Mutual-Assistance Learning for Standalone Mono-Modality Survival Analysis of Human Cancers

Zhenyuan Ning, *Graduate Student Member, IEEE*, Zhangxin Zhao, Qianjin Feng, *Member, IEEE*, Wufan Chen, *Senior Member, IEEE*, Qing Xiao, *Graduate Student Member, IEEE*, and Yu Zhang, *Member, IEEE*

Abstract—Current survival analysis of cancers confronts two key issues. While comprehensive perspectives provided by data from multiple modalities often promote the performance of survival models, data with inadequate modalities at the testing phase are more ubiquitous in clinical scenarios, which makes multi-modality approaches not applicable. Additionally, incomplete observations (i.e., censored instances) bring a unique challenge for survival analysis, to tackle which, some models have been proposed based on certain strict assumptions or attribute distributions that, however, may limit their applicability. In this paper, we present a mutual-assistance learning paradigm for standalone mono-modality survival analysis of cancers. The mutual assistance implies the cooperation of multiple components and embodies three aspects: 1) it leverages the knowledge of multi-modality data to guide the representation learning of an individual modality via mutual-assistance similarity and geometry constraints; 2) it formulates mutual-assistance regression and ranking functions independent of strong hypotheses to estimate the relative risk, in which a bias vector is introduced to efficiently cope with the censoring problem; 3) it integrates representation learning and survival modeling into a unified mutual-assistance framework for alleviating the requirement of attribute distributions. Extensive experiments on several datasets demonstrate our method can significantly improve the performance of mono-modality survival model.

Index Terms—Mutual-assistance learning, mono-modality test, censored data, survival analysis, human cancer

1 Introduction

Cancer poses a serious challenge to increase the life expectancy of human beings [1]. According to estimates from the World Health Organization (WHO), cancer is the second leading cause of death globally, accounting for one in six deaths [2]. Clinically, the prognosis evaluation of cancers, as a branch of survival analysis in the healthcare domain, can aid physicians significantly in risk-benefit assessment and clinical decision by predicting the occurrence time/risk of death or recurrence so as to improve the survival rate of patients [3]. Unfortunately, subjective evaluation by physicians on the basis of baseline characteristics (e.g., tumor grade and stage) may suffer from low accuracy and poor repeatability [4]. Therefore, it is desirable to construct a robust and accurate model over clinical data for the prognosis prediction of cancers.

With the explosive development of data acquisition and storage techniques, data from multiple modalities or sources can be collected to bring comprehensive insights into the mechanism of cancers. Current studies have demonstrated that incorporating abundant information from multi-modality data has great potential to promote the performance of model in most scenarios [5], [6], [7], [8], [9], [10]. However, data with inadequate modalities (especially mono-modality data) at the testing phase are more ubiquitous in clinical practices due to the high collection costs or the poor data quality, which makes existing well-established multi-modality approaches not applicable. Under such circumstances, some work has utilized imputation techniques to complement incomplete multi-modality data and make them applicable for the multi-modality models, which, however, would introduce unnecessary noises especially when the original multi-modality feature space contains outliers or has the shifted feature distribution and might result in a sub-optimal solution [11], [12], [13], [14], [15]. To take full advantage of multi-modality information and generalize it to generic clinical scenarios, an appealing way is to utilize additional information from multi-modality data available in the training phase to reinforce the

- *The authors are with the School of Biomedical Engineering, the Guangdong Provincial Key Laboratory of Medical Image Processing and the Guangdong Province Engineering Laboratory for Medical Imaging and Diagnostic Technology, Southern Medical University, Guangzhou 510515, China. E-mail: {jonnyning, zhangxin_zhao, felicity_shaw}@foxmail.com, {fengqj99, chenwf, yuzhang}@smu.edu.cn.*

Manuscript received 19 April 2022; revised 5 August 2022; accepted 13 November 2022. Date of publication 16 November 2022; date of current version 5 May 2023.
This work was supported in part by the National Natural Science Foundation of China under Grants U22A20350 and 61971213, in part by the Basic and Applied Basic Research Foundation of Guangdong Province under Grant 2019A1515010417, and in part by the Guangdong Provincial Key Laboratory of Medical Image Processing under Grant 2020B1212060039.
This work involved human subjects or animals in its research. The author(s) confirm(s) that all human/animal subject research procedures and protocols are exempt from review board approval.
(Corresponding authors: Qing Xiao and Yu Zhang.)
Recommended for acceptance by G. Carneiro.
This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2022.3222732, provided by the authors.
Digital Object Identifier no. 10.1109/TPAMI.2022.3222732

standalone survival representation learning of mono-modality data and then improve the performance of the mono-modality model in the testing phase.

As a unique phenomenon in survival data, the incomplete observation (also referred to as censoring) issue makes the survival prediction task different from regression or classification tasks [16], [17]. The censoring issue means that partial instances do not experience the occurrence of the event of interest during the observation so the observed time is not consistent with the true event occurrence time, which is usually attributed to time limitation and subject withdrawal [18]. On the basis of the true event occurrence time of instances, the censoring issue typically occurs in one of the following three ways, i.e., left-censoring, right-censoring, and interval-censoring [19]. In general, most cancer survival data belong to the right-censoring issue, which means that the real survival time of the censored instances is longer than their recorded time [20], [21], [22]. To tackle this issue, many survival analysis methods have been proposed and widely used for cancer prognosis evaluation [23], [24], [25], [26]. Depending on the target of interest (i.e., survival time or event order of instances) output by the model, these survival analysis techniques can be roughly classified into two categories, namely, regression-based and ranking-based approaches [27]. However, most existing methods typically require prior knowledge of attribute distributions or make relatively strong assumptions, which would limit their applicability since the distribution and assumption are generally violated in various clinical scenarios [28]. A desirable way is to integrate both representation learning and survival modeling into a unified framework, so that the distribution of latent representations is adaptive for the survival model in a data-driven manner while relaxing restriction condition for survival model construction. Additionally, recent studies have demonstrated that the incorporation of ranking and regression constraints can further improve the performance of survival models [27], [29], on which, to the best of our knowledge, research has rarely been conducted.

In this paper, we propose a mutual-assistance learning (MaL) paradigm for effectively promoting standalone mono-modality survival analysis of human cancers. The mutual assistance implies the cooperation of multiple components and embodies three aspects. 1) To make use of multi-modality information and perform standalone mono-modality prediction, the proposed method first learns a low-rank latent subspace for mono-modality data, and then leverages the knowledge of multi-modality data to reinforce the subspace learning of an individual modality via mutual-assistance similarity and geometry (i.e., distance and angle) constraints such that mono-modality latent representations can preserve the globally structural and locally geometric characteristics of multi-modality data. 2) For wide applications, the proposed method formulates mutual-assistance regression and ranking functions independent of strong hypotheses to estimate the relative risk of survival. The regression function focuses on survival time estimation, in which a bias vector is introduced to efficiently cope with the censoring problem. And the estimated result is calibrated by the ranking function which takes the event order into consideration. 3) To strengthen the connection between representation learning and survival analysis and ease the requirement of attribute distributions, the proposed method integrates representation learning and survival modeling into a unified mutual-assistance framework that can be efficiently solved by the proposed optimization algorithm.

The main contributions of this paper lie in the following aspects:

- We propose a mutual-assistance learning paradigm for standalone mono-modality cancer survival analysis and its effectiveness is demonstrated via extensive experiments on several public datasets.
- We devise mutual-assistance similarity and geometry (i.e., distance and angle) constraints to assist mono-modality latent representation learning, through which the holistically structural and locally geometric characteristics of multi-modality data can be preserved so as to improve the performance of the mono-modality model.
- We design a mutual-assistance regression-ranking survival model without the constraint of strong hypotheses, which focuses on survival time and event order simultaneously. Additionally, a simple and efficient strategy (i.e., the bias vector) is introduced to deal with the censoring issue.
- We integrate representation learning and survival analysis into a unified mutual-assistance framework so as to make the distribution of latent representations adaptive for the survival model in a data-driven manner, and propose an optimization algorithm to solve it efficiently.

The remaining portion of this paper is organized as follows. In Section 2, we briefly review the related work. The proposed method and its optimization algorithm are introduced in Section 3 and Section 4, respectively. Section 5 presents experimental settings and results. The discussion and conclusion are drawn in Sections 6 and 7, respectively.

2 Related Work

2.1 Survival Representation Learning With Mono-Modality Data

Representation learning aims to seek a new feature subspace with specific characteristics (e.g., low-rank property and sparseness) from raw data for the downstream tasks, based on which many survival models with mono-modality data have been proposed for cancer prognosis prediction [30], [31], [32], [33], [34], [35], [36], [37]. For example, Yu et al. [32] selected top features from histopathological images using the regularized machine learning method for lung cancer prognosis. Zhu et al. [33] presented an effective framework to exploit all discriminative patterns in histopathological images for survival prediction. Huang et al. [34] constructed a variety of deep learning models to learn low-dimensional representations from genomic data for 12 cancers. However, such approaches only utilize information from a view (modality) and can not benefit from additional information inherent in other available views.

2.2 Survival Representation Learning With Multi-Modality Data

Recently, many studies have demonstrated that multi-modality data can advance in cancer prognosis analysis by providing complementary information, and multi-modality representation learning has rapidly become a hot research topic [5], [6], [7], [8], [9], [10]. For example, Shao et al. [7] proposed an ordinal multi-modal feature selection framework to conduct early-stage cancer prognosis prediction based on multi-dimensional genomic data and histopathological images. Zhang et al. [8] made an attempt to incorporate multi-omics data and histopathological images via a multiple kernel learning strategy to improve the performance of the survival model for glioblastoma multiforme. In addition, as an advanced technology, deep representation learning is also wisely used for multi-modality cancer survival analysis and achieves promising performance [38], [39], [40], [41], [42], [43], [44], [45]. For instance, Vale-Silva et al. [39] presented a multi-modal deep learning method for long-term pan-cancer survival analysis. Azher et al. [40] developed a hybrid multi-modal modeling approach to integrate comprehensive information from several modalities data for cancer prognosis prediction. Cheerla et al. [41] constructed a multi-modal deep neural network for prognosis evaluation, in which an unsupervised encoder was utilized to compress several modalities into the fused feature representations. However, due to high collection cost or poor data quality (e.g., stain color variations [46]), available samples with complete multi-modality data are often limited at the testing phase. Our work focuses on an especially extreme case where only a single modality is available in the testing phase. In order to make existing multi-modality survival approaches applicable to such a case, some work has proposed to impute incomplete multi-modality data by adopting some distance-based (e.g., hot-deck and k-nearest neighbor) and low-rank-based methods (e.g., expectation maximization and singular value decomposition (SVD)) [11], [12], [14], [15], [47]. Although these imputation techniques make the existing multi-modality models applicable to incomplete-modality data, they are likely to introduce unnecessary noises especially when the original multi-modality feature space contains outliers or has the shifted feature distribution. From another perspective, since these methods generally performed data imputation and multi-modality feature learning separately, it might result in a sub-optimal solution since some useful information for data imputation may be compromised in the feature learning stage. To this end, Ning et al. [48] presented a partial mapping strategy to deal with incomplete data during representation learning, in which projection/reconstruction matrices were partially trained by the incomplete multi-modality data. Even though this method does not introduce extra noise, it learns multi-modality shared representations that are globally optimal but not necessarily optimal for individual modality when the mono-modality representations are obtained by the corresponding partial projection matrix. We argue that incorporating additional information from multi-modality data available in the training phase can reinforce the standalone survival representation learning of mono-modality data and then improve the performance of mono-modality model in the testing phase.

2.3 Survival Model

Many survival methods have been proposed to handle right-censored data in the health care domain, which can be typically grouped into two categories: ranking-based and regression-based approaches according to the output type of the model [27]. The former aims at predicting the order of the event of interest, while the goal of the latter is to estimate the time to the event of interest along with its probability [49]. Although these methods have been successfully applied to cancer prognosis evaluation, they generally require prior knowledge of attribute distributions and parameter assumptions [28]. For instance, the Cox proportional hazard model [23] and its variants (e.g., Lasso-Cox [50] and EN Cox [51]) follow the proportional hazard assumption that the hazard ratio of two individuals is constant over time (independent of time) and the attributes are assumed to have an exponential influence on the outcome. Linear regression models [24], [25], [52] presume that the observed time can be regressed by a linear combination of the attributes, and some of them (e.g., Tobit model [24]) require extra condition that the error term complies with a specific distribution (e.g., Gaussian distribution). The Accelerated Failure Time (AFT) model [25] and its variants suppose that the relationship between the covariate and the logarithm of the observed time is linear, additionally requiring that the error variable adheres to a certain parameter distribution. However, if attribute distributions and parameter assumptions are violated, the estimates generated by these methods might be inconsistent and result in performance degradation [28]. Although the clinical attributes (e.g., gene data or image features) in the original space commonly fail to meet the distribution condition, a potentially feasible strategy is to learn latent representations tailored for survival model, in which representation learning and survival modeling are integrated into a unified stage so that the distribution of latent representations is adaptive for the survival model in a data-driven manner. Meanwhile, softening the restriction of parameter assumptions in survival models will be also beneficial to efficiently coping with the censored data in various clinical scenarios. On the other hand, recent studies have demonstrated that the incorporation of ranking and regression constraints can further improve the performance of survival models [27], [29]. For example, in [27], the SVM-based survival model imposed both ranking and regression constraints on the soft threshold operation construction and outperformed the model with either ranking or regression constraints. Nevertheless, it relies on a relatively complex basic form (i.e., SVM), which may suffer from time-consuming computation and difficult optimization. Also, it requires the same assumption as the basic form and some important predefined parameters (e.g., kernel type).

3 Mutual-Assistance Learning for Standalone Mono-Modality Survival Analysis

In this part, we detailedly introduce our proposed MaL paradigm for standalone mono-modality survival analysis, including 1) **MA-1**: unified mutual-assistance framework of representation learning and survival modeling (Section 3.1); 2) **MA-2**: multi-modality reinforced mono-modality representation learning with mutual-assistance constraints (Section 3.2); 3) **MA-3**: mutual-assistance regression-ranking survival modeling (Section 3.3). Main notations used in this paper are summarized in Table S1 (in *Supplementary Materials*) for reference, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2022.3222732.

3.1 MA-1: Unified Mutual-Assistance Framework of Representation Learning and Survival Modeling

Let the triplet $\mathcal{N} = (\mathbf{X}, \mathbf{T}, \boldsymbol{\delta})$ be the sample collection, where $\mathbf{X} \in \mathbb{R}^{n\times m_x}$ represents the mono-modality feature matrix with n samples and m_x features, and $\mathbf{T} \in \mathbb{R}^{n\times 1}$ and $\boldsymbol{\delta} \in \mathbb{R}^{n\times 1}$ ($\boldsymbol{\delta}^{(i)} = 1$ means uncensoring, or $\boldsymbol{\delta}^{(i)} = 0$) denote the observed time and event indicator of samples, respectively. We further suppose that $\mathbf{Z} \in \mathbb{R}^{n\times m_z}$ (where $m_z \gg m_x$) represents the available multi-modality feature matrix. For example, given data from two modalities, $\mathbf{X}^{\{1\}} \in \mathbb{R}^{n\times m_{x_1}}$ and $\mathbf{X}^{\{2\}} \in \mathbb{R}^{n\times m_{x_2}}$, we have $\mathbf{Z} = [\mathbf{X}^{\{1\}}, \mathbf{X}^{\{2\}}]$ and $m_z = m_{x_1} + m_{x_2}$. To perform prognosis prediction, a typical strategy is to seek a feature subset in the original feature space and use them to estimate the observed time. However, such a strategy is easily affected by the noises contained in the original feature space and the feature distribution commonly does not meet the requirement of prognosis models. Therefore, it is expected to project the original mono-modality feature space into a clean latent space (i.e., $\mathcal{R}(\mathbf{X} \rightarrow \mathbf{Y})$, where $\mathbf{Y} \in \mathbb{R}^{n\times m_y}$ denotes the latent representations and m_y is the corresponding dimension), and integrate it with the prognosis process (i.e., $\mathcal{P}(\mathbf{Y} \rightarrow \mathbf{T})$) to form a unified mutual-assistance framework (**MA-1**) so that the learned representations can be tailored for the prognosis model in a data-driven manner, vice versa. To make full use of available multi-modality information, the mutual-assistance geometry and similarity constraints (**MA-2**) are devised to assist mono-modality latent representation learning so as to preserve the characteristics of multi-modality data, i.e., $\mathcal{R}\big(\mathbf{X}\xrightarrow[\textbf{MA-2}]{\mathbf{Z}}\mathbf{Y}\big)$. Additionally, a mutual-assistance regression-ranking survival model (**MA-3**) is proposed to simultaneously focus on the survival time and event order, i.e., $\mathcal{P}\big(\mathbf{Y}\xrightarrow[\textbf{MA-3}]{}\mathbf{T}\big)$. More formally, the general unified framework can be formulated as follows:

$$\underbrace{\mathcal{R}\Big(\mathbf{X}\xrightarrow[\textbf{MA-2}]{\mathbf{Z}}\mathbf{Y};\Psi_1) + \mathcal{P}(\mathbf{Y}\xrightarrow[\textbf{MA-3}]{}\mathbf{T};\Psi_2\Big)}_{\textbf{MA-1}}, \tag{1}$$

where Ψ_1 and Ψ_2 are the balancing parameters. Intuitively, the first term establishes the connection between the original feature space and the latent space, while the second term bridges the latent space and the target space.

Remark 1. *The* **MA-1** *presents a unified mutual-assistance framework of representation learning and survival modeling, through which the survival model's dependency on the original clinical attribute distribution is reduced since the learnable latent representations are tailored for the model during the joint optimization.*

3.2 MA-2: Multi-Modality Reinforced Mono-Modality Representation Learning With Mutual-Assistance Constraints

To obtain clean and compact representations for mono-modality data, we project the original feature space into a low-dimensional latent subspace that should be approximately of low-rank, which can be formulated as follows:

$$\min_{\mathbf{Y},\mathbf{W}} \frac{\alpha}{2}\|\mathbf{Y} - \mathbf{X}\mathbf{W}\|_F^2 + \epsilon * \mathrm{rank}(\mathbf{Y}), \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{m_x\times m_y}$ is the projection matrix, and α and ϵ are the nonnegative weighted coefficients. The squared Frobenius norm in the first term aims to minimize the fitting error and alleviate the influence of random noises, while the second term regularizes the latent representations $\mathbf{Y}$ to be low-rank and informative. In view of the discrete characteristic of the function rank($\cdot$), we relax it using the nuclear norm, resulting in the following formula:

$$\min_{\mathbf{Y},\mathbf{W}} \frac{\alpha}{2}\|\mathbf{Y} - \mathbf{X}\mathbf{W}\|_F^2 + \epsilon\|\mathbf{Y}\|_*, \tag{3}$$

where the nuclear norm $\|\mathbf{Y}\|_*$ equals to the sum of all singular values of the matrix $\mathbf{Y}$. Apparently, it is hard to directly utilize the comprehensive information of multi-modality data for mono-modality representation reinforcement. Therefore, we first propose mutual-assistance similarity and geometry constraints to assist mono-modality latent representation learning for the preservation of the holistically structural and locally geometric characteristics of multi-modality data.

3.2.1 Global Similarity Preservation

To ensure the structural invariance between the latent space and the original feature space, the similarity regularization term is often applied to latent representations as follows:

$$\min_{\mathbf{Y}} \sum_{j,k}^{n} \mathbf{S}_{\mathbf{X}}^{(j,k)}\|\mathbf{Y}^{(j,:)} - \mathbf{Y}^{(k,:)}\|_2^2, \tag{4}$$

where $\mathbf{S}_{\mathbf{X}}^{(j,k)} = \exp\Big(-\frac{\|\mathbf{X}^{(j,:)}-\mathbf{X}^{(k,:)}\|_2^2}{2\sigma_x^2}\Big)$ is the element (of the similarity matrix $\mathbf{S}_{\mathbf{X}}$) that measures the similarity between the jth and kth samples in the original space and σ_x is the kernel width. In this scenario, the similarity regularization term is inherently based on a single modality so that it does not benefit from the information available in other modalities. A feasible way of utilizing the multi-modality information is to encourage the similarity between the mono-modality latent space and the multi-modality feature space. To this end, the similarity matrix is redefined as $\mathbf{S}_{\mathbf{Z}}$, where each element is computed by $\mathbf{S}_{\mathbf{Z}}^{(j,k)} = \exp\Big(-\frac{\|\mathbf{Z}^{(j,:)}-\mathbf{Z}^{(k,:)}\|_2^2}{2\sigma_z^2}\Big)$. Intuitively, if the jth and kth samples in the multi-modality feature space (i.e., $\mathbf{Z}^{(j,:)}$ and $\mathbf{Z}^{(k,:)}$) are close, their corresponding mono-modality latent representations $\mathbf{Y}^{(j,:)}$ and $\mathbf{Y}^{(k,:)}$ are also expected to be close to each other, vice versa. In this way, the holistically similarity characteristics inherent in multi-modality data can be potentially preserved. For convenience, we utilize $R_S(\mathbf{Y})$ to refer to the refined similarity constraint that can be simplified with the algebraic computation as follows:

$$R_S(\mathbf{Y}) = \min_{\mathbf{Y}} tr(\mathbf{Y}^T\mathbf{M}\mathbf{Y}) - tr(\mathbf{Y}^T\mathbf{S}_{\mathbf{Z}}\mathbf{Y}) = \min_{\mathbf{Y}} tr(\mathbf{Y}^T\mathbf{L}\mathbf{Y}), \tag{5}$$

where $\mathbf{L} = \mathbf{M} - \mathbf{S}_{\mathbf{Z}}$ and $\mathbf{M}$ is a diagonal matrix whose entries are the sum of corresponding rows of $\mathbf{S}_{\mathbf{Z}}$.

3.2.2 Local Geometry Preservation

The above-mentioned similarity is global-based and mainly focuses on the relation between each sample and all other samples. But it ignores the local characteristics which have been demonstrated to be more stable and robust to local changes and thus perform better than holistic ones [53], [54]. Additionally, it builds on the connection of the multi-modality original space and the mono-modality latent space, but it may still not take full consideration of the underlying shared representations in multi-modality data. To deeply enhance the guidance of multi-modal data to mono-modal latent representation learning, we devise the local geometry constraint (including distance and angle) to cooperate with the global similarity term. Specifically, similar to [48], the shared representations (denoted as $\mathbf{U} \in \mathbb{R}^{n\times m_u}$ and m_u is the corresponding dimension) of multi-modality data are first learned by minimizing the following projection error:

$$\min_{\mathbf{U},\mathbf{P}} \frac{\eta}{2}\|\mathbf{U}-\mathbf{Z}\mathbf{P}\|_F^2, \tag{6}$$

where $\mathbf{P} \in \mathbb{R}^{m_z\times m_u}$ is the projection matrix and η is the non-negative weighted coefficient. Before building up a connection between $\mathbf{Y}$ and $\mathbf{U}$, we make the following definition.

Definition 1 (Reference Point Definition). The sample collection is partitioned into C groups according to the samples' survival time in ascending order. For each group, its reference point is constructed by the linear combination of all samples within this group. Taking matrix $\mathbf{X}$ for example, its reference points can be represented by

$$\min_{\mathbf{R}_\mathbf{X},\{\mathbf{G}_c\}_{c=1}^C} \sum_{c=1}^{C}\|\mathbf{R}_\mathbf{X}^{(c,:)}-\mathbf{G}_c\mathbf{X}_c\|_F^2 \quad s.t. \sum_{i=1}^{n_c}\mathbf{G}_c^{(i)}=1, \tag{7}$$

where $\mathbf{R}_\mathbf{X} \in \mathbb{R}^{C\times m_x}$ is the feature matrix of the reference points, $\mathbf{G}_c \in \mathbb{R}^{1\times n_c}$ is the linear weighted vector (n_c denotes the number of samples within the cth group), and $\mathbf{X}_c \in \mathbb{R}^{n_c\times m_x}$ represents the sample collection of the cth group.

The reference point acts as an intermediate attribute for local characteristic alignment between $\mathbf{Y}$ and $\mathbf{U}$, since both of them have heterogeneous spaces with diverse feature dimensions. Note that the grouping operator is commonly used in clinical studies to cluster the patients into several risk subgroups for identifying group-specific biomarkers. In this paper, we consider the special case of Definition 1 by assigning $\mathbf{G}_c = \frac{1}{n_c}\mathbf{I}_c$, where all entries in $\mathbf{I}_c \in \mathbb{R}^{1\times n_c}$ equal to 1. The group-specific local manifold is described by the geometric relations (including distance and angle relations) between the sample and the corresponding reference point. It is expected that the learned mono-modality latent representations can preserve the group-specific local manifold of multi-modality data.

1) Distance Preservation: With the reference points, we define a pair of the weighted-distance matrices $\mathbf{F}_\mathbf{Y}$ and $\mathbf{F}_\mathbf{U}$ for $\mathbf{Y}$ and $\mathbf{U}$. For the jth sample, its weighted-distance vectors with respect to the representations $\mathbf{Y}^{(j,:)}$ and $\mathbf{U}^{(j,:)}$ are calculated by

$$\begin{cases} \mathbf{F}_\mathbf{Y}^{(j,:)} &= [\mathbf{H}_\mathbf{X}^{(j,1)}\mathbf{D}_\mathbf{Y}^{(j,1)},\ldots,\mathbf{H}_\mathbf{X}^{(j,c)}\mathbf{D}_\mathbf{Y}^{(j,c)},\ldots,\mathbf{H}_\mathbf{X}^{(j,C)}\mathbf{D}_\mathbf{Y}^{(j,C)}] \\ \mathbf{F}_\mathbf{U}^{(j,:)} &= [\mathbf{H}_\mathbf{Z}^{(j,1)}\mathbf{D}_\mathbf{U}^{(j,1)},\ldots,\mathbf{H}_\mathbf{Z}^{(j,c)}\mathbf{D}_\mathbf{U}^{(j,c)},\ldots,\mathbf{H}_\mathbf{Z}^{(j,C)}\mathbf{D}_\mathbf{U}^{(j,C)}] \end{cases}, \tag{8}$$

where $\mathbf{H}_\star^{(j,c)}$ ($\star$ denotes $\mathbf{X}$ or $\mathbf{Z}$) is a weight coefficient determined by the Gaussian distance between the jth sample and the cth reference point in the original feature space to maintain original manifold information, while $\mathbf{D}_{\star\star}^{(j,c)}$ ($\star\star$ denotes $\mathbf{Y}$ or $\mathbf{U}$) is defined as the Euclidean distance between the jth sample and the cth reference point within the latent feature subspace. The mathematical forms of $\mathbf{H}_\star^{(j,c)}$ and $\mathbf{D}_{\star\star}^{(j,c)}$ are presented as follows:

$$\begin{cases} \mathbf{H}_\mathbf{X}^{(j,c)} &= \exp\Big(-\frac{\|\mathbf{X}^{(j,:)}-\mathbf{R}_\mathbf{X}^{(c,:)}\|_2^2}{2\sigma_x^2}\Big) \\ \mathbf{H}_\mathbf{Z}^{(j,c)} &= \exp\Big(-\frac{\|\mathbf{Z}^{(j,:)}-\mathbf{R}_\mathbf{Z}^{(c,:)}\|_2^2}{2\sigma_z^2}\Big) \end{cases}, \tag{9}$$

$$\begin{cases} \mathbf{D}_\mathbf{Y}^{(j,c)} &= \|\mathbf{Y}^{(j,:)}-\mathbf{R}_\mathbf{Y}^{(c,:)}\|_2^2 \\ \mathbf{D}_\mathbf{U}^{(j,c)} &= \|\mathbf{U}^{(j,:)}-\mathbf{R}_\mathbf{U}^{(c,:)}\|_2^2 \end{cases}, \tag{10}$$

where $\mathbf{R}_\star^{(c,:)}$ and $\mathbf{R}_{\star\star}^{(c,:)}$ are the feature vectors of reference points of the cth group in the original space and the latent space, respectively. Furthermore, we apply Eq. (8) to all samples to form the matrix pair of $\mathbf{F}_\mathbf{Y}$ and $\mathbf{F}_\mathbf{U}$ and minimize the following formula for the local manifold preservation:

$$R_D(\mathbf{Y},\mathbf{U}) = \min_{\mathbf{Y},\mathbf{U}} \|\mathbf{F}_\mathbf{Y}-\mathbf{F}_\mathbf{U}\|_F^2. \tag{11}$$

2) Angle Preservation: In addition to the distance preservation, we also concern about the angle invariance. Based on the reference points, we construct paired auxiliary matrices $\mathbf{A}_\mathbf{Y}$ and $\mathbf{A}_\mathbf{U}$, whose entry connects the sample and its reference points in mono-modality latent space and multi-modality shared space, respectively. For the jth sample, its auxiliary vectors as regards the representations $\mathbf{Y}^{(j,:)}$ and $\mathbf{U}^{(j,:)}$ are defined as

$$\begin{cases} \mathbf{A}_\mathbf{Y}^{(j,:)} &= [\Big\langle\mathbf{Y}^{(j,:)},\mathbf{R}_\mathbf{Y}^{(1,:)}\Big\rangle,\ldots,\Big\langle\mathbf{Y}^{(j,:)},\mathbf{R}_\mathbf{Y}^{(C,:)}\Big\rangle] \\ \mathbf{A}_\mathbf{U}^{(j,:)} &= [\Big\langle\mathbf{U}^{(j,:)},\mathbf{R}_\mathbf{U}^{(1,:)}\Big\rangle,\ldots,\Big\langle\mathbf{U}^{(j,:)},\mathbf{R}_\mathbf{U}^{(C,:)}\Big\rangle] \end{cases}, \tag{12}$$

where $\Big\langle\mathbf{Y}^{(j,:)},\mathbf{R}_\mathbf{Y}^{(c,:)}\Big\rangle$ (or $\Big\langle\mathbf{U}^{(j,:)},\mathbf{R}_\mathbf{U}^{(c,:)}\Big\rangle$) represents the inner product of $\mathbf{Y}^{(j,:)}$ and $\mathbf{R}_\mathbf{Y}^{(c,:)}$ (or $\mathbf{U}^{(j,:)}$ and $\mathbf{R}_\mathbf{U}^{(c,:)}$). Let $\Theta_\mathbf{Y}^{(j,:)}$ and $\Theta_\mathbf{U}^{(j,:)}$ denote the angles between the jth sample and all reference points in the mono-modality latent space and the multi-modality shared space, respectively. It can be proved (in *Supplementary Materials*, available online) that maximizing the following optimization function (when $\mathbf{Y}$ and $\mathbf{U}$ are fixed) is equivalent to enforcing $\Theta_\mathbf{Y}^{(j,:)} = \Theta_\mathbf{U}^{(j,:)}$,

$$\max_{\Phi} \sum_{j=1}^{n} \mathbf{A}_\mathbf{Y}^{(j,:)}\cdot\mathbf{A}_\mathbf{U}^{(j,:)} \quad s.t. \mathbf{U}=\mathbf{Y}\Phi, \tag{13}$$

where $\Phi \in \mathbb{R}^{m_y\times m_u}$ is the angle-preserving projection that links $\mathbf{Y}$ and $\mathbf{U}$. For example, we take a special case of orthogonal projection (i.e., $\Phi\Phi^T = \mathbf{I}$, $\mathbf{I} \in \mathbb{R}^{m_y\times m_y}$ is an identity matrix) into consideration in that we can get $\Big\langle\mathbf{U}^{(j,:)},\mathbf{R}_\mathbf{U}^{(1,:)}\Big\rangle = \mathbf{U}^{(j,:)}\mathbf{R}_\mathbf{U}^{(1,:)T} = \mathbf{Y}^{(j,:)}\Phi\Phi^T\mathbf{R}_\mathbf{Y}^{(1,:)T} = \Big\langle\mathbf{Y}^{(j,:)},\mathbf{R}_\mathbf{Y}^{(1,:)}\Big\rangle$. The term $\sum_{j=1}^{n}\mathbf{A}_\mathbf{Y}^{(j,:)}\cdot\mathbf{A}_\mathbf{U}^{(j,:)}$ has the expanded form as follows:

$$\sum_{j=1}^{n}\sum_{c=1}^{C}\mathbf{Y}^{(j,:)}\mathbf{R}_{\mathbf{Y}}^{(c,:)T}\mathbf{U}^{(j,:)}\mathbf{R}_{\mathbf{U}}^{(c,:)T}. \tag{14}$$

With some algebraic operation, the angle constraint (denoted as $R_A(\mathbf{Y},\mathbf{U},\Phi)$) can be written as

$$R_A(\mathbf{Y},\mathbf{U},\Phi)=\max_{\mathbf{Y},\mathbf{U},\Phi} tr(\mathbf{R}_{\mathbf{U}}^T\mathbf{R}_{\mathbf{Y}}\mathbf{Y}^T\mathbf{U}) \quad s.t.\mathbf{U}=\mathbf{Y}\Phi. \tag{15}$$

Remark 2. *Within* **MA-2***, the similarity preservation cooperates with the geometry (both distance and angle) preservation to make full use of multi-modality data to reinforce mono-modality representation learning. The former focuses on the influence of multi-modality data in the original space and captures the inherent relations (i.e., similarity) among samples from a global view, while the latter considers the multi-modality representations in the latent space and attempts to maintain local relations (via reference points) within latent representations.*

3.3 MA-3: Mutual-Assistance Regression-Ranking Survival Modeling

In general, the survival data contain uncensored and censored samples. The recorded survival time of each censored sample provides the lower bound of real survival (event) time, which is considered to be potentially useful for survival modeling. The proposed mutual-assistance regression-ranking survival model mainly focuses on 1) an efficient strategy to deal with the censored issue and 2) joint estimation of survival time and event order.

3.3.1 Regression Formulation

Different from the conventional regression task, survival regression is hindered by the right-censored samples that are labeled with the truncated time information. We first model the relationship between the true survival time $\mathcal{T}\in\mathbb{R}^{n\times 1}$ and the recorded survival time $\mathbf{T}\in\mathbb{R}^{n\times 1}$ as follows:

$$\mathcal{T}=\mathbf{T}-(\boldsymbol{\delta}-\mathbf{1})\odot\mathbf{E} \quad s.t.\mathbf{E}^{(i)}\geq 0, \tag{16}$$

where $\odot$ denotes the Hadamard product, $\mathbf{1}\in\mathbb{R}^{n\times 1}$ is the all-ones vector, $\mathbf{E}\in\mathbb{R}^{n\times 1}$ is a bias vector with nonnegative elements, and $\boldsymbol{\delta}\in\mathbb{R}^{n\times 1}$ is an indicator vector (viz, its element equals to 1 and 0 for the uncensored and censored sample, respectively). Intuitively, the bias vector $\mathbf{E}$ only compensates for the observed time of right-censored samples. That is, for the uncensored sample collection $\mathcal{N}_u$ (i.e., $\boldsymbol{\delta}_u=\mathbf{1}$), their observed time $\mathbf{T}_u$ is equal to $\mathcal{T}_u$, while for the censored sample collection $\mathcal{N}_c$ (i.e., $\boldsymbol{\delta}_c=\mathbf{0}$), their true survival time is longer than the observed time, i.e., $\mathcal{T}_c=\mathbf{T}_c+\mathbf{E}_c$. With the mono-modality latent representations $\mathbf{Y}$, we predict true survival time by a linear regression model $\hat{\mathcal{T}}=\mathbf{YK}$, where $\mathbf{K}\in\mathbb{R}^{m_y\times 1}$ is the coefficient vector. However, since $\mathbf{E}$ in Eq. (16) is unknown, we jointly estimate $\mathbf{E}$ and $\hat{\mathcal{T}}$ by optimizing the following formula:

$$\begin{aligned}R_R(\mathbf{Y},\mathbf{K},\mathbf{E})&=\min_{\mathbf{Y},\mathbf{E},\hat{\mathcal{T}}}\|\mathcal{T}-\hat{\mathcal{T}}\|_F^2 \quad s.t.\mathbf{E}^{(i)}\geq 0\\&=\min_{\mathbf{Y},\mathbf{K},\mathbf{E}}\|(\mathbf{T}-(\boldsymbol{\delta}-\mathbf{1})\odot\mathbf{E})-\mathbf{YK}\|_F^2 \quad s.t.\mathbf{E}^{(i)}\geq 0.\end{aligned} \tag{17}$$

For convenience, we denote the term $\mathbf{YK}+(\boldsymbol{\delta}-\mathbf{1})\odot\mathbf{E}$ as $\hat{\mathbf{T}}$, namely the predicted recorded survival time.

3.3.2 Ranking Calibration

Actually, the regression model often makes the mistake of comparison as it may not distinguish two close cases. Taking a pair of uncensored samples for example, we denote their survival time as $\mathcal{T}^{(i)}$ and $\mathcal{T}^{(j)}$, respectively, and assume that $\mathcal{T}^{(j)}-\mathcal{T}^{(i)}=\xi>0$. Given the regression formula in Eq. (17), we consider the error gaps between the true survival time and the predicted ones for ith and jth samples, i.e., $\|\mathcal{T}^{(i)}-\hat{\mathcal{T}}^{(i)}\|_F^2=\zeta_1^2$ and $\|\mathcal{T}^{(j)}-\hat{\mathcal{T}}^{(j)}\|_F^2=\zeta_2^2$ (where $\zeta_1,\zeta_2>0$). Under this circumstance, the predicted survival time of ith and jth samples can be $\hat{\mathcal{T}}^{(i)}=\mathcal{T}^{(i)}-\zeta_1$ or $\mathcal{T}^{(i)}+\zeta_1$, and $\hat{\mathcal{T}}^{(j)}=\mathcal{T}^{(j)}-\zeta_2$ or $\mathcal{T}^{(j)}+\zeta_2$. Accordingly, the error gap between the predicted time for these two samples, i.e., $\hat{\xi}=\hat{\mathcal{T}}^{(j)}-\hat{\mathcal{T}}^{(i)}$, equals to $\xi-\zeta_2-\zeta_1$, $\xi-\zeta_2+\zeta_1$, $\xi+\zeta_2-\zeta_1$, and $\xi+\zeta_2+\zeta_1$. Apparently, the regression model cannot always guarantee that $\hat{\xi}>0$ (namely $\hat{\mathcal{T}}^{(j)}>\hat{\mathcal{T}}^{(i)}$) holds on. However, the pair comparison is significantly crucial in the prognosis evaluation and sample stratification. Therefore, it is desirable to utilize the relative position or ranking of the observed time of samples to calibrate the learned representations and regression model.

From the Bayesian perspective [36], it can utilize the maximization of the $\mathbb{P}(\Omega|\Lambda)$ to ensure the consistency purpose of pairwise ranking for all cases as follows:

$$\mathbb{P}(\Omega|\Lambda)=\frac{\mathbb{P}(\Omega)\mathbb{P}(\Lambda|\Omega)}{\mathbb{P}(\Lambda)}\propto\mathbb{P}(\Omega)\mathbb{P}(\Lambda|\Omega), \tag{18}$$

where Λ is the representations of samples as the input of the model, while Ω represents the parameter collection of the model. When each pair of data are independent of each other and the general prior density $\mathbb{P}(\Omega)$ is fixed, the main issue focuses on the pair comparison definition, i.e., $\mathbb{P}(\Lambda|\Omega)$. In the prevalent concordance index (C-index) [55], it defines a measure of goodness-of-fit for a pair of outcomes. However, the concordance index is nondifferentiable, which can not be embedded into the proposed method. Therefore, we expect to formulate a differentiable ranking constraint term on the basis of the regression model. Towards the uncensored sample collection, we have the following definition:

$$\mathbb{P}(\Lambda|\Omega):=\sum_{\{i,j\}\in\mathcal{N}_u(\mathcal{T}^{(j)}>\mathcal{T}^{(i)})}(\hat{\mathcal{T}}^{(j)}-\hat{\mathcal{T}}^{(i)}), \tag{19}$$

where $\{i,j\}\in\mathcal{N}_u(\mathcal{T}^{(j)}>\mathcal{T}^{(i)})$ denotes the sample pair $\{i,j\}$, satisfying $\mathcal{T}^{(j)}>\mathcal{T}^{(i)}$, from the uncensored sample collection.

Note that the ranking calibration should take full account of censored data so as to explore the information as much as possible. Even though the true survival time of censored samples is truncated, we can at least confirm and utilize the information between censored samples (denoted as $\mathbf{T}^{(j)}$) and those uncensored ones (denoted as $\mathbf{T}^{(i)}$) that have shorter observed survival time, namely, $\mathbf{T}^{(i)}<\mathbf{T}^{(j)}$. For uncensored samples, we have $\mathcal{T}_u=\mathbf{T}_u$ and $\hat{\mathcal{T}}_u=\hat{\mathbf{T}}_u$. Thus we can further extend Eq. (19) for dealing with the censored samples and design its equivalent optimization formulation as follows:

$$\min_{\mathbf{Y},\mathbf{E},\mathbf{K}}\sum_{\mathcal{L}(i,j)=1}(\hat{\mathbf{T}}^{(i)}-\hat{\mathbf{T}}^{(j)}), \tag{20}$$

where $\mathcal{L}(i,j)$ is defined as follows:

$$\mathcal{L}(i,j)=\begin{cases}1, & \text{if } \boldsymbol{\delta}^{(i)}=1 \text{ and } \mathbf{T}^{(i)} < \mathbf{T}^{(j)}\\ 0, & \text{otherwise}\end{cases}. \tag{21}$$

Intuitively, Eq. (20) can penalize incorrect forecast pairs even if the predicted time is closed to the observed time but the relative order is wrong. In order to use matrix operator for computation acceleration, we first give an example to illustrate how the objective loss of ranking calibration is constructed. Support that there is a collection containing five samples with the ascending sorted observation time $\mathbf{T}=[\mathbf{T}^{(1)};\mathbf{T}^{(2)};\mathbf{T}^{(3)};\mathbf{T}^{(4)};\mathbf{T}^{(5)}]$ and the censored status $\boldsymbol{\delta}=[1;0;1;0;1]$. Accordingly, we denote the estimated results of the regression model on this collection as $\hat{\mathbf{T}}=[\hat{\mathbf{T}}^{(1)};\hat{\mathbf{T}}^{(2)};\hat{\mathbf{T}}^{(3)};\hat{\mathbf{T}}^{(4)};\hat{\mathbf{T}}^{(5)}]$. Based on the regulation of Eq. (20), the objective loss Δ contains three segments which equal to the number of uncensored samples. For the first segment Δ_1, we need to take the pairs of $\{1,2\}$, $\{1,3\}$, $\{1,4\}$, and $\{1,5\}$ into account and we can get $\Delta_1=\sum_{j>1}^{5}(\hat{\mathbf{T}}^{(1)}-\hat{\mathbf{T}}^{(j)})$. Likewise, we also have $\Delta_2=\sum_{j>3}^{5}(\hat{\mathbf{T}}^{(3)}-\hat{\mathbf{T}}^{(j)})$ and $\Delta_3=0$. Finally, the objective loss can be computed by $\Delta=\Delta_1+\Delta_2+\Delta_3=(4-0)\hat{\mathbf{T}}^{(1)}+(0-1)\hat{\mathbf{T}}^{(2)}\quad+(2-1)\hat{\mathbf{T}}^{(3)}+(0-2)\hat{\mathbf{T}}^{(4)}+(0-2)\hat{\mathbf{T}}^{(5)}$. Interestingly, the above-mentioned process can be further summarized into a concise version as follows:

$$R_K(\mathbf{Y},\mathbf{K},\mathbf{E})=\min_{\mathbf{Y},\mathbf{K},\mathbf{E}}\mathbf{O}(\mathbf{YK}+(\boldsymbol{\delta}-\mathbf{1})\odot\mathbf{E}), \tag{22}$$

where $\mathbf{O}\in\mathbb{R}^{1\times n}$ is an auxiliary vector whose entry is defined as $\mathbf{O}^{(i)}=n_b^i-n_a^i$. If $\boldsymbol{\delta}^{(i)}=1$, the n_b^i is the number of samples whose observation time is longer than the ith sample, while $n_b^i=0$ for $\boldsymbol{\delta}^{(i)}=0$. And n_a^i counts the number of uncensored samples whose survival time is shorter than the ith sample regardless of the status of $\boldsymbol{\delta}^{(i)}$.

Remark 3. *Since the regression function can estimate specific length of survival time and the ranking model can penalize the wrong event order, we combine them together and expect positive assistance between them so as to deal with censored samples for efficiently predicting the relative survival risk.*

3.4 The Objective Function

Expecting the reciprocal interaction of representation learning and survival modeling will further improve the model's performance, we integrate these two parts into a mutual-assistance unified framework (i.e., **MA-1**) as follows:

$$\begin{aligned}\min_{\mathcal{Q}}\ &\frac{\eta}{2}\|\mathbf{U}-\mathbf{ZP}\|_F^2+\frac{\alpha}{2}\|\mathbf{Y}-\mathbf{XW}\|_F^2+\epsilon\|\mathbf{Y}\|_*\\ &+\rho R_S(\mathbf{Y})+\gamma R_D(\mathbf{Y},\mathbf{U})-\theta R_A(\mathbf{Y},\mathbf{U},\Phi)\\ &+\frac{\beta}{2}R_R(\mathbf{Y},\mathbf{K},\mathbf{E})+\kappa R_K(\mathbf{Y},\mathbf{K},\mathbf{E}),\end{aligned} \tag{23}$$

where $\mathcal{Q}=\{\mathbf{Y},\mathbf{U},\mathbf{W},\mathbf{P},\Phi,\mathbf{K},\mathbf{E}\}$ is the variable set and $\eta,\alpha,\kappa,\gamma,\theta,\rho$, and ϵ are all nonnegative tuning parameters. Accordingly, the first six terms and the last two terms in Eq. (23) constitute $\mathcal{R}(\mathbf{X}\xrightarrow[\textbf{MA-2}]{\mathbf{Z}}\mathbf{Y};\Psi_1)$ and $\mathcal{P}(\mathbf{Y}\xrightarrow[\textbf{MA-3}]{}\mathbf{T};\Psi_2)$ in Eq. (1), respectively.

4 Optimization Algorithm

To solve the above target function, we first introduce an auxiliary variable $\mathbf{J}$ to replace matrix $\mathbf{Y}$ in the nuclear norm so that it can be easy to separate and optimize the term, and then utilize Augmented Lagrange Multiplier (ALM) to fuse all constraints into the main target function. Note that the nonnegative constraint for $\mathbf{E}$ is additionally considered when $\mathbf{E}$ is updated (details can be found in *5) Updating* $\mathbf{E}$). Thus, Eq. (23) can be transformed into the following form:

$$\begin{aligned}\min_{\mathcal{Q}}\ &\frac{\eta}{2}\|\mathbf{U}-\mathbf{ZP}\|_F^2+\frac{\alpha}{2}\|\mathbf{Y}-\mathbf{XW}\|_F^2+\epsilon\|\mathbf{J}\|_*\\ &+\rho tr(\mathbf{Y}^T\mathbf{LY})+\gamma\|\mathbf{F_Y}-\mathbf{F_U}\|_F^2-\theta tr(\mathbf{R_U}^T\mathbf{R_Y}\mathbf{Y}^T\mathbf{U})\\ &+\frac{\beta}{2}\|\mathbf{T}-\mathbf{YK}-(\boldsymbol{\delta}-\mathbf{1})\odot\mathbf{E}\|_F^2\\ &+\kappa\mathbf{O}(\mathbf{YK}+(\boldsymbol{\delta}-\mathbf{1})\odot\mathbf{E})\\ &+\psi(\mathbf{V}_1,\mathbf{U}-\mathbf{Y}\Phi)+\psi(\mathbf{V}_2,\mathbf{Y}-\mathbf{J}),\end{aligned} \tag{24}$$

where $\psi(\mathbf{V},\mathbf{N})=(\sigma/2)\|\mathbf{N}\|_F^2+tr(\mathbf{V}^T\mathbf{N})$, and $\mathbf{V}$ denotes a Lagrangian multiplier having the same size as matrix $\mathbf{N}$, and σ is a positive scalar, and the variable collection is further expanded as $\mathcal{Q}=\{\mathbf{Y},\mathbf{U},\mathbf{W},\mathbf{P},\Phi,\mathbf{K},\mathbf{J},\mathbf{E},\mathbf{V}_1,\mathbf{V}_2\}$. We next apply inexact Alternating Direction Method of Multipliers (ADMM) [56] to solve the target function effectively by updating the variables in $\mathcal{Q}$. Fixing other variables, we update each target variable iteratively, which leads to several sub-problems below.

1) Updating $\mathbf{Y}$: The optimization problem with respect to the latent representations of mono-modality data can be written as

$$\begin{aligned}\min_{\mathbf{Y}}\ &\frac{\alpha}{2}\|\mathbf{Y}-\mathbf{XW}\|_F^2+\rho tr(\mathbf{Y}^T\mathbf{LY})+\gamma\|\mathbf{F_Y}-\mathbf{F_U}\|_F^2\\ &-\theta tr(\mathbf{R_U}^T\mathbf{R_Y}\mathbf{Y}^T\mathbf{U})+\frac{\beta}{2}\|\mathbf{T}-\mathbf{YK}-(\boldsymbol{\delta}-\mathbf{1})\odot\mathbf{E}\|_F^2\\ &+\kappa\mathbf{O}(\mathbf{YK}+(\boldsymbol{\delta}-\mathbf{1})\odot\mathbf{E})\\ &+\psi(\mathbf{V}_1,\mathbf{U}-\mathbf{Y}\Phi)+\psi(\mathbf{V}_2,\mathbf{Y}-\mathbf{J}).\end{aligned} \tag{25}$$

For convenience, we denote Eq. (25) as $G(\mathbf{Y})$, which can be optimized by the Gradient Descent strategy as follows:

$$\mathbf{Y}^{(k+1)}=\mathbf{Y}^{(k)}-\frac{\partial G(\mathbf{Y})}{\partial\mathbf{Y}}, \tag{26}$$

and the first derivative with respect to $\mathbf{Y}$ can be formulated as

$$\begin{aligned}\frac{\partial G(\mathbf{Y})}{\partial\mathbf{Y}}=\ &\alpha(\mathbf{Y}-\mathbf{XW})+2\rho\mathbf{LY}\\ &+4\gamma\sum_{m=1}^{C}[(\mathbf{F_Y}^{(:,m)}-\mathbf{F_U}^{(:,m)})\mathbf{H_X}^{(:,m)^T}\odot\mathbf{I}](\mathbf{Y}-\mathbf{R_Y^m})\\ &-\theta\mathbf{U}\mathbf{R_U}^T\mathbf{R_Y}-\beta(\mathcal{T}-\mathbf{YK})\mathbf{K}^T+\kappa\mathbf{O}^T\mathbf{K}^T-\mathbf{V}_1\Phi^T\\ &+\sigma_1(\mathbf{Y}\Phi-\mathbf{U})\Phi^T+\mathbf{V}_2+\sigma_2(\mathbf{Y}-\mathbf{J}),\end{aligned} \tag{27}$$

where $\mathcal{T}=\mathbf{T}-(\boldsymbol{\delta}-\mathbf{1})\odot\mathbf{E}$, $\mathbf{I}\in\mathbb{R}^{n\times n}$ is the identity matrix, and $\mathbf{R_Y^m}=[\mathbf{R_Y}^{(m,:)};\mathbf{R_Y}^{(m,:)};\ldots;\mathbf{R_Y}^{(m,:)}]\in\mathbb{R}^{n\times q}$ contains n copies of the mth reference point in $\mathbf{R_Y}$.

2) Updating $\mathbf{U}$: To update $\mathbf{U}$, we need to solve the following objective function:

$$\begin{aligned}\min_{\mathbf{U}}\ &\frac{\eta}{2}\|\mathbf{U}-\mathbf{ZP}\|_F^2+\gamma\|\mathbf{F_Y}-\mathbf{F_U}\|_F^2-\theta tr(\mathbf{R_U}^T\mathbf{R_Y}\mathbf{Y}^T\mathbf{U})\\ &+\psi(\mathbf{V}_1,\mathbf{U}-\mathbf{Y}\Phi).\end{aligned} \tag{28}$$

We mark Eq. (28) as $G(\mathbf{U})$ and get the derivative with respect to $\mathbf{U}$ as follows:

$$\frac{\partial G(\mathbf{U})}{\partial \mathbf{U}} = \eta(\mathbf{U}-\mathbf{ZP}) - \theta\mathbf{Y}\mathbf{R}_{\mathbf{Y}}^{T}\mathbf{R}_{\mathbf{U}} + \mathbf{V}_1 + \sigma_1(\mathbf{U}-\mathbf{Y}\Phi) + 4\gamma\sum_{m=1}^{C}[(\mathbf{F}_{\mathbf{U}}^{(:,m)} - \mathbf{F}_{\mathbf{Y}}^{(:,m)})\mathbf{H}_{\mathbf{Z}}^{(:,m)^T} \odot \mathbf{I}](\mathbf{U}-\mathbf{R}_{\mathbf{U}}^{m}), \quad (29)$$

where $\mathbf{R}_{\mathbf{U}}^{m} \in \mathbb{R}^{n\times q}$ contains n copies of the mth reference point in $\mathbf{R}_{\mathbf{U}}$. Thus, we can update $\mathbf{U}$ by

$$\mathbf{U}^{(k+1)} = \mathbf{U}^{(k)} - \frac{\partial G(\mathbf{U})}{\partial \mathbf{U}}. \quad (30)$$

3) Updating $\mathbf{W}$: Fixing other variables, we can obtain the following sub-problem and closed-form solution with respect to $\mathbf{W}$:

$$\mathbf{W}^{*} = \arg\min_{\mathbf{W}} \frac{\alpha}{2}\|\mathbf{Y}-\mathbf{XW}\|_F^2 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \quad (31)$$

4) Updating $\mathbf{P}$: Fixing other variables, the sub-problem and closed-form solution of $\mathbf{P}$ are as follow:

$$\mathbf{P}^{*} = \arg\min_{\mathbf{P}} \frac{\eta}{2}\|\mathbf{U}-\mathbf{ZP}\|_F^2 = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{U}. \quad (32)$$

5) Updating Φ: We fix other variables and obtain the sub-problem and update rule of Φ as follow:

$$\begin{aligned}\Phi^{*} &= \arg\min_{\Phi} \psi(\mathbf{V}_1^T, \mathbf{U}-\mathbf{Y}\Phi)\\ &= (\mathbf{Y}^T\mathbf{Y})^{-1}\left(\frac{\mathbf{Y}^T\mathbf{V}_1}{\sigma_1} + \mathbf{Y}^T\mathbf{U}\right). \end{aligned} \quad (33)$$

6) Updating $\mathbf{K}$: Similarly, we can obtain the following sub-problem and closed-form solution with respect to $\mathbf{K}$:

$$\begin{aligned}\mathbf{K}^{*} &= \arg\min_{\mathbf{K}} \frac{\beta}{2}\|\mathcal{T}-\mathbf{YK}\|_F^2 + \kappa\mathbf{O}(\mathbf{YK} + (\boldsymbol{\delta}-1)\odot\mathbf{E})\\ &= (\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\left(\mathcal{T} - \frac{\kappa}{\beta}\mathbf{O}^T\right). \end{aligned} \quad (34)$$

7) Updating $\mathbf{J}$: Fixing other variables, we get the following target function with respect to $\mathbf{J}$:

$$\begin{aligned}&\arg\min_{\mathbf{J}} \epsilon\|\mathbf{J}\|_* + \psi(\mathbf{V}_2^T, \mathbf{Y}-\mathbf{J})\\ &= \arg\min_{\mathbf{J}} \epsilon\|\mathbf{J}\|_* + tr(\mathbf{V}_2^T(\mathbf{Y}-\mathbf{J})) + \frac{\sigma_2}{2}\|\mathbf{Y}-\mathbf{J}\|_F^2\\ &= \arg\min_{\mathbf{J}} \frac{\epsilon}{\sigma_2}\|\mathbf{J}\|_* + \frac{1}{2}\|\mathbf{J} - (\mathbf{Y} + \frac{1}{\sigma_2}\mathbf{V}_2)\|_F^2. \end{aligned} \quad (35)$$

According to the singular value thresholding (SVT) algorithm [57], the above function has the solution as follows:

$$\mathbf{J} = \ \ D_{\frac{\epsilon}{\sigma_2}}(\mathbf{Y} + \tfrac{1}{\sigma_2}\mathbf{V}_2) = \tilde{\mathbf{U}}\mathrm{diag}(\{\mathbf{\Sigma}^{(i,i)} - \tfrac{\epsilon}{\sigma_2}\}_+)\tilde{\mathbf{V}}^T, \quad (36)$$

where $\mathbf{Y} + \frac{1}{\sigma_2}\mathbf{V}_2 = \tilde{\mathbf{U}}\mathbf{\Sigma}\tilde{\mathbf{V}}^T$ is the singular value decomposition (SVD) results of $\mathbf{Y} + \frac{1}{\sigma_2}\mathbf{V}_2$ and $\{b\}_+ = \max(0, b)$.

8) Updating $\mathbf{E}$: With other variables fixed, we obtain target function with respect to $\mathbf{E}$ as follows:

$$\begin{aligned}&\min_{\mathbf{E}} \frac{\beta}{2}\|\mathbf{T} - \mathbf{YK} - (\boldsymbol{\delta}-1)\odot\mathbf{E}\|_F^2\\ &\quad + \kappa\mathbf{O}(\mathbf{YK} + (\boldsymbol{\delta}-1)\odot\mathbf{E}). \end{aligned} \quad (37)$$

We update $\mathbf{E} \in \mathbb{R}^{n\times 1}$ element by element by solving the following element-specific problem:

$$\begin{aligned}&\min_{\mathbf{E}^{(i)}} \frac{\beta}{2}\|\mathbf{T}^{(i)} - \mathbf{Y}^{(i,:)}\mathbf{K} - (\boldsymbol{\delta}^{(i)}-1)\odot\mathbf{E}^{(i)}\|_F^2\\ &\quad + \kappa\mathbf{O}^{(i)}(\mathbf{Y}^{(i,:)}\mathbf{K} + (\boldsymbol{\delta}^{(i)}-1)\odot\mathbf{E}^{(i)}). \end{aligned} \quad (38)$$

Note that the entry of $\mathbf{E}$ depends on the censoring status of the sample. Since the survival time of the censored samples is longer than their observed time, the corresponding entries of $\mathbf{E}$ provide positive compensations for time estimation of censored samples, while no compensation for uncensored samples. That is, only when $\boldsymbol{\delta}^{(i)} = 0$, we introduce $\mathbf{E}^{(i)} = \mathbf{B}^{(i)^2}$ to make sure the updated $\mathbf{E}^{(i)}$ for the censored sample is effective, and then Eq. (38) is rewritten as

$$\min_{\mathbf{B}^{(i)}} \frac{\beta}{2}\|\mathbf{T}^{(i)} - \mathbf{Y}^{(i,:)}\mathbf{K} + \mathbf{B}^{(i)^2}\|_F^2 + \kappa\mathbf{O}^{(i)}(\mathbf{Y}^{(i,:)}\mathbf{K} - \mathbf{B}^{(i)^2}). \quad (39)$$

Taking the derivative of Eq. (39) with respect to $\mathbf{B}^{(i)}$ and making it equal to zero as follows:

$$\begin{aligned}\frac{\partial G(\mathbf{B}^{(i)})}{\partial \mathbf{B}^{(i)}} &= 2\beta\mathbf{B}^{(i)}(\mathbf{T}^{(i)} - \mathbf{Y}^{(i,:)}\mathbf{K} + \mathbf{B}^{(i)^2}) - 2\kappa\mathbf{O}^{(i)}\mathbf{B}^{(i)}\\ &= 0, \end{aligned} \quad (40)$$

we can get the optimization of $\mathbf{E}$ as follows:

$$\mathbf{E}^{(i)^*} = \begin{cases} \boldsymbol{\tau}^{(i)} & \text{if } \boldsymbol{\delta}^{(i)} = 0 \text{ and } \boldsymbol{\tau}^{(i)} \geq 0\\ 0 & \text{others} \end{cases}, \quad (41)$$

where $\boldsymbol{\tau}^{(i)} = \frac{\kappa}{\beta}\mathbf{O}^{(i)} - \mathbf{T}^{(i)} + \mathbf{Y}^{(i,:)}\mathbf{K}$.

Algorithm 1. Optimization Algorithm for MaL

```
Training Stage
  Input: Mono-modality data matrix $\mathbf{X}$, multi-modality data matrix $\mathbf{Z}$, predefined parameters.
  Initialize each component in the variable collection $\mathcal{Q}$ with random values between [0,1]. Calculate $\mathbf{L}$, $\mathbf{R}_{\mathbf{X}}$, $\mathbf{R}_{\mathbf{Z}}$, $\mathbf{H}_{\mathbf{X}}$, $\mathbf{H}_{\mathbf{Z}}$, and $\mathbf{O}$.
1: repeat
2:     calculate $\mathbf{R}_{\mathbf{Y}}$, $\mathbf{R}_{\mathbf{U}}$, $\mathbf{D}_{\mathbf{Y}}$, $\mathbf{D}_{\mathbf{U}}$;
3:     update $\mathbf{Y}$ by using Eqs. (26) and (27);
4:     update $\mathbf{U}$ by using Eqs. (30) and (29);
5:     update $\mathbf{W}$, $\mathbf{P}$, $\Phi$, $\mathbf{K}$, $\mathbf{J}$, $\mathbf{E}$ by using Eqs. (31), (32), (33), (34), (36), (41);
6:     update $\mathbf{V}_1$, $\mathbf{V}_2$ by using Eqs. (42) and (43);
7:     update $\sigma_1$, $\sigma_2$ by $\sigma = \min(\mu\sigma, \sigma_{\max})$;
8: until converges
  Output: Projection matrix $\mathbf{W}$ and coefficient vector $\mathbf{K}$
Testing Stage
  Input: Mono-modality testing samples $\mathbf{X}_{\text{test}}$
  Output: Predictive value $\hat{\mathcal{T}}_{\text{test}} = \mathbf{X}_{\text{test}}\mathbf{W}\mathbf{K}$
```

9) Updating $\mathbf{V}_1$, $\mathbf{V}_2$: The following equations calculate the iterative process of multipliers:

$$\mathbf{V}_1^{(k+1)} = \sigma_1(\mathbf{U}^{(k+1)} - \mathbf{Y}^{(k+1)}\Phi^{(k+1)}) + \mathbf{V}_1^{(k)}, \quad (42)$$

$$\mathbf{V}_2^{(k+1)} = \sigma_2(\mathbf{Y}^{(k+1)} - \mathbf{J}^{(k+1)}) + \mathbf{V}_2^{(k)}. \quad (43)$$

Therefore, the objective function of MaL can be solved by conducting the above steps iteratively until it converges. Algorithm 1 gives supplementary details.

5 Experiments and Results

In this section, we first introduce the studied datasets as well as experimental settings, including competing methods, parameter settings and evaluation metrics. Then, we thoroughly compare our method with various survival models and show experimental results. Finally, we elaborate on the mutual-assistance properties of our proposed model. Note that we mainly focuses on two modalities, i.e., gene expression sequences and histopathological data. Additionally, the extension of the proposed method on more modalities data is also discussed in Section 6.1 Extension on More-Modality Setup.

5.1 Datasets and Experimental Settings

We carried out all experiments on ten typical cancer datasets (i.e., KIRC, LIHC, LUAD, SKCM, BLCA, OV, LUSC, LGG, THCA, and UCEC) derived from TCGA,[1] a publicly available data consortium. The basic characteristics of these datasets are briefly listed in Table 1. For each dataset, both gene expression data (RNA sequences) and histopathological data (whole slide H&E stained images) were collected and preprocessed by the same pipeline as that in [48] and [58]. For genomic data, we applied Principal Component Analysis (PCA) to the raw 19,841 protein coding mRNA sequences of all input samples and finally got a 128-D genomic feature matrix. The histopathological images were quantified by a 162-D Parameter Free Threshold Adjacency Statistics (PFTAS) image feature operator, which was then also reduced to 128-D image features by PCA. Finally, both genomic and histopathological feature matrices were normalized to the range of [0,1] via the min-max normalization. The core codes of MaL are available.[2]

We compared the proposed method with previous survival models. On the one hand, six classical statistical methods were enrolled: 1) three semi-parametric models, including Cox [23], Lasso-Cox [50], and EN-Cox [51]; 2) three parametric models, including Tobit model [24], Buckley-James (BJ) regression model [52], and Accelerated Failure Time (AFT) model [25]. Note that we chose the Exponential and Gaussian distributions for Tobit and AFT models, respectively. On the other hand, we compared our model with several recent machine learning algorithms for survival analysis: 1) SVM-based Ranking and Regression method for survival analysis (SVM-RR) [27]; 2) Multi-Task Learning formulation for Survival Analysis (MTLSA) [59]; 3) Random Survival Forests (RSF) [60]; 4) Multi-task Multi-modal for joint Diagnosis and Prognosis (M2DP) [61]; 5) Relation-aware Shared Representation learning (RaSR) [48]; 6) Multi-constraint Latent Representation learning (McLR) [58]. Considering that the last three models (M2DP, RaSR, and McLR) were originally designed for multi-modality data, we slightly adjusted them to adapt certain scenarios (e.g., mono-modality training) if necessary. We reported the best performance of all competing methods achieved by adopting source codes and parameter suggestions from their studies.

In this paper, we utilized 10-fold cross-validation strategy to evaluate the performance of all methods. To determine the values of eight nonnegative parameters (namely $\eta, \alpha, \beta, \kappa, \gamma, \theta, \rho,$ and ϵ) in our model, we randomly selected 10% of the training samples as a validation set and fine-tuned these parameters within the specified range (i.e., $\eta, \alpha, \beta, \kappa, \gamma, \theta \in$ {1E-05, 1E-04, 1E-03} and $\rho, \epsilon \in$ {0.01, 0.1, 1}). According to the best results achieved by the model on the validation set, we set $\eta = \alpha = \kappa =$ 1E-05, $\beta = \gamma =$ 1E-03, $\theta =$ 1E-04, $\rho = 1$, and $\epsilon = 0.01$ for all datasets due to their slight impacts on model's performance. As a significant parameter, the dimension of latent representations is detailedly analyzed in Discussion section. We adopted two quantitative metrics, i.e., C-index and AUC [7], to evaluate the prognostic performance at the patient-level and event-time level, respectively. And the specific definitions of these two metrics are provided in *Supplementary Materials*, available online.

TABLE 1
The Basic Characteristics of the Studied TCGA Datasets, Including Instance Number, Gender (Male/Female), Follow-Up Time (Mean ± Standard Deviation, Unit: Month), and Status (Alive/Dead)

Dataset	Instance	Gender	Follow-up	Status
KIRC	405	267/138	46.27±32.26	266/139
LIHC	333	226/107	29.33±23.96	217/116
LUAD	438	206/232	30.91±30.68	283/155
SKCM	367	218/149	61.64±67.60	232/135
BLCA	318	238/80	29.03±29.31	222/96
OV	295	0/295	39.92±32.40	132/163
LUSC	403	298/105	33.16±32.91	276/127
LGG	502	279/223	32.08±31.88	379/123
THCA	499	135/364	40.30±32.83	483/16
UCEC	539	0/539	37.80±30.17	448/91
TOTAL	4,099	1,867/2,232	37.97±36.91	2,938/1,161

5.2 Comparison With State-of-The-Art Methods in Mono-Modality Scenario

In this subsection, we initially compared our proposed method with all competing methods that were tested on mono-modality data. We show all comparison results with statistical methods in Tables 2 (C-index) and S2 (AUC). For comparison with machine learning approaches, the experimental results are exhibited in Tables 3 (C-index) and S3 (AUC). From Tables 2-3 and S2-S3, available online, we can have the following observations: I) Except SVM-RR and RSF, nearly all machine learning methods perform better than statistical methods on all datasets. Statistical methods, SVM-RR, and RSF attempted to construct the connection between the original feature matrix and the observed time for prognosis evaluation. But the prognostic performance of such models may be affected by the noise interference and data redundancy in the the original feature space. Instead, those methods like M2DP, RaSR, and McLR learned compact feature subspace by discovering and maintaining the intrinsic structure of original data, which benefits to accurate model construction. Additionally, the high-dimensional original feature space is likely to bring overfitting issue that results in performance degradation of models. As we expected, these methods (e.g., EN-Cox, Lasso-Cox, M2DP, and RaSR) with feature selection/dimension reduction mechanisms generally outperform those models (e.g., Cox, AFT, and BJ) without ones. II) Although the BJ model that used the KM estimator to deal with censored samples works better than Tobit and AFT

[1] https://portal.gdc.cancer.gov/
[2] https://github.com/YuZhang-SMU/Cancer-Prognosis-Analysis/tree/main/MaL%20Code

TABLE 2
Comparison Results With Statistical Methods in Mono-Modality Scenario in Terms of C-Index (Mean ± Standard Deviation) on Genomic (X1) and Histopathological (X2) Data of All Datasets

Dataset	Modality	MaL	Cox	Lasso-Cox	EN-Cox	Tobit	AFT	BJ
KIRC	X1	**0.699±0.075**	0.594±0.066	0.606±0.086	0.608±0.087	0.601±0.110	0.605±0.073	0.628±0.107
	X2	**0.673±0.059**	0.531±0.132	0.537±0.129	0.547±0.073	0.531±0.138	0.537±0.117	0.541±0.048
LIHC	X1	**0.719±0.067**	0.543±0.136	0.544±0.126	0.545±0.126	0.550±0.135	0.558±0.121	0.563±0.077
	X2	**0.664±0.042**	0.526±0.049	0.526±0.106	0.528±0.109	0.541±0.113	0.542±0.061	0.579±0.081
LUAD	X1	**0.695±0.047**	0.570±0.065	0.577±0.054	0.588±0.072	0.553±0.075	0.562±0.099	0.555±0.099
	X2	**0.681±0.058**	0.529±0.062	0.543±0.075	0.548±0.063	0.519±0.102	0.548±0.083	0.537±0.052
SKCM	X1	**0.653±0.053**	0.493±0.112	0.503±0.107	0.505±0.111	0.508±0.115	0.513±0.154	0.542±0.103
	X2	**0.701±0.061**	0.569±0.094	0.574±0.081	0.575±0.082	0.569±0.101	0.570±0.094	0.575±0.106
BLCA	X1	**0.721±0.055**	0.566±0.092	0.568±0.120	0.586±0.096	0.560±0.113	0.568±0.133	0.572±0.098
	X2	**0.702±0.043**	0.581±0.090	0.592±0.092	0.598±0.134	0.560±0.077	0.563±0.106	0.545±0.091
OV	X1	**0.682±0.065**	0.549±0.099	0.550±0.097	0.555±0.048	0.543±0.101	0.549±0.080	0.537±0.072
	X2	**0.644±0.043**	0.472±0.073	0.486±0.098	0.487±0.100	0.434±0.062	0.445±0.084	0.456±0.062
LUSC	X1	**0.708±0.064**	0.536±0.106	0.539±0.112	0.540±0.112	0.522±0.088	0.532±0.094	0.544±0.095
	X2	**0.671±0.032**	0.534±0.078	0.537±0.081	0.555±0.126	0.532±0.067	0.535±0.100	0.505±0.088
LGG	X1	**0.784±0.071**	0.676±0.133	0.698±0.106	0.712±0.088	0.676±0.094	0.695±0.077	0.724±0.092
	X2	**0.725±0.059**	0.553±0.094	0.561±0.105	0.567±0.128	0.573±0.108	0.573±0.117	0.577±0.115
THCA	X1	**0.862±0.082**	0.577±0.308	0.661±0.207	0.664±0.156	0.621±0.305	0.664±0.202	0.685±0.229
	X2	**0.944±0.036**	0.709±0.196	0.721±0.183	0.746±0.218	0.700±0.124	0.748±0.137	0.779±0.251
UCEC	X1	**0.711±0.087**	0.566±0.085	0.590±0.081	0.603±0.116	0.577±0.100	0.594±0.077	0.604±0.087
	X2	**0.705±0.078**	0.511±0.113	0.526±0.108	0.548±0.138	0.520±0.112	0.521±0.135	0.530±0.082

TABLE 3
Comparison Results With Machine Learning Methods in Mono-Modality Scenario in Terms of C-Index (Mean ± Standard Deviation) on Genomic (X1) and Histopathological (X2) Data of All Datasets

Dataset	Modality	MaL	SVM-RR	MTLSA	RSF	M2DP	RaSR	McLR
KIRC	X1	**0.699±0.075**	0.632±0.114	0.665±0.065	0.630±0.086	0.658±0.078	0.652±0.033	0.668±0.102
	X2	**0.673±0.059**	0.565±0.086	0.627±0.067	0.543±0.082	0.638±0.091	0.652±0.073	0.643±0.059
LIHC	X1	**0.719±0.067**	0.601±0.108	0.649±0.077	0.595±0.135	0.648±0.056	0.695±0.039	0.658±0.092
	X2	**0.664±0.042**	0.554±0.102	0.644±0.090	0.533±0.106	0.595±0.068	0.652±0.065	0.656±0.067
LUAD	X1	**0.695±0.047**	0.577±0.101	0.594±0.070	0.579±0.044	0.606±0.090	0.649±0.055	0.670±0.063
	X2	**0.681±0.058**	0.555±0.102	0.620±0.068	0.555±0.075	0.601±0.073	0.663±0.053	0.636±0.050
SKCM	X1	**0.653±0.053**	0.538±0.070	0.617±0.075	0.515±0.096	0.593±0.086	0.648±0.061	0.648±0.097
	X2	**0.701±0.061**	0.583±0.100	0.613±0.074	0.586±0.100	0.630±0.100	0.659±0.082	0.672±0.077
BLCA	X1	**0.721±0.055**	0.598±0.123	0.611±0.065	0.585±0.088	0.649±0.077	0.701±0.073	0.696±0.109
	X2	**0.702±0.043**	0.599±0.039	0.623±0.094	0.550±0.096	0.662±0.078	0.680±0.074	0.676±0.089
OV	X1	**0.682±0.065**	0.556±0.047	0.610±0.090	0.571±0.063	0.611±0.065	0.655±0.061	0.638±0.075
	X2	**0.644±0.043**	0.494±0.070	0.623±0.061	0.504±0.076	0.596±0.051	0.630±0.052	0.634±0.068
LUSC	X1	**0.708±0.064**	0.568±0.061	0.603±0.083	0.557±0.095	0.604±0.063	0.672±0.065	0.665±0.080
	X2	**0.671±0.032**	0.559±0.096	0.628±0.058	0.560±0.092	0.625±0.053	0.649±0.056	0.653±0.108
LGG	X1	**0.784±0.071**	0.759±0.054	0.778±0.056	0.728±0.143	0.761±0.067	0.753±0.096	0.762±0.079
	X2	**0.725±0.059**	0.582±0.102	0.659±0.077	0.591±0.137	0.651±0.085	0.687±0.075	0.679±0.070
THCA	X1	**0.862±0.082**	0.705±0.231	0.835±0.165	0.655±0.301	0.847±0.160	0.833±0.123	0.845±0.158
	X2	**0.944±0.036**	0.788±0.221	0.836±0.152	0.780±0.186	0.908±0.107	0.902±0.116	0.871±0.127
UCEC	X1	**0.711±0.087**	0.626±0.142	0.647±0.073	0.614±0.118	0.613±0.124	0.688±0.079	0.664±0.067
	X2	**0.705±0.078**	0.557±0.080	0.641±0.101	0.560±0.154	0.634±0.096	0.649±0.069	0.677±0.060

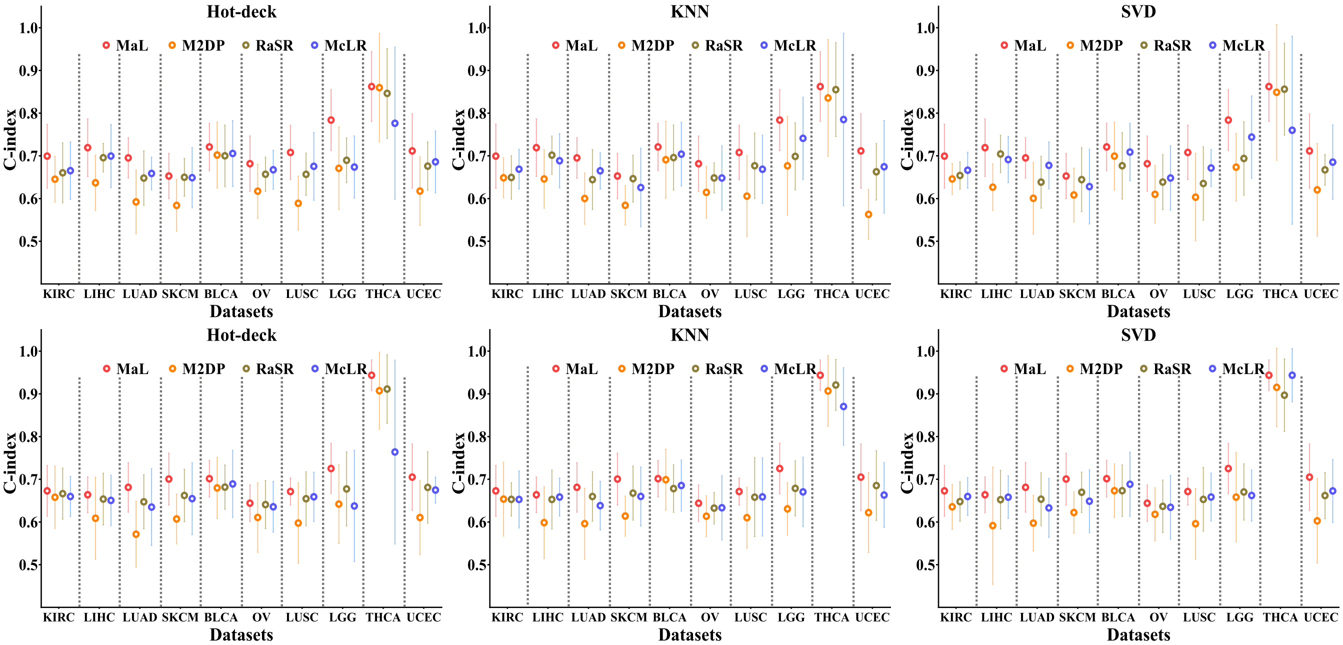


Fig. 1. Box plots of C-index (mean ± standard deviation) achieved by the combinations of multi-modality methods and imputation techniques in incomplete-modality scenario on genomic (Top) and histopathological (Bottom) data of all datasets.

models, these parametric (regression) methods show competing performance with semi-parametric methods, which indicates that only considering the regression strategy is still not enough. Utilizing the regression and ranking constraints, SVM-RR outperforms these parametric methods in most cases, which demonstrates the potential of the combination of regression and ranking constraints. However, SVM-RR is still inferior to other machine learning methods, probably because it is not easy to find the optimal kernel function from predefined alternatives for data in the original space. Additionally, RSF can model nonlinear effects of variables directly. Correspondingly, results obtained by RSF are superior to those of statistical methods in most cases, especially the basic Cox model. However, it has been pointed out that RSF is sensitive to outliers, making it not competitive among machine learning methods [62]. III) Our proposed MaL model achieves the best performance on all datasets and metrics, which may benefit from several potential advantages in MaL: 1) unlike most survival methods that require prior of original feature distribution, mutual-assistance unified framework (**MA-1**) reduces the survival model's dependency on the original attribute distribution; 2) it can preserve the comprehensive perspectives from multi-modality data for mono-modality testing via mutual-assistance similarity and geometry constraints (**MA-2**); 3) mutual-assistance regression-ranking survival model (**MA-3**) can efficiently deal with censored data without strong hypotheses and is able to simultaneously focus on the survival time and event order.

For further comparison, we plotted the KM curves (along with p-values) of all competing methods on all datasets (Figs. S1-S10 in *Supplementary Materials*), available online. Specifically, we used the median of the prognostic index as the threshold, and divided the entire dataset into high-risk and low-risk subgroups. The farther apart the two curves are, the better the model performance is. As we can observe from Figs. S1-S10, available online, similar conclusions as above-mentioned contents can be drawn. Most methods can not efficiently stratify patients or do not meet the statistical significance ($p \leq 0.05$). Compared with all competing methods, the proposed model not only shows good stratification performance, but also generally achieves the best significant p-value, which may own to the reference point proposal and ranking calibration. Note that, relatively speaking, all methods do not perform very well on THCA dataset (at least its genomic data), possibly because THCA dataset contains a higher censoring ratio that would increase the uncertainty of KM estimator and degrade the stratification performance. More explanations can refer to *Supplementary Materials*, available online.

5.3 Comparison With Multi-Modality Methods in Incomplete-Modality Scenario

In this part, we focus on the comparison between the proposed method and recent multi-modality methods (i.e., M2DP, RaSR, and McLR) on incomplete-modality scenario, in which the multi-modality data are partially missing and the multi-modality models are trained and tested with the imputed complete-modality data. We utilized three imputation techniques (i.e., Hot-deck [11], KNN [12], and SVD [15]) to complement missing-modality data according to available multi-modality data so that multi-modality models can be directly trained and tested on those data. Experimental results are exhibited in Fig. 1 (C-index) and Fig. S11 (AUC), available online. According to results in Fig. 1 and Fig. S11, available online, we can find these multi-modality methods achieve better performance on complemented multi-modality data for certain datasets (e.g., BLCA, KIRC, BLCA, OV, and THCA). This finding suggests that the imputation approaches do provide a feasible way to make multi-modality methods applicable for incomplete-modality data, but the performance improvement shows limited benefit. Also, the imputation approaches do not always work (e.g., on SKCM, BLCA, and LGG) since they may introduce additional noises

TABLE 4
Comparison Results With Multi-Modality Methods in Complete-Modality Scenario in Terms of C-Index (mean ± standard Deviation) on all Datasets

Method	Metric	KIRC	LIHC	LUAD	SKCM	BLCA	OV	LUSC	LGG	THCA	UCEC
MaL	C-index	**0.699**	**0.719**	0.695	**0.701**	**0.721**	**0.682**	**0.708**	**0.784**	**0.944**	**0.711**
	(STD)	(0.075)	(0.067)	(0.047)	(0.061)	(0.055)	(0.065)	(0.064)	(0.071)	(0.036)	(0.087)
M2DP	C-index	0.661	0.656	0.629	0.634	0.715	0.620	0.636	0.773	0.923	0.663
	(STD)	(0.100)	(0.079)	(0.069)	(0.081)	(0.080)	(0.101)	(0.067)	(0.080)	(0.051)	(0.104)
RaSR	C-index	0.667	0.721	**0.698**	0.661	0.707	**0.678**	**0.696**	**0.783**	0.938	**0.704**
	(STD)	(0.070)	(0.047)	(0.037)	(0.058)	(0.035)	(0.031)	(0.050)	(0.049)	(0.061)	(0.038)
McLR	C-index	**0.679**	**0.723**	**0.703**	**0.706**	**0.730**	0.672	0.694	0.736	**0.954**	0.692
	(STD)	(0.060)	(0.059)	(0.041)	(0.081)	(0.066)	(0.045)	(0.071)	(0.088)	(0.046)	(0.075)

For MaL, the best results on mono-modality data are reported. The first two best results on each dataset are marked in boldface.

and inconsistent distribution. From another perspective, these methods generally performed data imputation and multi-modality feature learning separately, and some useful information for data imputation may be compromised in the feature learning stage. Our proposed method still outperforms all combinations of multi-modality methods and imputation approaches, which uses mutual-assistance similarity and geometry constraints (**MA-2**) to reinforce mono-modality latent representation learning so as to preserve the multi-modality information. It is worth mentioning that the proposed MaL not only is free from the disturbance of extra noises (that may be) brought by imputation strategies, but also works in especially extreme case where only mono-modality data is available in the testing phase. We also compare RaSR with MaL in the incomplete-modality scenario without any imputation technologies, which can be found in *Supplementary Materials*, available online.

5.4 Comparison With Multi-Modality Methods in Complete-Modality Scenario

We further make comparison between the proposed method (only on mono-modality data) and the multi-modality methods on complete multi-modality data. As we can see from Table 4 (C-index) and Table S4 (AUC), available online, we can see that the performance of M2DP is relatively inferior to other methods, probably because it conducts feature learning and prognosis prediction separately, in which learned features may not be appropriate for the prognosis model. Moreover, McLR and MaL generally do a better job than RaSR due to the fact that RaSR focuses on the global column-wise attributes of latent representations via the min-redundancy and max-relevancy regularizers, but ignores the local properties which are concerned by the similarity constraint in McLR and the geometry preservation in MaL. Interestingly, MaL achieves comparable performance with McLR (e.g., on C-index, No. of Rank1, MaL : McLR = 5 : 5; No. of Rank2, MaL : McLR = 9 : 6), in which the largest difference between MaL and McLR is equal to 0.01 (on THCA) when the McLR shows the best performance. The results are supported by two possible reasons as follows: I) The **MA-2** in MaL can efficiently reinforce the mono-modality representation learning by preserving both global and local characteristics of multi-modality data. II) As for the survival model, all three competing methods apply the Cox model to predict the relative event order while MaL establishes a mutual-assistance regression-ranking model for the fitting of both survival time and event order. Also, MaL has benefited from the bias vector which estimates and utilizes the rich information from censored samples. In this way, it can achieve a fairly close or even better performance to the multi-modality models.

5.5 Efficacy of MA-1

The **MA-1** in MaL refers to the cooperation of representation learning and survival modeling. To explore the interaction between these two components, we developed four variants: 1) removing the representation learning in MaL and directly conducting prognosis prediction (denoted as MaL_RL); 2) replacing the survival modeling in MaL with Cox model (denoted as MaL_Cox); 3) replacing the survival modeling in MaL with AFT model (denoted as MaL_AFT); 4) detaching the survival modeling from the representation learning (denoted as MaL_SM). Note that for the last three variants, the representation learning and the prognosis analysis were conducted sequentially, namely as two separate processes. From Figs. 2a-2b (C-index) and Fig. S13 (AUC), available online, we can observe that: I) The performance of MaL is superior to that of MaL_RL, and similar results can be also

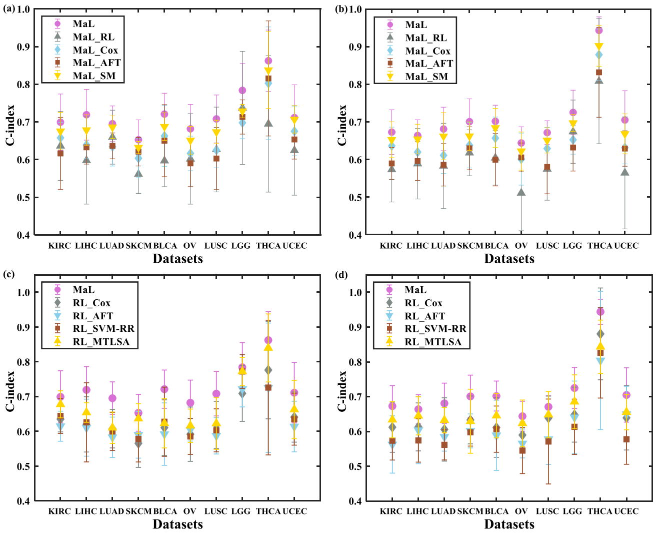


Fig. 2. Box plots of C-index (mean ± standard deviation) achieved by MaL and its variants (for validating the efficacy of **MA-1**) on genomic ((a) and (c)) and histopathological ((b) and (d)) data of all datasets.

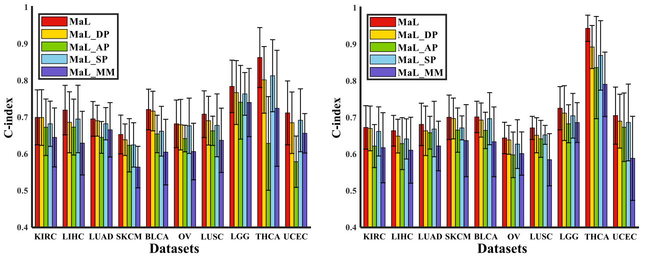


Fig. 3. Bar plots of C-index (mean ± standard deviation) achieved by MaL and its variants (for validating the efficacy of **MA-2**) on genomic (Left) and histopathological (Right) data of all datasets.

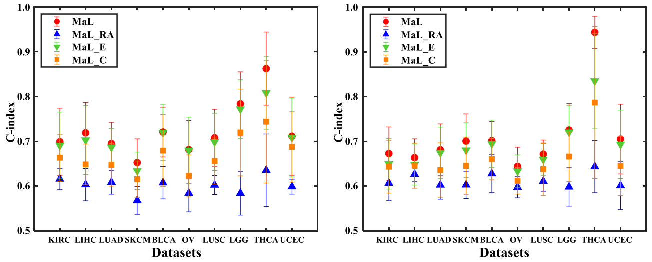


Fig. 4. Box plots of C-index (mean ± standard deviation) achieved by MaL and its variants (for validating the efficacy of **MA-3**) on genomic (Left) and histopathological (Right) data of all datasets.

found between MaL_Cox (MaL_AFT) and its baseline Cox (AFT) model, implying that representation learning is important to survival model by providing a clean and compact subspace. II) MaL_SM obtains higher results than both MaL_Cox and MaL_AFT, indicating that the proposed mutual-assistance prognosis model which contains both ranking and regression functions can boost the accuracy of prognosis prediction, and further validating the conclusion from [27]. III) MaL outperforms MaL_SM and other variants. Although representation learning can provide an informative latent feature subspace for the variants, it still cannot ensure that the latent representations meet the requirement of feature distribution in survival models. MaL integrates the representation learning and survival modeling into a mutual-assistance unified framework which benefits from joint optimization so that latent representations are tailored for the model in a task-driven manner.

For further comparison, we introduced another four variants, namely, RL_Cox, RL_AFT, RL_SVM-RR, and RL_MTLSA, in which representation learning is jointly trained with corresponding survival models. The experimental results can be found in Figs. 2c-2d (C-index) and Figs. S13c-S13d (AUC), available online. *First*, all variants with the corresponding representation learning strategy outperform their baseline models, which further demonstrates the efficacy of the representation learning strategy. *Second*, our proposed MaL, which integrates **MA-3** and representation learning into the unified mutual-assistance framework, works better than all variants, and it suggests that **MA-3** is an effective component for improving the performance of the proposed model.

5.6 Efficacy of MA-2

In MaL, **MA-2** consists of similarity and geometry (i.e., distance and angle) preservation. To corroborate the effectiveness of each part, we developed four variants as follows: 1) MaL_DP (without distance preservation), 2) MaL_AP (without angle preservation), 3) MaL_SP (without similarity preservation), and 4) MaL_MM (without reinforcement of multi-modality data, namely without geometry and similarity preservation). From Fig. 3 (C-index) and Fig. S14 (AUC), available online, we have the following observations: I) MaL_MM is considerably inferior to MaL and its variants. It is not surprising about that since MaL and its variants can utilize the comprehensive perspectives provided by multi-modality data. II) MaL consistently exhibits better results than MaL_DP, MaL_AP, and MaL_SP, which verifies the effectiveness of mutual-assistance similarity and geometry preservation that helps preserve the structural and geometric characteristics of multi-modality data. III) Generally speaking, the total performance degradation of MaL_DP and MaL_AP is larger than that of MaL_SP, suggesting that the geometry constraint is more efficient than the similarity constraint in the multi-modality information preservation. There are two potential reasons: 1) the latent space is more suitable to information preservation than the original feature space since it can capture more inherent characteristics of multi-modality data. 2) local characteristics are more robust to local changes than holistic ones, which has also been demonstrated in [53].

5.7 Efficacy of MA-3

We further investigate the efficacy of **MA-3** which consists of mutual-assistance regression formulation and ranking calibration by designing three variants: 1) MaL_RA: removing the ranking calibration; 2) MaL_E: removing the bias vector; and 3) MaL_C: replacing the regression-ranking function with Cox model. Experimental results are plotted in Fig. 4 (C-index) and Fig. S15 (AUC), available online. From the figures (MaL versus MaL_RA), we can see that the performance of model sharply drops when the ranking calibration is taken away from MaL. A possible reason for this observation is that only using regression model can not distinguish two samples with close predictive risk indexes, which can be remedied by the ranking calibration in MaL. For MaL versus MaL_C, MaL outperforms MaL_C on all datasets, probably because Cox model only emphasizes the chronological order but not the length of survival time, while MaL benefits from the mutual-assistance regression-ranking function so as to focus on the survival time and event order simultaneously. Additionally, as to MaL versus MaL_E, the performance of MaL_E declines (as expected) when the bias vector is removed from MaL, which can be attributed to the fact that the bias vector can efficiently deal with the censoring issue in survival analysis and improve the prediction performance by fully mining the information of censored samples. There are three possible reasons to explain why the bias vector works. *First*, the bias vector is learnable and only works for censored samples (via the indicator vector δ), by which we can directly regress the observed time of all samples. *Second*, the non-negativity of the bias vector places a lower bound on the estimated time of censored samples. In other word, the estimated time is at least longer than the observed one, which provides a feasible way to make full use of information from censored samples. *Finally*, the bias vector appears in both regression formulation and ranking calibration such that it can help improve the accuracy of survival time regression and event order estimation simultaneously. We also summarize the results of the main ablation experiments in Table S5, available online. Please refer to *Supplementary Materials* for more details, available online.

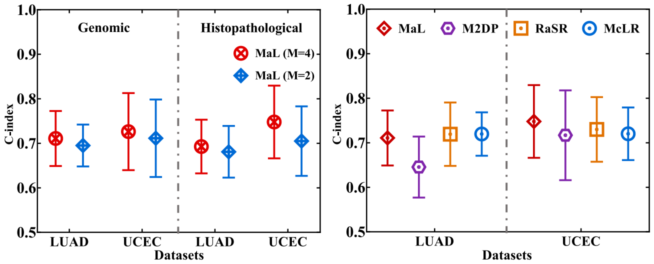


Fig. 5. Box plots of C-index (mean ± standard deviation) of MaL with the more-modality setup on LUAD and UCEC datasets. The left exhibits the performance of MaL with different setups (i.e., M = 2 or 4) on the genomic and histopathological data, while the right shows the comparison results of all competing methods with four-modality setup.

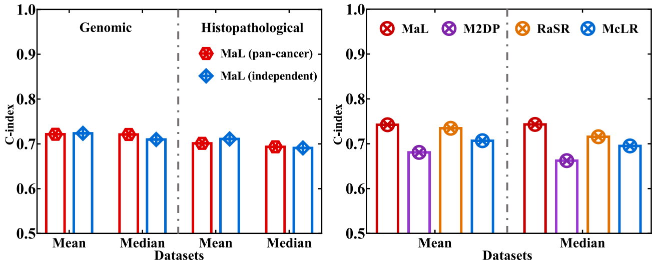


Fig. 6. Bar plots of C-index of MaL on the larger-scale dataset. The left exhibits the performance of MaL on the pan-cancer dataset and the independent-cancer datasets in terms of the mean and median values of their results, while the right shows the comparison results of all competing methods on the pan-cancer dataset.

6 Discussion

In this section, we first extend MaL to a more-modality setup and a larger-scale dataset. Then, we analyze the convergence, complexity and running time of MaL, followed by the model interpretability and prognostic biomarkers. We also make a comparison with multi-modality deep learning methods, and discuss the limitations and future work.

6.1 Extension on More-Modality Setup

Considering that it is not easy to collect the matched multi-modality (especially more than two modalities) data in clinical practices, we only take two modalities into consideration in the previous content, namely, gene expression sequences and histopathological data, which are commonly included in the used TCGA datasets. In fact, the proposed method can also be applied to the more-modality scenario by taking their feature matrices as input. For text-type data (such as DNA methylation, miRNA, and copy number variation sequences), they can be directly fed into the proposed method after the same preprocessing with the protein coding mRNA sequences. As to image-type data (such as CT, PET, and MR images), we first extract various disease-related features (e.g., texture features and intensity features) to quantify the images, and then use those extracted features as the input of the model, which is also similar to the preprocessing of histopathological data.

In this experiment, we additionally enrolled another two modalities, i.e., miRNA and the copy number variation (CNV), for both LUAD and UCEC datasets and tested the performance of the proposed MaL in the scenario where it is trained with four modalities data and tested with mono-modality data. From the left subfigures shown in Fig. 5 (C-index) and Fig. S16 (AUC), available online, we can observe that the performance of MaL on either the genomic data or the histopathological data is further improved when more modalities are enrolled, revealing that more modalities might provide more comprehensive information for the model. Meanwhile, we also test multi-modality methods (i.e., M2DP, RaSR, and McLR) on complete four modalities data. As shown in the right subfigures of Fig. 5 (C-index) and Fig. S16 (AUC), available online, MaL on mono-modality testing data still achieves comparable performance with both RaSR and McLR on four-modality testing data, which further demonstrates the efficacy of its extension on more-modality setup. Note that although the proposed MaL performs better on genomic data than histopathological data of most datasets as shown in Tables 2-3 and S2-S3, it does not always occur (e.g., on THCA) and certain methods also show better performance on histopathological data. It may be associated with specific cancer types. Therefore, it is difficult to give a definite conclusion of which data type is the best for survival analysis. Interestingly, the genomic data may be a better choice for MaL and other representation learning-based methods (e.g., M2DP and RaSR) when compared with histopathological data, which is consistent with the observations in [48], [63].

6.2 Extension on Larger-Scale Dataset

In view of cancer heterogeneity, we have conducted and compared all competing methods on each independent dataset with respect to a certain specific cancer type. In fact, one advantage of our method is that our proposed model essentially belongs to a traditional machine learning method that contains relatively fewer parameters, and thus it does not require very large-scale datasets for training. To investigate the scalability of our proposed method, we conducted experiments on a larger-scale dataset. Since the number of the matched multi-modality data is limited, similar to [39][41],we directly combined all studied datasets and obtained a large-scale pan-cancer dataset (called TOTAL) which includes 4,099 samples, as Table 1 described. We compare the performance of MaL on the pan-cancer dataset and the independent-cancer datasets. The results in the left subfigures of Fig. 6 (C-index) and Fig. S17 (AUC), available online, show that MaL still exhibits promising performance when it is applied to the large-scale pan-cancer dataset. However, the heterogeneity of different cancer types will increase the difficulty of pan-cancer analysis so that some testing results on the pan-cancer dataset are likely not as good as those on independent-cancer datasets. Considering that three multi-modality methods (i.e., M2DP, RaSR, and McLR) achieve good performance among all comparison methods, we also compare MaL (with mono-modality testing) with them (with multi-modality testing) on the pan-cancer dataset. From the right subfigures of Fig. 6 (C-index) and Fig. S17 (AUC), available online, the MaL works better than three multi-modality methods, indicating a good scalability of MaL on the scale of dataset. Notably, in right subfigures, the best results of MaL on mono-modality data are reported.

6.3 Convergence, Complexity and Running Time

The convergence of ALM and ADMM strategy is theoretically analyzed in [64] and [56], respectively. As expected,

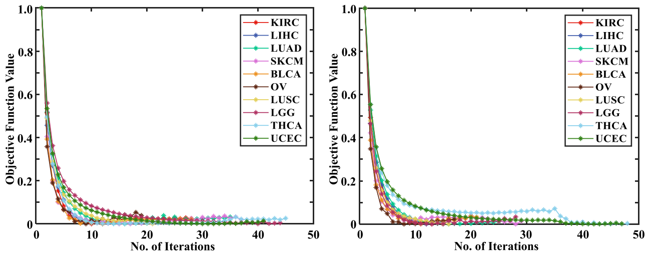


Fig. 7. Convergence curves of the objective function on genomic (Left) and histopathological (Right) data of all datasets.

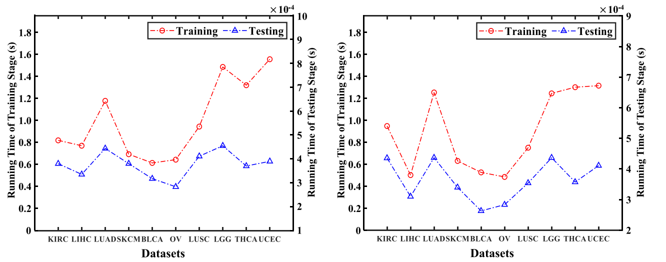


Fig. 8. Running time of the proposed method on genomic (Left) and histopathological (Right) data of all datasets.

when we get the optimal solutions of all variables, the objective function of MaL has also reached its lower bound, as shown in Fig. 7. It can be observed that the optimization algorithm converges within 20 iterations for almost all datasets, regardless of gene data or histopathological images. The maximum tangent slope of most convergence curves appears in the first iteration and is with the scope of $[-0.440, -0.652]$, which suggests that MaL can converge quickly.

With regard to the time complexity in the training phase, we consider the update of all variables, which results in a total complexity of $\mathcal{O}((C + m_y + m_u)n^2 + (m_x m_y + m_y + m_u + m_u C + m_y C + m_y m_u + m_z m_u + r m_y + m_x^2 + m_y^2 + m_z^2) n + m_x^3 + m_y^3 + m_z^3 + m_y^2 m_u)$ in each iteration, where r is the rank of variable $\mathbf{J}$. Furthermore, we derive an approximation of the foregoing complexity, since m_u, m_y, and C are rather small when compared with m_z and m_x. Therefore, the overall training complexity of each iteration is simplified to $\mathcal{O}((C + m_y + m_u)n^2 + (m_x^2 + m_z^2)n + m_x^3 + m_z^3)$. As for testing, given n_{test} samples, the corresponding time complexity is $\mathcal{O}(n_{\text{test}} m_x m_y)$. Fig. 8 shows the average running time of our proposed method on all studied datasets, with the left subfigure corresponding to the genomic data while the right one the histopathological data. From Fig. 8, we can find that our proposed method is efficient in either the training or testing phase, especially the latter which is at the millisecond level. Additionally, the running time is prominently related to the number of samples, which also supports our theoretical complexity analysis. In *Supplementary Materials*, available online, we also compare the proposed method with other competing ones in terms of running time. Please refer to the Section S3.9 and Table S6 for more details.

6.4 Model Interpretability and Prognostic Biomarkers

The interpretability of the model is one of the most important aspects for survival analysis methods [40]. To interpret the proposed model, we first explain the meaning of the outputs, namely, the projection matrix $\mathbf{W}$ and the coefficient vector $\mathbf{K}$. Since the projection matrix $\mathbf{W}$ connects the original mono-modality data $\mathbf{X}$ and the latent representations $\mathbf{Y}$, it is inherently a weight matrix with each element $\mathbf{W}^{(i,j)}$ representing the importance of the ith feature in $\mathbf{X}$ to the construction of the jth feature in $\mathbf{Y}$. Furthermore, each element in $\mathbf{K}$ denotes the importance of the corresponding feature vector in $\mathbf{Y}$ to the final prognostic prediction.

Based on the above-mentioned content, we can identify potential biomarkers related to the prognosis of cancers. For genomic data, we first pick out the element with the biggest value in $\mathbf{K}$, and then trace back to the projection matrix $\mathbf{W}$ to find out the corresponding column and select the important elements of the column thereafter. The selected elements connect the contributive features in the original space $\mathbf{X}$. Since the PCA algorithm is essentially a linear transformation, in which the projection matrix bridges the original genomic data and feature matrix $\mathbf{X}$. Similar to [40], after finding the most contributive feature vector in $\mathbf{X}$, we can link it to a specific column vector of the projection matrix in PCA, through which we can select the biggest 20 values in the column vector and these values indicate the most important 20 genomic biomarkers among 19,841 candidates.

Different from genomic biomarker selection, in our preprocessing pipeline for histopathological images, features extracted from all patches of whole slide images (WSI) were integrated into the patient-level features by a patch pooling strategy so as to reduce computational cost, which makes it difficult to automatically determine patch biomarkers. Therefore, to give an intuitive illustration, we randomly select two subjects from high-risk and low-risk subgroups (identified by the proposed method and consistent with the observed survival time) for each dataset. We then locate the representative regions on the histopathological images of these two subjects, and randomly select patches from these regions to exhibit the differences between high-risk and low-risk subgroups.

For each cancer type, we exhibit some representative patches of histopathological images and the top 20 genes selected by our model in Tables S7-S16. Taking UCEC as an example (Table S16), the subject with high-risk (or low-risk) value has a relatively shorter (or longer) survival time, which is also found in other cancers and coincides with the basic definition of survival risk, suggesting that our model achieves high accuracy in stratifying patients. Through the contrastive analysis of histopathological patches from low- and high-risk subgroups, we can observe that a large number of irregular nucleus and chromatin condensation commonly appear in high-risk patients. And it has been widely recognized that the change of nuclear structure is relevant to cancer progression [65]. As to gene sequences, the mitochondrial DNA (mtDNA) is selected by MaL for all studied cancers (i.e., MT), which is consistent with the conclusion in [66] that mitochondrial genome may be one of critical factors in carcinogenesis.

6.5 Comparison With Deep Learning Methods

Deep learning methods are known for the ability to learn deep representations from the original data and model the nonlinear relationship. Recently, various deep learning methods, based on either multi-modality data [38], [39],

TABLE 5
Rough Comparison With Deep Learning Methods on the TCGA Datasets

Method	Modality	Partition	KIRC	LIHC	LUAD	SKCM	BLCA	OV	LUSC	LGG	THCA	UCEC
Vale-Silva et al. [39]	X1-6	HO	>0.699	>0.719	<0.695↑	>0.701	>0.721	<0.682↑	<0.708↑	<0.784↑	>0.944	<0.711↑
Azher et al. [40]	X1-2, X5	HO	-	0.701↑	-	0.645↑	0.681↑	-	-	-	-	-
Cheerla et al. [41]	X1-4	HO	0.78	0.78	0.72	0.54↑	0.60↑	0.54↑	0.63↑	0.70↑	0.53↑	0.67↑
Chai et al. [67]	X1, X4-6	CV	-	0.710↑	0.629↑	0.644↑	0.646↑	-	0.597↑	0.823	-	-
Herrmann et al. [68]	X3-4, X6-8	CV	0.721	0.566↑	0.636↑	0.580↑	0.618↑	0.575↑	0.501↑	0.695↑	-	0.646↑
Ours	X1 or X2	CV	0.699	0.719	0.695	0.701	0.721	0.682	0.708	0.784	0.944	0.711

Method	Modality	Partition	KIRC	LUAD	LUSC	GBM&LGG	GBM	BRCA
Chen et al. [43]	X2, X6, X9	CV	0.720	-	-	0.826↑	-	-
Chang et al. [35]	X2	CV	-	-	0.668↑	-	-	-
Di et al. [36]	X2	CV	-	-	0.661↑	-	0.663↑	-
Chen et al. [37]	X2	CV	0.642↑	0.538↑	-	-	-	-
Wang et al. [44]	X1-2	CV	-	-	-	-	-	0.723↑
Li et al. [45]	X1-3, X6	HO	-	-	-	-	-	0.766↑
Ours	X1 or X2	CV	0.699	0.695	0.708	0.834	0.764	0.774

Results that are inferior to ours are marked with ↑. Abbreviations: X1, mRNA; X2, WSI; X3, clinical data; X4, miRNA; X5, DNAm; X6, CNV; X7, RNA; X8, mutation; X9, RNA-Seq; CV, cross-validation; HO, hold-out.

[40], [41], [42], [43], [44], [45], [67], [68] or mono-modality data [35], [36], [37] have been proposed for cancer survival analysis. Considering that these deep learning methods may contain many parameters, we cannot tune all potentially key parameters for each dataset, which may result in performance reduction and is unfair for comparison. Fortunately, we can make a rough comparison with several deep learning methods that used TCGA datasets. Apart from ten datasets used in our method, we also tested MaL on three extra datasets from TCGA, i.e., GBM (157 samples), GBM&LGG (the union set of the GBM and the LGG datasets, 659 samples), and BRCA (a relatively large dataset with 1,065 samples). We provide the comparison results in Table 5. While these competing methods have used different modalities and evaluation strategies, we can still draw some conclusions: 1) As shown in Table 5, all deep learning models have achieved promising results on one or more ($\geq$ two) modalities data for cancer survival analysis. And the performance of our method on mono-modality testing data, not only outperforms that of mono-modality deep learning models, but also is comparable to and even better than that of certain deep learning methods on complete multi-modality testing data. 2) Compared with most of these deep learning methods, our model is more interpretable and can help identify potential prognostic biomarkers. 3) Relatively speaking, our model refers to fewer parameters and thus it does not require very large-scale datasets for training.

6.6 Limitations and Future Work

This work still has several limitations and some potential interesting directions for future work. *First*, the proposed method mainly concentrates on the prediction of mono-modality data in the testing phase. For partially incomplete multi-modality data, a straightforward way is to use mono-modality data for standalone testing. Alternatively, we made an attempt by using the ensemble strategy (i.e., voting) to integrate the predictive results of two standalone models for those samples with matched multi-modal data in testing phase. The experimental results show this strategy helps improve the performance of MaL in most cases (7 of 10 datasets) when compared with the highest results on mono-modality data. Although it exhibits slight improvement (overall average performance: 0.739 *versus* 0.736 *versus* 0.723 *versus* 0.711 [the ensemble results *versus* the highest results on mono-modality data *versus* the results on genomic data *versus* the results on histopathological data] for C-index, 0.776 *versus* 0.772 *versus* 0.756 *versus* 0.743 for AUC), it implies that the ensemble strategy is feasible and better built-in ensemble mechanisms can be considered in the future to integrate the predictive results of models. *Second*, this work independently investigates the underlying information of diverse cancers, while commonalities and relationships among different cancer types can also provide new insights into the survival analysis of cancers, which makes the pan-cancer analysis one of the possible directions. Moreover, the influence of the censoring ratio deserves deeper investigation in the future.

7 Conclusion

The main concern of this paper is to enhance the performance of standalone mono-modality survival analysis of human cancers. To this end, we propose a mutual-assistance learning paradigm that contains three kinds of mutual-assistance strategies: 1) mutual-assistance similarity and geometry constraints utilize the knowledge of multi-modality data to reinforce the mono-modality representation learning; 2) mutual-assistance regression and ranking functions simultaneously estimate survival time and event order while are independent of strong hypotheses; 3) a mutual-assistance unification of representation learning and survival modeling alleviates the requirement of attribute distribution. Experiments on ten datasets from TCGA demonstrate its superiority.

Acknowledgments

We would like to express our gratitude to The Cancer Genome Atlas (TCGA) database for the open-source data and to anonymous reviewers for their valuable and insightful comments.

References

[1] H. Sung et al., "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[2] World Health Organization, "WHO report on cancer: setting priorities, investing wisely and providing care for all." Accessed: Feb. 3, 2020. [Online]. Available: https://www.who.int/publications/i/item/who-report-on-cancer-setting-priorities-investing-wisely-and-providing-care-for-all

[3] J. Emmerson and J. Brown, "Understanding survival analysis in clinical trials," *Clin. Oncol.*, vol. 33, no. 1, pp. 12–14, 2021.

[4] Z. Ahmed, K. Mohamed, S. Zeeshan, and X. Dong, "Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine," *Database*, vol. 2020, 2020, Art. no. baaa010.

[5] C. M. Tempany et al., "Multimodal imaging for improved diagnosis and treatment of cancers," *Cancer*, vol. 121, no. 6, pp. 817–827, 2015.

[6] M. Fan, P. Xia, R. Clarke, Y. Wang, and L. Li, "Radiogenomic signatures reveal multiscale intratumour heterogeneity associated with biological functions and survival in breast cancer," *Nature Commun.*, vol. 11, no. 1, pp. 1–12, 2020.

[7] W. Shao et al., "Integrative analysis of pathological images and multi-dimensional genomic data for early-stage cancer prognosis," *IEEE Trans. Med. Imag.*, vol. 39, no. 1, pp. 99–110, Jan. 2020.

[8] Y. Zhang, A. Li, J. He, and M. Wang, "A novel MKL method for GBM prognosis prediction by integrating histopathological image and multi-omics data," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 1, pp. 171–179, Jan. 2020.

[9] Q. He et al., "Feasibility study of a multi-criteria decision-making based hierarchical model for multi-modality feature and multi-classifier fusion: Applications in medical prognosis prediction," *Inf. Fusion*, vol. 55, pp. 207–219, 2020.

[10] Z. Ning et al., "Integrative analysis of cross-modal features for the prognosis prediction of clear cell renal cell carcinoma," *Bioinformatics*, vol. 36, no. 9, pp. 2888–2895, 2020.

[11] R. R. Andridge and R. J. Little, "A review of hot deck imputation for survey non-response," *Int. Statist. Rev.*, vol. 78, no. 1, pp. 40–64, 2010.

[12] X. Dong et al., "TOBMI: Trans-omics block missing data imputation using a k-nearest neighbor weighted approach," *Bioinformatics*, vol. 35, no. 8, pp. 1278–1283, 2019.

[13] M. S. Santos, J. P. Soares, P. H. Abreu, H. Araújo, and J. Santos, "Influence of data distribution in missing data imputation," in *Proc. Conf. Artif. Intell. Med. Europe*, 2017, pp. 285–294.

[14] P. J. García-Laencina, P. H. Abreu, M. H. Abreu, and N. Afonoso, "Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values," *Comput. Biol. Med.*, vol. 59, pp. 125–133, 2015.

[15] T. Hastie, R. Tibshirani, G. Sherlock, M. Eisen, P. Brown, and D. Botstein, "Imputing missing data for gene expression arrays," Division Biostatist., Stanford Univ., Stanford, CA, USA, Tech. Rep., 1999.

[16] Y. Li, K. S. Xu, and C. K. Reddy, "Regularized parametric regression for high-dimensional survival analysis," in *Proc. SIAM Int. Conf. Data Mining*, 2016, pp. 765–773.

[17] M. Sloma, F. Syed, M. Nemati, and K. S. Xu, "Empirical comparison of continuous and discrete-time representations for survival prediction," in *Proc. Survival Prediction-Algorithms, Challenges Appl.*, 2021, pp. 118–131.

[18] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman, "Survival analysis Part I: Basic concepts and first analyses," *Brit. J. Cancer*, vol. 89, no. 2, pp. 232–238, 2003.

[19] D. G. Kleinbaum et al., *Survival Analysis: A Self-Learning Text*, vol. 3. Berlin, Germany: Springer, 2012.

[20] S. Prinja, N. Gupta, and R. Verma, "Censoring in clinical trials: Review of survival analysis techniques," *Indian J. Community Med.: Official Publication Indian Assoc. Prev. Social Med.*, vol. 35, no. 2, 2010, Art. no. 217.

[21] H. Soleimani, J. Hensman, and S. Saria, "Scalable joint models for reliable uncertainty-aware event prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 8, pp. 1948–1963, Aug. 2018.

[22] S. G. Zadeh and M. Schmid, "Bias in cross-entropy-based training of deep survival networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 9, pp. 3126–3137, Sep. 2021.

[23] D. R. Cox, "Regression models and life-tables," *J. Roy. Statist. Soc.: Ser. B. (Methodol.)*, vol. 34, no. 2, pp. 187–202, 1972.

[24] J. Tobin, "Estimation of relationships for limited dependent variables," *Econometrica: J. Econometric Soc.*, vol. 26, pp. 24–36, 1958.

[25] J. D. Kalbfleisch and R. L. Prentice, *The Statistical Analysis of Failure Time Data*. Hoboken, NJ, USA: Wiley, 2011.

[26] C. Berzuini and C. Larizza, "A unified approach for modeling longitudinal and failure time data, with application in medical monitoring," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 2, pp. 109–123, Feb. 1996.

[27] V. Van Belle, K. Pelckmans, S. Van Huffel, and J. A. K. Suykens, "Support vector methods for survival analysis: A comparison between ranking and regression approaches," *Artif. Intell. Med.*, vol. 53, no. 2, pp. 107–118, 2011.

[28] P. Wang, Y. Li, and C. K. Reddy, "Machine learning for survival analysis: A survey," *ACM Comput. Surv.*, vol. 51, no. 6, pp. 1–36, 2019.

[29] F. Kiaee, H. Sheikhzadeh, and S. E. Mahabadi, "Relevance vector machine for survival analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 3, pp. 648–660, Mar. 2016.

[30] P. Raman et al., "A comparison of survival analysis methods for cancer gene expression RNA-sequencing data," *Cancer Genet.*, vol. 235, pp. 1–12, 2019.

[31] M. N. Wright, T. Dankowski, and A. Ziegler, "Unbiased split variable selection for random survival forests using maximally selected rank statistics," *Statist. Med.*, vol. 36, no. 8, pp. 1272–1284, 2017.

[32] K.-H. Yu et al., "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features," *Nature Commun.*, vol. 7, no. 1, pp. 1–10, 2016.

[33] X. Zhu, J. Yao, F. Zhu, and J. Huang, "WSISA: Making survival prediction from whole slide histopathological images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7234–7242.

[34] Z. Huang et al., "Deep learning-based cancer survival prognosis from rna-seq data: Approaches and evaluations," *BMC Med. Genomic.*, vol. 13, no. 5, pp. 1–12, 2020.

[35] J.-R. Chang, C.-Y. Lee, C.-C. Chen, J. Reischl, T. Qaiser, and C.-Y. Yeh, "Hybrid aggregation network for survival analysis from whole slide histopathological images," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2021, pp. 731–740.

[36] D. Di, S. Li, J. Zhang, and Y. Gao, "Ranking-based survival prediction on histopathological whole-slide images," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2020, pp. 428–438.

[37] R. J. Chen et al., "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16144–16155.

[38] Y. Wu, J. Ma, X. Huang, S. H. Ling, and S. W. Su, "DeepMMSA: A novel multimodal deep learning method for non-small cell lung cancer survival analysis," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2021, pp. 1468–1472.

[39] L. A. Vale-Silva and K. Rohr, "Long-term cancer survival prediction using multimodal deep learning," *Sci. Rep.*, vol. 11, no. 1, pp. 1–12, 2021.

[40] Z. L. Azher, L. J. Vaickus, L. A. Salas, B. C. Christensen, and J. J. Levy, "Development of biologically interpretable multimodal deep learning model for cancer prognosis prediction," in *Proc. 37th ACM/SIGAPP Sympos. Appl. Comput.*, 2022, pp. 636–644.

[41] A. Cheerla and O. Gevaert, "Deep learning with multimodal representation for pancancer prognosis prediction," *Bioinformatics*, vol. 35, no. 14, pp. i446–i454, 2019.

[42] L. A. Vale-Silva and K. Rohr, "Pan-cancer prognosis prediction using multimodal deep learning," in *Proc. IEEE 17th Int. Symp. Biomed. Imag.*, 2020, pp. 568–571.

[43] R. J. Chen et al., "Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis," *IEEE Trans. Med. Imag.*, vol. 41, no. 4, pp. 757–770, Apr. 2022.

[44] Z. Wang, R. Li, M. Wang, and A. Li, "GPDBN: Deep bilinear network integrating both genomic data and pathological images for breast cancer prognosis prediction," *Bioinformatics*, vol. 37, no. 18, pp. 2963–2970, 2021.

[45] R. Li, X. Wu, A. Li, and M. Wang, "HFBSurv: Hierarchical multimodal fusion with factorized bilinear models for cancer survival prediction," *Bioinformatics*, vol. 38, no. 9, pp. 2587–2594, 2022.

[46] J.-R. Chang et al., “Stain mix-up: Unsupervised domain generalization for histopathology images,” in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2021, pp. 117–126.

[47] R. Mazumder, T. Hastie, and R. Tibshirani, “Spectral regularization algorithms for learning large incomplete matrices,” *J. Mach. Learn. Res.*, vol. 11, pp. 2287–2322, 2010.

[48] Z. Ning, D. Du, C. Tu, Q. Feng, and Y. Zhang, “Relation-aware shared representation learning for cancer prognosis analysis with auxiliary clinical variables and incomplete multi-modality data,” *IEEE Trans. Med. Imag.*, vol. 41, no. 1, pp. 186–198, Jan. 2022.

[49] B. Jing et al., “A deep survival analysis method based on ranking,” *Artif. Intell. Med.*, vol. 98, pp. 1–9, 2019.

[50] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *J. Roy. Statist. Soc.: Ser. B. (Methodol.)*, vol. 58, no. 1, pp. 267–288, 1996.

[51] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, “Regularization paths for cox’s proportional hazards model via coordinate descent,” *J. Statist. Softw.*, vol. 39, no. 5, 2011, Art. no. 1.

[52] J. Buckley and I. James, “Linear regression with censored data,” *Biometrika*, vol. 66, no. 3, pp. 429–436, 1979.

[53] N. Armanfard, J. P. Reilly, and M. Komeili, “Local feature selection for data classification,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 6, pp. 1217–1227, Jun. 2016.

[54] M. Yu, L. Liu, and L. Shao, “Structure-preserving binary representations for RGB-D action recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1651–1664, Aug. 2016.

[55] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, “Evaluating the yield of medical tests,” *Jama*, vol. 247, no. 18, pp. 2543–2546, 1982.

[56] X. Zhen, M. Yu, X. He, and S. Li, “Multi-target regression via robust low-rank learning,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 497–504, Feb. 2018.

[57] J.-F. Cai, E. J. Candès, and Z. Shen, “A singular value thresholding algorithm for matrix completion,” *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.

[58] Z. Ning et al., “Multi-constraint latent representation learning for prognosis analysis using multi-modal data,” *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 01, 2021, doi: 10.1109/TNNLS.2021.3112194.

[59] Y. Li, J. Wang, J. Ye, and C. K. Reddy, “A multi-task learning formulation for survival analysis,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1715–1724.

[60] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, “Random survival forests,” *Ann. Appl. Statist.*, vol. 2, no. 3, pp. 841–860, 2008.

[61] W. Shao et al., “Multi-task multi-modal learning for joint diagnosis and prognosis of human cancers,” *Med. Image Anal.*, vol. 65, 2020, Art. no. 101795.

[62] F. Miao, Y.-P. Cai, Y.-T. Zhang, and C.-Y. Li, “Is random survival forest an alternative to cox proportional model on predicting cardiovascular disease?,” in *Proc. 6th Eur. Conf. Int. Federation Med. Biol. Eng.*, 2015, pp. 740–743.

[63] B. He et al., “Integrating spatial gene expression and breast tumour morphology via deep learning,” *Nature Biomed. Eng.*, vol. 4, no. 8, pp. 827–834, 2020.

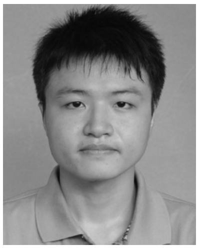
[64] Z. Lin, R. Liu, and Z. Su, “Linearized alternating direction method with adaptive penalty for low-rank representation,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 612–620.

[65] M. Reis-Sobreiro et al., “Emerin deregulation links nuclear shape instability to metastatic potential,” *Cancer Res.*, vol. 78, no. 21, pp. 6086–6097, 2018.

[66] E. Mambo et al., “Tumor-specific changes in mtDNA content in human cancer,” *Int. J. Cancer*, vol. 116, no. 6, pp. 920–924, 2005.

[67] H. Chai, X. Zhou, Z. Zhang, J. Rao, H. Zhao, and Y. Yang, “Integrating multi-omics data through deep learning for accurate cancer prognosis prediction,” *Comput. Biol. Med.*, vol. 134, 2021, Art. no. 104481.

[68] M. Herrmann, P. Probst, R. Hornung, V. Jurinovic, and A.-L. Boulesteix, “Large-scale benchmark study of survival prediction methods using multi-omics data,” *Brief. Bioinf.*, vol. 22, no. 3, 2021, Art. no. bbaa167.



Zhenyuan Ning (Graduate Student Member, IEEE) received the BS degree from Southern Medical University, China, in 2018, where he is currently working toward the PhD degree with the School of Biomedical Engineering. He was also a visiting scholar (2019-2020) with the Department of Radiology and BRIC, University of North Carolina at Chapel Hill. His research interests include survival analysis, medical image analysis, and machine learning.

Zhangxin Zhao received the BS degree in biomedical engineering from Southern Medical University, China, in 2022. And she is currently working toward the MS degree with the School of Biomedical Engineering. Her research interests include survival analysis, medical image analysis, and machine learning.

Qianjin Feng (Member, IEEE) received the MS and PhD degrees in biomedical engineering from First Military Medical University, China, in 2000 and 2003, respectively. From 2003 to 2004, he was a faculty member with the School of Biomedical Engineering, First Military Medical University. Since 2004, he has been with Southern Medical University, China, where he is currently a professor and the dean of the School of Biomedical Engineering. His research interests include medical image analysis, pattern recognition, and computerized-aided diagnosis.

Wufan Chen (Senior Member, IEEE) received the BS and MS degrees both from Beihang University, in 1975 and 1981, respectively. He is currently a full professor with the Guangdong Provincial Key Laboratory of Medial Image Processing, School of Biomedical Engineering, Southern Medical University. His research interests include biomedical imaging principle and image processing.

Qing Xiao (Graduate Student Member, IEEE) received the BS degree in biomedical engineering from Southern Medical University, China, in 2020. She is currently working toward the PhD degree with Southern Medical University. Her research interests include machine learning, medical image analysis, and survival analysis.

Yu Zhang (Member, IEEE) received the PhD degree in biomedical engineering from First Military Medical University, China, in 2003. He now is a professor and the vice dean with the School of Biomedical Engineering, Southern Medical University, China. His research interests include medical image processing and analysis, survival analysis, and machine learning.