Carnegie Mellon University

Survival Stratification for Colon Cancer via Multi-omics Integration

02718: Computational Medicine

Wanzi Xiao, Xueke Jin
MS Computational Biology
wanzix, xuekej@andrew.cmu.edu

Introduction

- Colon adenocarcinoma (COAD) is the second most commonly diagnosed cancer worldwide and is the second leading cause of cancer-related deaths

- Prognosis stratification in colon cancer helps to address cancer heterogeneity

- Multi-omics integration using deep learning algorithms has shown great promise for the survival prediction of breast cancer and neuroblastoma, but only few studies have been published in colon cancer

- Recent work including
  - OmiVAE: Variational Autoencoder for integrated multi-omics (di-omics) analysis of pancancer
  - OmiEmbed: A Unified Multi-Task Deep Learning Framework for Multi-Omics Data

Data Resource

Dataset Info	Colon Cancer	
Source	UCSC Xena data portal, TCGA	
Additional label	Vital_status, days_to_birth, days_to_death, stage	
Omics type	Gene expression RNA-seq	miRNA expression
Feature number	54186	1881
Sample number	513	462

Table 1. Overview information of the colon-cancer dataset

Data Preprocess

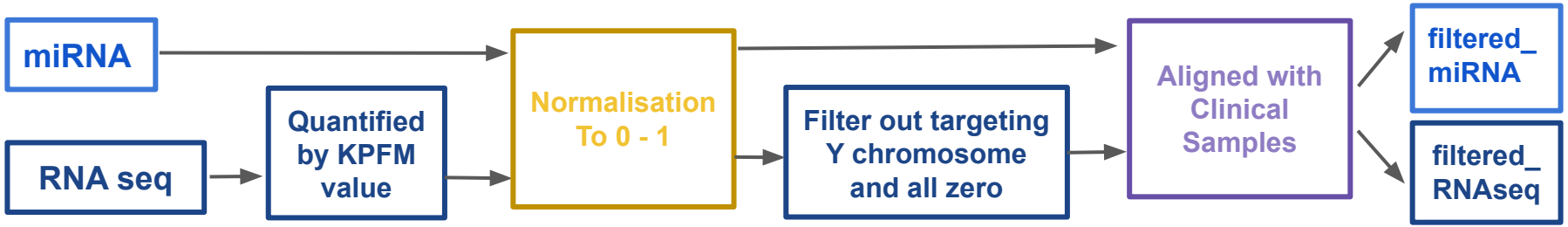
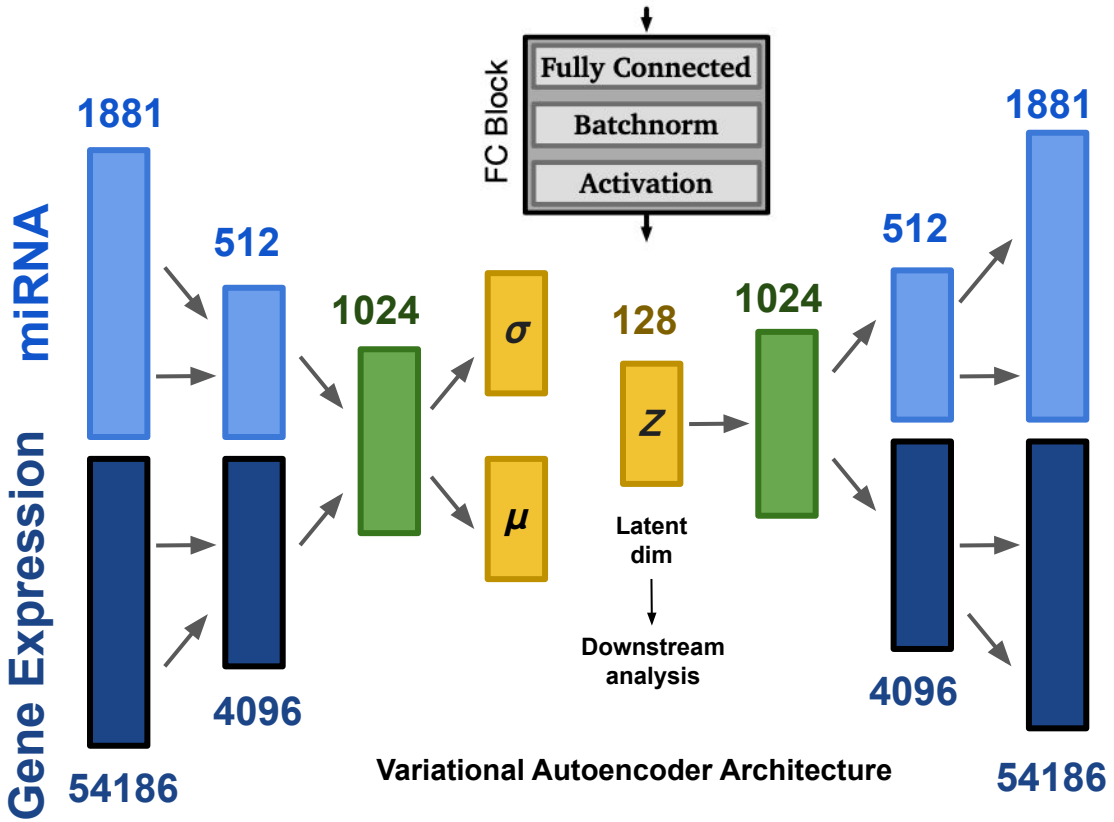


Figure 2. Data preprocess workflow

Methods



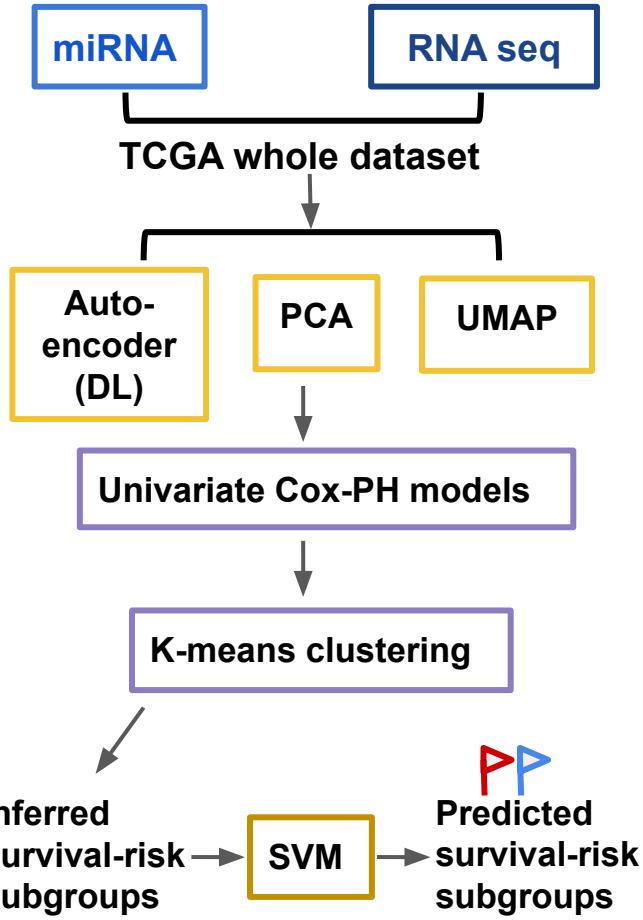


Figure 1. Overview of method workflow

Downstream Analysis

1.Feature selection

- Cox Proportional Hazards Model: Used to select significant features based on their p-values.

$$h(t, X) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)$$

2. Clustering

- K-means Clustering: Identifies subgroups within the data.
- Silhouette Score and Calinski-Harabasz Score: metrics help determine the optimal number of clusters.

$$\text{WCSS} = \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2$$

3. Classification

- 5 fold cross-validation with Support Vector Machine (SVM) with GridSearchCV: GridSearchCV helps find the best parameters for the SVM, optimizing its performance.

4. Model Evaluation

- Concordance Index(c- index): This metric evaluates the model's predictive accuracy, particularly in survival analysis.

Downstream Analysis

5. Survival Analysis

- Kaplan-Meier Survival Curves: Used to estimate the survival function from lifetime data: $S(t) = \prod_{t_i \leq t} \frac{n_i - d_i}{n_i}$
- Log-rank Test: Assesses the survival distribution of two or more groups.
- Cox Proportional Hazards Model: Fits the final survival model to the data, allowing for the estimation of hazard ratios.

6. Risk Scoring and Grouping

- Patients are assigned risk scores and categorized into risk groups: $\text{Risk Score} = \exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)$

7. ROC Curve and AUC

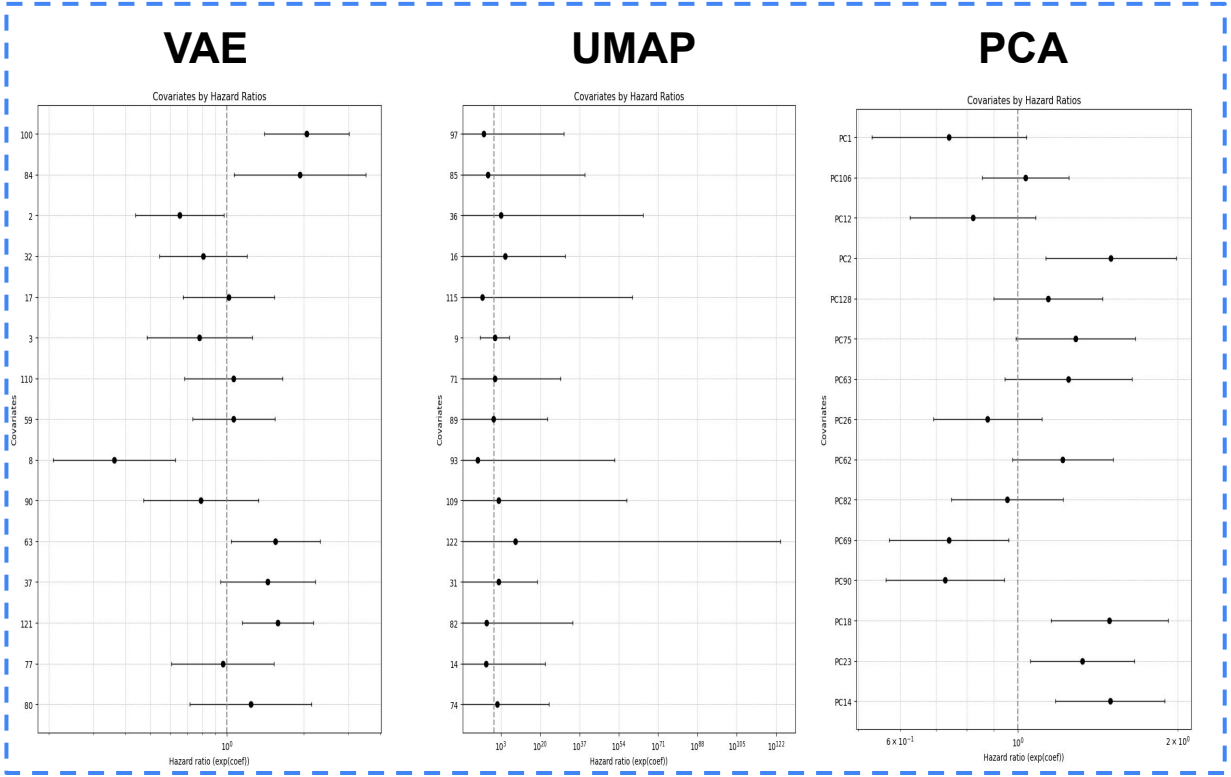
- The Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) provide a measure of the model's diagnostic ability, used to evaluate the binary classification performance.

8. Correlation Analysis

- Analyzing the correlation matrix of gene expressions (or other features) can uncover relationships between variables, important in understanding disease mechanisms.

Results and Analysis

1. Hazard Ratio Forest



- The PCA appears to provide the most stable hazard ratios with relatively narrow confidence intervals, suggesting a more reliable estimation of risk.
- VAE and PCA offer more interpretable results, with hazard ratios on a more conventional scale.

Figure 3. Hazard ratio forest comparison of three models

2. Kaplan-Meier Curve

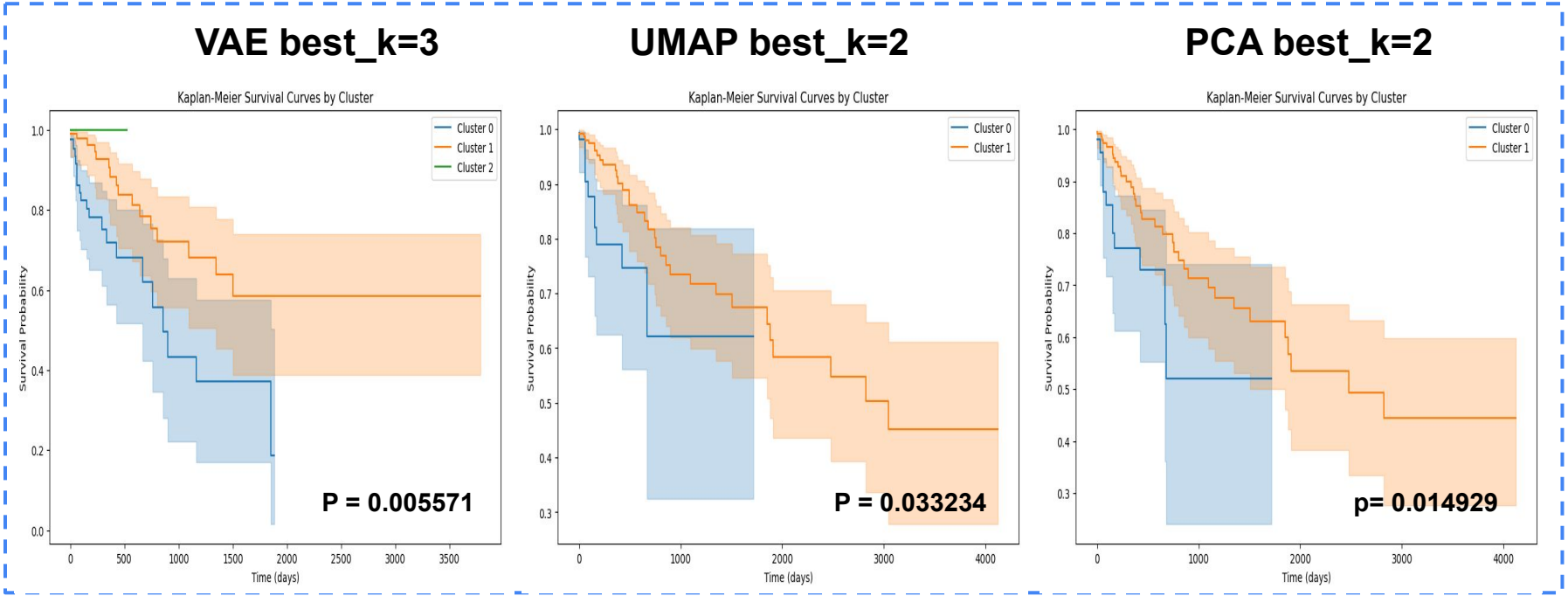


Figure 4. Kaplan-Meier Curve comparison of three models

- Cluster Number: The presence of an additional cluster in the VAE model could mean it has captured a more nuanced stratification of the data.
- Survival Probability: a cluster that has a better survival probability than the others.
- Statistical Significance: The VAE model has the lowest p-value, indicating the strongest statistical evidence for differences in survival between clusters.

3. Risk Score and Group Visualization

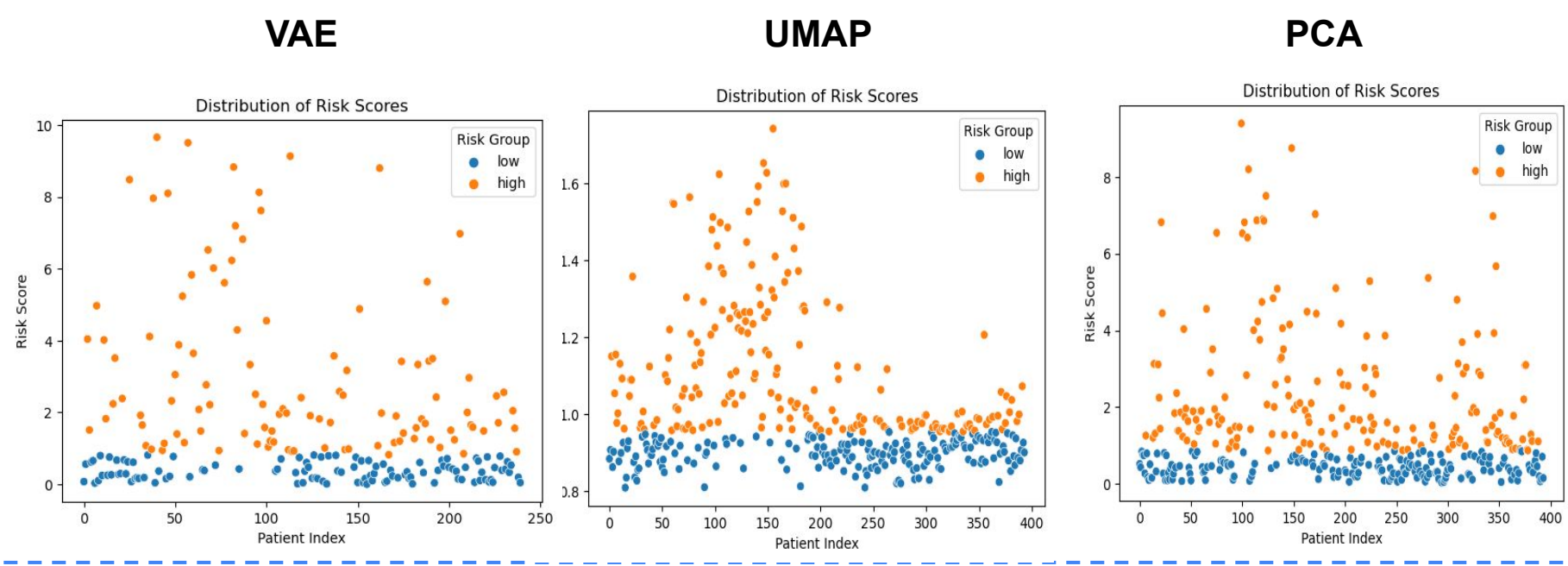


Figure 5. Risk score comparison of three models

- Risk Score Range: The VAE and PCA models show a similar range of risk scores, whereas the UMAP model presents a narrower range.
- Risk Group Separation: VAE and PCA seem to allow for a broader spectrum of risk within the high-risk group, which could be useful for identifying patients at very high risk.

4. ROC Curve

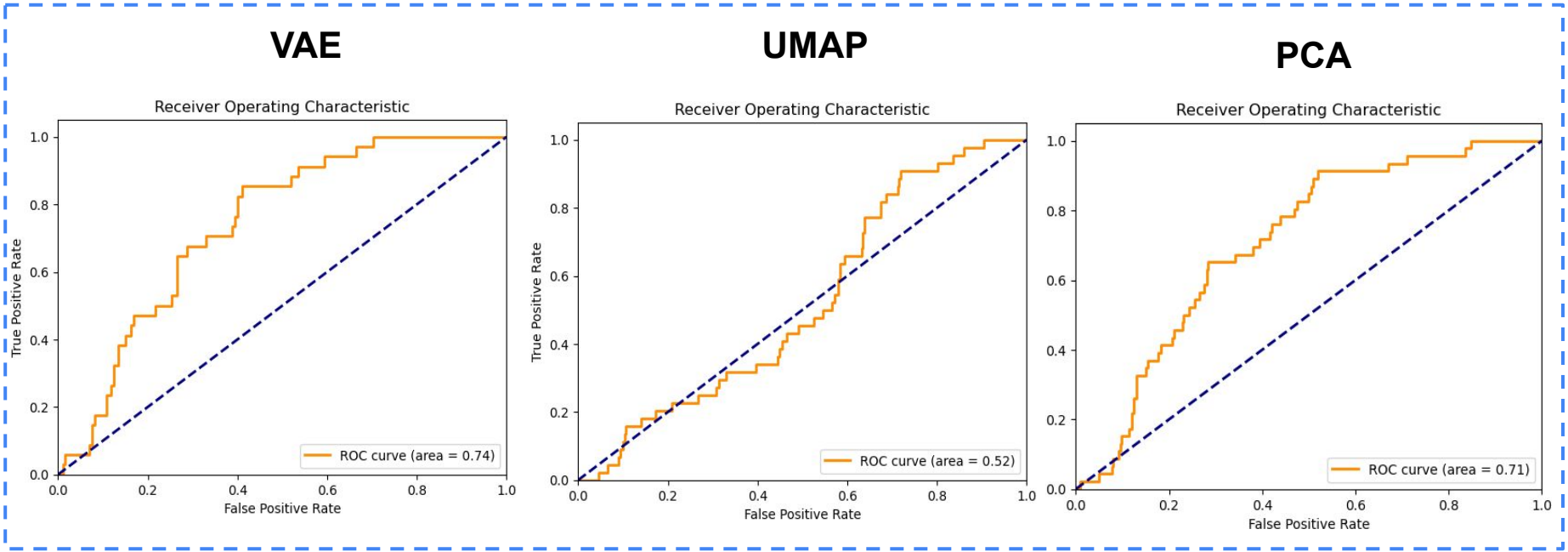


Figure 6. ROC curve comparison of three models

- Model Discrimination: The VAE model appears to have the best performance with the highest AUC, followed closely by the PCA model.
- Classification Thresholds: The shape of the ROC curves implies that different thresholds for classification could be optimized for each model to potentially improve their performance.

5. Correlation Matrix

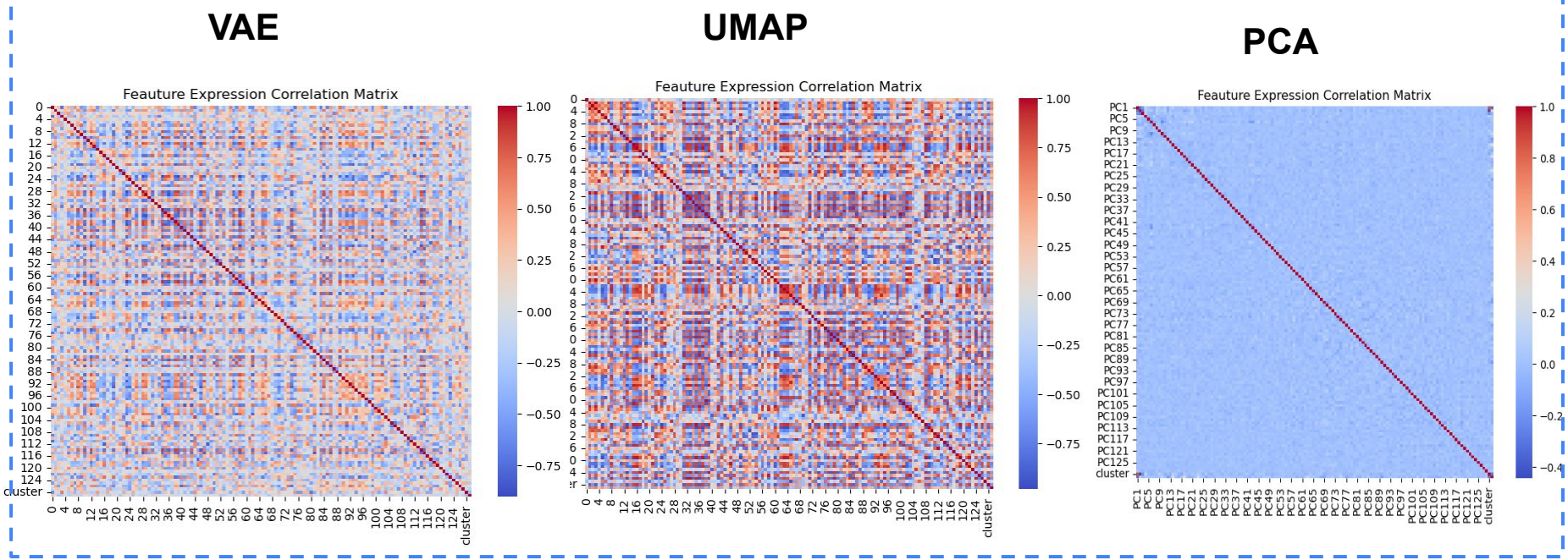


Figure 7. Correlation matrix comparison of three models

Complexity of Feature Relationships:
The VAE and UMAP models preserve complex relationships within the data, as indicated by the variety of correlation strengths. PCA, by design, reduces this complexity by transforming the data into principal components that are uncorrelated.

Thank you!

Reference

- Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. Clin Cancer Res. 2018 Mar 15;24(6):1248-1259. doi: 10.1158/1078-0432.CCR-17-0853. Epub 2017 Oct 5. PMID: 28982688; PMCID: PMC6050171.
- Zhang X, Xing Y, Sun K, Guo Y. OmiEmbed: A Unified Multi-Task Deep Learning Framework for Multi-Omics Data. Cancers (Basel). 2021 Jun 18;13(12):3047. doi: 10.3390/cancers13123047. PMID: 34207255; PMCID: PMC8235477.
- Yu T. AIME: Autoencoder-based integrative multi-omics data embedding that allows for confounder adjustments. PLoS Comput Biol. 2022 Jan 26;18(1):e1009826. doi: 10.1371/journal.pcbi.1009826. PMID: 35081109; PMCID: PMC8820645.
- X. Zhang, J. Zhang, K. Sun, X. Yang, C. Dai and Y. Guo, "Integrated Multi-omics Analysis Using Variational Autoencoders: Application to Pan-cancer Classification," 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 2019, pp. 765-769, doi: 10.1109/BIBM47256.2019.8983228.