Integrative Analysis of Pathological Images and Multi-Dimensional Genomic Data for Early-Stage Cancer Prognosis

Wei Shao, Zhi Han, Jun Cheng, Liang Cheng, Tongxin Wang, Liang Sun, Zixiao Lu, Jie Zhang, Daoqiang Zhang, and Kun Huang

***Abstract*— The integrative analysis of histopathological images and genomic data has received increasing attention for studying the complex mechanisms of driving cancers. However, most image-genomic studies have been restricted to combining histopathological images with the single modality of genomic data (e.g., mRNA transcription or genetic mutation), and thus neglect the fact that the molecular architecture of cancer is manifested at multiple levels, including genetic, epigenetic, transcriptional, and post-transcriptional events. To address this issue, we propose a novel ordinal multi-modal feature selection (OMMFS) framework that can simultaneously identify important features from both pathological images and multi-modal genomic data (i.e., mRNA transcription, copy number variation, and DNA methylation data) for the prognosis of cancer patients. Our model is based on a generalized sparse canonical correlation analysis framework, by which we also take advantage of the ordinal survival information among different patients for survival outcome prediction. We evaluate our method on three early-stage cancer datasets derived from The Cancer Genome Atlas (TCGA) project, and the experimental results demonstrated that both the selected image and multi-modal genomic markers are strongly correlated with survival enabling effective stratification of patients with distinct survival than the comparing methods, which is often difficult for early-stage cancer patients.**

***Index Terms*— Histopathological images, multi-modal genomic data, survival analysis, early-stage cancer, ordinal multi-model feature selection.**

I. Introduction

CANCER is one of the leading causes of death worldwide. It is reported that the number of affected people in developing country will reach 20 million annually as early as 2025 [1]. Thus, effective and accurate prognosis of human cancer, as well as effective stratification of cancer patients into subgroups with different predicted outcomes has attracted much more attention than ever before [2].

So far, a large amount of biomarkers have been developed and applied to cancer prognosis, including the histopathological images, genetic mutations, gene expression signatures, and protein markers. Of all these types of biomarkers, histopathology image is generally considered to be the gold standard for cancer diagnosis and prognosis since it can provide morphological attributes of cells that are highly related to the degree of the aggressiveness of cancers [3]. With the help of ever-increasing computing resources, many computational histopathologic systems have been proposed to extract morphological biomarkers for the prognosis of considerable number of cancers such as lung [4], breast [5] and kidney cancers [7]. Besides histopathological image, it is known that cancer is caused by genomic mutations that alter normal functions and biological processes in cells. Accordingly, many researchers [8], [9], [18] also use the patients' molecular profiles such as genetic alteration and gene expression signatures to drive diagnostic, prognostic, and therapeutic practices.

In many clinical and research studies, it is common to acquire multiple biomarkers for a more accurate assessment of disease status and progression stages [39]. Recently, researches also explored to combine both imaging and genomic markers for the prognosis of cancers. For instance, Cheng *et al.* [6] constructed a novel framework that can predict the survival outcomes of patients with renal cell carcinoma by using a combination of quantitative image features extracted from histopathological images and eigengenes extracted from the gene expression data. Yuan *et al.* [10] integrated both image data and genomic data to improve the survival prognosis for breast cancer patients. Other efforts

Manuscript received April 29, 2019; revised May 23, 2019; accepted May 26, 2019. Date of publication June 3, 2019; date of current version December 27, 2019. The work of W. Shao, L. Sun, and D. Zhang was supported in part by the National Natural Science Foundation of China under Grant 61876082, Grant 61861130366, and Grant 61703301, in part by the National Key Research and Development Program of China under Grant 2018YFC2001602, and in part by the Royal Society–Academy of Medical Sciences Newton Advanced Fellowship under grant NA\R1\180371. The work of J. Cheng was supported in part by the Shenzhen Peacock Plan. *(Corresponding authors: Zhi Han; Daoqiang Zhang; Kun Huang.)*

W. Shao, L. Sun, and D. Zhang are with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing 211106, China (e-mail: dqzhang@nuaa.edu.cn).

Z. Han, L. Cheng, Z. Lu, J. Zhang, and K. Huang are with the School of Medicine, Indiana University, Indianapolis, IN 46202 USA (e-mail: zhihan@iu.edu; kunhuang@iu.edu).

J. Cheng is with the School of Biomedical Engineering, Shenzhen University, Shenzhen 518073, China.

T. Wang is with the Department of Computer Science, Indiana University Bloomington, Bloomington, IN 47405 USA.

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2019.2920608

includes [11] have developed a workflow that can integrate morphological features with genomic feature for biomarker discovery in ovarian cancer. These studies suggested that the imaging and genetic data complement with each other and can provide better performance of patient stratification when used together.

Although integrating imaging and genomic features can better predict the clinical outcome than individual type of biomarkers, the existing integrative studies have been restricted to combine pathological images with only one modality of genomic activity for survival analysis (*e.g.*, mRNA expression or genetic mutation). As a matter of fact, eukaryotic gene regulation is a complex process controlled at multiple levels [12], [13], involving various genetic, epigenetic, transcriptional, translational and post-translational modification events. Intuitively, the joint investigation of pathological images and multi-modal genomic data can bring new insight into the mechanism of cancer progression, as well as providing more comprehensive and effective biomarkers for survival outcome prediction. For multi-modal data (*e.g.*, various types of genomic data and histopathological image data), some features may be redundant when combining them together for survival analysis. Feature selection is commonly used to remove the redundant and irrelevant features. Most of the existing methods [6], [10] concatenate the histopathological image features with genomic data at the beginning, and then apply traditional feature selection methods such as the least absolute shrinkage and selection operator (LASSO) [10] to select important features that are strongly associated with cancer prognosis. However, these methods often treat individual patient independently, and thus miss the strong ordinal relationship among the survival time of different patients (*i.e.*, patient A has longer survival time than patient B), which intuitively can benefit the following survival analysis. In addition, the direct combination of multi-modal data did not take the intrinsic relationship between the features across different modalities into consideration. As a matter of fact, the exploitation of multi-modal association has been widely accepted as a key component of current multi-modality based machine learning approaches [14], [15].

In this study, we take advantage of the ordinal survival information among different patients, and propose a novel ordinal multi-modality feature selection (OMMFS) method that can simultaneously identify important features from both pathological imaging data and multi-modal genomic data, (*i.e.*, mRNA expression, copy number variation and DNA methylation data), for the prognosis of cancer patients. Specifically, we formulate the proposed OMMFS method under generalized canonical correlation analysis (GCCA) framework, which aims at seeking linear transformation to ensure that the projection of each modality is approximated to a common subspace, and thus is capable of capturing the inter-correlation among different modalities. In addition, we also add linear constrains to preserve the ordinal survival information among different groups of patients, *i.e.*, for each data modality, the average projection of the patients from the long-term survival groups should be larger than that of short-term survival groups.

To validate the effectiveness of the proposed method, we test it on three early-stage cancer datasets derived from The Cancer Genome Atlas (TCGA), including kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP) and lung squamous cell carcinoma (LUSC), since all corresponding pathological image and multi-modal genomic data (*i.e.*, mRNA expression, copy number variation and DNA methylation data) of these three datasets are available in TCGA. Since prognosis is particularly important for early stage cancer patients, we focus on the patients in stages I and II whose prognosis usually cannot be effectively predicted from clinical phenotypes. The experimental results demonstrate that the combination of histopathological images with multi-dimensional genomic data can better predict patient outcome when comparing with several existing algorithms.

Our preliminary work that only combines image data with gene expression data for survival output prediction was reported on MICCAI 2018 [22]. In this journal paper, we have offered new contributions in the following aspects: 1) Extending the model that can combine pathological image data with multi-dimensional genomic data for survival analysis, which can better stratify patients with distinct survival than our preliminary work; 2) evaluating the effectiveness of the proposed method on two additional datasets (*i.e.*, KIPC, LUSC); 3) providing more comprehensive biomarkers for survival outcome prediction than our preliminary work; 4) comparing our method with the existing learning approaches on another two measurements (*i.e.*, concordance index and AUC); and 5) investigating the influence of the group number in the proposed OMMFS model.

II. Materials and Methods

A. Dataset

The Cancer Genome Atlas (TCGA) has generated multi-modal genome, epigenome, transcriptome, and imaging data for thousands of tumor samples across more than 30 types of cancers [16]. In this paper, we collect three early-stage (*i.e.*, stage I and stage II) cancer datasets from TCGA including kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP) and lung squamous cell carcinoma (LUSC) datasets. For each dataset, only subjects with completely matched copy number variation (CNV), DNA methylation, mRNA, histopathological image, and clinic outcome (*i.e.*, survival status and time) data are selected. The details of the cohort information is shown in Table. I, where censored patients mean that the death events of these patients were not observed during their follow-up, and their exact survival times are longer than their recorded data, while the non-censored category corresponds to the patients whose recorded survival times are the exact time from initial diagnosis to death.

B. Overview of Our Method

Fig. 1 shows the flowchart of our method, which consists of four steps. First, features are extracted from CNV, DNA methylation, mRNA expression, and histopathological image data; Secondly, for each data modality, log-rank test is adopted

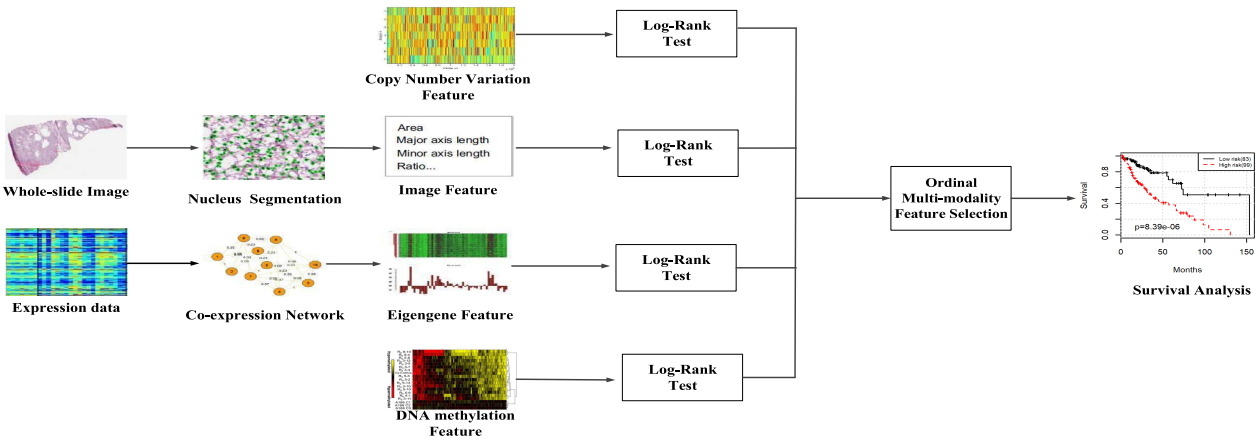

Fig. 1. The flowchart of the proposed method.

TABLE I
PATIENT DEMOGRAPHICS FOR DIFFERENT CANCER DATASETS

	KIRC	KIRP	LUSC
Censored/Non-Censored	196/45	165/16	220/145
Stage I / Stage II	200/41	163/18	215/150
Total	241	181	365

to pre-select candidate features that are relevant to survival analysis; Thirdly, a novel ordinal multi-modal feature selection (OMMFS) algorithm selects survival-associating features from the pre-selected features of different modalities; and lastly, the Cox proportional hazard model is applied for the prognostic prediction of early-stage cancer patients.

C. Feature Extraction and Pre-Selection

For cancer prognosis, different types of biomarkers (*i.e.*, histopathological image, mRNA, DNA methylation, copy number variation) are expected to provide complementary information. Therefore, features from all of the above four modalities are extracted to predict the clinical outcome. Specifically, for whole-slide images of each patient, $4-6$ regions of size 5K by 5K pixels are extracted by the experts. Based on the consideration that the progression of tumor is also supported and nurtured by the microenvironment around the cancer regions, the selected regions not only contain the tumor cells but also include the microenvironment cells including inflammation cells (*e.g.*, lymphocytes) and fibroblasts. Then, we apply an unsupervised segmentation algorithm [17] to segment cell nuclei from the images. After that, for each segmented nucleus, we extract 10 different types of features, *i.e.*, nuclear area (denoted as area), lengths of the major and minor axes of cell nucleus, and the ratio of major axis length to minor axis length (major, minor, and ratio), mean pixel values of nucleus in RGB these three channels (rMean, gMean, and bMean), and mean, maximum, and minimum distances (distMean, distMax, and distMin) to its neighboring nuclei. Next, for each type of feature, a 10-bin histogram and five statistic measurements (*i.e.*, mean, SD, skewness, kurtosis, and entropy) are used to aggregate the cell-level features into patient-level features, and thus a 150-dimensional imaging feature for each patient can be obtained. We use the same naming rules for both cell-level features and patient-level features. For instance, Area_bin1 represents the percentage of very small nuclei, while Area_bin10 indicates the percentage of very large nuclei in the patient sample. For gene expression data, to overcome the large number of genes which pose a challenge to the statistical power, we use the co-expression network analysis algorithms [9] to cluster genes into co-expressed modules and summarize each module into an eigengene using singular value decomposition. As to CNV data, we download its level-3 files, by which we can obtain the copy number estimation of each gene for different patients. In addition, we also use the level-3 DNA Methylation (DME) data, by which we can directly extract the methylation level of each methylation site (*i.e.*, beta-value) and its regulated gene symbol. After the feature extraction step, we finally identify 888 features (including 150 image features, 53 eigengene features, 339 DNA methylation (DME) features, and 346 copy number variation (CNV) features) for KIRC dataset; 750 features (including 150 image features, 66 eigengene features, 239 DME features, and 295 CNV features) for KIRP dataset; and 2011 features (including 150 image features, 58 eigengene features, 701 DME features, and 1102 CNV features) for LUSC dataset.

For the derived multi-view data, a standard log-rank test [6], which has been widely used in survival analysis, is performed to pre-select features in the training set. Specifically, for each feature, we divide patients into two groups (*i.e.*, low and high groups) according to its median value. Then, Kaplan-Meier estimator is used for patient stratification, and the p-value is calculated via log-rank test [6]. Finally, we have applied the multiple-testing compensation method described in [48] to keep the features whose False Discovery Rate (FDR) is small than 0.05 for the following integration step. After the pre-selection step via log-rank test, we derive 77 features (including 12 image features, 10 eigengenes features, 20 DME features and 35 CNV features) for the KIRC dataset; 67 features (including 10 image features, 10 eigengene features, 19 DME features and 28 CNV features) for the KIRP dataset; and 133 features (including 17 image features, 4 eigengene features, 87 DME features and 25 CNV features) for the

LUSC dataset. Moreover, we also compare the stratification performance of different feature pre-selection strategies (*i.e.*, log-rank test and univariate Cox regression method) in Supplementary Section S2, and the experimental results shown in Fig.S4 validates the advantage of the log-rank based feature pre-selection strategy for patient stratification.

D. Generalized Sparse Canonical Correlation Analysis (GSCCA)

Although the above log-rank based method could pre-select the survival-associated feature, it selects the features derived from different modalities independently and thus neglects to take into account the intrinsic relationship across different modailities. As a matter of fact, the exploitation of multi-modal association has been widely accepted as a key component of current multi-modality based machine learning approaches [22]. Accordingly, we further implement the second-stage feature selection model under Generalized Sparse Canonical Correlation Analysis (GSCCA) framework [44] that can capture the intrinsic relationship among multiple views. For the purpose of clarity, we begin by defining our notation below. Suppose $\boldsymbol{X_I} \in R^{N\times p}, \boldsymbol{X_M} \in R^{N\times q}, \boldsymbol{X_E} \in R^{N\times r}$ and $\boldsymbol{X_V} \in R^{N\times s}$ represent the image, DNA methylation, eigengene and copy number variation data, where N is the number of participants, and p, q, r, s are the dimensionalities of different modalities. The GSCCA model [44] solves the optimization problem in Eq. (1) by finding a shared representation $\boldsymbol{G} \in R^N$ of different modalities.

$$\begin{aligned} &\min_{\boldsymbol{G},\boldsymbol{\omega_i}} \quad \|\boldsymbol{G} - \boldsymbol{X_i}\boldsymbol{\omega_i}\|_2^2 + \sum_i r_i \|\boldsymbol{\omega_i}\|_1 \\ &s.t. \quad \|\boldsymbol{G}\|_2 = 1, \quad i \in \{I, M, E, V\} \end{aligned} \tag{1}$$

where the first term is used to approximate the projections of each modality, *i.e.*, $\boldsymbol{X}_i\boldsymbol{\omega}_i$, to the shared representation $\boldsymbol{G}$, and thus the inter-correlations across different modalities can be captured. The second L1-norm regularized terms are used to select a small number of feature from the pre-selected feature of each modality. Here, r_i $(i \in I, M, E, C)$ are regularization parameters and larger r_i value implies that fewer features in the i-th modality will be preserved for the following prognosis prediction.

E. Ordinal Multi-Modality Feature Selection (OMMFS)

In the GSCCA model, we only consider the inter-correlation among different modalities, and thus ignore survival information of different patients. To address this problem, we propose an ordinal multi-modality feature selection (OMMFS) algorithm to identify informative features under GSCCA framework. Specifically, we divide $\boldsymbol{X} = [\boldsymbol{X}_I, \boldsymbol{X}_M, \boldsymbol{X}_E, \boldsymbol{X}_V] \in R^{(N\times(p+q+r+s))}$ into $\boldsymbol{X}^C$ and $\boldsymbol{X}^{NC}$, where $\boldsymbol{X}^C \in R^{(k\times(p+q+r+s))}$ and $\boldsymbol{X}^{NC} \in R^{((N-k)\times(p+q+r+s))}$ correspond to the multi-modal features for censored and non-censored patients, respectively, and k denotes the number of censored patients. We also define $Y = [\boldsymbol{Y}^C; \boldsymbol{Y}^{NC}]$, where $\boldsymbol{Y}^C \in R^k$ and $\boldsymbol{Y}^{NC} \in R^{(N-k)}$ indicate the recorded last follow-up time and the survival time for censored and non-censored patients, respectively. Here, the exact survival times for censored patients are longer than the recorded last follow-up time since we have missed their death events. In the proposed OMMFS method, we divide all the patients (include both censored and non-censored patients) into $m = 4$ groups with equal size basing on the quartiles of $\boldsymbol{Y}$(include both the last follow-up time for censored patients and the survival time for non-censored patients), where the recorded time in group i is larger than that in group j if $i > j$, and vice versa. We denote the mean imaging, DNA methylation, eigengene, and copy number variation feature for censored patients in group j as $\boldsymbol{u}_i^j (i \in \{I, M, E, V\})$ and those for non-censored patients as $\boldsymbol{v}_i^j (i \in \{I, M, E, V\})$. Then, the objective function of the OMMFS model is defined as:

$$\begin{aligned} \min_{\boldsymbol{G},\boldsymbol{\omega_i},\theta_i^j,\varepsilon_i^j} & \sum_i \|G - \boldsymbol{X_i}\omega_i\|_2^2 + \alpha \sum_i \sum_j (\theta_i^j + \varepsilon_i^j) \\ & + \sum_i r_i \|\boldsymbol{\omega_i}\|_1 + \beta \left\| (\sum_i \boldsymbol{X}_i^{NC}\boldsymbol{\omega_i}) - \boldsymbol{Y}^{NC} \right\|_2^2 \end{aligned} \tag{2}$$

$$s.t. \quad (v_i^{j+1} - v_i^j)\omega_i > 1 - \theta_i^j \tag{3}$$

$$\begin{aligned} & (u_i^{j+1} - v_i^j)\boldsymbol{\omega_i} > 1 - \varepsilon_i^j \\ & \theta_i^j > 0, \quad \varepsilon_i^j > 0, \ i \in \{I, M, E, V\}, \ j = 1, 2, 3. \end{aligned} \tag{4}$$

where the first and third terms in Eq. (2) are the same as they are stated in the generalized sparse canonical correlation analysis (*i.e.*, GSCCA) model. The forth part in Eq. (2) is used to estimate the relationship between the multi-modal data and the survival time for non-censored patients, where $\boldsymbol{X}_i^{NC}$ denotes the data of non-censored patients for modality i. We add the linear inequalities in (3) to ensure that the ordinal survival information of different groups of non-censored patients is preserved after the projections being adopted on each modality of the multi-modal data. In addition, since the genuine survival time for censored patients are longer than their recorded data, we can also preserve the group-level ordinal relationship by adding the constrains (shown in Eq.(4)) that the average projections for the censored patients in groups $i + 1$ should be larger than that for non-censored patients in group i. Here, θ_i^j and ϵ_i^j are slack variables, and we can maximize the margin between the projections of different survival groups by minimizing the sum of these slack variables (second term in Eq.(2)).

F. Optimization Algorithm

For the objective function of the proposed OMMFS model, we adopt the alternating direction method of multipliers (ADMM) [19] algorithm to solve it. To change the problem in Eq. (3) into the ADMM form, we introduce the variables $\boldsymbol{J_i} (i \in \{I, M, E, V\})$ and $\boldsymbol{Q}$, and then the OMMFS model can be transformed as:

$$\begin{aligned} \min_{\boldsymbol{G},\boldsymbol{Q},\boldsymbol{\omega_i},\boldsymbol{J_i}} & \sum_i \|\boldsymbol{Q} - \boldsymbol{X_i}\boldsymbol{\omega_i}\|_2^2 + \alpha(\sum_i \sum_j ([b_{ij}]_+ + [c_{ij}]_+)) \\ & + \sum_i r_i \|\boldsymbol{J_i}\|_1 + \beta \left\| (\sum_i \boldsymbol{X_i^{NC}}\boldsymbol{\omega_i}) - \boldsymbol{Y^{NC}} \right\|_2^2 \\ s.t. \quad & \boldsymbol{Q} = \boldsymbol{G}, \boldsymbol{J_i} = \boldsymbol{\omega_i} \end{aligned} \tag{5}$$

where $[b_{ij}]_+ = \max(0, 1 - (\boldsymbol{v}_i^{j+1} - \boldsymbol{v}_i^j)\boldsymbol{\omega_i})$ and $[c_{ij}]_+ = \max(0, 1 - (\boldsymbol{u}_i^{j+1} - \boldsymbol{v}_i^j)\boldsymbol{\omega_i})$. Then, the augmented Lagrangian form of Eq. (5) can be written as:

$$\begin{aligned} &\mathrm{L}(\boldsymbol{\omega_i}, \boldsymbol{J_i}, \boldsymbol{G}, \boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{O_i}) \\ &= \sum_i \|\boldsymbol{G} - \boldsymbol{X_i}\boldsymbol{\omega_i}\|_2^2 + \sum_i r_i \|\boldsymbol{J_i}\|_1 \\ &\quad + \alpha \sum_i \sum_j ([b_{ij}]_+ + [c_{ij}]_+) + \beta \left\| \left(\sum_i \boldsymbol{X_i^{NC}}\boldsymbol{\omega_i}\right) - \boldsymbol{Y^{NC}} \right\|_2^2 \\ &\quad + \ < \boldsymbol{R}, \boldsymbol{G} - \boldsymbol{Q} > + \frac{\rho_1}{2} \|\boldsymbol{G} - \boldsymbol{Q}\|_2^2 + \sum_i < \boldsymbol{O_i}, \boldsymbol{\omega_i} - \boldsymbol{J_i} > \\ &\quad + \frac{\rho_2}{2} \sum_i \|\boldsymbol{\omega_i} - \boldsymbol{J_i}\|_2^2 \end{aligned} \tag{6}$$

where $\boldsymbol{R}$ and $\boldsymbol{O_i}$ are Lagrange multipliers. A general ADMM scheme for Eq. (6) repeats the following 4 main steps until convergence:

1) $\boldsymbol{\omega_i} \leftarrow \arg\min_{\omega_i} \mathrm{L}(\boldsymbol{\omega_i}, \boldsymbol{J_i}, \boldsymbol{G}, \boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{O_i})$*:* It is convex with respect to $\boldsymbol{\omega_i}$. However, $[b_{ij}]_+$ and $[c_{ij}]_+$ are not differentiable to $\boldsymbol{\omega_i}$. Instead, we calculate their sub-gradients $\nabla[b_{ij}]_+, \nabla[c_{ij}]_+$ according to the method in [45]. Then, the optimal $\boldsymbol{\omega_i}$ can be obtained by applying the gradient descent algorithm, which are shown in Eq. (7)

$$\begin{aligned} \boldsymbol{\omega_i} = &(2(\boldsymbol{X_i})^T \boldsymbol{X_i} + \rho_2 \boldsymbol{I} + 2\beta(\boldsymbol{X_i^{NC}})^T \boldsymbol{X_i^{NC}})^{-1} [2(\boldsymbol{X_i})^T \boldsymbol{G} \\ &- \alpha \sum_j (\nabla b_{ij} + \nabla c_{ij}) - \boldsymbol{O_i} + \rho_2 \boldsymbol{J_i} - 2\beta(\boldsymbol{X_i^{NC}})^T \\ &\left(\sum_{k \neq i} \boldsymbol{X_k^{NC}} \boldsymbol{\omega_k} - \boldsymbol{Y^{NC}}\right)] \end{aligned} \tag{7}$$

2) $\boldsymbol{J_i} \leftarrow \arg\min_{J_i} \mathrm{L}(\boldsymbol{\omega_i}, \boldsymbol{J_i}, \boldsymbol{G}, \boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{O_i})$*:* This optimization problem can be reformulated as:

$$\min_{J_i} \ \frac{\rho_2}{2} \|\boldsymbol{J_i} - \boldsymbol{\omega_i}\|_2^2 + r_i \|\boldsymbol{J_i}\|_1 - (\boldsymbol{O_i})^T \boldsymbol{J_i} \tag{8}$$

Since the $L1$-norm is non-differentiable at zero, a smooth approximation has been estimated for $L1$ term by including an extremely small value. Then, by taking the derivative regarding to $\boldsymbol{J_i}$ and let it to be zero, we can obtain:

$$\boldsymbol{J_i} = (r_i \boldsymbol{D_i} + \rho_2 \boldsymbol{I})^{-1} (\boldsymbol{O_i} + \rho_2 \boldsymbol{\omega_i}) \tag{9}$$

where $\boldsymbol{D_i}$ is a diagonal matrix with the $k-th$ element as $1/\left\|J_i^k\right\|_1$. Here, J_i^k denotes the $k-th$ element in $\boldsymbol{J_i}$.

3) $\boldsymbol{G} \leftarrow \arg\min_{G} \mathrm{L}(\boldsymbol{\omega_i}, \boldsymbol{J_i}, \boldsymbol{G}, \boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{O_i})$*:* We can solve the optimal $\boldsymbol{G}$ by applying the gradient descent algorithm as follow:

$$\boldsymbol{G} = \boldsymbol{Q} + \frac{1}{\rho_1} \boldsymbol{R} \tag{10}$$

4) $\boldsymbol{Q} \leftarrow \arg\min_{Q} \mathrm{L}(\boldsymbol{\omega_i}, \boldsymbol{J_i}, \boldsymbol{G}, \boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{O_i})$*:* It also has a close form solution:

$$\boldsymbol{Q} = (8 + \rho_1 \boldsymbol{I})^{-1} (2 \sum_i \boldsymbol{X_i}\boldsymbol{\omega_i} + \boldsymbol{R} + \rho_1 \boldsymbol{G}) \tag{11}$$

We show the optimization algorithm for the proposed OMMFS method in Table. II.

TABLE II
THE OPTIMIZATION ALGORITHM FOR THE PROPOSED OMMFS METHOD

The optimization procedure for OMMFS method.

Input:
$\boldsymbol{X_i}(i \in \{I, M, E, V\})$ *% The data matrix for multi-modal data.*
$\boldsymbol{Y} = [\boldsymbol{Y^C}, \boldsymbol{Y^{NC}}] \in R^N$ *% The recorded survival time.*
Initialize:
$\boldsymbol{\omega_i} = \boldsymbol{J_i} = \boldsymbol{O_i} = \boldsymbol{0}(i \in \{I, M, E, V\})$ *%Initialized as zero vector.*
$\boldsymbol{G} = \boldsymbol{Q} = \boldsymbol{R} = \boldsymbol{0}$ *%Initialized as zero vector.*
While not converge do:
Update $\boldsymbol{\omega_i}$ according to Eq. (7);
Update $\boldsymbol{J_i}$ according to Eq. (9);
Update $\boldsymbol{G}$ according to Eq. (10);
Update $\boldsymbol{Q}$ according to Eq. (11);
Update $\boldsymbol{R} = \boldsymbol{R} + \rho_1(\boldsymbol{G} - \boldsymbol{Q}), \boldsymbol{O_i} = \boldsymbol{O_i} + \rho_2(\boldsymbol{\omega_i} - \boldsymbol{J_i})$
End While
Output:
$\boldsymbol{\omega_i}(i \in \{I, M, E, V\})$ *% The non-zero elements in* $\boldsymbol{\omega_i}$ *correspond to the selected feature in each modality*

G. Prognostic Prediction

One important aspect of prognostic prediction is to effectively stratify cancer patients into subgroups with different prediction outcomes, and it is based on the survival analysis of different patients. The goal of survival analysis is to predict the time duration until the death event occurs and proportional hazards models [42] are a class of survival models in statistics, which assess the instantaneous rate of death at time t. We show the definition of hazard function as follows:

$$h(t|x) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq O \leq t + \Delta t | O \geq t; \boldsymbol{x})}{\Delta t} \tag{12}$$

where $\boldsymbol{x} = (x_1, x_2, \ldots x_n)$ coressponds to the covariate variable (*i.e.*, selected feature) of dimensionality n. In the hazards modeling methods, Cox proportional hazard model [21] is among the most popular ones. It models the hazards function as $h(t|x) = h_0(t) exp(\boldsymbol{\beta^T x})$. Here, $h_0(t)$ is the baseline hazard, and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots \beta_n)$ is the vector of regression parameters, which can be estimated by minimizing its corresponding negative log partial likelihood function [21]. It is worth noting that $f(\boldsymbol{x}) = \boldsymbol{\beta}^T \boldsymbol{x}$ is also being called as risk function and we follow the method in [6], which uses the median risk score predicted by the risk function as a threshold to stratify patients in the testing set into low-risk and high-risk groups. Finally, we test if these two groups has distinct survival outcome by using the Kaplan-Meier estimator and log-rank test [7].

III. EXPERIMENTAL RESULTS

A. Experimental Settings

To evaluate the performance of our proposed OMMFS method, we test it on 3 public available datasets (*i.e.*, KIRC, KIRP, and LUSC) derived from The Cancer Genome Atlas (*i.e.*, TCGA) database. For all of these three datasets, we randomly split all the patients into two parts (*i.e.*, training part and testing part). From the training set, we split 20% of them as the validation set. For parameter settings, we tune the parameters α and β in the OMMFS model from $\{75, 100, 125\}$ and $\{0.1, 0.5\}$ respectively, and the regularization parameters r_i $(i \in \{I, M, E, V\})$ are tuned from $\{5, 10\}$, $\{20, 25, 35\}$, $\{1, 5\}$ and $\{10, 20\}$, respectively.

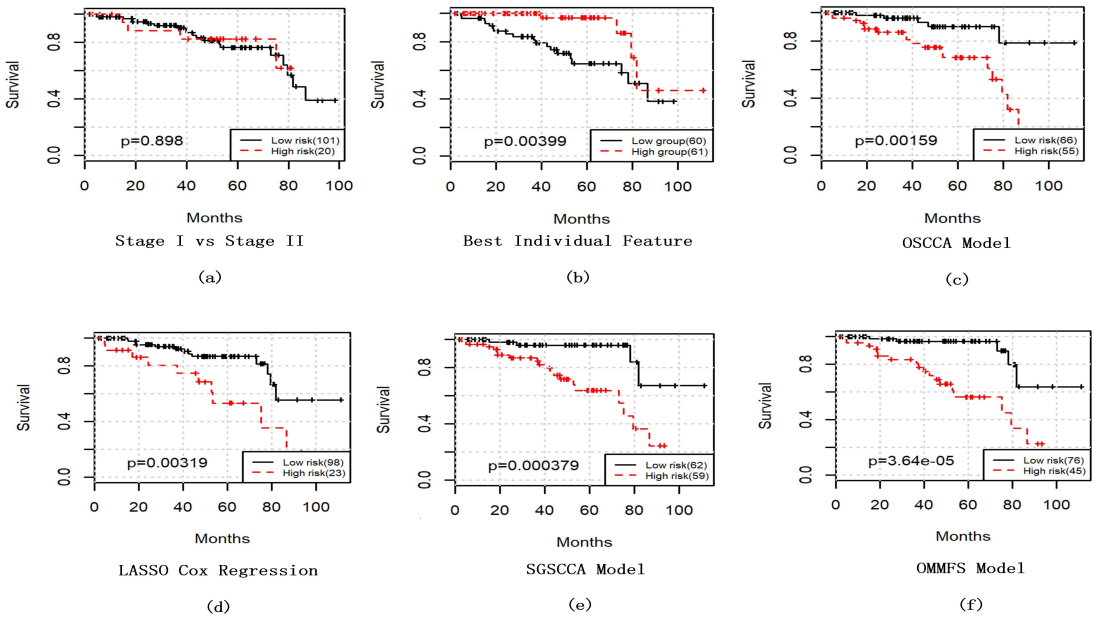

Fig. 2. The survival curves by applying different methods on KIRC dataset.

B. Comparison of OMMFS With Other Methods for Patient Stratification on KIRC Dataset

The prognostic power of our proposed OMMFS method was compared with the following five univariate or multivariate algorithms on early-stage KIRC (*i.e.*, kidney clear renal cell carcinoma) dataset. 1) Best Individual Feature: the most significant single feature (*i.e.*, with the smallest p-value) by applying log-rank test. 2)Stage: Stratify patients according to the stage (*i.e.*, stage I and stage II) information derived from TCGA. 3) Lasso-Cox regression [6]: Simply concatenate the image, eigengene, DNA methylation, and copy number variation feature together, and then use Lasso method for feature selection. 4) SGSCCA: Compared to OMMFS model, it contains the same objective function (shown in Eq.(2)), but dose not take the ordinal survival information (shown in the inequalities in (3) and (4)) into consideration. 5) OSCCA [22]: Our previous work, which considers the ordinal information for survival analysis, but only incorporate image and eigengene these two modalities. We show the survival curves of different methods in Fig. 2.

As we can observe from Fig. 2(a), the Kaplan-Meier curves for tumor stage I and stage II are intertwined (p-value= 0.898), which indicates that the stratification of early-stage KIRC patients is difficult. However, the proposed OMMFS method (Fig. 2(f)) could easily separate the patients into low and high risk groups, and achieve significantly superior stratification (P-value $= 3.64e-5$) than the best individual feature (Fig. 2(b), p-value= 0.00399) and the OSCCA method (Fig. 2(c), P-value= 0.00159), which combines pathological images with only one aspect of genomic activity (*i.e.*, eigengene) for survival analysis. These results clearly demonstrate the advantage of the integrative analysis of image and multi-dimensional genomic data for patient outcome prediction. Moreover, it is worth noting that the OMMFS model also can provide better prognostic prediction than the SGSCCA (Fig.2(e), p-value= $3.79e-4$) and Lasso methods (Fig.2(d), p-value= 0.00319), since the ordinal survival information and the correlation among different modalities for feature selection are preserved in OMMFS. More comparisons between OMMFS and other methods *i.e.*, SAM [46] and individual modality based methods that can verify the superiority of our method are reported in Supplementary Section S1 and S4, respectively.

TABLE III
THE SELECTED FEATURES BY APPLYING OMMFS MODEL ON KIRC DATASET

Feature Type	Selected Feature
Eigengene Feature	Eigengene 12, Eigengene 20, Eigengene 38
Image Feature	Area_bin9, rMean_bin1, rMean_mean
Methylation Feature	cg04797323(SOCS2),cg07935264(1L1B), cg16326979(OXR1)
CNV Feature	FGF7P3, GSTP1, RPS4Y2

Next, we also evaluate the prognostic power of the selected 12 features (shown in Table. III) using the proposed OMMFS method, with the false discovery rate (*i.e.,* FDR) are shown in Table S1 in the Supplementary Material. Specifically, for the selected image features, Area_bin9 is related to large nuclei. It has been demonstrated that the renal cell carcinoma patients with large values of nuclei size had worse prognosis than other patients [23]. As to eigengene features, three co-expression modules (corresponding to Module 12, Module 20 and Module 38) are identified (details are shown in Table. S7 in Supplementary Material). The gene ontology enrichment analysis for the selected co-expression modules using Ingenuity Pathway Analysis (IPA) (https://www.qiagen.com) show that Module 38 contains 13 genes in total and 11 of them are enriched with biological processes to cancer (P-value= $3.41e-4$). While the gene enrichment analysis by Toppgene (https://toppgene.cchmc.org/) indicates that Module 20 consists of 26 genes, and three genes (CNTN3,

TABLE IV
THE SELECTED FEATURES BY APPLYING OMMFS MODEL ON KIRP DATASET

Feature Type	Selected Feature
Eigengene Feature	Eigengene 1, Eigengene 50
Image Feature	Ratio_bin8, Minor_bin2
Methylation Feature	cg00839584(1L1A), cg27067621(NOX1), cg20358834(LRFN4)
CNV Feature	CNTNAP3B, GOLGA8B, NBPF20, PIP

TABLE V
THE SELECTED FEATURES BY APPLYING OMMFS MODEL ON LUSC DATASET

Feature Type	Selected Feature
Eigengene Feature	Eigengene 13, Eigengene 15, Eigengene 36
Image Feature	disMax_bin2, gMean_bin3
Methylation Feature	cg08047457(RASSF1), cg21634602(APC), cg16319578(HSPA2), cg19358493(EMX2), cg09851465(clorf87)
CNV Feature	EXPNP, FAM133B, GSTTP2 SLC25A18, ZIC4, ZNF630

ASTN1 and EPHA5) in it are Receptor Tyrosine Kinases (P-value= $2.81e-4$), which are proven to be linked to the pathophysiology of different types of cancers [24]. For DNA methylation feature and CNV feature, two genes are involved in cytokine receptor binding (SOCS2 [25], 1L1B [40]), which plays critical roles in tumor progression [41], and the single nucleotide polymorphism (SNP) in gene GSTP1 is also known to cause genetic damage, which increases cancer risk [26].

C. Comparison of OMMFS With Other Methods for Patient Stratification on KIRP Dataset

We also tested our OMMFS method on early-stage KIRP (kidney renal papillary cell carcinoma) dataset. Fig.3 presents the survival curves for the stratified patient groups by the proposed OMMFS model and several existing methods mentioned in Section 3.1 for comparison purposes. As can be seen from Fig.3, the proposed OMMFS method still achieves superior patients stratification performance (p-value= $2.16e-5$) than all of the other methods. Here, the better prognosis prediction of our method is mainly due to the same reasons as mentioned in the KIRC study.

Identified feature were listed in Table. IV, with their corresponding FDR values are shown in Table S2 in Supplementary Materials. As shown in Table. IV, the selected image features (*e.g.*, Ratio_bin8, Minor_bin2) correspond to the cells that are characterized by relatively large ratio between long and minor axes, which are the spindle-shaped cells. In fact, stromal cells such as fibroblasts are spindle-shaped, and it is known that the percentage of stromal tissues plays an important role in renal tumor prognosis [6]. For eigengene features, two co-expression modules are identified (corresponding to Module 1 and Module 50, details are shown in Table. S8 in the Supplementary Materials). IPA gene ontology enrichment analysis shows that Module 1 with 661 genes are significantly enriched with five canonical pathways (*i.e.*,71 genes in EIF2 signaling pathway (p-value= $7.74e-55$), 51 genes in oxidative phosphorylation (p-value= $1.72e-49$), 57 genes in mitochondrial dysfunction (p-value= $3.93e-45$), 43 genes in sirtuin signaling pathway (p-value= $7.61e-19$), and 32 genes in the regulation of eIF4 and p70S6K signaling (p-value= $1.36e-18$)). Among these 661 genes, 542 are related to cancer (p-value= $4.47e-2$) and 565 associated with organismal injury and abnormalities (p-value= $6.79e-3$). In Module 50 (12 genes in total), four genes (DDX3Y, KDM5D, USP9Y, UTY) involve in cellular growth and proliferation (p-value= $9.98e-3$), which is the key process in the cancer cell development. As to DNA methylation features, the genetic instability of IL1A contribute most to the development of the renal cell carcinoma [27], and multiple studies have revealed that the increased expression of NOX1 is associated with different cancers, including breast cancer [28] and colon cancer [29]. For CNV features, four genes (CNTNAP3B, GOLGA8B, NBPF20, PIP) have been identified in our OMMFS workflow, the enrichment analysis by Toppogene shows that PIP can negatively regulate lymphocyte apoptotic process (p-value= $2.57e-3$), and it has been previously shown that the renal tumors can escape immune recognition and destruction through the induction of apoptosis in activated T lymphocytes [30].

D. Comparison of OMMFS With Other Methods for Patient Stratification on LUSC Dataset

We also evaluate survival prediction power of our OMMFS method on early-stage LUSC (lung squamous cell carcinoma) dataset. Fig. 4 presents the Kaplan-Meier curves by applying different methods. As can be seen from Fig. 4, similar to the results in KIRC and KIRP datasets, our method still achieves the best patient stratification performance (p-value= $8.39e-6$) than the comparing methods. Besides, it is worth noting that, compared with the best individual variables (p-value= $9.27e-4$), the direct combination of multi-modal data (p-value= $3.02e-3$) cannot improve the accuracy of prognostic prediction for early-stage patients, which validates the importance of capturing the correlation among different modalities for patient outcome prediction.

After examining the survival-associated features (listed in Table. V), whose FDR values are shown in Table. S3 in the Supplementary Materials, two image features were identified (disMax_bin2, gMean_bin3). Among these two features, disMax_bin2 corresponds to the maximum distance between each segmented cell and its neighboring cells. The cells with smaller neighboring distance (disMax_bin2) usually refer to the cancer cells or lymphocytes that cluster together in the tissue images, and this topological feature has been established previously for tumor grading [31] and survival analysis [7]. In addition to histopathologic image feature, our model also selects three genomic co-expression modules *i.e.*, Module 13, Module 15, and Module 36 (details are shown in Table. S9 in the Supplementary Materials). Gene ontology enrichment analysis by Toppgene shows that Module 13 is enriched with

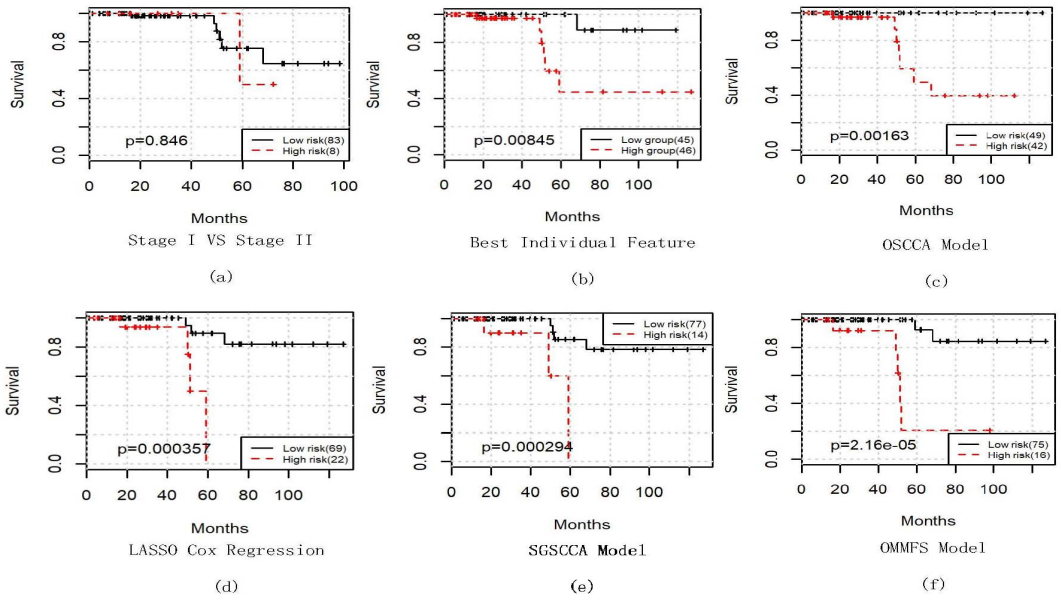

Fig. 3. The survival curves by applying different methods on KIRP dataset.

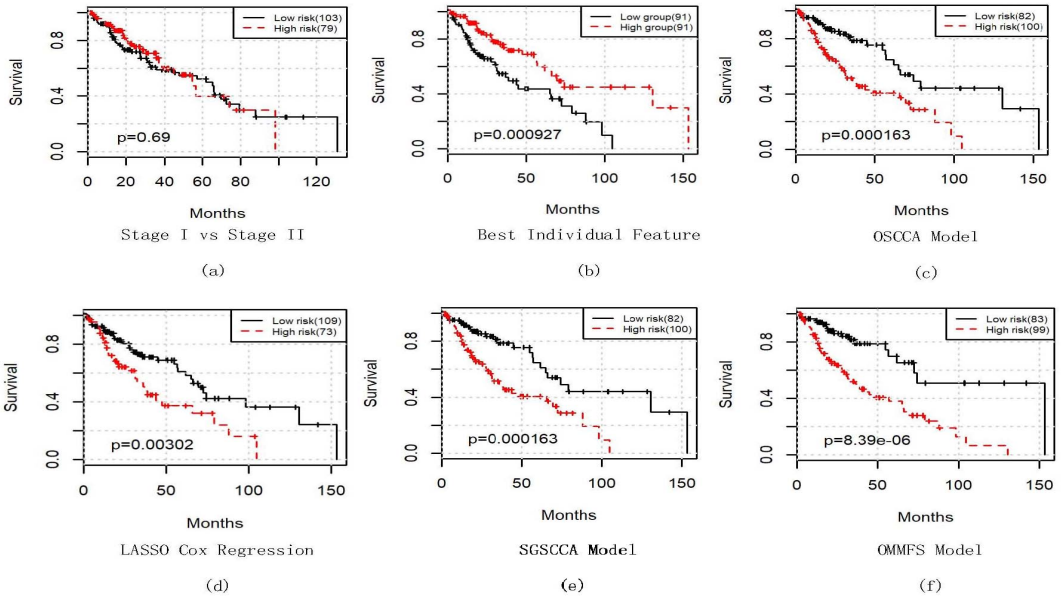

Fig. 4. The survival curves by applying different methods on LUSC dataset.

DNA binding transcription factors (p-value= $5.979e-7$), which is consistent with our knowledge that the regulation of transcription factor [32] plays an important role in cancer progression. Module 15 is enriched with genes that are associated with immune response (p-value= $1.21e-5$), and it is reported that the deregulation of the immune response genes are associated with the initiation and progression of cancers [33]. Also, the enrichment analysis shows that three genes (DLX5, DLX6, ZKSCAN5) involved in Module 13 are located at *7q22* (p-value= $2.977e-6$) and the structural abnormalities of *7q22* is the most frequently appeared phenomenon in lung cancer [34]. While four genes (ZNF567, ZFP14, ZNF260, ZNF566) in Module 36 are located at *19q13.12* (p-value= $6.244e-9$) and it is verified that the Genetic polymorphisms in chromosome *19q13* are in relation to lung cancer [35]. For DNA methylation feature, we totally derive five features. Specifically, the aberrant methylation of gene RASSF1 has been found in lung cancer [36] and it is proved that the gene HSPA2 correlates with clinical features in non-small cell lung carcinoma patients [37]. As for the modality of copy number variation, six genes are identified. The study in [38] has verified that 92% percentage of patients with ZIC4 antibody had small-cell lung cancer. These results again validate the advantage of the OMMFS method that can identify meaningful imaging and multi-dimensional genomic biomarkers for clinical outcomes prediction.

TABLE VI
COMPARISON OMMFS WITH OTHER METHODS ON THE MEASUREMENTS OF CONCORDANCE INDEX AND AUC

Method	KIRC		KIPC		LUSC	
	CI	AUC	CI	AUC	CI	AUC
LASSO	0.633	0.672	0.623	0.641	0.686	0.701
OSCCA	0.667	0.699	0.667	0.689	0.721	0.696
SGSCCA	0.723	0.743	0.723	0.742	0.716	0.734
OMMFS	0.767	0.774	0.743	0.768	0.773	0.793

IV. DISCUSSION

A. Comparison OMMFS With Other Methods on the Measurements of Concordance Index and AUC

Besides the comparison of different methods by the effectiveness of stratifying patients into high-risk and low-risk groups, we also take the Concordance Index (*i.e.*, CI) and AUC as evaluation metrics [43]. Here, CI is used to quantify the ranking quality at patient level, and is calculated as follows:

$$CI = \frac{1}{n} \sum_{i \in \{1 \ldots N\}} \sum_{y_j > y_i} I(X_i \beta > X_j \beta) \tag{13}$$

where n denotes the number of comparable pairs, y_i corresponds to the actual survival observation for the *i-th* patient, and $I(.)$ is the indicator function. On the other hand, the AUC index quantifies the ranking quality at event-time level, which can be calculated below:

$$AUC = \frac{1}{num} \sum_{t \in T} \sum_{y_i < t} \sum_{y_j > t} I(X_i \beta > X_j \beta) \tag{14}$$

where t represents the set of all possible event times in the dataset, and num represents the cumulative number of comparable pairs calculated over all event times. Both the value of CI and AUC range from 0 to 1. The larger CI and AUC value means the better prediction performance of the model and vice versa. The results are shown in Table. VI. As can be seen from Table. VI, the proposed method OMMFS method could achieve to the CI of $0.767, 0.743, 0.773$, the AUC of $0.774, 0.768, 0.793$ for the datasets of KIRC, KIPC and LUSC, respectively, which are higher than the comparing methods. These results again validate the superiority of the proposed method for survival prediction. More comparisons between OMMFS and other methods on the measurement of AIC are discussed in Section S.3 of the Supplementary Materials, which can also show the advantage of our method.

B. The Association Among the Selected Feature of Different Modalities

In order to investigate the association among the selected feature of different modalities. The average Pearson Correlation Coefficients between the projections of any two modalities $\boldsymbol{X_k \omega_k}$, $\boldsymbol{X_j \omega_j}((k, j \in \{I, M, E, V\}))$ on the testing datasets are shown in Fig. 5. From Fig. 5, we observe that the proposed OMMSF algorithm is capable of identifying higher correlation among the multi-modal data on all the three datasets (*i.e.*, KIRC, KIPC, LUSC). In addition, we also find that the selected imaging and genomic features (shown in Table.III, Table.IV and Table.V) are biologically correlated. Specifically, for KIRC dataset, the image feature Area_bin9 correlates with cancer cells, and the selected genomic features are also associated with the development of cancer cells. For instance, the Androgen-induced IncRNA in SOCS2 will inhibit apoptosis in cancer cells [49], GSTP1 [26] is a driver for cancer cell metabolism and pathogenicity, and the selected eigengene feature (*e.g.*, eigengene 20, eigengene 38) consists of genes that are also associated with the pathophysiology or biological process to cancers. As to KIRP dataset, the selected image features (Ratio_bin8, Minor_bin2) usually refer to the stromal cells like fibroblasts. For the selected genomic features, the expression level of NOX1 in stromal fibroblasts will regulate inflammation reaction, cell proliferation and migration [50], and the PIP gene has been shown to express in stromal cells that will control metabolic processes within cells [51]. Finally, for LUSC dataset, the selected image feature disMax_bin2 is associated with lymphocytes or cancer cells. For the selected genomic feature, Eigengene 15 is strongly enriched with immune response genes, which are closely associated with the lymphocytes function, and the study in [52] have verified that the over-expression of RASSF1C isoform exists in a subset of lung cancer cell lines. In conclusion, the more accurate capture of the inherent correlations among multi-modal data with biological interpretation may be the main reason that the OMMFS outperforms other currently available methods for survival prediction.

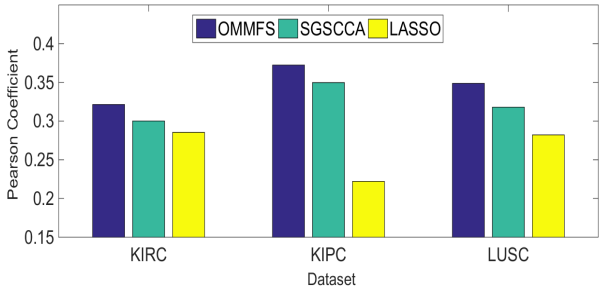

Fig. 5. The average Pearson Correlation Coefficients between the projections of any two modalities.

C. The Effects of Group Number in OMMFS Model

In this section, we evaluate the influence of the number of divided groups (*i.e.*, m) in the OMMFS model. Specifically, we vary the group number in the OMMFS from $\{3, 4, 6, 8, 10\}$, and record their corresponding p-value for patient stratification via log-rank test. The experimental results are shown in Table. VII. As can be seen from Table. VII, on the one hand, the p-value of OMMFS model fluctuates in a small range with the variance of the group number in all of the three datasets, which demonstrate that our model is robust to group number. On the other hand, when comparing the p-value shown in Fig. 2, Fig. 3, Fig. 4 and Table. VII, we find that no matter how many groups are divided, using Eq.(4) and Eq.(5) to preserve the ordinal information among different groups of patients could achieve to better stratification performance (*i.e.*, with smaller p-value) than the comparing methods (*i.e.*, LASSO, OSCCA, SGSCCA) for all of the three datasets.

These results again verify the advantage of adding ordinal survival information for patient outcome prediction.

D. The Effectiveness of the Two-Stage Bio-Marker Discovery Strategy

As introduced in Section II, we have developed a two-stage approach to select the optimal candidate features for patient stratification from multi-view data. To further validate the effectiveness of the above two-stage bio-marker discovery strategy, we evaluate the prognosis performance if we overlook applying the OMMFS algorithm or the pre-selection step for biomarker discovery, with the experimental results are shown in Fig. S2 and Fig. S3 in the Supplementary Materials, respectively.

As can be seen from Fig. S2 and Fig. S3, on one hand, the prognosis performance can be substantially improved by applying the proposed OMMFS method for further feature selection(shown in Fig. S2). These results clearly demonstrate that the consideration about the correlation among multi-modal data can improve the performance of patient outcome prediction. On the other hand, the proposed method can better stratify patients with different survival outcomes on all the datasets if the pre-selection step is included (shown in Fig. S3). This is because the initially extracted features may contain elements that are not significant for discrimination between different survival groups. After screening out these survival-uncorrelated features in the pre-selection step, the remaining feature candidates are all individually associated with cancer prognosis that will benefit to stratify patients into different survival groups.

E. Visualization of the Selected Image Features by OMMFS

In this section, we visualize the sample patches that are associated with the selected image features in both high and low survival risk groups, and the results are shown in Fig. 6.

As can be seen from Fig. 6, the first row visualizes sample patches corresponding to the selected image feature *i.e.,* Area_bin9, on KIRC dataset. Here, Area_bin9 indicates the percentage of large nuclei cells in the patient samples, and the patches in high survival risk group have higher percentage of large nuclei than that in low survival risk group. These results are consistent with the existing discovery that the renal cell carcinoma patients with large values of nuclei size have worse prognosis than the patients with small nuclei size [23]. The second row visualizes the sample patches that are related to the image features *i.e.*, Ratio_bin8, Minor_bin2 on KIRP dataset. As introduced in Section II.C, Ratio_bin8 corresponds to the spindle-shaped stromal cells that are characterized by relatively large ratio between long and minor axes, and Minor_bin2 also refers to the cells with small size, *e.g.,* lymphocytes. As can be observed from the second row of Fig.6, more lymphocytes in high survival risk group infiltrate into the stromal cells than that in low survival risk group, which is in accordance with the fact that the wide-spread of stromal–infiltrating lymphocytes is associated with short–term survival outcome in cancers [47]. The third row visualizes the sample patches with regards to the image feature disMax_bin2. Here, disMax_bin2 corresponds to the maximum distance between each segmented cell and its neighboring cells. The cells with smaller neighboring distance (disMax_bin2) usually refer to the cancer cells or lymphocytes that cluster together in the tissue images. As can be seen from the third row of Fig.6, the low density of lymphocytes or high cancer cell density will lead to the high survival risk and vice versa. These observations are consistent with the common knowledge that higher density of lymphocytes was associated with the strong immune systems that can fight infection, while the higher density of cancer cells usually comes with severe clinical disease [6].

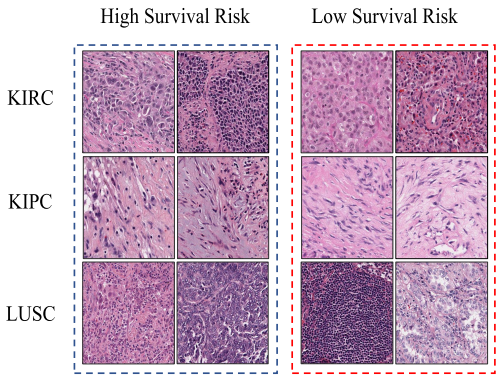

Fig. 6. Sample patches that are associated with the selected image features in both high and low survival risk groups.

F. The Robustness of the Cox Regression Model

As shown in Tables III, IV and V, we eventually select 12, 11 and 16 features for KIRC, KIPC and LUSC datasets, respectively. In order to verify the robustness of the Cox regression model, we test its prognosis performance after adding some noise to the selected features. Specifically, since each feature is normalized with center 0 and deviation 1 in the preprocessing step, we firstly add the Gaussian Noise that follows the distribution $N(0, \sigma I_{n\times n})$ to the selected features and σ is tuned from 0.0005,0.005 and 0.025. Then, based on the noisy data, we build Cox model and use it to stratify patients in the testing set into different survival risk groups, with the stratification performance evaluated via log-rank test. For each σ, we repeat the experiment 1000 times and report the average p-value in Table S4 in Supplementary Materials. As can be seen from Table S4, the Cox model is stable when adding Gaussian Noise with smaller variance ($\sigma = 0.0005, \sigma = 0.005$). After the variance increases to 0.025, the stratification performance only decreases in a small range, these results clearly verify the robustness of the Cox model based on our selected features.

G. Prognostic Groups Discovery and Stratification Performance Comparison with iCluster

Based on the multi-modal features that are selected by our OMMFS method, we also apply the iCluster algorithm [53] to discover novel prognostic groups on KIRC, KIRP and LUSC datasets. Here, we divide each dataset into three clusters,

TABLE VII
LOG-RANK TEST FOR PATIENT STRATIFICATION WITH DIFFERENT GROUP NUMBER IN OMMFS MODEL

Dataset	$m = 3$	$m = 4$	$m = 6$	$m = 8$	$m = 10$
KIRC	$2.25e-6$	$3.64e-5$	$7.33e-5$	$3.64e-5$	$8.64e-5$
KIPC	$6.98e-5$	$2.16e-5$	$2.16e-5$	$4.74e-5$	$8.62e-5$
LUSC	$8.98e-6$	$8.36e-6$	$1.39e-5$	$9.92e-6$	$7.24e-6$

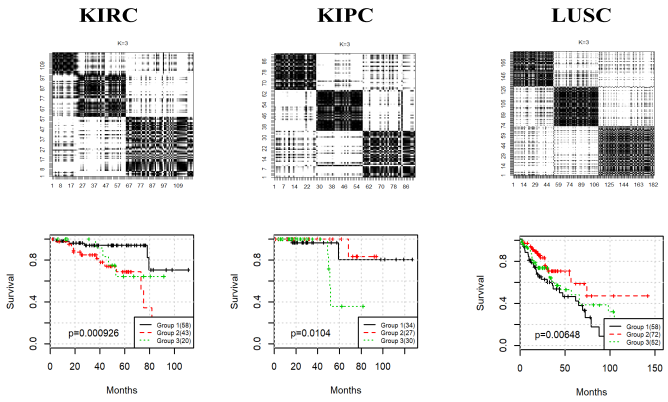

Fig. 7. The cluster separability plots and survival curves based on the clustering algorithm of iCluster [53].

with the cluster separability plots and the survival curves of different clusters are shown in Fig.7. As can be seen from Fig.7, the diagonal block structure are clearly visible in the cluster separability plots, which indicates that using the selected feature by our OMMFS method can lead to meaningful patient groups. In addition, we also find that the Kaplan-Mejer curves of different clustering groups show significantly different survival times (p-value<0.05), which also demonstrates the capacity of OMMFS for finding prognostic factors for patient outcome prediction.

In addition, we also compare the stratification performance of the proposed OMMFS method with the iCluster algorithm that is based on the pre-selected feature, with the results are shown in Fig. S6 in Supplementary Materials. As can be seen from Fig. S6, the proposed OMMFS algorithm achieve to better stratification performance than the iCluster approach. This could be partially due to that OMMFS incorporates the supervised survival information to build the prognosis prediction model, while the iCluster is an unsupervised technique without using survival information. In addition, as shown in Fig.2, Fig.3 and Fig.4, our method could also achieves better prognosis performance than the comparing supervised prediction models (*i.e.*, Lasso-Cox, OSCCA, SGSCCA) since it considers the ordinal survival information and the correlation among histopathological images and multi-dimensional genomic data that can benefit the prediction of clinical outcome.

V. CONCLUSION

In this study, we develop a feature selection method and workflow OMMFS which can effectively extract multi-modal features and identify survival associated biomarkers from both histopathological image and multi-dimensional genomic data. The proposed OMMFS algorithm is under generalized sparse canonical correlation analysis framework, in which we take patients ordinal survival information into consideration which has been generally ignored in survival studies. The results on three early-stage cancer datasets (*i.e.*, KIRC, KIRP, LUSC) have demonstrated that our method can achieve significantly better patient stratification performance than the currently available methods that can handle image and molecular features. Its superior performance on early-stage cancer patients means clinicians can intervene with personalized treatment plan in the early stage of tumor development, which can greatly increase the patient survival chance. OMMFS is a general framework and can be used to find multi-modal biomarkers for other types of cancers or predict response of specific treatment, which can contribute significantly to personalized treatment development in precision medicine. In addition, our feature pre-selection strategy is based on the median value of each feature, which is arbitrary without incorporating any prior distribution information. In future, we plan to determine the threshold according to the distribution of each feature, so that we can pre-select features that are more related to cancer prognosis for patient stratification.

REFERENCES

[1] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, and A. Jemal, "Global cancer statistics,2012," *Cancer J. Clinicians*, vol. 65, pp. 87–108, Mar. 2015.
[2] A. Calon *et al.*, "Stromal gene expression defines poor-prognosis subtypes in colorectal cancer," *Nature Genet.*, vol. 47, pp. 320–329, Apr. 2015.
[3] J. Xu *et al.*, "Stacked sparse Autoencoder (SSAE) for nuclei detection on breast cancer histopathology images," *IEEE Trans. Med. Imag.*, vol. 35, no. 1, pp. 119–130, Jan. 2016.
[4] K. H. Yu *et al.*, "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features," *Nature Commun.*, vol. 7, Aug. 2016, Art. no. 12474.
[5] Z. Huang *et al.*, "SALMON: Survival analysis learning with multi-omics neural networks on breast cancer," *Frontiers Genet.*, vol. 10, p. 166, Mar. 2019.

[6] J. Cheng *et al.*, "Integrative analysis of histopathological images and genomic data predicts clear cell renal cell carcinoma prognosis," *Cancer Res.*, vol. 77, pp. 91–100, Nov. 2017.

[7] J. Cheng, X. Mo, X. Wang, A. Parwani, Q. Feng, and K. Huang, "Identification of topological features in renal tumor microenvironment associated with patient survival," *Bioinformatics*, vol. 34, pp. 1024–1030, Mar. 2017.

[8] W. Hankey *et al.*, "Mutational mechanisms that activate Wnt signaling and predict outcomes in colorectal cancer patients," *Cancer Res.*, vol. 78, pp. 617–630, Feb. 2018.

[9] J. Zhang and K. Huang, "Pan-cancer analysis of frequent DNA co-methylation patterns reveals consistent epigenetic landscape changes in multiple cancers," *BMC Genomics*, vol. 18, pp. 1045–1055, Jan. 2017.

[10] Y. Yuan *et al.*, "Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling," *Sci. Transl. Med.*, vol. 4, pp. 143–157, Oct. 2012.

[11] F. C. Martins *et al.*, "Combined image and genomic analysis of high-grade serous ovarian cancer reveals PTEN loss as a common driver event and prognostic classifier," *Genome Biol.*, vol. 15, pp. 526–538, Dec. 2014.

[12] W. Li, S. Zhang, C. C. Liu, and X. J. Zhou, "Identifying multi-layer gene regulatory modules from multi-dimensional genomic data," *Bioinformatics*, vol. 28, pp. 2458–2466, Oct. 2012.

[13] Z. Yang and G. Michailidis, "A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data," *Bioinformatics*, vol. 32, no. 1, pp. 1–8, 2016.

[14] C. Xu, D. Tao, and C. Xu, "Multi-view intact space learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 12, pp. 2531–2544, Dec. 2015.

[15] M. Liu, D. Zhang, and D. Shen, "Relationship induced multi-template learning for diagnosis of Alzheimer's disease and mild cognitive impairment," *IEEE Trans. Med. Imag.*, vol. 35, no. 6, pp. 1463–1474, Jun. 2016.

[16] Y. Zhu, P. Qiu, and Y. Ji, "TCGA-Assembler: Open-source software for retrieving and processing TCGA data," *Nature Methods*, vol. 11, pp. 599–600, Jun. 2014.

[17] H. A. Phoulady, D. B. Goldgof, L. O. Hall, and P. R. Mouton, "Nucleus segmentation in histology images with hierarchical multilevel thresholding," *Med. Imag. Digit. Pathol. Int. Soc. Opt. Photon.*, vol. 23, Mar. 2016, Art. no. 979111.

[18] S. Huang *et al.*, "Prognostic significance of mixed-lineage leukemia (MLL) gene detected by real-time fluorescence quantitative PCR assay in acute myeloid leukemia," *Med. Sci. Monitor: Int. Med. J. Exp. Clin. Res.*, vol. 22, no. 1, pp. 3009–3017, 2016.

[19] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Comput. Optim. Appl.*, vol. 2, no. 1, pp. 17–40, 1976.

[20] M. Gustafsson, D. Tayli, C. Ehrenborg, M. Cismasu, and S. Nordebo, "Antenna current optimization using MATLAB and CVX," *FERMAT*, vol. 15, no. 5, pp. 1–29, 2016.

[21] D. Y. Lin and L. J. Wei, "The robust inference for the Cox proportional hazards model," *J. Amer. Stat. Assoc.*, vol. 84, pp. 1074–1078, Dec. 1989.

[22] W. Shao *et al.*, "Ordinal multi-modal feature selection for survival analysis of early-stage renal cancer," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Sep. 2018, pp. 648–656.

[23] H. Kanamaru, H. Akino, Y. Suzuki, S. Noriki, and K. Okada, "Prognostic value of nuclear area index in combination with the world health organization grading system for patients with renal cell carcinoma," *Urology*, vol. 57, pp. 257–261, Feb. 2001.

[24] M. K. Paul and A. K. Mukhopadhyay, "Tyrosine kinase—Role and significance in cancer," *Int. J. Med. Sci.*, vol. 1, pp. 101–107, Jun. 2004.

[25] J. U. Kazi and L. Rönnstrand, "Suppressor of cytokine signaling 2 (SOCS2) associates with FLT3 and negatively regulates downstream signaling," *Mol. Oncol.*, vol. 7, pp. 693–703, Jun. 2013.

[26] P. Zimniak *et al.*, "Naturally occurring human glutathione S-transferase GSTP1-1 isoforms with isoleucine and valine in position 104 differ in enzymic properties," *Eur. J. Biochem.*, vol. 224, pp. 893–899, Sep. 1994.

[27] E. Diakoumis, G. Sourvinos, H. Kiaris, D. Delakas, A. Cranidis, and D. A. Spandidos, "Genetic instability in renal cell carcinoma," *Eur. Urol.*, vol. 33, pp. 227–232, Feb. 1998.

[28] J. A. Choi *et al.*, "Pro-survival of estrogen receptor-negative breast cancer cells is regulated by a BLT2—Reactive oxygen species-linked signaling pathway," *Carcinogenesis*, vol. 31, pp. 543–551, Sep. 2009.

[29] E. Laurent *et al.*, "Nox1 is over-expressed in human colon cancers and correlates with activating mutations in κ-Ras," *Int. J. Cancer*, vol. 123, pp. 100–107, Jul. 2018.

[30] R. G. Uzzo *et al.*, "Mechanisms of apoptosis in T cells from patients with renal cell carcinoma," *Clin. Cancer Res.*, vol. 5, pp. 1219–1229, May 1999.

[31] C. Demir and B. Yener, "Automated cancer diagnosis based on histopathological images: A systematic survey," Rensselaer Polytech. Inst., Tech. Rep. 5, 2005.

[32] A. S. Baldwin, "Control of oncogenesis and cancer therapy resistance by the transcription factor NF-κB," *J. Clin. Invest.*, vol. 107, pp. 241–246, Feb. 2001.

[33] S. Y. Kim *et al.*, "Deregulation of immune response genes in patients with Epstein-Barr virus-associated gastric cancer and outcomes," *Gastroenterology*, vol. 14, pp. 137–147, Jan. 2010.

[34] S. Berker-Karaüzüm, G. Lüleci, G. Özbilim, A. Erdogan, A. Kuzucu, and A. Demircan, "Cytogenetic findings in thirty lung carcinoma patients," *Cancer Genet. Cytogenetics*, vol. 100, pp. 114–123, Jan. 1998.

[35] M. N. Timofeeva *et al.*, "Genetic polymorphisms in 15q25 and 19q13 loci, cotinine levels, and risk of lung cancer in EPIC," *Cancer Epidemiol. Prevention Biomarkers*, vol. 20, pp. 2250–2261, Oct. 2011.

[36] H. Chen *et al.*, "Aberrant methylation of RASGRF2 and RASSF1A in human non-small cell lung cancer," *Oncol. Rep.*, vol. 15, pp. 1281–1285, May 2006.

[37] D. Scieglinska *et al.*, "HSPA2 is expressed in human tumors and correlates with clinical features in non-small cell lung carcinoma patients," *Anticancer Res.*, vol. 34, pp. 2833–2840, Jun. 2014.

[38] L. Bataller, D. F. Wade, F. Graus, H. D. Stacey, M. R. Rosenfeld, and J. Dalmau, "Antibodies to Zic4 in paraneoplastic neurologic disorders and small-cell lung cancer," *Neurology*, vol. 62, pp. 778–782, Mar. 2004.

[39] D. Zhang, Y. Wang, L. Zhou, H. Yuan, and D. Shen, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," *NeuroImage*, vol. 55, no. 3, pp. 856–867, 2011.

[40] A. J. P. Smith and S. E. Humphries, "Cytokine and cytokine receptor gene polymorphisms and their functionality," *Cytokine Growth Factor Rev.*, vol. 20, pp. 43–59, Feb. 2009.

[41] M. Lee and I. Rhee, "Cytokine signaling in tumor progression," *Immune Netw.*, vol. 17, pp. 214–227, Aug. 2017.

[42] D. Kumar and D. Klefsjoe, "Proportional hazards model," *Int. J. Rel. Qual. Saf. Eng.*, vol. 1, pp. 337–352, May 2014.

[43] H. C. Chen, R. L. Kodell, K. F. Cheng, and J. J. Chen, "Assessment of performance of survival prediction models for cancer prognosis," *Med. Res. Methodol.*, vol. 12, p. 102, Dec. 2012.

[44] D. M. Witten and R. J. Tibshirani, "Extensions of sparse canonical correlation analysis with applications to genomic data," *Stat. Appl. Genet. Mol. Biol.*, vol. 8, no. 1, pp. 1–27, 2009.

[45] X. Qin, X. Tan, and S. Chen, "Mixed bi-subject kinship verification via multi-view multi-task learning," *Neurocomputing*, vol. 214, pp. 350–357, Nov. 2016.

[46] V. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc. Nat. Acad. Sci.*, vol. 98, pp. 5116–5121, Apr. 2001.

[47] T. Khoury, V. Nagrale, M. Opyrchal, X. Peng, D. Wang, and S. Yao, "Prognostic significance of stromal versus intratumoral infiltrating lymphocytes in different subtypes of breast cancer treated with cytotoxic neoadjuvant chemotherapy," *Appl. Immunohistochem. Mol. Morphology*, vol. 26, pp. 523–532, Sep. 2018.

[48] J. D. Storey, J. E. Taylor, and D. Siegmund, "Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach," *J. Roy. Stat. Soc., Ser. B*, vol. 66, pp. 187–205, Feb. 2004.

[49] A. Misawa, K. Takayama, T. Urano, and S. Inoue, "Androgen-induced long noncoding RNA (lncRNA) SOCS2-AS1 promotes cell growth and inhibits apoptosis in prostate cancer cells," *J. Biol. Chem.*, vol. 291, pp. 17861–17880, Aug. 2016.

[50] W. O'Brien, T. Heimann, and F. Rizvi, "NADPH oxidase expression and production of superoxide by human corneal stromal cells," *Mol. Vis.*, vol. 15, pp. 2535–2543, Dec. 2009.

[51] Q. Wang *et al.*, "Suppression of OSCC malignancy by oral glands derived-PIP identified by iTRAQ combined with 2D LC-MS/MS," *J. Cellular Physiol.*, vol. 28, pp. 1–10, Jan. 2019.

[52] M. E. Reeves, M. Firek, S. T. Chen, and Y. G. Amaar, "Evidence that RASSF1C stimulation of lung cancer cell proliferation depends on IGFBP-5 and PIWIL1 expression levels," *PloS one*, vol. 9, Jul. 2014, Art. no. e101679.

[53] R. Shen, A. B. Olshen, and M. Ladanyi, "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis," *Bioinformatics*, vol. 25, no. 22, pp. 2906–2912, 2009.