.7704–0.8206)
SVM	0.7839(0.7398–0.8280)	0.7892(0.7631–0.8154)

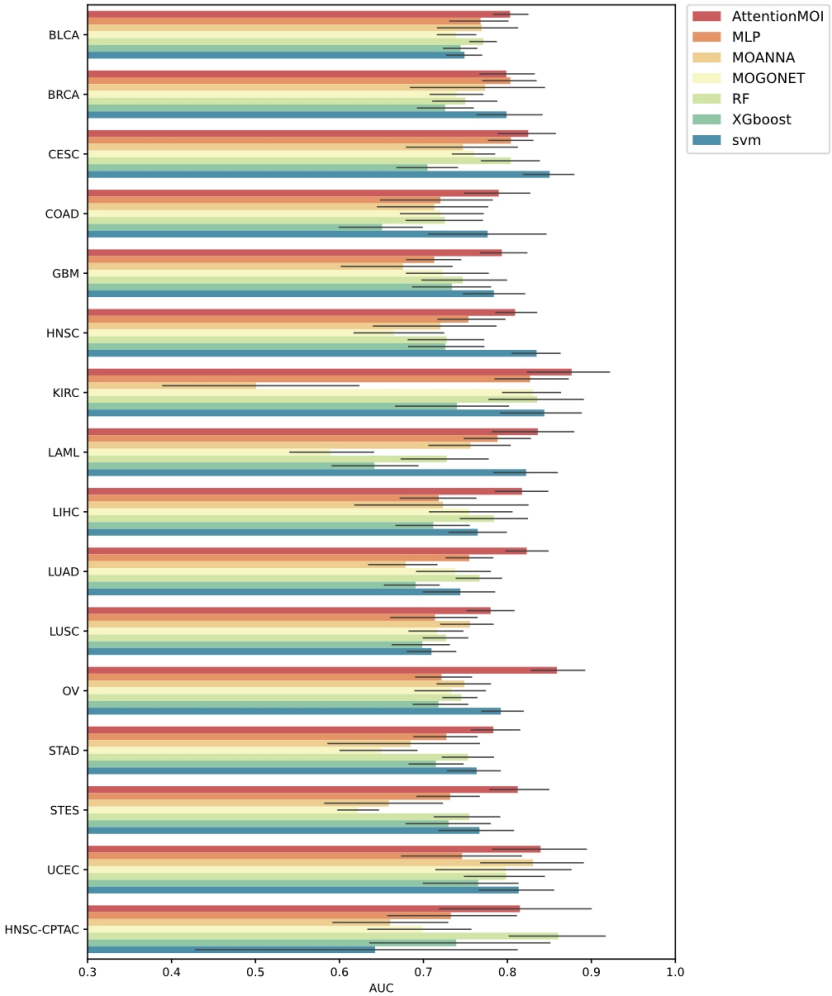
- When FSD threshold became smaller, the number of features became larger
- when only hundreds of features were selected for the model, traditional machine learning methods tend to perform better
- **AttentionMOI performed better when feature number is large**

Results: 5.AttentionMOI improved model performance II



- To further evaluate the performance of AttentionMOI, compared it with ML algorithms and current DL multi-omics integration algorithms, MOANNA and MOGONET(under threshold=0.2).
  - MOANNA is an Autoencoder-based framework.(2023)
  - MOGONET is an GCN-based framework.(2021)

Results: 5.AttentionMOI improved model performance III



Among 15 TCGA cancer types, AttentionMOI outperformed other models with higher AUC in 12 cases

Figure: AUCs obtained from different models under FSD threshold=0.2

Conclusion

Conclusion I

The key contributions of this framework

1. **Addressing the Curse of Dimensionality and noisy features with FSD module:** by selectively choosing biologically relevant features

2. **Robust Multi-Omics Integration with AttentionMOI:** improves the integration process by weighting the influence of different omics data.

Conclusion II

Limitations of this framework

1. **Modality missing:** If the DNA module or RNA module is missing, the framework will resemble an MLP model, losing interactions of omics data.

2. **Interpretation of the model:** Features from different omics are interpreted separately, but they may contribute to certain phenotype together while this model does not take the joint functions into consideration.

Conclusion III

Future Research Prospects

1. **Modality missing:** stagewise pretraining by single modality data with masked data modeling or crossmodality representation of multi-omics data may reduce the dependence of the model on the data

2. **Interpretation of the model:** as biological functions are realized in regulatory networks, it's possible to represent the associations among omics using network to develop more interpretable models for multi-omics integration

References

[1] Girish Chandrashekar and Ferat Sahin. "A survey on feature selection methods". In: *Computers & Electrical Engineering* 40.1 (2014), pp. 16–28.

[2] Frank J Massey Jr. "The Kolmogorov-Smirnov test for goodness of fit". In: *Journal of the American statistical Association* 46.253 (1951), pp. 68–78.

[3] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).

[4] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic Attribution for Deep Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, June 2017, pp. 3319–3328.

Thank you