Survival-Related Clustering of Cancer Patients by Integrating Clinical and Biological Datasets

Xinming Wei
Marquette University

SURVIVAL-RELATED CLUSTERING OF CANCER PATIENTS BY INTEGRATING CLINICAL AND BIOLOGICAL DATASETS

By

Xinming Wei

A Thesis Submitted to the Faculty of the Graduate School,
Marquette University,
in Partial Fulfillment of the Requirements for
the Degree of Master of Science

Milwaukee, Wisconsin
August 2020

ABSTRACT

SURVIVAL-RELATED CLUSTERING OF CANCER PATIENTS BY INTEGRATING CLINICAL AND BIOLOGICAL DATASETS

Xinming Wei

Marquette University, 2020

Subtype-based treatments and drug therapies are essential aspects to be considered in cancer patients' clinical trials to provide appropriate personalized therapies. With the advancement of the next-generation sequencing technology, several computational models, integrating genomic and transcriptomic datasets (*i.e.,* multi-omics) in the prediction of subtype-based classification in cancer patients, were emerged. However, integration of the prognostic features from the clinical data, related to survival risks with the multi-omics datasets in the prediction of different subtypes, is limited and an important research area to be explored. In this study, we proposed a data integration pipeline with the prognostic features from the clinical data and multi-omics datasets to predict the survival-risk-based subtypes in Kidney Renal Clear Cell Carcinoma (KIRC) patients from The Cancer Genome Atlas (TCGA) database. Firstly, we applied an **unsupervised clustering algorithm** on KIRC patients and clustered them into two survival-risk-based subgroups, *i.e.,* subtypes. Then, using the clustering-based subtype labels as class labels for cancer patients, we trained a **supervised classification** model to determine the class label of un-labeled patients.

In our clustering step, we applied **multivariate** Cox Proportional Hazard (Cox-PH) model to select the survival-related prognostically significant features (*p-value < 0.05*) from the patients' multivariate clinical data. Then, we used the *Silhouette Coefficient* to determine the optimal number (*k*) of the clusters. In our classification step, we integrated high dimensional multi-omics datasets with three different data modalities (such as *gene expression, microRNA expression, and DNA methylation*). We utilized a dimension-reduction approach, followed by a **univariate** Cox-PH for each reduced data modality with patients' survival status. Then, we selected the survival-related reduced-omics-features in our classification model. In this step, we applied a supervised classification method with 10-fold cross-validation to check our survival-based subtype prediction accuracy. We tested multiple machine learning and deep learning algorithms in different steps of the pipeline for clustering (*K-means, K-modes and, Gaussian mixture model*), dimension-reduction (*Denoising Autoencoder and Principal Component Analysis*) and classification

(*Support Vector Machine and Random Forest*) purposes. We proposed an optimized model with the highest survival-specific-subtype classification accuracy as the final model.

ACKNOWLEDGMENTS

Xinming Wei

This work would not have been possible without the financial support of a grant from the National Institutes of Health. I would especially like to thank my advisor, Dr. Serdar Bozdag, the Director of Bioinformatics in the Department of Computer Science, for giving me a lot of help and guidance during my two-year Marquette career. In this long and short study, his patience and professionalism became an important reason for achieving this goal. As my advisor and mentor, he taught me more than academics. I wish all the best for him in his new position.

I also want to thank the people who have taught me in other courses and projects. I want to thank the people who helped me in these two years of study. I would like to thank Dr. Maadooliat, as my academic advisor, who gave me useful advice and recommendation during studying at Marquette. I would particularly thank Banabithi, who have given me much guidance and help in my project and other studies with her professional knowledge. And I would also appreciate Ziynet, Cagatay, and Jubair, as my friends more than lab members.

No one is more important in my life and in pursuing this degree than my family. Without the support of my parents, there is no way I could have achieved this goal. Their love and understanding are my biggest motivation. Most importantly, I wish to thank my loving and supportive girlfriend, Xinrui, who provides unending inspiration and companionship.

TABLE OF CONTENTS

CHAPTERS

I. INTRODUCTION

A. General Introduction

Cancer is a complex genetic disease that is the most intractable medical and health problem in the world [1]. With the advances in high throughput sequencing technologies, nowadays, we could collect much data to study the diagnosis and prognosis of cancer. Cancer can be divided into different subtypes according to cell morphology and cell period; different subtypes of cancer might have different survival characteristics in addition to different molecular expressions. For decades, one of the most popular areas in cancer studies has been to explore patterns of molecular data sets to cluster and treat different subtypes of specific cancer. There are several studies on clustering cancer patients, integrating multi-omics data, specifically, gene expression, DNA methylation, and micro RNA (miRNA) expression [2, 3, 24]. However, studies on clustering cancer patients with prognostic variables present in the clinical datasets are limited. Therefore, we propose a clustering model to apply clinical data that was survival-related to cluster cancer patients, followed by a classification approach using multi-omics datasets to determine the cluster label of the new cancer patients.

In our clustering step, we applied a multivariate Cox-PH model to select the survival-risk-related prognostically significant clinical features and determined the survival-related clusters. We tested multiple clustering algorithms, such as K-means, K-modes, and Gaussian mixture model (See Chapter I.C).

In our classification step, we integrated the high dimensional multi-omics datasets with three different data modalities (such as *gene expression, miRNA expression, and DNA methylation*). We utilized a dimension-reduction approach, followed by a univariate Cox-PH for each reduced data modality with patients' survival status.

To reduce the dimension of the multi-omics datasets, we tested different algorithms, such as denoising autoencoder [5], Principal components analysis (PCA), and PCA-surv [6] (See Chapter I.C). For the classification purpose, we tested two different machine learning methods, such as support vector machine (SVM) and random forest [7] (See chapter I.C).

Among these different algorithms, we selected K-modes for clustering, PCA-surv for dimension reduction, and SVM for classification, respectively, that produce an optimized model in terms of classification accuracy of our survival- specific subtypes prediction.

We tested our pipeline on Kidney Renal Clear Cell Carcinoma (KIRC) cancer from the TCGA database [8]. We integrated KIRC clinical data along with gene expression, DNA methylation, and miRNA expression datasets in the pipeline (See Chapter III.B for details). The datasets were downloaded using the *TCGAbiolinks* package in the R programming language [9].

The purpose of this work was to establish the importance of applying clinical data in the prediction of cancer subtypes, along with multi-omics datasets. We also assessed the predicted survival-related subtypes under different clinical conditions. Our results suggested that cancer stages of the patients play a more prominent role in clustering analysis than patients' age.

B. Research Problems

Research by integrating multi-omics datasets have been increasingly applied in the diagnosis and prognosis of cancer [24]. However, integration of the prognostic features from the clinical data, related to survival risks with the multi-omics datasets in the prediction of different subtypes, is limited and an important research area to be explored. Therefore, in this study, we proposed a clustering model applying clinical data that was survival-related to cluster cancer patients, followed by a classification approach using multi-omics datasets to predict the cluster label of new patients. We utilized Cox-PH to find prognostic clinical variables and clustered cancer patients based on these variables. After integrating multi-omics datasets, we used multi-omics features to determine the cluster label of un-labeled patients.

C. Definition and Explanation of Key Terminology

TCGA KIRC data: The Cancer Genome Atlas is a database that provides multiple cancers patients' genomic and clinical data [10, 14]. Large-scale genome sequencing of over 10,000 samples in more than 30 types of cancers was performed. The KIRC data is one of those datasets to record the data of kidney renal clear cell carcinoma (KIRC) patients.

Univariate Cox-PH: Univariate analysis using the Cox regression technique is applied when there is a single, independent, potentially survival-related variate. The variables that are related to survival are selected respectively as independent features.

Multivariate Cox-PH: Multivariate analysis using the Cox regression technique is applied when there are multiple, potentially survival-related covariates [11]. DR Cox

proposed it in 1984, and the concept of multivariate Cox-PH was proposed in 1996 [40]. This concept can be applied to select survival-related data from multiple variables.

K-means: The K-means is an iterative clustering algorithm [41]. First, K data points are chosen randomly as cluster centers. Each data is divided into clusters around the center at the minimum distance. In each iteration, the distance between the sample and the center is recalculated, and the sum of distances to the center reaches the minimum at the final stage.

GMM: The Gaussian mixture model (GMM) is a probabilistic model that assumes that all the data points are generated from a mixture of Gaussian distributions. In probabilistic modeling, the probability distribution over all the discovered clusters is inferred for each observation. [42].

K-modes: The K-modes algorithm is an extension of the K-means algorithm. It is a multi-iteration clustering algorithm. It is widely used when applying categorical data in clustering, which is not available in K-means [43]. For example, in our case, genders, races, cancer stages in the clinical data, are such categorical variables that are not available in K-means to compute the distance. K-modes uses dissimilarity measure as the distance measure for clustering. First, K samples were chosen randomly as cluster centers, and each sample was divided into clusters around the center at the minimum number of mismatched features. The fewer the mismatched features between the sample and the cluster center indicates the smaller the distance between them. Through multiple iterations, the total number of mismatched values between the sample and the cluster center is minimized to determine which cluster the samples belong to.

Rank normalization: Rank normalization is a pre-processing method on data for quantile normalization. It is based on the premise of not losing parameter information and avoiding the impact of specific extreme values on the entire data. The normalization replaced each observation with a ranking in the matrix, divided by the total number of features [12, 13]. We applied the rank normalization on the gene expression, DNA methylation, and miRNA expression data matrices downloaded from TGCA.

Denoising autoencoder: The denoising autoencoder can reduce the noise in the original data by reconstructing the input data. The function of the denoising autoencoder is to learn from the original data with the superimposed noise. The features it learns can be decoded to the output almost the same as the original input data. The obtained feature from the hidden nodes is robust by reducing the noise in the data from extreme values and learning the same feature value.

PCA: Principal Component Analysis (PCA) is a statistical method to reduce the dimension of data that contains sufficient information. By transforming data that may have a linear correlation into linearly uncorrelated components, all the components selected according to the proportion of variance are called principal components. A few principal components reveal the internal structure among multiple variables with a threshold of with the Proportion of Variance. By performing this step, a few principal components are derived from the original variables. They retain as much information as possible from the original variables and are not linearly correlated.

PCA-surv: The components obtained by principal component analysis are utilized in survival analysis. By using Cox-PH, components related to survival were selected as new features.

SVM: Support Vector Machine (SVM) is a widely used classification model. It divides the feature values in space by the maximum distance; its decision method uses maximum-margin hyperplane to maximize the distance between the features. SVM can be used for nonlinear classification by kernel method, one of the conventional kernel-learning methods [7, 14].

RF: Random Forest (RF) is a classification model based on the decision tree. Essentially, the random forest is an integrated learning model that uses decision trees for multiple iterations. Through multiple decision combinations to solve the prediction problem brought by every single decision tree, the model can generate multiple classifiers, each of which can independently predict and learn.

D. Context of Research Study within the Greater Discipline

The future challenge of bioinformatics is to comprehensively understand the systematic role of molecular information, which can be achieved by studying multi-omics data simultaneously. Integrating multiple types of data of cancer patients into deep-learning algorithms and other analysis algorithms may resolve our current knowledge gap in molecular mechanisms, the interaction between genes and the environment, and the vertical effects of cancer development [15, 16]. Multi-omics data integration methods can improve the understanding of genetic diseases and cancer and may lead to new strategies for early diagnosis, prognosis, and treatment of human diseases. At present, there are many studies on multi-omics, and as well as comprehensive methods and frameworks in integrative analyses [18, 19, 20]. Moreover, Sun and Hu summarized the analysis methods of high-throughput multi-omics data. They provided an updated, comprehensive method and

framework to integrate genome, epigenome, transcriptome, proteome, and metabolome data into the emerging field of multi-omics research in human diseases [37]. These findings reveal the diversity of biological systems and can identify biomarkers and gene loci to a certain extent. The expressions of multi-omics data are similar, and it is possible to discover similarities by integrating features through deep-learning algorithms. High-throughput experimental methods can provide various datasets to study multi-omics data simultaneously.

II. HYPOTHESIS (THEORY)

A. Brief Overview of Theoretical Foundations Utilized in the Study

Diseases are expressed differently at different stages and in different populations [21]. In survival analysis, the samples by integrating prognostic clinical data of different populations may have patterns in different clusters. For example, different age ranges, different stages of the disease, different races, different expression levels of physical indicators are factors that affect prognostic characteristics. Therefore, clinical data related to survival were selected through Cox-PH analysis, and these variables were used for clustering samples into different survival clusters with survival differences.

On the other hand, multi-omics data, such as gene expression data, DNA methylation data, microRNA expression data also have similar survival-related characteristics. In previous studies, Hao et al. proved that patients could be clustered by integrating multi-omics data by conducting different dimensionality reduction processing and extracting corresponding features [22].

We integrated the multi-omics data into low-dimensional, representative, new features related to survival, classifying them according to the previous clustering, and verifying the reliability of clustering based on prognostic clinical data. Since many previous and future studies cluster based on multi-omics data, there is evidence to verify clustering results by multi-omics data [23].

Although it is not always possible to achieve complete agreement by comparing the two types of data, we expect to find correlations between prognostic clinical data and multi-omics data, such as gene expression, through this verification process. Through further exploration and improvement of the algorithm, it will better handle the clustering, diagnosis, treatment, and prognosis of cancer patients and provide a more comprehensive understanding of the analysis of cancer.

B. Literature Reviewed, Discussed and Applied

Overview of cancer

Many cancers (i.e., BRCA, KIRC, AML, etc.) can be divided into several subtypes based on characteristics of the development stage and cancer cell morphology. Therefore, different cancer subtypes may have different levels of gene expression or survival patterns. In a recent study, Rappoport et al. have given a comprehensive comparison among different methods (i.e., iCluster, SNF, rMKL-LPP, K-means) to cluster cancer patients [24]. Similarly, we can cluster cancers into subtypes based on different survival characteristics and look for associations between specific groups and different subtypes.

Integrating prognostic clinical data

As early as 1976, Solberg et al. began to explore the use of cancer clinical trial data to perform cluster analyses on patients [26]. They found that the clustering results based on the results of clinical trials had a steady correspondence with the results of the biopsy. Studies have also shown that gene expression analysis can be used to predict the clinical outcome of cancer [27], which indicates that the clustering features of genomics are related to clinical data and prognostic signatures. In the research of other diseases, cluster analysis of clinical data is also used to identify the subgroup of fibromyalgia [28]. However, we proposed a pipeline based on Poirion et al. in 2018 while not widely discussed to cluster patients related to survival analysis, especially the clustering results derived from clinical data related to survival.

Survival analysis

The predictive model of survival time is a standard tool for cancer prognosis survival analysis [29], while the Cox-PH model is the most commonly used survival prediction model for cancer-survival prediction [4]. The survival risk and actual survival time of each patient can be determined through different performance indicators and clustering different indicators can find the consistency of the survival risk and survival time. We clustered the patients to determine their survival risk subtypes according to their survival status and time.

Research has shown that in the field of deep learning, identifying complex multi-omics data interactions related to patient survival time and risk at the molecular level is important. It is not only for the development of new diagnostic and therapeutic methods but also for accurate survival predictions [31]. Hao et al. developed a new path-based sparse deep neural network (PASNet) for cancer survival analysis. Besides, integrating survival-

related clinical data from cancer samples is expected to improve cancer survival prediction and diagnosis. Specifically, the integration of multi-omics data and clinical data can be used for survival prediction and diagnosis in cancer research. It also provides an in-depth understanding of cancer and multi-omics data by comparing the performance of the current state-of-the-art model with TCGA cancer data and statistically evaluating the outstanding performance. There is also relevant biological literature that supports the biological interpretation of the PASNet.

Furthermore, the integration of multi-omics data in another study by Hao et al. have demonstrated an understanding of the complex mechanisms of human genetic diseases and cancer and provides excellent help for precise medication and treatment [22]. In this study, they proposed a deep neural network based on genes and pathways for multi-omics data integration (MiNet) to predict cancer-survival outcomes. The study integrates multi-omics regulation(i.e., genomics, epigenomics, and transcriptomics) in a neural network based on genes and pathways. This provides a specific classification model for 10-fold cross-validation in the survival-related analysis.

A recent study [32] described the challenges of deep learning in cancer survival analysis, such as dealing with different types of multi-omics data and hugely different sample sizes. Improving the deep-learning model will integrate multiple types of data, predict the survival results and survival time of samples, and maximize the correlation among multi-omics data, clinical data, and survival analysis. This research can provide a better perspective on cancer drug treatment and survival analysis [25].

Integrating multi-omics data

Human diseases and cancer involve the influence of multi-omics interaction and expression and are affected by environmental factors, as well. Many studies have focused on multi-omics research [37]. At the molecular level (i.e., genetics, epigenetics, and transcriptomics), recent technological advances have allowed integrated analysis of the human genome, epigenome, and metabolome at the population level. Complex and dynamic molecular networks are involved in human diseases. High-throughput technology enables omics research to allow us to obtain evidence of disease diagnosis in different omics data. However, single omics research can only provide a limited understanding of the molecular mechanisms of cancers.

In recent years, the diagnosis of cancer has not only been limited to a single omic study, while the diagnosis of a single omic can only provide limited information of the disease. Multi-omics research has gradually been applied to the diagnosis and treatment of cancer. However, multi-omics research applied to prognosis work is still minimal. In a recent study [30], somatic mutations, DNA copy number, DNA methylation, gene expression, and miRNA expression were used in prognostic studies of multi-omics features. Zhu et al. found that mRNA and miRNA expression profiles are the best in prognosis and diagnosis, followed by DNA methylation. They also stated that kernel machine-learning methods always outperform prognostic signatures. In another study [33], although nothing was summarized about disease clustering and prediction, a framework for the biological and functional roles of multi-omics was built. In addition to genomics, transcriptomics, epigenomics, proteomics, and metabolomics data have also been discussed. However, multi-omics research can broaden the complexity of cancer research and improve the accuracy of cancer diagnosis and prognosis.

The complexity of tumor genes has proven to be an enormous challenge to the diagnosis and prognosis of cancer. Single omics cannot provide a complete and comprehensive evaluation basis. As the most commonly used genomic and transcriptomics study in prognosis, tens of thousands of genetic variations, including SNPs and DNA copies, are linked to various human diseases through GWAS [34]. In genomic and transcriptomics studies, next-generation sequencing technology [36], including RNA-seq [35] methods, is the most commonly used means of providing genetic variation. In the evaluation of cancer prognosis, the epigenetic group, which refers to heritable molecular modification and mainly includes DNA methylation, is also a critical evaluation characteristic [3].

As far as GWAS is concerned, although tens of thousands of SNPs have been identified for disease diagnosis and identification, the functional meaning and mechanism of related loci are still unclear. Besides, mutations in the genome alone cannot predict disease risk throughout the disease cycle [38]. Based on the success of single-omics discovery research, a multi-omics approach integrates data obtained from different omics to understand their interactions and impact on the disease process. Sun et al, summarized the main omics methods available in population studies and reviews the deep-learning methods that integrate multi-omics layers, which provide a better perspective for gene discovery and functional analysis of human diseases.

In another study, Zhang et al. clustered patients with high-risk neuroblastoma according to clinical information and prognostic results [39]. However, there is still a lack of survival risk analysis for high-risk neuroblastoma. To make up for the gap in survival-risk analysis, they used a deep-learning algorithm, autoencoder, to compare with PCA to

understand clustering based on multi-omics data integration. They used the K-means clustering method to identify two with distinct survival subtypes of risk differences. The results showed that with the methods and datasets they examined, classification based on autoencoders was superior. Zhang et al. also verified the feature selection of autoencoder for high-risk neuroblastoma through two independent data sets, which helped to control survival risk better and proved that deep-learning-based algorithms are very useful in multi-omics integration.

C. Hypotheses and Justifications Tied to Prior Sections or Statements

By studying the previous research results, we understand that the integration of multi-omics research plays an important role in the clustering of cancer patients. We expect the integration of multi-omics data to verify the characteristics of survival analysis in clinical data. We hypothesize that there are similar survival characteristics in clinical data and multi-omics data. We selected variables closely related to survival through Cox-PH, clustered the cancer samples into different survival risks subtypes, and then used the integrated multi-omics data in the classification model to verify our clustering results. We tested multiple machine learning and deep learning algorithms in different steps of the pipeline for clustering, dimension-reduction, and classification purposes and proposed an optimized model with the highest survival-specific-subtype classification accuracy as the final model.

D. Theoretical Assumptions and Limitations

Based on previous studies and known theories, we assumed that the clinical data of cancer patients would show specific characteristics related to survival and particular

clustering phenomenon in different ages, disease stages, and other clinical data. The survival-related clustering is somewhat similar in the integration of multi-omics data, such as gene expression, DNA methylation, and microRNA expression.

It is very promising to obtain high compliance results by using multi-omics data to verify the clustering results based on clinical data. Meanwhile, the results may not be high consistency between clinical and biological datasets. The processing of multi-omics data may have more aspects to improve. Therefore, based on the existing processing methods and algorithms, the results verified by two kinds of data could be improved in advance.

III. METHODS

A. Introduction and General Description, Study Methods, and Study Design

In this project, we clustered cancer patients by integrating survival-related clinical data to predict survival-specific clusters. The multivariate Cox-PH algorithm was used to select variables significant to survival, meaning *p*-values are less than 0.05. By applying those variables, multiple clusters can be predicted by various clustering methods, such as K-means, K-modes, and GMM. Thus, we can cluster patients into different groups based on their survival characteristics. Classification is done by integrating multi-omics datasets, such as gene expression, DNA methylation, and microRNA expression.

Gene expression, DNA methylation, and miRNA expression data are enormous datasets. They contain a large amount of information. Moreover, studying on such datasets is very meaningful and promising to obtain useful information for diagnosis. However, analyzing such a large amount of data requires much time and large calculation memory. Through dimensionality reduction of the data sets, we can reduce the required storage space,

speed up the calculation (such as in machine learning algorithms), remove redundant features, and avoid overfitting. Lower dimensions require less calculations. Lower dimensions data can be applied to the algorithms that are not suitable for high dimensions data. There are many alternative methods available to reduce the dimension of the multi-omics data. For example, denoising autoencoder (DAE) is a neural network that can apply multiple layers and multi-omics datasets as original parameter settings. Moreover, PCA is also an alternative method for dimension reduction. By using these algorithms, we could reduce the dimensions of the massive amounts of data in mRNA, methylation, and miRNA datasets from TCGA. For DAE, by using an autoencoder, we built a network with 100 hidden nodes (h = 100) to limit every single omic data to 100 new features. To use the PCA method for dimension reduction, components with Proportion of Variance greater than 0.01 were selected. On the other hand, an alternative option was to select components from the PCA method as new features that were survival-related, while p-values were less than 0.01 by using the Cox-PH analysis, which means there is a probability of 0.01 that the features are not related to survival.

For each omic, we built an individual model to select new features related to survival. For this step, individual Cox-PH was used to select those features related to survival from new matrices produced by dimension-reduction methods. The patients were then classified with these features by the clusters inferred to clinical survival-related variables. K-fold cross-validation was utilized for classification by using SVM with linear kernels and RF; and thus, to build a supervised classification model to verify whether there were common patterns with clinical data. Accuracy, specificity, and sensitivity were used

to evaluate the results. A confusion matrix was used to combine these parameters to have a comprehensive view of the results.

B. Datasets

We obtained the three omics datasets and a clinical dataset, including mRNA, miRNA, and DNA methylation, and clinical data of TCGA KIRC by using the TCGAbiolinks package in the R programming tool. In this project, 534 samples in clinical data were also obtained from TGCA for clustering. 317 samples of KIRC were obtained from TCGA for integrating and pre-processing. For mRNA expression data, we used Fragments Per Kilobase of transcript per Million mapped reads (FPKM) produced by high-throughput sequencing data (HT-seq) platform. For DNA methylation, the average methylation value of all the CpG sites was calculated from the Illumina Human Methylation 450K platform. For miRNA expression sequencing data, we used reads per million (RPM) normalized quantification values from British Columbia Genome Sciences Centre (BCGSC) miRNA profiling of the miRNA expression quantification dataset. For clinical data, we used clinical supplement data from TCGA KIRC projects. In 317 samples in multi-omics data and 534 samples in clinical data, we finally selected the overlapped samples, which contained 315 samples, with all types of data, including multi-omics data and clinical data.

C. Samples in TCGA Project

The TCGA project was launched in the United States in 2005 and aimed to apply genomic analysis techniques to study genomic changes in cancer. Large-scale genome sequencing of over 10,000 samples in more than 30 types of cancers was performed. It is

especially valuable that these samples have very detailed prognostic follow-up information [10].

TCGA contains the following data: 1) Clinical sample information: Biospecimen, Clinical; and 2) Sequencing data: The five methods of RNA-Seq, WXS, miRNA-Seq, Genotyping Array, and Methylation Array are mainly used to analyze samples.

RNA-Seq data in TCGA is transcriptome sequencing. The transcriptome data on TCGA uses full transcriptome sequencing, which contains various non-coding RNAs, so the generally downloaded RNA-Seq data contains lncRNA, mRNA, pseudogenes, etc. RNA-Seq quantitative expression data is currently available for public download in three forms: HT-Seq-FPKM, HT-Seq-UQ-FPKM, HT-Seq-Counts, where FPKM is used to measure the abundance of transcript expression, counts is counted on a gene in the reads sequenced, and UQ-FPKM is standardized FPKM by the upper quartile.

miRNA is a type of non-coding single-stranded RNA molecule that is approximately 22 nucleotides (nt) in length and is encoded by an endogenous gene. It is a significant type of non-coding small RNA in biology, and the regulation of organisms; about one-third of genes in humans are regulated by miRNA [44]. TCGA provided miRNA-Seq sequencing data results, using a database of miRBase v21. There are two main types of data currently available for public download: miRNA Expression Quantification and Isoform Expression Quantification, where the Isoform Expression Quantification data contains mature miRNA. TCGA provides quantitative data in Counts and FPKM formats.

In miRNA, pre-miRNA is a precursor miRNA, about 70–90 bases in length; pre-miRNA is digested by Dicer enzyme and becomes mature miRNA about 20–24 nt long. miRNA is generally mature miRNA, about 20–24 nt in length, developed from various

precursor miRNAs. The relationship between miRNA and target genes is, generally, miRNA regulates target genes and reduces the expression of the target genes.

DNA methylation is a reversible and inheritable process that can lead to changes in chromatin structure, DNA conformation, DNA stability, and the way DNA interacts with proteins to control gene expression. TCGA provides methylation chip data. There are two main types of DNA methylation: 450K and 27K. Generally, 450K is the most used. The methylation is mainly located on CpG sites. The methylation will regulate the expression of the genes.

CpG site is a site on the DNA sequence with a base of C or G. Methylated cytosines are primarily found at CpG sites. As for gene-promoter region. At present, there is no unified expression of the gene-promoter region. Generally, we think that the transcription starts site (TSS) of a gene is between 2 kb upstream and 500 bp downstream. CpG islands are generally considered that the regions where CpG sites are significant. The overall methylation level of CpG island regions is often low and frequently appears in the promoter region and exon region of genes. The relationship between methylation and genes is that hypermethylation in the promoter of a gene will downregulate the expression of its downstream genes, which is mostly negatively correlated.

TCGA provides a wealth of clinical follow-up information, including medication, relapse, age, survival, etc., which contains more than 100 variables. Commonly used clinical information includes: 1) age; 2) gender; 3) stages; 4) time of relapse; 5) overall survival, 6) race; and 7) ethnicity.

TCGA has a separate ID, also called barcode, for each patient, such as TCGA-02-0001. This ID is universal in the TCGA database. According to this ID, we can find the same patient in different types of data, including clinical follow-up information.

Different sampling sites of patients have different codes; for example, 01 indicates cancer tissue, 10 indicates adjacent tissue, 01 to 09 generally indicates tumor site, and 10 and above indicates normal control. Portion means different components of the same organization. Analyte represents the type of molecule analyzed; D represents DNA. The Center represents the detection center.

D. Algorithms and Methodology

Identification of the survival subtypes

We used Cox-PH regression to select survival-related variables through the multivariate method. In the survival analysis, the Cox proportional hazard model was used to select survival-related variables with a *p*-value that is less than 0.05 [4]. The optimal cluster number *k* is estimated by calculating the Silhouette score and the plot. And we clustered the obtained samples applying different clustering methods, such as K-means, K-modes, and GMM (Figure 1). We used the survival package in R [46] to analyze the survival difference among different survival risk subtypes in the TCGA KIRC dataset. We use a “survdiff” function that calculates the difference in survival between the two subtypes and plots a Kaplan-Meier curve with a log-rank *p*-value.

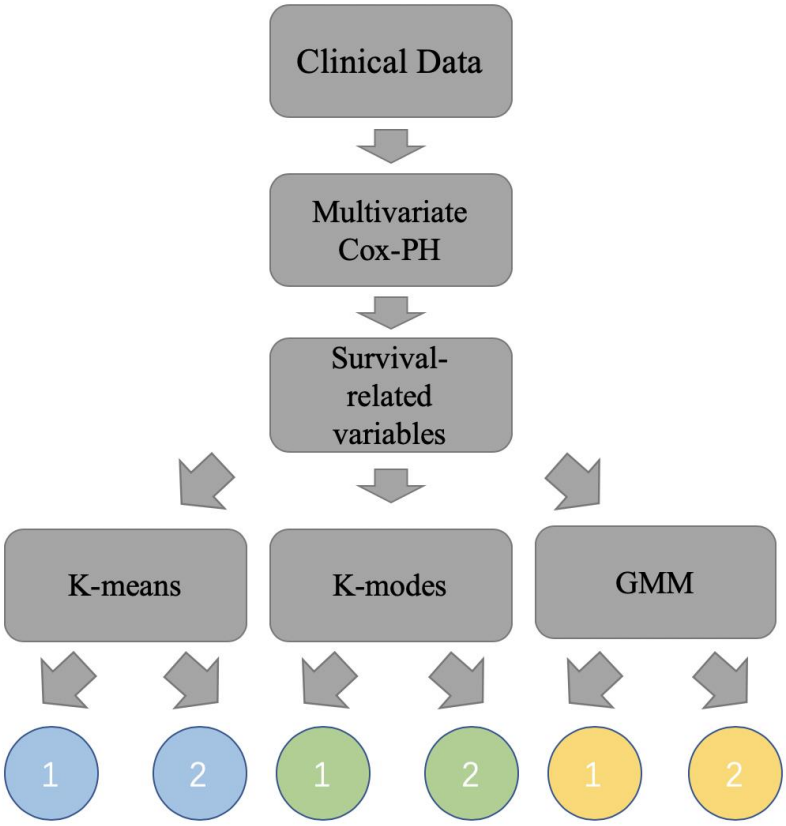


Figure 1. Clustering samples by using three clustering methods.

Regarding the K-means clustering algorithm with multiple iterations, the data is first randomly defined as K-centers, where K is the number of subtypes determined by silhouette score. The basic K-means algorithm flow is as follows; for the samples of the KIRC data, the sum of the squared error (SSE) is used as the objective function of clustering, so the clustering results can also be measured among sample with different variables.

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} dist(x, c_i)$$

This represents the sum of center c_i distances from sample x to cluster C_i; the clustering result should cause *SSE* to reach the minimum value.

The Gaussian mixture model is a probabilistic clustering method that assumes that all data samples x are generated by a mixture of k multivariate Gaussian distributions.

$$p(x) = \sum_{i=1}^{k} \alpha_i \cdot p(x|\mu_i, \sum_i)$$

where $p(x|\mu_i, \sum_i)$ is the probability density function of the n-dimensional random vector x following the Gaussian distribution.

$$p(x) = \frac{1}{(2\pi)^{\frac{n}{2}} |\sum|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

For discrete attribute data sets, calculating the cluster mean and the Euclidean distance between points becomes inappropriate. As an extension of K-means, K-modes is suitable for categorical attribute datasets.

Suppose there are n samples, m attributes are all discrete, and the number of clusters is k, which are randomly determined k clustering centers; and C_i is a vector of length m. When comparing the distance between each sample and k centers, this distance is the number of mismatched attribute values. All samples are divided into clusters that minimize the sum of all distances; after each sample is clustered, the clustering center is also re-determined and iterates in turn until each sample is clustered into k modes. Repeat the above steps until the total distance no longer decreases and the final clustering result is returned.

Normalization procedure

We first applied rank normalization to each omics data. For a given omic, we defined the input matrix $M = (v_1, \ldots, v_m)$ as a list of *m* sample vectors *v*, having *n* features in each matrix. For a given sample vector $v = (x_1, \ldots, x_n)$, the function *rank(x_i)* represents the ranking of each feature x_i in *v* (*n* represents the feature *x* as the highest value, and 1 represents the feature *x* as the lowest value). v_{rank} is defined by:

$$v_{rank} = (rank(x_1), \ldots, rank(x_n)) \cdot \frac{1}{n}$$

We then normalized $M_{rank} = (v_{rank1}, \ldots, v_{rankm})$ by computing the Pearson correlation coefficient between each pair of samples.

$$M_{corr}(i,j) = d_{pearson}(v_{rank\,i}, v_{rank\,j})$$

Thus, $M_{corr} = \{M_{corr}(i,j) \mid i,j \in m\}$ is an *m*×*m* matrix. Finally, for each sample vector of $M_{corr} = (m_1, \ldots, m_m)$, we again applied the rank normalization to the Pearson correlation coefficient matrix, where *corr* is the Pearson correlation coefficient between samples:

$$m_{rank} = (rank(corr_1), \ldots, rank(corr_n)) \cdot \frac{1}{n}$$

$$M_{normalized} = (m_{rank\,1}, \ldots, m_{rank\,m})$$

New features selection

We selected the common samples from every single omics data, that is each sample has all types of data utilized in clustering and classification. Then, each omic data was used as a separate matrix for rank normalization. Then we computed the Pearson correlation coefficient between every pair of samples. Rank normalization was applied again to the Pearson correlation matrix. Thus, three input matrices were obtained. These three matrices were input into the trained denoising autoencoder and PCA model to obtain new features [17]. Through the R language survival package for univariate function for these features, the features related to survival were selected. Finally, we selected individual features related to survival, stacked them into a single matrix, and inputted the features in the classifier (Figure 2).

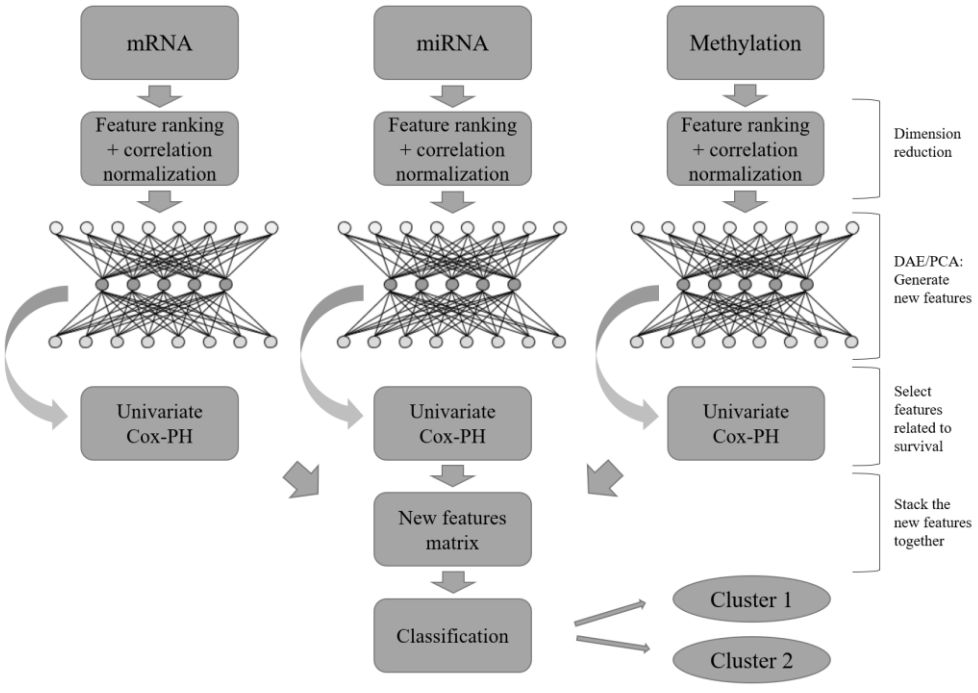


Figure 2. New features selection and classification with multi-omics data.

Denoising autoencoder construction

For each input matrix $M_{normalized}^{OMIC}$, we constructed a denoising autoencoder with one hidden layer. A denoising autoencoder can be defined as a function $f(v) = v'$, where *v* is an input vector of size *m*, and $size(v') = size(v) = m$. *f* is a function that encodes and reconstructs *m* features by encoding *m* features into *h* hidden nodes and decoding to restore *h* nodes to *m* features. This is done by encoding and decoding, so we define *f* by:

$$f(v) = \sigma(W' \cdot \sigma(W \cdot v + b) + b')$$

Transpose the decoded weight matrix *W'* into the encoded weight matrix transpose: $W' = W^T$, which is referred to as "tied weights." By optimizing the parameters of the model, the average reconstruction error of the network is minimized. *b* and *b'* are two biases vectors of sizes *h* and *m*, respectively. σ is a nonlinear activation function, such as the *sigmoid* or *tanh* functions. The autoencoder uses the *adam* optimization algorithm to iteratively find the best *W*, *W'*, *b*, and *b'* minimize a loss function loss (*v*, *v'*). *Adam* uses *Momentum* and *Adaptive Learning Rates* to converge faster. *n* is each training iteration, and a specific percentage of the weight matrix coefficients is randomly set to 0 (decreasing) to train the denoising autoencoder to reduce overfitting. We define Z^{OMIC} as the transformed version of the $M_{normalized}^{OMIC}$ matrix. After training the autoencoder, the transformation *Z* of *v* is given by $f(v) = \sigma(W \cdot v + b)$.

We used the *ruta* package and the *Keras* framework to build denoising autoencoders. For each DAE, we used *h* = 100 (hidden nodes), the *tanh* as activation function, the *binary-crossentropy* as loss function, *sigmoid* as output function, and the

adam optimizer minimize the loss. Finally, we trained the autoencoders with 50 epochs and a 50% dropout rate (Figure 3). We used the hidden nodes as new features from the denoising autoencoder.

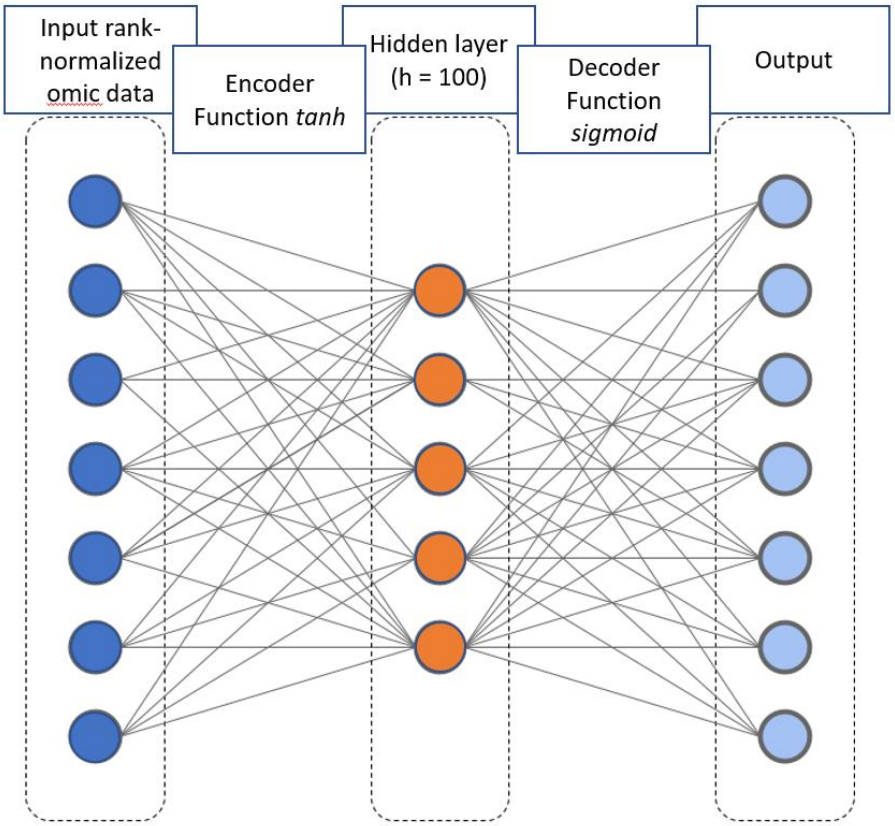


Figure 3. The network structure of the one hidden layer autoencoder.

To use the PCA method for dimension reduction, we selected the components with Proportion of Variance (PoV) greater than 0.01. PoV means a part of variance as a whole. This means that we use fewer components to represent 99% of the overall variance. On the other hand, an alternative option is to select components from the PCA method as new

features that are survival-related while *p*-values are less than 0.01 by using Cox-PH analysis. For each omic, we built an individual model to select new features related to survival. At this time, individual Cox-PH was used to select those features related to survival from new matrices that were produced by dimension-reduction methods.

Identification of features related to survival

We selected the feature matrices obtained from the autoencoder: Z^{mRNA}, Z^{miRNA}, and $Z^{Met^{h}ylation}$ were the features related to survival. For each feature of these matrices, we use a univariate Cox-PH model to select those with a log-rank *p*-value < 0.01. We stacked the new features selected from each omics to get a new feature matrix, Z_{com}. We used the univariate Cox-PH functions of the survival package in R to select features and compute the *p*-values.

Classifiers construction

For each normalized matrix—$M_{normalized}^{mRNA}$, $M_{normalized}^{miRNA}$, and $M_{normalized}^{Methylation}$—we applied the new feature matrix Z_{com} to the classification. We also stacked all the features related to survival to construct a "multi-omics" SVM/RF model. Each model was built using SVM/RF through 10-fold cross-validation. By predicting the survival risk subtype of each sample, the accuracy of the clustering results was obtained.

E. Assumptions with Implied Limitations

Based on previous studies and known theories, we assumed that the clinical data of cancer patients would show specific characteristics related to survival and certain clustering patterns in different ages, disease stages, and other clinical data. This survival-

related clustering patterns may have a similar expression in multi-omics data, such as gene expression, DNA methylation, and microRNA expression.

Using multi-omics data to verify the clustering results based on clinical data is very promising to be able to obtain similar survival-related patterns between two types of datasets. At the same time, the results may not show a perfect consistency between them, as the selection of clinical data and the processing of multi-omics data may have further need to improve. Therefore, based on existing processing methods and algorithms, results verified by two kinds of data could be improved in advance.

IV. RESULTS

A. Brief Overview of Results

Our project obtained training and testing results related to the survival analysis by clustering 534 samples in clinical data and classifying 317 multi-omics samples. The samples were from the TCGA KIRC project: Three omics datasets and clinical dataset, including mRNA, miRNA, DNA methylation, and clinical data of TCGA KIRC, were downloaded using the TCGAbiolinks package in R programming tool. We applied a multivariate Cox-PH method to select survival-related clinical variables, and age and stage were selected as the features in the survival analysis. Using the Silhouette Coefficient (top Silhouette score: 0.53 for k = 2), we determined that the optimal solution for the number of clusters k related to survival was 2. By applying these variables, two clusters were predicted by various clustering methods, such as K-means, K-modes, GMM, to indicate different survival risks. K-fold cross-validation was used to evaluate the classification by using SVM with linear kernels and RF; thus, to build a supervised classification model to

verify whether there are common patterns with clinical data. In the classification, we observed the best results when we labeled samples using the K-modes clustering results, classified samples with the features obtained from the principal components selected by the univariate Cox-PH method. We performed a classification algorithm using RF validation in 10-fold cross-validation (Accuracy = 0.7503) while also with high sensitivity and specificity. The classification result with p-value < $2e$ - 16 also proved that the result had high reliability.

We further discussed which samples were misclassified among different omics features. To summarize, while 161 samples were clustered in cluster 1, and 154 samples were clustered in cluster 2 using the K-modes clustering method. Furthermore, 21 samples were misclassified in Cluster 1, and 26 samples were misclassified in Cluster 2. We compared the feature heat maps of multi-omics and labeled the misclassified samples. We found that in the features heatmaps of mRNA and methylation, there were partial sub-clustering behaviors of the misclassified samples. In the features heatmaps of miRNA, the sample distribution was more scattered.

B. Findings (Results) of the Clustering and Classification

We first downloaded the clinical data of KIRC patients from the *TCGAbiolinks* package. We selected clinical variables that are significantly related to survival through multivariate Cox-PH (Table 1), age and stage. By using the Silhouette Coefficient (top Silhouette score: 0.53 for k = 2), we determined that the optimal solution for the number of clusters k related to survival was 2, indicating that among all KIRC patients, two subtypes related to survival were determined (Figure 4).

Table 1. *Log*-rank *p*-value of multivariate Cox-PH for the TCGA KIRC patients' clinical data to select significantly related to survival analysis for clustering. By this method, age (*p*-value = 7.37*e* - 05) and stage (*p*-value = 5.18*e* - 15) were selected as the variables to cluster the cancer samples.

	Coefficient	***z***	***p*-value**	
Gender	0.04900	0.228	0.819	
Age	0.04176	3.964	7.37*e* - 05	***
Stage	0.71874	7.823	5.18*e* - 15	***
Race	-0.05056	-0.191	-0.191	
Ethnicity	0.51587	1.394	0.163	

Significant codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

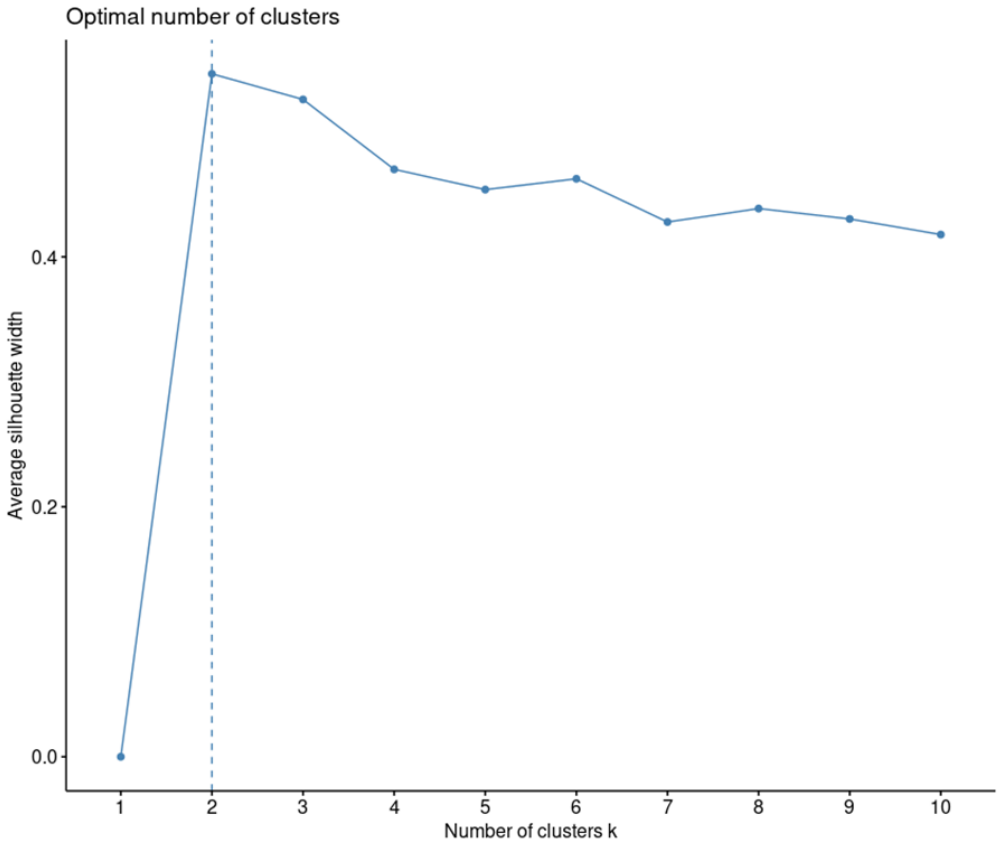


Figure 4. Silhouette score plot for the TCGA KIRC patients with clinical data that are significantly related to survival analysis.

Fewer dimensions mean less calculations, and fewer dimensions can allow the use of algorithms that are not suitable for large dimensions. So, we use DAE, PCA, and survival-related principal components. For DAE, we built a network with 100 hidden nodes (h = 100) to limit each omic data to 100 new features. Then survival-related features were computed. There were 52 survival-related features in DNA methylation, 25 survival-related features in mRNA, and 21 survival-related features in miRNA. To use the PCA method for dimension reduction, components with Proportion of Variance greater than 0.01 were selected. There were 13 components selected in miRNA, and 11 components were selected in mRNA and methylation separately. On the other hand, an alternative option is to select components from the PCA method as new features that are survival-related while p-values are less than 0.01 by using Cox-PH analysis. Here, there were 13 components in miRNA and mRNA of each omic were selected that were survival related. And there were 11 components selected to be significantly related to survival in DNA methylation (Figure 5). For each omic, we built an individual model to select new features related to survival. For this purpose, univariate Cox-PH was used to select the features that are related to survival from new matrices that were produced by dimension-reduction methods. In summary, there were 98 survival-related features selected by DAE from multi-omics data. There were 35 principal components selected by PCA and 37 survival-related components selected by PCA.

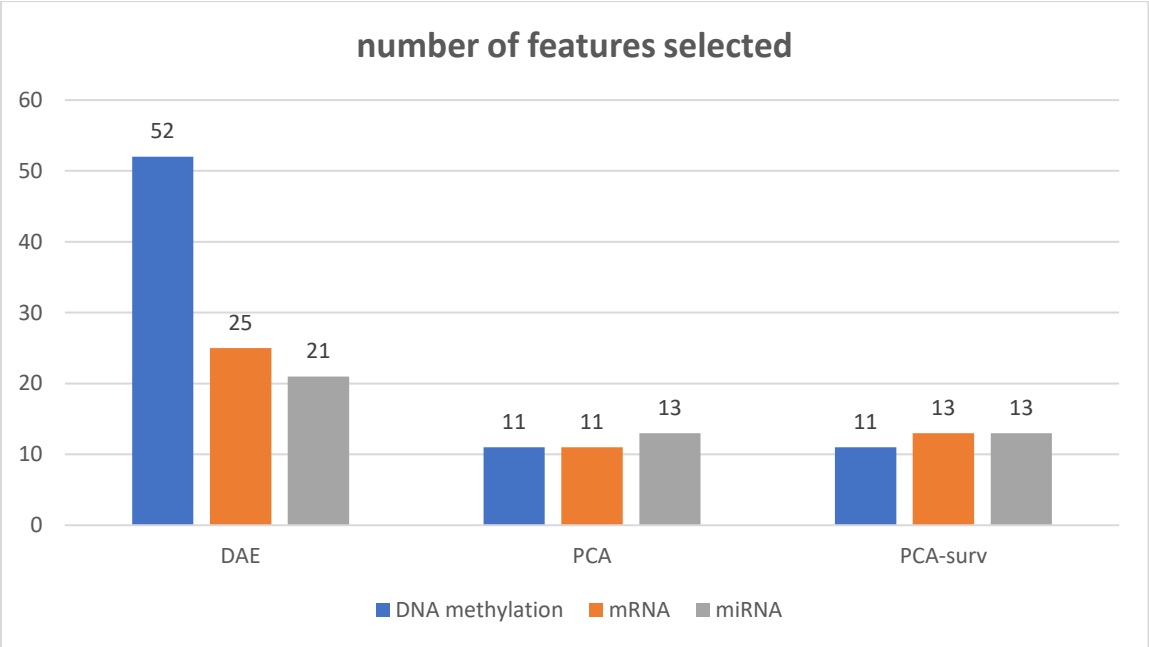


Figure 5. The number of features selected by different methods, DAE, PCA, and PCA-surv, separately.

Table 2. The classification in the 10-fold cross-validation was performed. The samples were labeled by K-means clusters with new features obtained by DAE, PCA, and PCA-surv respectively. The samples were classified using random forest and SVM methods. The results were obtained through a confusion matrix of classification.

	Random forest			**Support vector machine**		
	DAE	**PCA**	**PCA-surv**	**DAE**	**PCA**	**PCA-surv**
Accuracy	0.5862	0.5397	0.5492	0.6143	0.6243	0.5598
Sensitivity	0.6986	0.6567	0.6687	0.6238	0.6437	0.5998
Specificity	0.4595	0.4077	0.4144	0.6036	0.6025	0.5146
Classification *p*-value	0.0003	0.29	0.127	1.014*e* - 13	< 2*e* - 16	0.0052

Table 3. The classification in the 10-fold cross-validation was performed. The samples were labeled by GMM clusters with new features obtained by DAE, PCA, and PCA-surv respectively. The samples were classified using random forest and SVM methods. The results were obtained through a confusion matrix of classification.

	Random forest			**Support vector machine**		
	DAE	**PCA**	**PCA-surv**	**DAE**	**PCA**	**PCA-surv**
Accuracy	0.627	0.6127	0.6302	0.6074	0.5958	0.636
Sensitivity	0.8036	0.7577	0.7823	0.7857	0.8486	0.7866
Specificity	0.3361	0.3739	0.3796	0.3137	0.1793	0.3880
Classification *p*-value	0.344	0.8101	0.2461	< 2*e* - 16	< 2*e* - 16	0.113

Table 4. The classification in the 10-fold cross-validation was performed. The samples were labeled by K-modes clusters with new features obtained by DAE, PCA, and PCA-surv respectively. The samples were classified using random forest and SVM methods. The results were obtained through a confusion matrix of classification.

	Random forest			**Support vector machine**		
	DAE	**PCA**	**PCA-surv**	**DAE**	**PCA**	**PCA-surv**
Accuracy	0.7481	0.7429	0.7503	0.7148	0.7296	0.7254
Sensitivity	0.7723	0.7785	0.7598	0.8833	0.9049	0.9243
Specificity	0.7229	0.7256	0.7403	0.1756	0.1689	0.0888
Classification *p*-value	< 2*e* - 16	< 2*e* - 16	< 2*e* - 16	0.9999	0.9995	0.9999

Among these different algorithms(Tables 2–4), we selected K-modes for clustering, PCA-surv for dimension reduction, and SVM for classification, respectively in 10-fold

cross-validation, that produce an optimized model in terms of classification accuracy (Accuracy = 0.7503) of our survival- specific subtypes prediction. The result of *p*-value < 2*e* - 16 also proves that the result had high reliability.

We compared the performance of the two patient clusters in the survival patterns because the results obtained by K-modes clustering in multi-omics classification performed best. We observed that the Kaplan-Meier plot that the two groups of patients have significant differences in survival analysis (*p*-value < 0.0001), which shows that our clustering can divide KIRC patients into high survival risk and low survival risk subtypes (Figure 6).

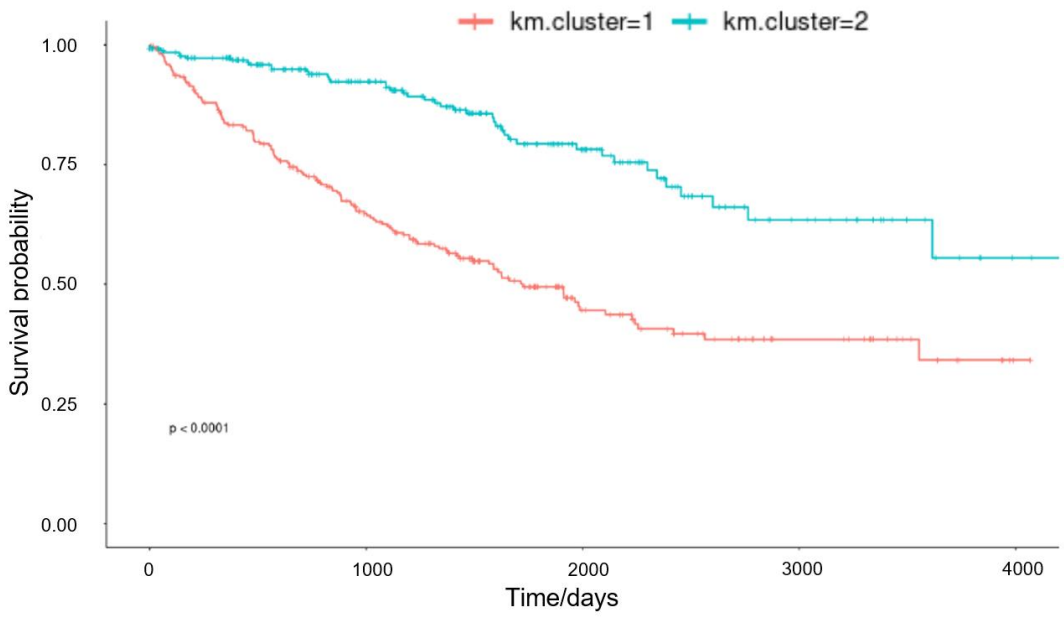


Figure 6. Survival profile with Kaplan-Meier plot of two subtypes for the TCGA KIRC patients.

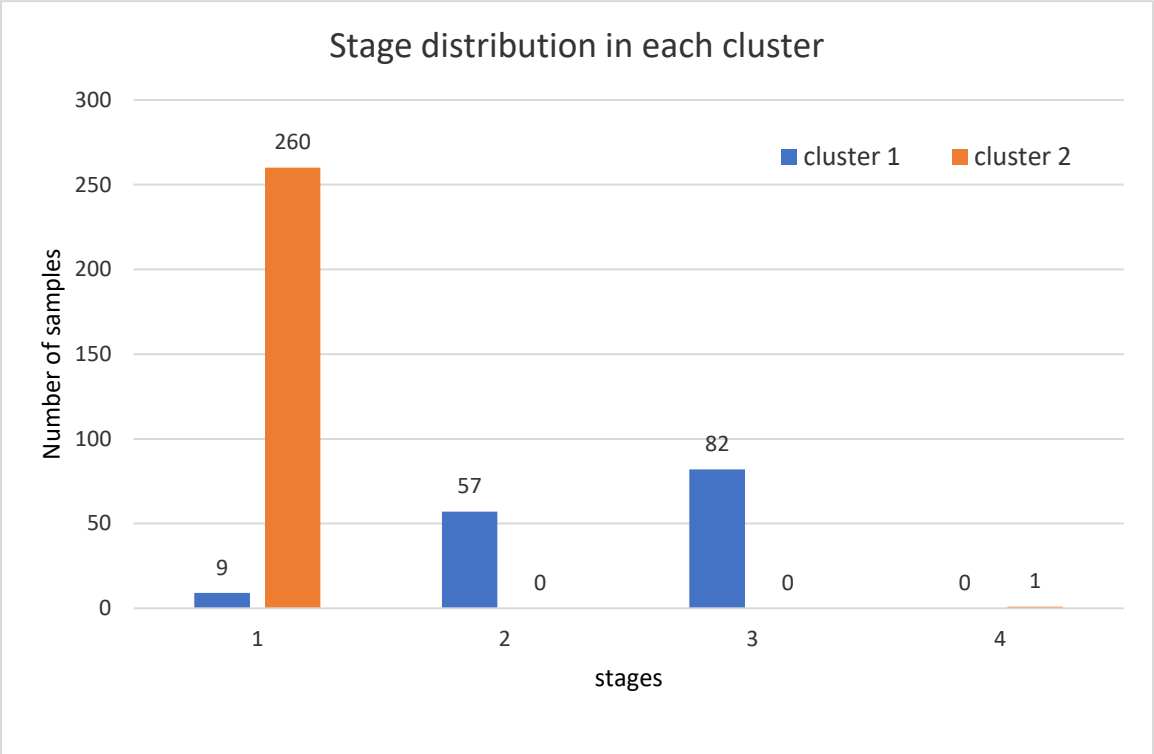


Figure 7. According to the results of K-modes clustering, the samples of different groups are distributed at different stages.

By comparing the distribution of the two groups of patients in different clinical data, we found that Cluster 1, which indicated the high survival risk subtype, were more distributed in stages 2–4. In Cluster 2, almost all the samples were distributed in the first stage, which indicated low survival-risk subtypes (Figure 7). We then discussed the role that stages played in clustering alone. We divided Stage 1 into an early stage, and we divided Stages 2–4 into an advanced stage. We used this as evidence for classification to cross-validate the survival-related features obtained by applying multi-omics data. The results obtained were very close to the best results obtained before (Table 5). We thought that stages played a more prominent role in sample clustering than age.

Table 5. The classification of samples labeled by K-modes clusters in the 10-fold cross-validation, and the new features obtained by PCA-surv were classified in cross-validation. The results were validated using the RF method. The results were obtained through a confusion matrix.

	Random forest
	PCA-surv
Accuracy	0.7545
Sensitivity	0.7419
Specificity	0.7667
***Classification p*-value**	$< 2e - 16$

We then further explored the misclassification of samples in each cluster. We got the following results:

In Cluster 1,

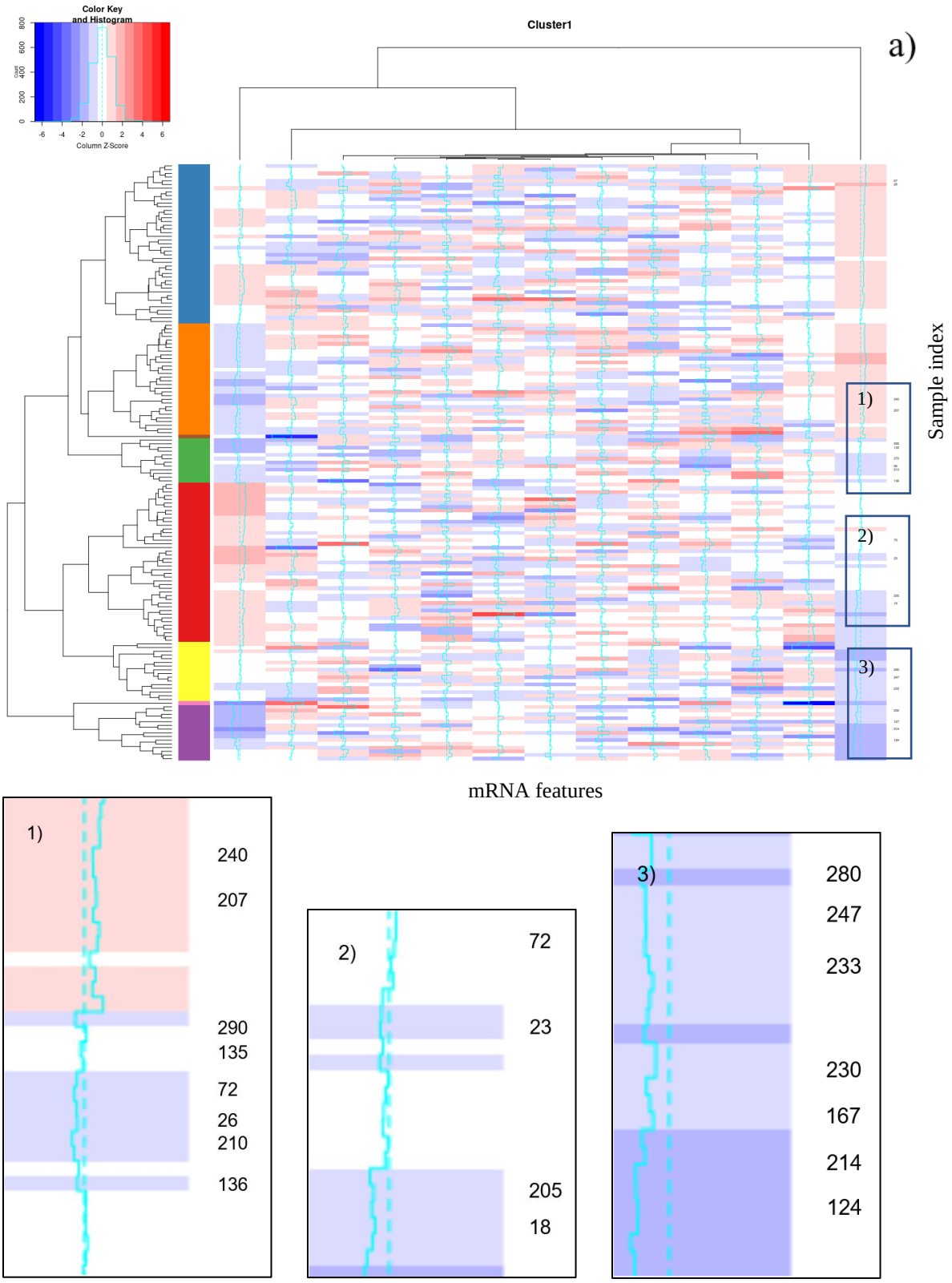
- 1 sample in Stage 1 was misclassified,
- 7 samples in Stage 2 were misclassified,
- 8 samples in Stage 3 were misclassified,
- 5 samples in Stage 4 were misclassified.

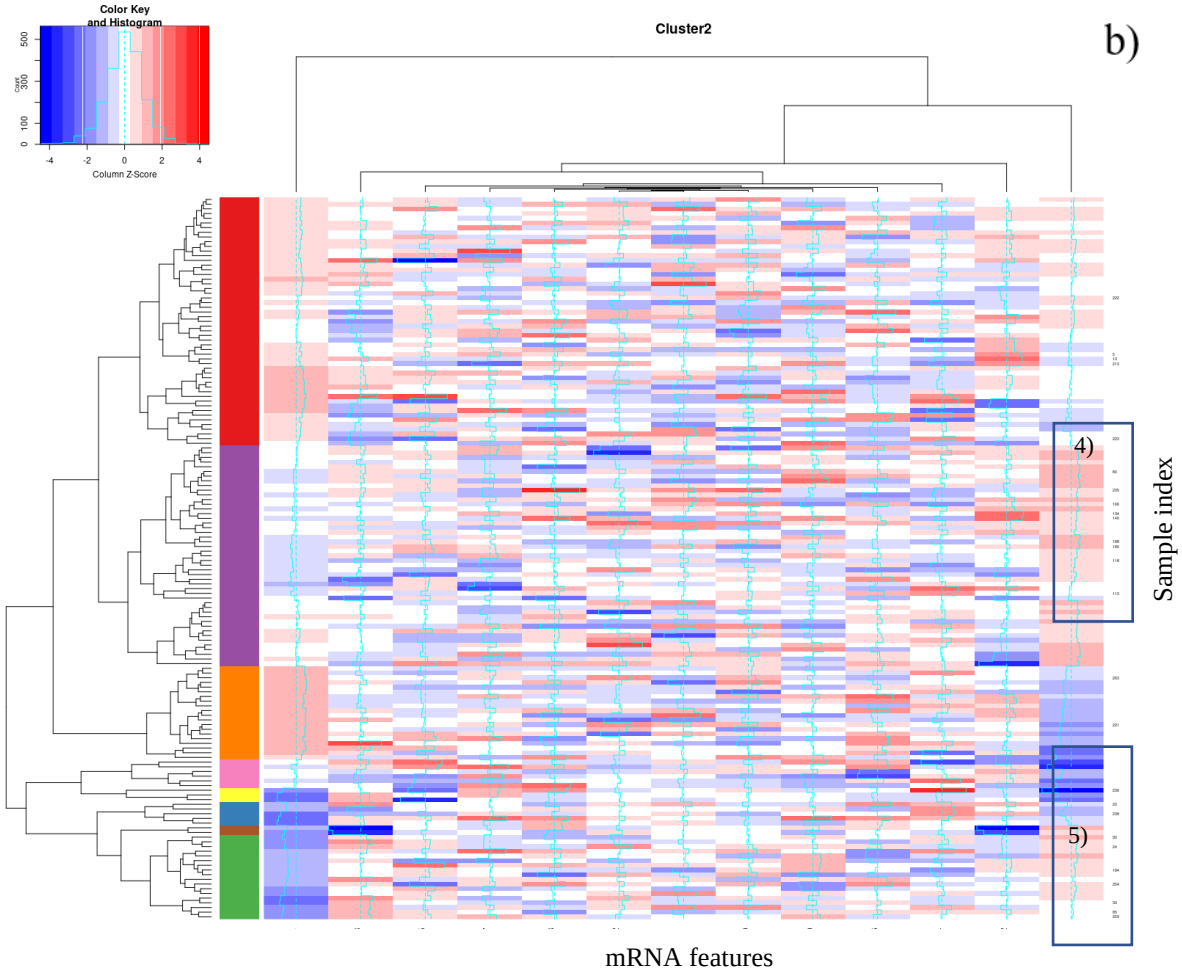
In Cluster 2,

- 25 samples in Stage 1 were misclassified,
- 1 sample in Stage 4 was misclassified.

To summarize, 21 samples were misclassified in Cluster 1, and 26 samples were misclassified in Cluster 2.

Finally, we compared the features heatmaps of multi-omics and labeled the misclassified samples. We can find that in the features heatmaps of mRNA and DNA methylation (Figures 8 and 9), there were partial sub-clustering behaviors of the misclassified samples. In the features heatmaps of miRNA (Figure 10), the sample distribution was more scattered.


Color Key and Histogram
Count
Column Z-Score
Cluster1
a)
Sample index
1)
2)
3)
mRNA features
1)
240
207
290
135
72
26
210
136
2)
72
23
205
18
3)
280
247
233
230
167
214
124


b)
Color Key and Histogram
Count
Column Z-Score
Cluster2
4)
5)
Sample index
mRNA features

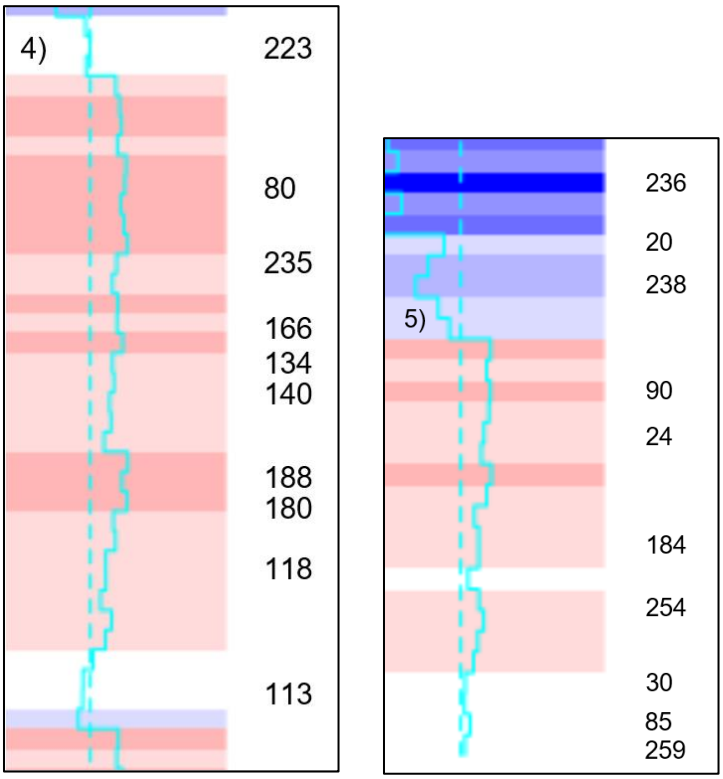
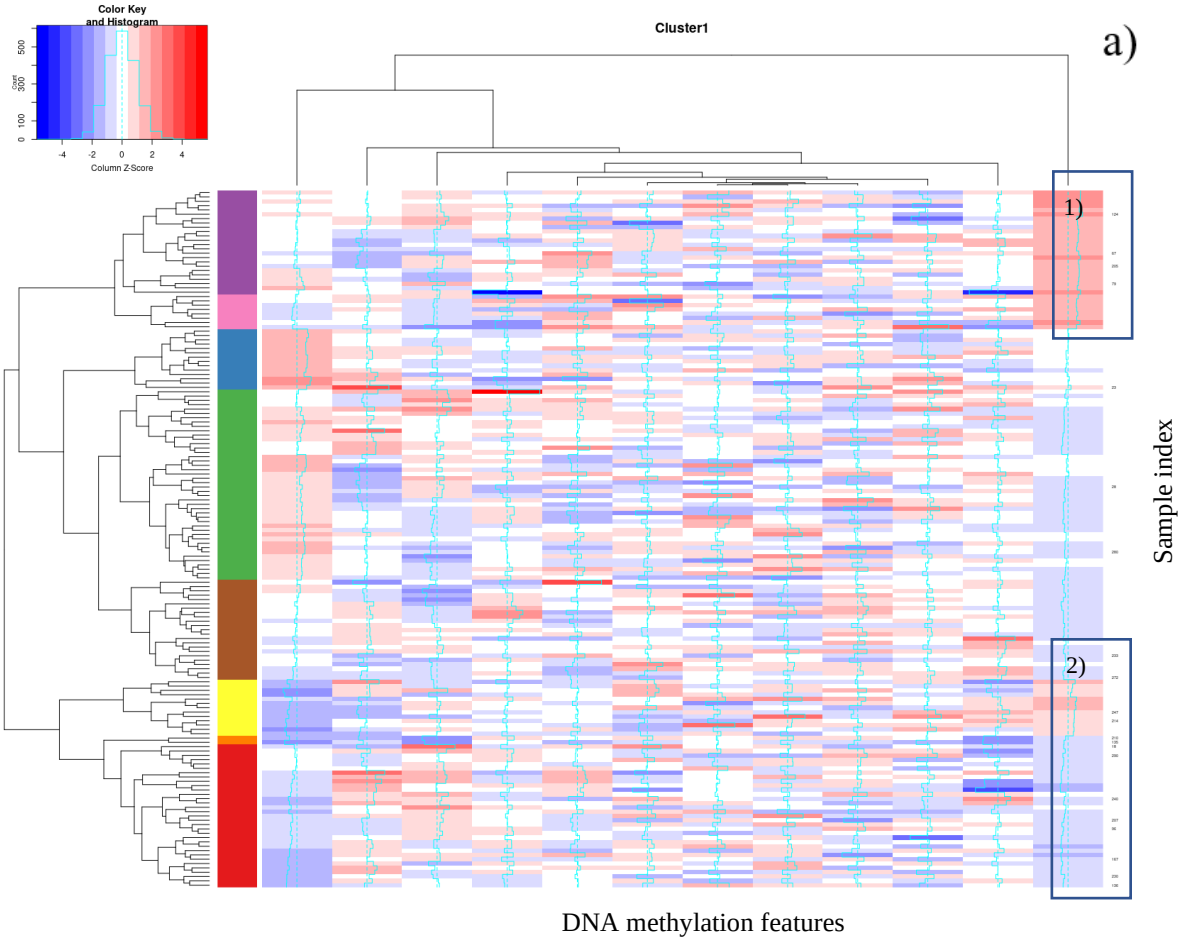
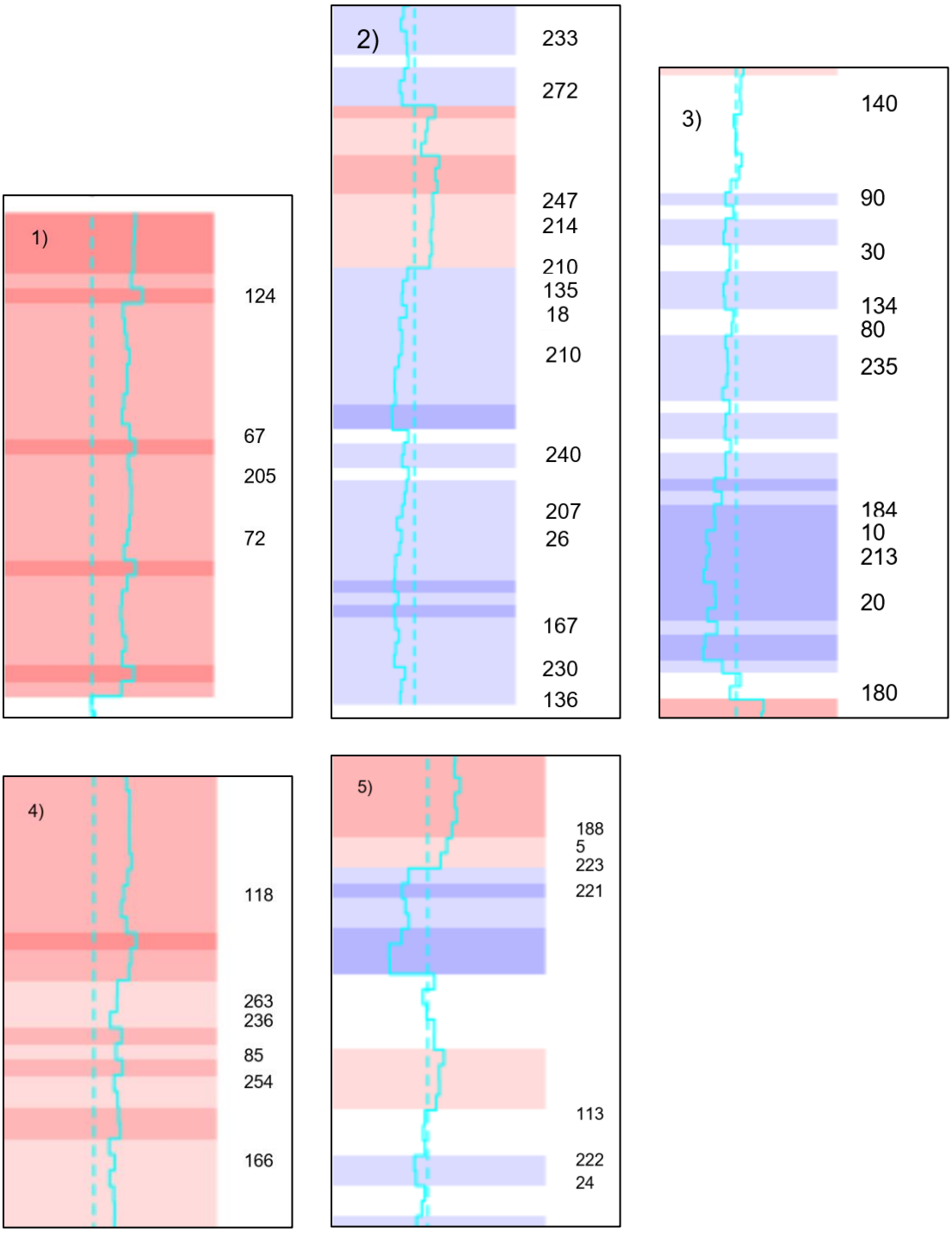


Figure 8. The heatmaps of mRNA features with samples that were misclassified in each cluster. a) The heatmap of Cluster 1 with mRNA features obtained by the PCA-surv method, and b) the heatmap of cluster 2 with mRNA features obtained by the PCA-surv method. 1-5) The sample index for each misclassified sample in the sub-cluster.


Color Key
and Histogram
Count
500
300
100
0
-4
-2
0
2
4
Column Z-Score
Cluster1
a)
1)
2)
Sample index
DNA methylation features


1)
124
67
205
72
2)
233
272
247
214
210
135
18
210
240
207
26
167
230
136
3)
140
90
30
134
80
235
184
10
213
20
180
4)
118
263
236
85
254
166
5)
188
5
223
221
113
222
24

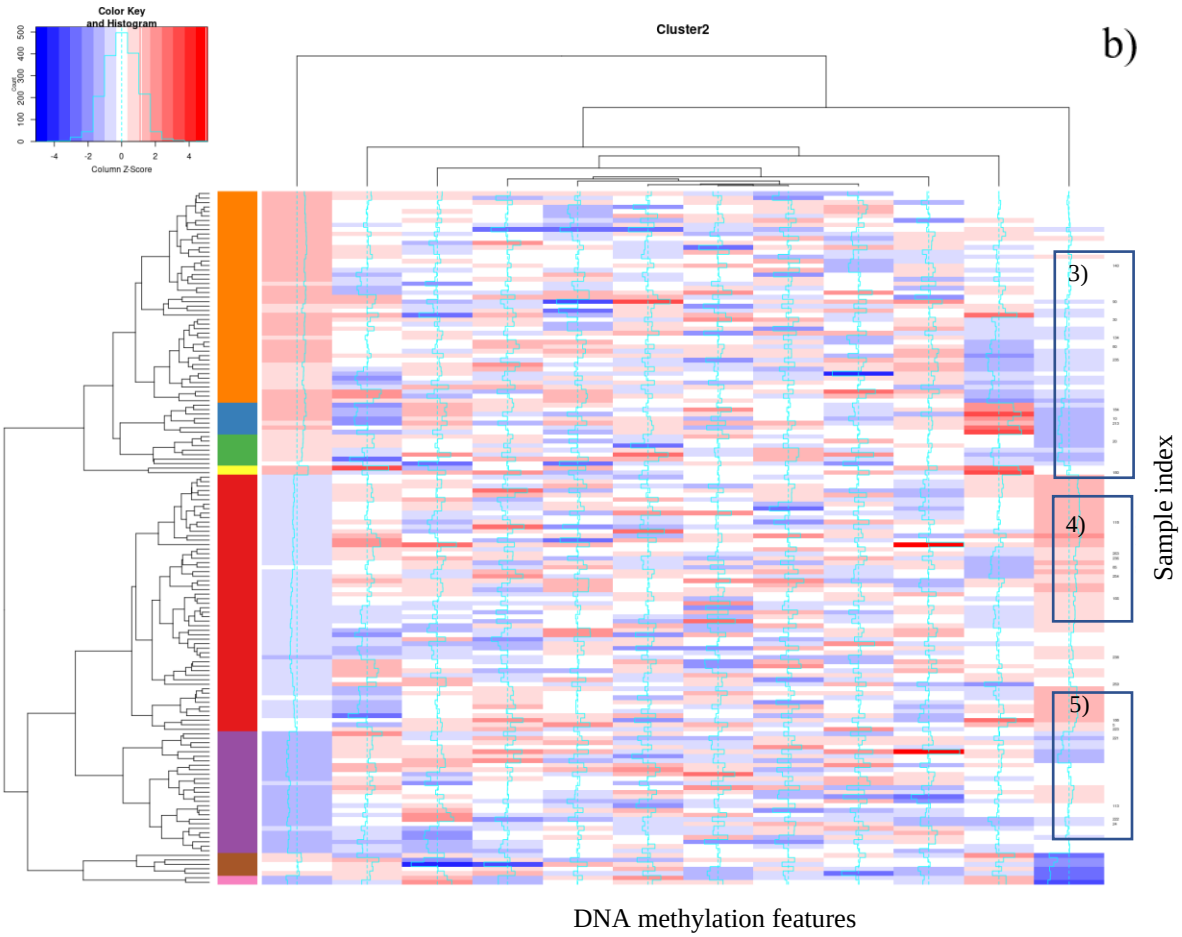
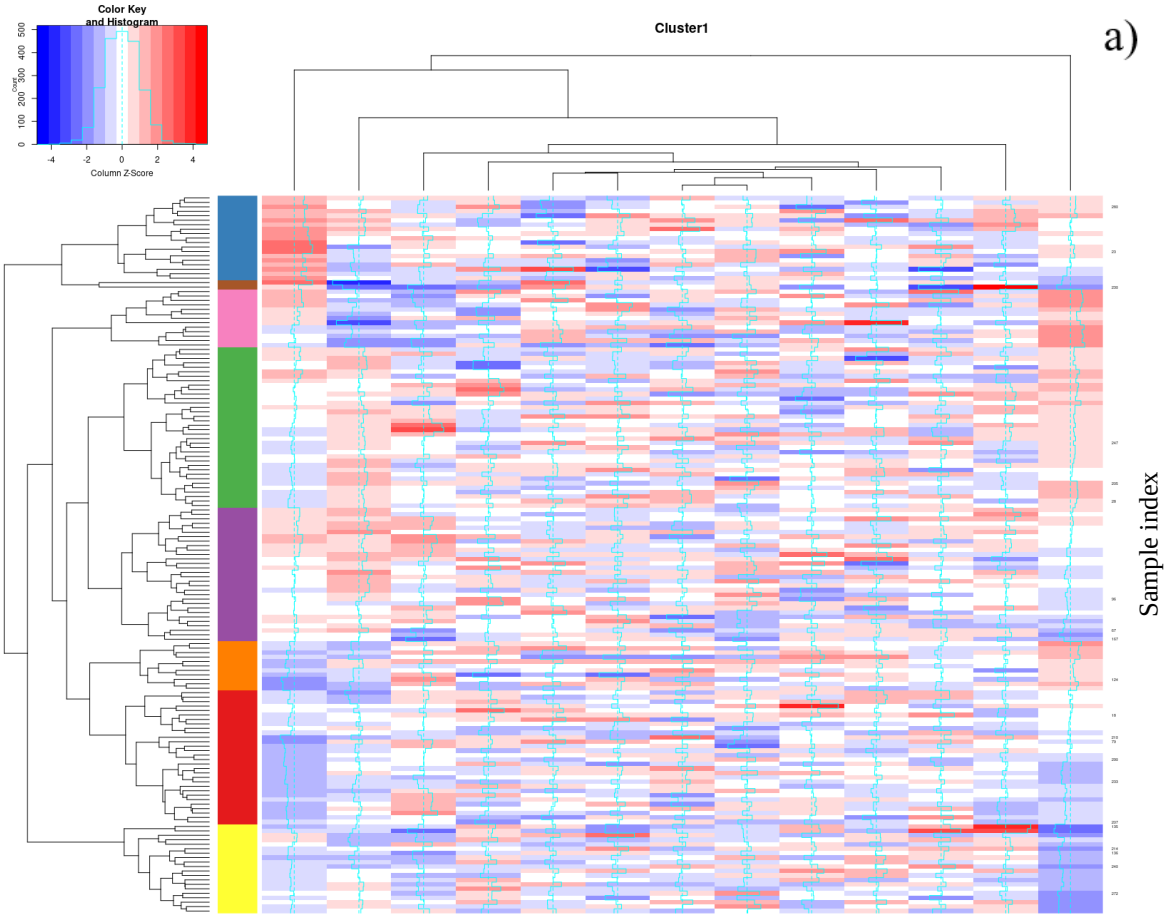


Figure 9. The heatmaps of DNA methylation features with samples that were misclassified in each cluster. a) The heatmap of Cluster 1 with methylation features obtained by the PCA-surv method, and b) the heatmap of Cluster 2 with methylation features obtained by the PCA-surv method. 1-5) The sample index for each misclassified sample in the sub-cluster.


Color Key
and Histogram
Count
0
100
200
300
400
500
-4
-2
0
2
4
Column Z-Score
Cluster1
a)
Sample index
miRNA features

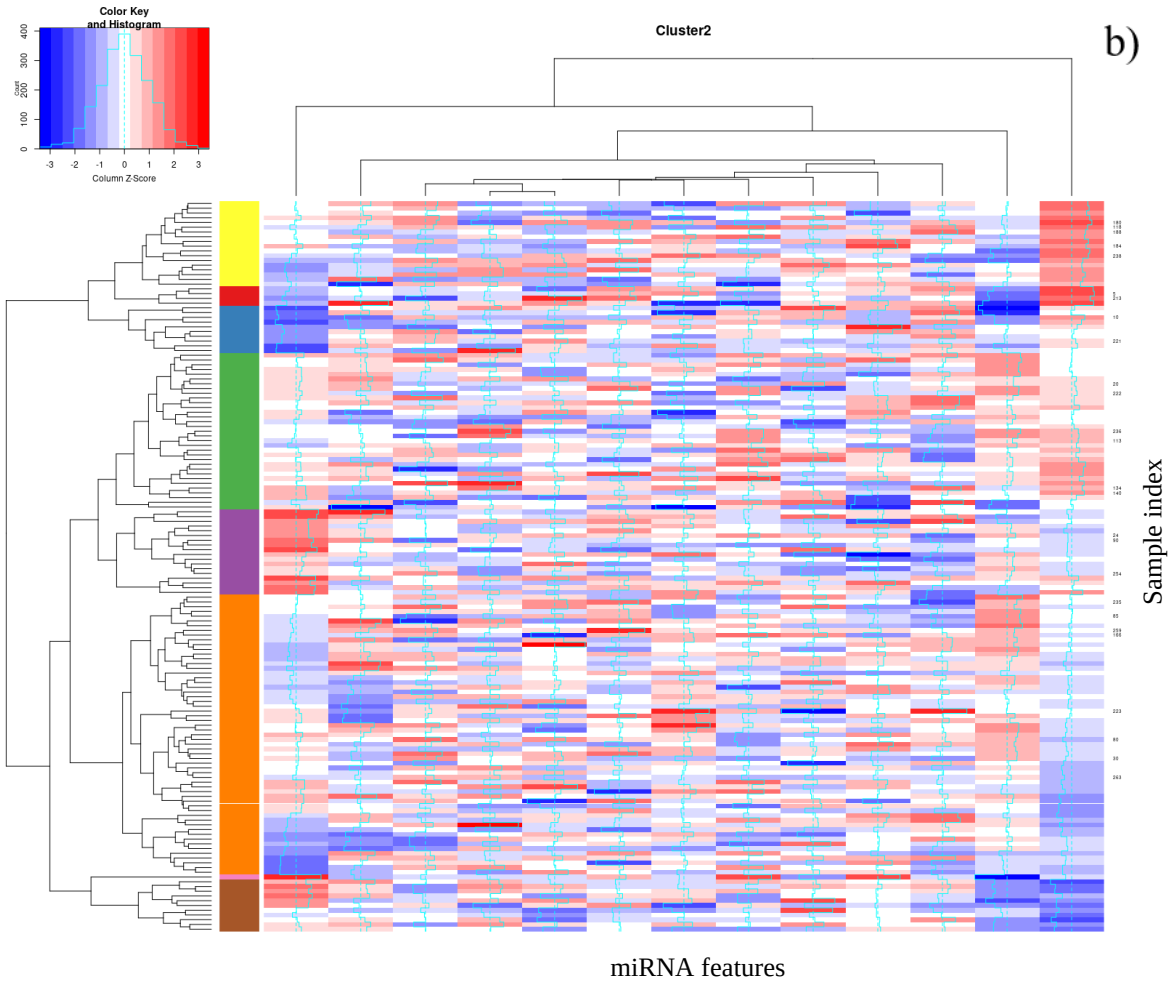


Figure 10. The heatmaps of miRNA features with samples that were misclassified in each cluster. a) The heatmap of Cluster 1 with miRNA features obtained by the PCA-surv method, and b) the heatmap of Cluster 2 with miRNA features obtained by the PCA-surv method.

V. CONCLUSION AND DISCUSSION

A. Brief Summary of the Research

We tested our pipeline on Kidney Renal Clear Cell Carcinoma (KIRC) cancer from the TCGA database. We integrated KIRC clinical data along with gene expression, DNA

methylation and miRNA expression datasets in the pipeline. The data for analysis was downloaded using TCGAbiolinks package in the R programming language.

In this project, 534 samples of clinical data and 317 samples of KIRC were obtained from TCGA for integrating and pre-processing. A total of 315 samples with all types of data were clustered into 2 clusters using K-means, GMM, and K-modes clustering methods. We used rank normalization to pre-process the multi-omics data and inputted the features into dimensional reduction models to select new features that were related to survival. To reduce the dimension of the multi-omics datasets, we tested different algorithms such as denoising autoencoder, Principal components analysis (PCA) and PCA-surv. For the classification purpose, we tested two different machine learning methods, such as, support vector machine (SVM) and random forest. Among these different algorithms, we selected K-modes, PCA-surv, and SVM for clustering, dimension reduction and classification, respectively, that produce an optimized model in terms of classification accuracy of our survival-risk-specific subtypes.

B. Findings (Results) and Implications

Through Cox-PH, we selected age and stage as two features related to survival. By using the Silhouette Coefficient, we determined that the optimal number of clusters of k is 2. After comparing Kaplan Meier-plot, these two subtypes were significantly different in survival analysis (p-value < 0.0001). We define these two subtypes as the high-risk group and the low-risk group.

We compared the autoencoder with the PCA method and performed two different treatments on the principal components obtained by the PCA method. For each DAE, we

used h = 100 (hidden nodes). Finally, we trained the autoencoders on 50 epochs with a 50% dropout rate. We defined Z^{OMIC} as the transformed version of the $M^{OMIC}_{normalized}$ matrix. To use the PCA method for dimension reduction, we selected components with Proportion of Variance greater than 0.01 and survival-related principal components from the PCA method as new features while *p*-values are less than 0.01 by using Cox-PH analysis. For each omic, we built an individual model to select new features related to survival. In this step, individual Cox-PH was used to select features that are related to survival from new matrices that were produced by dimension-reduction methods. New matrices containing 98, 35, and 37 features were obtained.

Through classification, the best accuracy value we obtained for different validation methods was 0.7503. This showed that samples with survival-related principal components could be more accurately classified in classification. At the same time, RF also performed better than SVM in classification.

We also discussed the results of different risk subtypes under unique clinical data. The results showed that stages play a more prominent role in clustering analysis than age. In addition, among the misclassified samples, 21 samples were misclassified in Cluster 1, and 26 samples were misclassified in Cluster 2. Finally, we compared the features heatmaps of multi-omics and labeled the misclassified samples. We found that in the features heatmaps of mRNA and methylation, there were partial sub-clustering behaviors of the misclassified samples. In the features heatmaps of miRNA, the sample distribution was more scattered.

C. Research Analysis of Findings

This was a preliminary exploration of the integration of clinical data and multi-omics data to find differences in survival risks in KIRC samples. Survival Clustering provides a new idea, and on this basis, we can consider more omics data and clinical data. Here, we have three omics datasets and one clinical dataset. We clustered two subtypes with significant differences in survival risk and applied multi-omics data to indicate risk subtypes through cross-validation. We can use different clinical data sets to determine the risk subtypes of KIRC patient samples. We need a more complete method to pre-process and select clinical data sets to consider more clinical data and prognostic signatures. Finally, through DAE and PCA, we performed dimensionality reduction and feature selection on multi-omics data, and we used more concise data to represent the characteristics of high-dimensional data. This can pre-process more omics data into verifiable feature matrices.

D. Reliability and Validity of Survival-Related Subtypes

By studying previous research results, we understand that the integration of multi-omics research plays an important role in the clustering of cancer patients. We expect the integration of multi-omics data to verify the characteristics of survival analysis in clinical data. We hypothesize that there are similar survival characteristics in clinical data and multi-omics data. We can select variables closely related to survival through Cox-PH and divide the cancer samples into different survival risks by clustering type.

E. Summary of Academic Study

We proposed a survival-related prognostic integration pipeline. Through Cox-PH analysis, we selected the clinical data of KIRC patients as survival-related clustering indicators, which gave overall significant *p*-values. We used the Silhouette Coefficient to determine the optimal number *k* of clusters. We obtained three omics datasets, including mRNA, miRNA, and methylation, and clinical data of the TCGA KIRC project from the Genomics Data Commons web portal (https://portal.gdc.cancer.gov); the data was downloaded by the TGCAbiolinks package in the R programming language.

Using survival analysis on clinical data, KIRC patients were divided into two survival subtypes. We then established a supervised classification model using the DAE and the PCA. For each omic, we built an individual model to select new features related to survival. At this time, individual Cox-PH was used to select features that are related to survival from new matrices that were produced by dimension-reduction methods. These features were then classified by the clusters inferred by the clinical survival-related variables. K-fold cross-validation was utilized for classification by using SVM with linear kernels and RF to build a supervised classification model to verify whether there are common patterns with clinical data. Accuracy, specificity, and sensitivity were used to evaluate the results. A confusion matrix was used to combine these parameters to gain a comprehensive view of the results.

Overall, we identified two subtypes of survival risk through cluster analysis. These two subtypes have significant survival differences. We used the K-modes clustering method to cluster KIRC samples into two different subtypes. We then used the DAE, PCA,

and PCA-surv methods to perform feature screening with the processed three omics data. We obtained matrixes containing 98, 35, and 37 new features, and we used the SVM/RF classification model to perform 10-fold cross-validation, the classification results obtained by the final RF method had the highest accuracy. Through classification, the best accuracy we obtained among different classification methods was 0.7503.

Among the misclassified samples, 21 samples were misclassified in Cluster 1, and 26 samples were misclassified in Cluster 2. We found that in the features heatmaps of mRNA and methylation, there were partial sub-clustering behaviors of the misclassified samples.

F. Limitations of the Theory or Method of Research

Although we can attain better verification results through this method, as it relates to the selection of clinical data related to survival, the method we used is still relatively straightforward, so there may be some data overfitting or some data are not correctly selected. We can consider as much clinical data, prognostic signatures, and biochemical indicators as possible. More normalization methods to integrate data belonging to different ranges into data of similar ranges could be considered, as the selection of data is not optimized enough. In addition, as it relates to dimensionality-reduction processing and feature selection of multi-omics data, no other methods were considered, resulting in some methods that had deviations in the selection of different omics data, so that some data were selected for many features while other data had few or no features. Therefore, there are still some shortcomings in data processing and algorithm optimization.

G. Future Study

In future research, we will try to use more clinical data, prognostic signatures, and biochemical indicators as the basis for clustering. We will use more normalization methods to process the data to a better selection to choose more comprehensive and reliable data. We hope to consider more data in survival analysis so that survival-related features can be more comprehensively applied to survival clustering algorithms.

In addition, as it relates to data processing of multi-omics, we expect to be able to refer to more data-processing methods, learn more deep-learning methods and neural network algorithms, and select new features related to survival after dimensional-reduction processing.

REFERENCES

[1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. *CA Cancer J Clin.* 2019 Jan;69(1):7-34. doi: 10.3322/caac.21551. Epub 2019 Jan 8. PMID: 30620402.

[2] Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. *Nat Rev Genet.* 2009 Mar;10(3):184-94. doi: 10.1038/nrg2537. PMID: 19223927; PMCID: PMC4550035.

[3] Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics*. 2016 Mar;8(3):389-99. doi: 10.2217/epi.15.114. Epub 2015 Dec 17. PMID: 26673039; PMCID: PMC4864062.

[4] Cox, D. R., & Oakes, D. (1984). *Analysis of Survival Data* (1st ed.). Chapman & Hall/CRC.

[5] Ng, A. (2011). Sparse autoencoder. *CS294A Lecture notes*, *72*(2011), 1–19.

[6] Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*(1–3), 37–52. https://doi.org/10.1016/0169-7439(87)80084-9

[7] Vapnik, V. N. (1998). *Statistical learning theory*. Wiley.

[8] Ricketts CJ, De Cubas AA, Fan H, Smith CC, Lang M, Reznik E, Bowlby R, Gibb EA, Akbani R, Beroukhim R, Bottaro DP, Choueiri TK, Gibbs RA, Godwin AK, Haake S, Hakimi AA, Henske EP, Hsieh JJ, Ho TH, Kanchi RS, Krishnan B, Kwiatkowski DJ, Lui W, Merino MJ, Mills GB, Myers J, Nickerson ML, Reuter VE, Schmidt LS, Shelley CS, Shen H, Shuch B, Signoretti S, Srinivasan R, Tamboli P, Thomas G, Vincent BG, Vocke CD, Wheeler DA, Yang L, Kim WY, Robertson AG; Cancer Genome Atlas Research Network, Spellman PT, Rathmell WK, Linehan WM. The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma. *Cell Rep*. 2018 Apr 3;23(1):313-326.e5. doi: 10.1016/j.celrep.2018.03.075. Erratum in: Cell Rep. 2018 Jun 19;23(12):3698. PMID: 29617669; PMCID: PMC6075733.

[9] Therneau, T. M., & Lumley, T. (2014). Package 'survival'. *Survival analysis Published on CRAN*, *2*, 3.

[10] NCI, & NHGRI. *The Cancer Genome Atlas Program*. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga. Retrieved 2009-04-28.

[11] Bradburn MJ, Clark TG, Love SB, Altman DG. Survival analysis part II: multivariate data analysis--an introduction to concepts and methods. *Br J Cancer*. 2003 Aug 4;89(3):431-6. doi: 10.1038/sj.bjc.6601119. PMID: 12888808; PMCID: PMC2394368.

[12] Tsodikov A, Szabo A, Jones D. Adjustments and measures of differential expression for microarray data. *Bioinformatics*. 2002 Feb;18(2):251-60. doi: 10.1093/bioinformatics/18.2.251. PMID: 11847073.

[13] Szabo A, Boucher K, Carroll WL, Klebanov LB, Tsodikov AD, Yakovlev AY. Variable selection and pattern recognition with gene expression data generated by the microarray technology. *Math Biosci*. 2002 Mar;176(1):71-98. doi: 10.1016/s0025-5564(01)00103-1. PMID: 11867085.

[14] Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS, Malta TM, Pagnotta SM, Castiglioni I, Ceccarelli M, Bontempi G, Noushmehr H. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*. 2016 May 5;44(8):e71. doi: 10.1093/nar/gkv1507. Epub 2015 Dec 23. PMID: 26704973; PMCID: PMC4856967.

[15] Civelek M, Lusis AJ. Systems genetics approaches to understand complex traits. *Nat Rev Genet*. 2014 Jan;15(1):34-48. doi: 10.1038/nrg3575. Epub 2013 Dec 3. PMID: 24296534; PMCID: PMC3934510.

[16] Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet*. 2015 Feb;16(2):85-97. doi: 10.1038/nrg3868. Epub 2015 Jan 13. PMID: 25582081.

[17] Poirion OB, Chaudhary K, Garmire LX. Deep Learning data integration for better risk stratification models of bladder cancer. *AMIA Jt Summits Transl Sci Proc*. 2018 May 18;2017:197-206. PMID: 29888072; PMCID: PMC5961799.

[18] Liu Y, Ding J, Reynolds LM, Lohman K, Register TC, De La Fuente A, Howard TD, Hawkins GA, Cui W, Morris J, Smith SG, Barr RG, Kaufman JD, Burke GL, Post W, Shea S, McCall CE, Siscovick D, Jacobs DR Jr, Tracy RP, Herrington DM, Hoeschele I. Methylomics of gene expression in human monocytes. *Hum Mol Genet*. 2013 Dec 15;22(24):5065-74. doi: 10.1093/hmg/ddt356. Epub 2013 Jul 29. PMID: 23900078; PMCID: PMC3836482.

[19] Petersen AK, Zeilinger S, Kastenmüller G, Römisch-Margl W, Brugger M, Peters A, Meisinger C, Strauch K, Hengstenberg C, Pagel P, Huber F, Mohney RP, Grallert H, Illig T, Adamski J, Waldenberger M, Gieger C, Suhre K. Epigenetics meets metabolomics: an epigenome-wide association study with blood serum metabolic traits. *Hum Mol Genet*. 2014 Jan 15;23(2):534-45. doi: 10.1093/hmg/ddt430. Epub 2013 Sep 6. PMID: 24014485; PMCID: PMC3869358.

[20] Shah S, Bonder MJ, Marioni RE, Zhu Z, McRae AF, Zhernakova A, Harris SE, Liewald D, Henders AK, Mendelson MM, Liu C, Joehanes R, Liang L; BIOS Consortium, Levy D, Martin NG, Starr JM, Wijmenga C, Wray NR, Yang J, Montgomery GW, Franke L, Deary IJ, Visscher PM. Improving Phenotypic Prediction by Combining Genetic and Epigenetic Associations. *Am J Hum Genet*. 2015 Jul 2;97(1):75-85. doi: 10.1016/j.ajhg.2015.05.014. Epub 2015 Jun 25. PMID: 26119815; PMCID: PMC4572498.

[21] Deer EL, González-Hernández J, Coursen JD, Shea JE, Ngatia J, Scaife CL, Firpo MA, Mulvihill SJ. Phenotype and genotype of pancreatic cancer cell lines. *Pancreas*. 2010 May;39(4):425-35. doi: 10.1097/MPA.0b013e3181c15963. Erratum in: Pancreas. 2018 Jul;47(6):e37. PMID: 20418756; PMCID: PMC2860631.

[22] Hao, J., Masum, M., Oh, J. H., & Kang, M. (2019). Gene- and Pathway-Based Deep Neural Network for Multi-omics Data Integration to Predict Cancer Survival Outcomes. In Z. Cai, P. Skums, & M. Li (Eds.), *Bioinformatics Research and Applications: 15th International Symposium, ISBRA 2019, Barcelona, Spain, June 3–6, 2019, Proceedings* (pp. 113–124). Cham: Springer. https://doi.org/10.1007/978-3-030-20242-2_10

[23] Miao R, Luo H, Zhou H, Li G, Bu D, Yang X, Zhao X, Zhang H, Liu S, Zhong Y, Zou Z, Zhao Y, Yu K, He L, Sang X, Zhong S, Huang J, Wu Y, Miksad RA, Robson SC, Jiang C, Zhao Y, Zhao H. Identification of prognostic biomarkers in hepatitis B virus-related hepatocellular carcinoma and stratification by integrative multi-omics analysis. *J Hepatol*. 2014 Oct;61(4):840-9. doi: 10.1016/j.jhep.2014.05.025. Epub 2014 May 22. PMID: 24859455.

[24] Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res*. 2018 Nov 16;46(20):10546-10562. doi: 10.1093/nar/gky889. Erratum in: Nucleic Acids Res. 2019 Jan 25;47(2):1044. PMID: 30295871; PMCID: PMC6237755.

[25] Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, Buettner F, Huber W, Stegle O. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol*. 2018 Jun 20;14(6):e8124. doi: 10.15252/msb.20178124. PMID: 29925568; PMCID: PMC6010767.

[26] Solberg HE, Skrede S, Elgjo K, Blomhoff JP, Gjone E. Classification of liver diseases by clinical chemical laboratory results and cluster analysis. *Scand J Clin Lab Invest*. 1976 Jan;36(1):81-5. doi: 10.1080/00365517609068022. PMID: 1257696.

[27] van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2002 Jan 31;415(6871):530-6. doi: 10.1038/415530a. PMID: 11823860.

[28] Docampo E, Collado A, Escaramís G, Carbonell J, Rivera J, Vidal J, Alegre J, Rabionet R, Estivill X. Cluster analysis of clinical data identifies fibromyalgia subgroups. *PLoS One*. 2013 Sep 30;8(9):e74873. doi: 10.1371/journal.pone.0074873. PMID: 24098674; PMCID: PMC3787018.

[29] Chen HC, Kodell RL, Cheng KF, Chen JJ. Assessment of performance of survival prediction models for cancer prognosis. *BMC Med Res Methodol*. 2012 Jul 23;12:102. doi: 10.1186/1471-2288-12-102. PMID: 22824262; PMCID: PMC3410808.

[30] Zhu B, Song N, Shen R, Arora A, Machiela MJ, Song L, Landi MT, Ghosh D, Chatterjee N, Baladandayuthapani V, Zhao H. Integrating Clinical and Multiple Omics Data for Prognostic Assessment across Human Cancers. *Sci Rep*. 2017 Dec 5;7(1):16954. doi: 10.1038/s41598-017-17031-8. PMID: 29209073; PMCID: PMC5717223.

[31] Hao J, Kim Y, Kim TK, Kang M. PASNet: pathway-associated sparse deep neural network for prognosis prediction from high-throughput data. *BMC Bioinformatics*. 2018 Dec 17;19(1):510. doi: 10.1186/s12859-018-2500-z. PMID: 30558539; PMCID: PMC6296065.

[32] Huang C, Zhang A, Xiao G. Deep Integrative Analysis for Survival Prediction. *Pac Symp Biocomput*. 2018;23:343-352. PMID: 29218895.

[33] Robertson AG, Shih J, Yau C, Gibb EA, Oba J, Mungall KL, Hess JM, Uzunangelov V, Walter V, Danilova L, Lichtenberg TM, Kucherlapati M, Kimes PK, Tang M, Penson A, Babur O, Akbani R, Bristow CA, Hoadley KA, Iype L, Chang MT; TCGA Research Network, Cherniack AD, Benz C, Mills GB, Verhaak RGW, Griewank KG, Felau I, Zenklusen JC, Gershenwald JE, Schoenfield L, Lazar AJ, Abdel-Rahman MH, Roman-Roman S, Stern MH, Cebulla CM, Williams MD, Jager MJ, Coupland SE, Esmaeli B, Kandoth C, Woodman SE. Integrative Analysis Identifies Four Molecular and Clinical Subsets in Uveal Melanoma. *Cancer Cell*. 2017 Aug 14;32(2):204-220.e15. doi: 10.1016/j.ccell.2017.07.003. Erratum in: Cancer Cell. 2018 Jan 8;33(1):151. PMID: 28810145; PMCID: PMC5619925.

[34] Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. *Nat Rev Genet*. 2009 Mar;10(3):184-94. doi: 10.1038/nrg2537. PMID: 19223927; PMCID: PMC4550035.

[35] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009 Jan;10(1):57-63. doi: 10.1038/nrg2484. PMID: 19015660; PMCID: PMC2949280.

[36] Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet*. 2010 Jan;11(1):31-46. doi: 10.1038/nrg2626. Epub 2009 Dec 8. PMID: 19997069.

[37] Sun YV, Hu YJ. Integrative Analysis of Multi-omics Data for Discovery and Functional Studies of Complex Human Diseases. *Adv Genet*. 2016;93:147-90. doi: 10.1016/bs.adgen.2015.11.004. Epub 2016 Jan 25. PMID: 26915271; PMCID: PMC5742494.

[38] Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, Montgomery GW, Goddard ME, Wray NR, Visscher PM, Yang J. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet*. 2016 May;48(5):481-7. doi: 10.1038/ng.3538. Epub 2016 Mar 28. PMID: 27019110.

[39] Zhang L, Lv C, Jin Y, Cheng G, Fu Y, Yuan D, Tao Y, Guo Y, Ni X, Shi T. Deep Learning-Based Multi-Omics Data Integration Reveals Two Prognostic Subtypes in High-Risk Neuroblastoma. *Front Genet*. 2018 Oct 18;9:477. doi: 10.3389/fgene.2018.00477. PMID: 30405689; PMCID: PMC6201709.

[40] Cox, D. R., & Wermuth, N. (1998). *Multivariate Dependencies: Models, Analysis and Interpretation*. Chapman & Hall/CRC.

[41] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* (pp. 281–297). Berkeley: University of California Press.

[42] Reynolds, D. (2009). Gaussian Mixture Models. *Encyclopedia of Biometrics*, 659–663. https://doi.org/10.1007/978-0-387-73003-5_196

[43] Chaturvedi, A., Green, P. E., & Caroll, J. D. (2001). K-modes Clustering. *Journal of Classification*, *18*(1), 35–55. https://doi.org/10.1007/s00357-001-0004-3

[44] Hammond SM. An overview of microRNAs. *Adv Drug Deliv Rev*. 2015 Jun 29;87:3-14. doi: 10.1016/j.addr.2015.05.001. Epub 2015 May 12. PMID: 25979468; PMCID: PMC4504744.