

MurreNet: Modeling Holistic Multimodal Interactions Between Histopathology and Genomic Profiles for Survival Prediction

Mingxin Liu¹, Chengfei Cai², Jun Li¹, Pengbo Xu³, Jinze Li¹,
Jiquan Ma⁴, and Jun Xu¹(✉)

¹ Jiangsu Key Laboratory of Intelligent Medical Image Computing,
School of Artificial Intelligence,
Nanjing University of Information Science and Technology, Nanjing, China
jxu@nuist.edu.cn

² College of Information Engineering, Taizhou University, Taizhou, China

³ College of Bioinformatics Science and Technology,
Harbin Medical University, Harbin, China

⁴ School of Computer and Big Data, Heilongjiang University, Harbin, China

Abstract. Cancer survival prediction requires integrating pathological Whole Slide Images (WSIs) and genomic profiles, a challenging task due to the inherent heterogeneity and the complexity of modeling both inter- and intra-modality interactions. Current methods often employ straightforward fusion strategies for multimodal feature integration, failing to comprehensively capture modality-specific and modality-common interactions, resulting in a limited understanding of multimodal correlations and suboptimal predictive performance. To mitigate these limitations, this paper presents a **Multimodal Representation Decoupling Network (MurreNet)** to advance cancer survival analysis. Specifically, we first propose a Multimodal Representation Decomposition (MRD) module to explicitly decompose paired input data into modality-specific and modality-shared representations, thereby reducing redundancy between modalities. Furthermore, the disentangled representations are further refined then updated through a novel training regularization strategy that imposes constraints on distributional similarity, difference, and representativeness of modality features. Finally, the augmented multimodal features are integrated into a joint representation via proposed Deep Holistic Orthogonal Fusion (DHOF) strategy. Extensive experiments conducted on six TCGA cancer cohorts demonstrate that our MurreNet achieves state-of-the-art (SOTA) performance in survival prediction.

Keywords: Whole slide images · Survival outcome prediction · Multimodal learning · Multimodal representation decoupling.

1 Introduction

Survival analysis, a cornerstone of clinical practice, aims to predict the relative risk of death in cancer prognosis and provide crucial references for the cancer prevention and treatment plan [22]. In current clinical practice, survival prediction

methods integrating gigapixel WSIs and genomic data are particularly promising [21]. WSIs provide qualitative morphological information and detailed spatial depiction of tumor tissues [16,17], and their interactions in the tumor micro-environment (TME); whereas genomics offer molecular information from the microcosmic perspective. Both two modalities have their unique strengths and constraints, and their integration is pivotal for advancing the precision of survival analysis. Therefore, the recent focus of survival analysis studies has shifted from single-modal predictions [4,7,11] to utilizing multimodal information [5,6,9,15,25] since their superior capabilities in patient survival prognosis [3,24].

The multimodal knowledge within integrative pathology-genomic data can be decomposed into distinct components: modality-common (inter-modal) and modality-specific (intra-modal) representations. Furthermore, these two divers features are extremely intricate, as each modality embodies wealthy information, yet only a small subset can be mutually correlated with cancer survival. Therefore, previous studies focused on integrating common information, highlighting the alignment of shared features across modalities [5,15,25,26,30]. For example, MGCT [15] uses mutual-guided cross-modality attention to model shared information and enhance modality-common interactions. While modality-common information often dominates multimodal fusion, resulting in the suppression of modality-specific information, thus neglecting the richness of distinct perspectives. Recently, several works [18,29,31] have attempted to improve multimodal interactions modeling by a *divide-and-conquer* manner, handling modality-specific and -common representations while preserving their unique contributions. For instance, CCL [31] presents a Cohort-individual Cooperative Learning framework for survival analysis by collaborating knowledge decomposition and cohort guidance. MuGI [18] leveraged a generator and a MLP to extract the shared and specific information respectively, then parallel cross-attentions to integrate the resulting features. Nevertheless, these methods still suffer from some limitations: simply decoupling the features into shared and specific parts then simple fusion, lacking further optimization on multimodal fusion or dedicated learning strategies for guiding highly divergent features through targeted constraints, which resulting in a suboptimal performance.

To tackle the challenges above, we propose the multimodal representation decoupling network (MurreNet), as depicted in Fig. 1, tailored for survival prediction by modeling holistic interactions between pathology and genomic data, which highlighting the importance of decomposed multimodal representation learning. Specifically, our key contributions are in three folds: (1) a Multimodal Representation Decomposition (MRD) module is devised to learn modality-specific and modality-common representations, providing a comprehensive and disentangled view of multimodal data. (2) a Deep Holistic Orthogonal Fusion (DHOF) module is proposed to integrate complementary modality-specific and modality-common information. (3) a robust loss combination of modality similarity, difference, reconstruction is introduced to provide comprehensive learning constraints. Extensive experimental results confirmed that our proposed model outperforms other cutting-edge cancer survival prediction methods significantly.

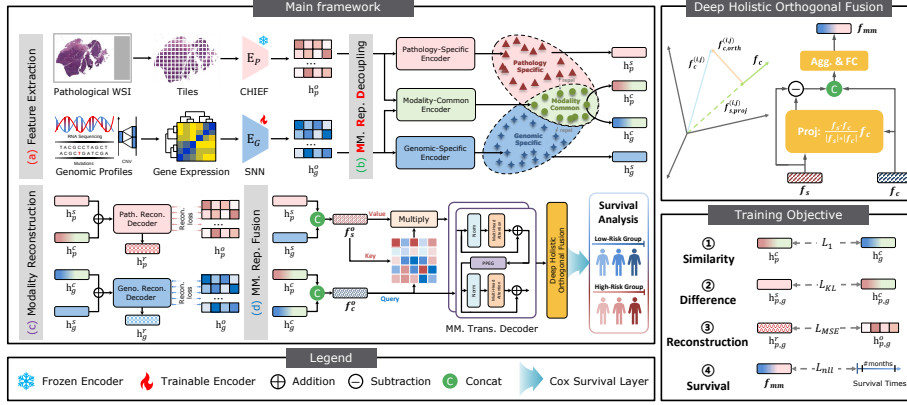


Fig. 1. Overview of the proposed MurreNet: (a) feature extraction for pathology and genomic features, (b) multimodal representation decoupling module, (c) modality reconstruction module, (d) multimodal representation fusion by deep holistic orthogonal fusion module and comprehensive training regularization strategy for cancer prognosis.

2 Methodology

2.1 Problem Formulation and Feature Extraction

Problem Formulation. In this multimodal survival analysis task, our objective is to learn a robust multimodal representation utilizing paired pathology P and genomic data G , and develop a survival prediction model to estimate the hazard function $f_{hazard}(t) = f_{hazard}(T = t | T \geq t, (P, G)) \in [0, 1]$, which represents the probability of death for a patient right after the discrete time point t rather than estimating the survival time directly. Then, an ordinal value can be obtained via integrating the negated hazard functions: $f_{surv}(t | (P, G)) = \prod_{u=1}^t (1 - f_{hazard}(u))$.

Pathological Feature Embedding. Due to the gigapixel nature of WSIs (e.g., $200,000 \times 100,000$ pixels), we adopt the CLAM [19] to automatically partition each WSI into a series of non-overlapping 256×256 patches at $20 \times$ magnification. Subsequently, the pathological foundation model, CHIEF [23], is employed in a frozen state to embed each patch into an instance-level feature vector. Finally, these patch-level features are converted into a 768-dimensional feature embedding h_p^o through a learnable fully-connected layer.

Genomic Feature Embedding. The genomic profiles consists of various 1×1 values, including gene mutation status, RNA Sequencing, DNA methylation, etc. However, the data exhibits a high-dimensional low-sample size characteristic. To mitigate this issue, we follow [5, 26] to group these high signal-to-noise ratio data into six genomic sequences (see 3.1). These grouped sequences are stacked then fed into a fully connected layer to obtain d -dimensional genomic embedding h_g^o .

2.2 Multimodal Representation Decoupling

To preform a comprehensive and disentangled perspective of multimodal representations, we develop three encoders within our MRD module. These encoders are designed to capture distinct types of knowledge: pathology-specific, genomic-specific, and modality-common information from pathological WSIs and genomic profiles. Specifically, we leverage MLP layers as the modality-specific encoders Γ_p and Γ_g to extract specific representations h_p^s and h_g^s , and a two-branch parallel MLP layer with a co-attention block as the modality-common encoder Γ_c , which bridges the multimodal interactions and combines input features to generate two modality-common representations h_p^c and h_g^c . The detailed operation of the modality-common encoder Γ_c can be formulated as:

$$h_p^c, h_g^c = \Gamma_c(h_p^o, h_g^o) = \text{MLP}(\mathbb{A}^\top) * h_p^o, \text{MLP}(\mathbb{A}) * h_g^o, \quad (1)$$

$$\mathbb{A} = \text{Linear}(h_p^o)^\top \cdot \text{Linear}(h_g^o), \quad (2)$$

where $\text{MLP}(\cdot)$ is composed of two fully-connected layers with layer normalization and GELU activation, $\mathbb{A}$ means the multimodal co-attention matrix, respectively. The $*$ is element-wise multiplication, while $\cdot$ denotes the matrix multiplication.

2.3 Multimodal Feature Integration

Multimodal Representation Learning. Inspired by previous work [21], we design a early-fusion based multimodal representation learning strategy to learn inter-modal interactions between histopathology and genomic profiles. Concretely, we first obtain the modality-specific representation f_s^o and -common representation f_c^o by concatenating their associated multimodal representations. Then, we develop a cross-attention module with two transformer decoders to integrate the modality-specific and -common representations into a joint embedding f_s' then mine their inner contextual information. This procedure can be formulated as:

$$f_s' = \text{CrossAttn}(f_c^o, f_s^o, f_s^o) = \text{Softmax}\left(\frac{\mathbf{W}_q \cdot f_c^o \cdot f_s^{o\top} \cdot \mathbf{W}_k^\top}{\sqrt{d_k}}\right) \mathbf{W}_v \cdot f_s^o \quad (3)$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{d_k \times d_k}$ are learnable weight matrices multiplied to the query f_c^o and key-value pair (f_c^o, f_s^o) , respectively. We then enhanced the resulting f_s' and f_c' by two transformer decoder $\mathcal{D}(\cdot)$ which can be formulated as:

$$\begin{aligned} f_s &= \mathcal{D}(f_s') = \text{MSA}\left(\text{LN}\left(f_s^{(2)}\right)\right) + f_s^{(2)}, \\ f_s^{(2)} &= \text{PPEG}\left(f_s^{(1)}\right), f_s^{(1)} = \text{MSA}\left(\text{LN}\left(f_s'\right)\right) + f_s', \end{aligned} \quad (4)$$

where MSA and LN are Multi-head Self-Attention and Layer Norm, and we use the Nystrom attention [30] for MSA to mitigate heavy computation burden. PPEG($\cdot$) is the Pyramid Position Encoding Generator in [20] to capture contextual information. Likewise, we can obtain enhanced modality-common representation f_c via $\mathcal{D}(\cdot)$.

Deep Holistic Orthogonal Fusion. We design a novel orthogonal fusion module to integrate modality-specific and -common representations, for modeling the holistic interactions across modalities. It takes f_s and f_c as inputs then calculates the projection $f_{s,proj}^{(i,j)}$ of each feature in specific representation onto the common representation f_c . Subsequently, we can compute the orthogonal component by:

$$f_{s,orth}^{(i,j)} = f_s^{(i,j)} - f_{s,proj}^{(i,j)}, f_{s,proj}^{(i,j)} = \frac{f_s^{(i,j)} \cdot f_c}{|f_c|^2} \cdot f_c = \frac{\sum_{k=1}^K f_{s,k}^{(i,j)} \cdot f_{c,k}}{\sum_{k=1}^K (f_{c,k})^2} \cdot f_c, \quad (5)$$

where $|f_c|^2$ represents the L_2 norm of f_c . Here, we then leverage the pooling functionality to aggregate the concatenated tensor $f_{s,orth}^{(i,j)}$ as illustrated in the upper-right section of Fig. 1. Lastly, a fully-connected layer is employed to produce the final multimodal representation f_{mm} for survival outcome prediction.

2.4 Training Objective

The overall learning of the model is performed by minimizing:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{sim} + \beta \mathcal{L}_{diff} + \gamma \mathcal{L}_{recon} + \mathcal{L}_{surv}, \quad (6)$$

Here, α , β , and γ are the positive hyper-parameters that determine the contribution of each regularization component to the overall training objective $\mathcal{L}_{total}$.

Similarity Loss. To reduce the discrepancy between the two modality-common representations and align them into the shared space, we introduce L_1 norm to measure the discrepancy between the distribution of common representations.

$$\mathcal{L}_{sim} = \mathcal{L}_1(h_p^o, h_p^c) = \frac{1}{n} \|h_p^o - h_p^c\|_1, \quad (7)$$

Difference Loss. We further employ a difference loss to ensure that the modality-common and -specific representations capture distinct aspects of the input. To this end, we introduce the Kullback-Leibler Divergence [12] for this purpose.

$$\mathcal{L}_{diff} = \mathcal{D}_{KL}(h_p^c, h_p^s) + \mathcal{D}_{KL}(h_g^c, h_g^s) = \sum_m^{\{p,g\}} \left[h_m^c \log \left(\frac{h_m^c}{h_m^s} \right) \right], \quad (8)$$

Reconstruction Loss. To prevent the learning of trivial representations, we first reconstruct the modality representations h_p^o and h_g^o using two MLP-based decoders to produce h_p^r and h_g^r two reconstructed representations. Then we apply mean squared error (MSE) loss as the modality reconstruction loss which ensures that the hidden representations capture the relevant details of each modality.

$$\mathcal{L}_{recon} = \frac{1}{2} (\mathcal{L}_{MSE}(h_p^o, h_p^r) + \mathcal{L}_{MSE}(h_g^o, h_g^r)) = \frac{1}{2} \sum_m^{\{p,g\}} (h_m^o - h_m^r)^2, \quad (9)$$

Survival Loss. Following previous works [5,15], we utilize NLL (negative log-likelihood) survival loss [27] as the $\mathcal{L}_{surv}$ for the survival prediction part.

3 Experiments and Results

3.1 Datasets

Following [30], we conducted extensive experiments on six cancer types from The Cancer Genome Atlas (TCGA) including Bladder Urothelial Carcinoma (BLCA), Breast Invasive Carcinoma (BRCA), Colon Adenocarcinoma & Rectum Adenocarcinoma (COADREAD), Lung Adenocarcinoma (LUAD), Uterine Corpus Endometrial Carcinoma (UCEC), and Stomach Adenocarcinoma (STAD). We utilized RNA sequencing (RNA-seq), Copy Number Variation (CNV), and mutation status, and further categorized them into six genomic sub-sequences (protein kinases, tumor suppressor genes, oncogenes, cell differentiation markers, transcription, and cytokines and growth) as the genomic profile input of models.

3.2 Implementation Details

Training settings. We select a diverse set of baseline methods, including those focused on pathological WSIs, genomic data, and both modalities. The methods for comparison include: MLP, SNN [11], Coxnet [7], ABMIL [8], DSMIL [13], CLAM [19], TransMIL [20], DTFD-MIL [28], M3IF [14], MCAT [5], CMTA [30], MoME [25], MOTCat [26], SurvPath [9], and PORPOISE [6]. Following standard practice, we performed 5-fold monte-carlo cross-validation on each dataset to mitigate batch effects. The mean Concordance index (C-index) with its standard deviation (std) are reported for a comprehensive performance comparison. Moreover, Kaplan-Meier [1] and T-test analyses are employed to assess the significance of differences in survival prediction between high- and low-risk groups.

Hyper-parameters. We trained all models with the Adam [10] optimizer for a total of 20 epochs. The learning rate and weight decay were set to 2×10^{-4} and 1×10^{-5} , respectively. During training of SurvPath [9], we randomly sample 4,096 patches per WSI to avoid memory overflow issues. We set the hyper-parameters α and β as 1×10^{-4} , γ as 1 for our training loss function in all cancer datasets.

3.3 Comparison with State-of-the-art Methods

We conducted extensive experiments compared with 15 bleeding-edge survival prediction methods using identical settings on six TCGA datasets. As shown in Table 1, MurreNet achieves superior performance on all cancer types. Notably, most multimodal methods steadily outperform unimodal methods, highlighting the crucial contribution of integrating information from both modalities. Against MCAT [5], the current SOTA multimodal method, our model achieves the performance increases of 4.87% on BLCA, 3.91% on BRCA, 11.71% on COADREAD,

Table 1. The performance of different approaches on six public TCGA datasets. “P.” indicates whether to use pathological images and “G.” indicates whether to use genomic profiles. The best and second best results are highlighted in **bold** and underlined.

Methods	Modality		BLCA	BRCA	COADREAD	LUAD	UCEC	STAD
	P.	G.	(n = 373)	(n = 956)	(n = 340)	(n = 453)	(n = 480)	(n = 349)
MLP		✓	0.611 ± 0.030	0.619 ± 0.053	0.675 ± 0.049	0.619 ± 0.044	0.675 ± 0.069	0.607 ± 0.045
SNN [11]		✓	0.625 ± 0.054	0.621 ± 0.051	0.606 ± 0.087	0.615 ± 0.044	0.703 ± 0.072	0.603 ± 0.052
Coxnet [7]		✓	0.591 ± 0.061	0.593 ± 0.067	0.601 ± 0.059	0.605 ± 0.044	0.535 ± 0.051	0.523 ± 0.055
ABMIL [8]	✓		0.640 ± 0.032	0.642 ± 0.030	0.703 ± 0.064	0.607 ± 0.041	0.701 ± 0.038	<u>0.638 ± 0.028</u>
DSMIL [13]	✓		0.636 ± 0.031	0.657 ± 0.037	0.684 ± 0.071	0.591 ± 0.026	0.729 ± 0.063	0.630 ± 0.032
CLAM [19]	✓		0.618 ± 0.042	0.593 ± 0.077	0.643 ± 0.108	0.580 ± 0.029	0.647 ± 0.037	0.634 ± 0.054
TransMIL [20]	✓		0.640 ± 0.039	0.590 ± 0.058	0.684 ± 0.024	0.592 ± 0.048	0.705 ± 0.027	0.604 ± 0.054
DTFD-MIL [28]	✓		0.610 ± 0.025	0.625 ± 0.060	0.633 ± 0.061	0.590 ± 0.033	0.663 ± 0.016	0.580 ± 0.033
M3IF [14]	✓	✓	0.662 ± 0.043	0.639 ± 0.062	<u>0.710 ± 0.043</u>	0.628 ± 0.050	0.724 ± 0.103	0.607 ± 0.044
MCAT [5]	✓	✓	0.677 ± 0.062	<u>0.691 ± 0.041</u>	0.649 ± 0.052	0.675 ± 0.039	0.658 ± 0.055	0.586 ± 0.027
CMTA [30]	✓	✓	0.683 ± 0.016	0.667 ± 0.033	0.678 ± 0.040	0.648 ± 0.036	0.740 ± 0.066	0.584 ± 0.023
MoME [25]	✓	✓	0.650 ± 0.033	0.659 ± 0.032	0.692 ± 0.038	0.650 ± 0.024	0.705 ± 0.041	0.630 ± 0.039
MOTCat [26]	✓	✓	<u>0.683 ± 0.025</u>	0.675 ± 0.021	0.618 ± 0.030	0.667 ± 0.027	0.685 ± 0.052	0.596 ± 0.028
SurvPath [9]	✓	✓	0.673 ± 0.011	0.685 ± 0.024	0.650 ± 0.024	<u>0.676 ± 0.036</u>	0.737 ± 0.049	0.622 ± 0.045
PORPOISE [6]	✓	✓	0.651 ± 0.046	0.624 ± 0.048	0.702 ± 0.045	0.642 ± 0.042	<u>0.737 ± 0.097</u>	0.612 ± 0.038
MurreNet	✓	✓	0.710 ± 0.030	0.718 ± 0.016	0.725 ± 0.044	0.691 ± 0.040	0.752 ± 0.075	0.651 ± 0.062

2.37% on LUAD, 14.29% on UCED, and 11.09% on STAD, respectively. This suggests that survival analysis using multimodal data should focus on the holistic interactions between pathology and genomic information, rather than merely exploiting the abundant shared correlations. Additionally, our model also outperforms CMTA [30], further emphasizing the importance of disentangling and reintegrating modality-specific and -common knowledge. Finally, our model consistently outperforms other SOTA multimodal methods by a large margin, including M3IF [14], MoME [25], MOTCat [26], SurvPath [9] and PORPOISE [6].

3.4 Ablation Study

To thoroughly validate the impact of each proposed component in our model, we perform an ablation study on six TCGA datasets by breaking down our model into several submodels. We begin with a multimodal survival baseline A, which simply concatenating pathology and genomic features for survival analysis. Next, we add the MRD module into model A as model B, followed by the incorporation of DHOF strategy as model C, and the addition of $\mathcal{L}_{sim}$ (D), $\mathcal{L}_{diff}$ (E), and $\mathcal{L}_{recon}$ training loss as our final model F. Quantitative results are presented in Table 2. By comparing models A and B, we observe that our MRD design significantly enhances modality-specific and -common representation learning, leading to performance improvements of 1.71% on BLCA, 2.07% on LUAD, 2.38% on UCEC, and 1.34% on STAD, respectively. The DHOF strategy in model C further improves the results, contributing to more accurate survival predictions. Additionally, the inclusion of the training strategies $\mathcal{L}_{sim}$, $\mathcal{L}_{diff}$, and $\mathcal{L}_{recon}$ continues to boost performance, as indicated

Table 2. Quantitative results for ablation study on four TCGA cancer datasets.

Model	Designs in our model					Datasets			
	MRD	DHOF	$\mathcal{L}_{sim}$	$\mathcal{L}_{diff}$	$\mathcal{L}_{recon}$	BLCA	LUAD	UCEC	STAD
A						0.645 ± 0.017	0.627 ± 0.053	0.673 ± 0.047	0.598 ± 0.026
B	✓					0.656 ± 0.031	0.640 ± 0.011	0.689 ± 0.010	0.606 ± 0.042
C	✓	✓				0.673 ± 0.024	0.652 ± 0.041	0.701 ± 0.052	0.617 ± 0.034
D	✓	✓	✓			0.695 ± 0.028	0.674 ± 0.039	0.733 ± 0.060	0.636 ± 0.038
E	✓	✓	✓	✓		0.701 ± 0.030	0.680 ± 0.042	0.746 ± 0.063	0.644 ± 0.040
F	✓	✓	✓	✓	✓	0.710 ± 0.030	0.691 ± 0.040	0.752 ± 0.075	0.651 ± 0.062

by the increase in C-index from model D to F. This suggests that the learning constraints on modality similarity, difference, and reconstruction play a crucial role in facilitating holistic multimodal interaction modeling for cancer prognosis.

3.5 Risk Stratification

To further substantiate the efficacy of our model for survival analysis, we stratify all patients into low- and high-risk groups on the median of the predicted risk scores generated by our model across six TCGA cohorts. Subsequently, we utilize Kaplan-Meier [1] survival curve to visually depict the survival events of all the patients. To evaluate the statistical significance of the differences between these groups, we apply the Log-rank test [2], where a p-value of 0.05 or below is deemed statistically significant. As shown in Fig. 2, the p-values for our model across six cohorts are consistently well below 0.05, underscoring the robustness of MurreNet in effective risk stratification and reinforcing its clinical potential.

4 Conclusion

In this paper, we propose the Multimodal Representation Decoupling framework (MurreNet) to integrate genomics and pathological WSIs for survival analysis. We innovatively propose MRD module which factorizes modalities into modality-common and modality-specific information to reduce the modality redundancy. To enhance learning constraints, we incorporate a combination of losses for modality similarity, difference, reconstruction and survival prediction task. Finally, we design the DHOF strategy to enable multimodal feature fusion and model the intricate inter- and intra-modality correlations. Extensive experiments on six TCGA datasets indicate effectiveness of our proposed MurreNet.

Acknowledgments. This work is supported by the National Key R&D Program of China (No.2023YFC3402800), National Natural Science Foundation of China (Nos.82441029, 62171230, 62101365, 92159301, 62301263, 62301265, 62302228, 82302291, 82302352, 62401272), Jiangsu Provincial Department of Science and Technology’s major project on frontier-leading basic research in technology (No.BK2023200).

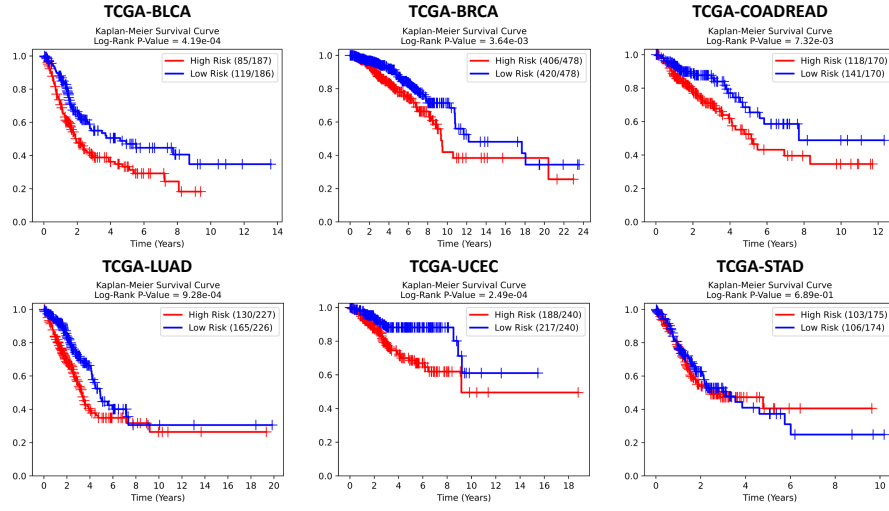


Fig. 2. Kaplan-Meier survival curves of the proposed model on six cancer datasets.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bland, J.M., Altman, D.G.: Survival probabilities (the kaplan-meier method). *Bmj* **317**(7172), 1572–1580 (1998)
2. Bland, J.M., Altman, D.G.: The logrank test. *Bmj* **328**(7447), 1073 (2004)
3. Cai, C., Li, J., Liu, M., Jiao, Y., Xu, J.: Seqfrt: Towards effective adaption of foundation model via sequence feature reconstruction in computational pathology. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 1808–1815. IEEE (2024)
4. Cai, C., Shi, Q., Li, J., Jiao, Y., Xu, A., Zhou, Y., Wang, X., Peng, C., Zhang, X., Cui, X., et al.: Pathologist-level diagnosis of ulcerative colitis inflammatory activity level using an automated histological grading method. *International Journal of Medical Informatics* **192**, 105648 (2024)
5. Chen, R.J., Lu, M.Y., Weng, W.H., Chen, T.Y., Williamson, D.F., Manz, T., Shady, M., Mahmood, F.: Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4025 (2021)
6. Chen, R.J., Lu, M.Y., Williamson, D.F., Chen, T.Y., Lipkova, J., Noor, Z., Shaban, M., Shady, M., Williams, M., Joo, B., et al.: Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell* **40**(8), 865–878 (2022)
7. Ching, T., Zhu, X., Garmire, L.X.: Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS computational biology* **14**(4), e1006076 (2018)
8. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)

9. Jaume, G., Vaidya, A., Chen, R.J., Williamson, D.F., Liang, P.P., Mahmood, F.: Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 11579–11590 (2024)
10. Kingma, D.P.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)
11. Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S.: Self-normalizing neural networks. *Advances in neural information processing systems* **30** (2017)
12. Kullback, S., Leibler, R.A.: On information and sufficiency. *The annals of mathematical statistics* **22**(1), 79–86 (1951)
13. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 14318–14328 (2021)
14. Li, H., Yang, F., Xing, X., Zhao, Y., Zhang, J., Liu, Y., Han, M., Huang, J., Wang, L., Yao, J.: Multi-modal multi-instance learning using weakly correlated histopathological images and tabular clinical information. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*. pp. 529–539. Springer (2021)
15. Liu, M., Liu, Y., Cui, H., Li, C., Ma, J.: Mgct: Mutual-guided cross-modality transformer for survival outcome prediction using integrative histopathology-genomic features. In: *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. pp. 1306–1312. IEEE (2023)
16. Liu, M., Liu, Y., Xu, P., Cui, H., Ke, J., Ma, J.: Exploiting geometric features via hierarchical graph pyramid transformer for cancer diagnosis using histopathological images. *IEEE Transactions on Medical Imaging* (2024)
17. Liu, M., Liu, Y., Xu, P., Ma, J.: Unleashing the infinity power of geometry: A novel geometry-aware transformer (goat) for whole slide histopathology image analysis. In: *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. pp. 1–5. IEEE (2024)
18. Long, L., Cui, J., Zeng, P., Li, Y., Liu, Y., Wang, Y.: Mugi: Multi-granularity interactions of heterogeneous biomedical data for survival prediction. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 490–500. Springer (2024)
19. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering* **5**(6), 555–570 (2021)
20. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems* **34**, 2136–2147 (2021)
21. Song, A.H., Chen, R.J., Jaume, G., Vaidya, A.J., Baras, A., Mahmood, F.: Multimodal prototyping for cancer survival prediction. In: *Proceedings of the 41st International Conference on Machine Learning*. vol. 235, pp. 46050–46073. PMLR (2024)
22. Song, A.H., Jaume, G., Williamson, D.F., Lu, M.Y., Vaidya, A., Miller, T.R., Mahmood, F.: Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering* **1**(12), 930–949 (2023)
23. Wang, X., Zhao, J., Marostica, E., Yuan, W., Jin, J., Zhang, J., Li, R., Tang, H., Wang, K., Li, Y., et al.: A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature* **634**(8035), 970–978 (2024)

24. Wang, Z., Zhang, Y., Xu, Y., Imoto, S., Chen, H., Song, J.: Histo-genomic knowledge association for cancer prognosis from histopathology whole slide images. *IEEE Transactions on Medical Imaging* (2025)
25. Xiong, C., Chen, H., Zheng, H., Wei, D., Zheng, Y., Sung, J.J., King, I.: Mome: Mixture of multimodal experts for cancer survival prediction. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 318–328. Springer (2024)
26. Xu, Y., Chen, H.: Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 21241–21251 (2023)
27. Zadeh, S.G., Schmid, M.: Bias in cross-entropy-based training of deep survival networks. *IEEE transactions on pattern analysis and machine intelligence* **43**(9), 3126–3137 (2020)
28. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y.: Dtf-d-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 18802–18812 (2022)
29. Zhang, Y., Xu, Y., Chen, J., Xie, F., Chen, H.: Prototypical information bottlenecking and disentangling for multimodal cancer survival prediction. *arXiv preprint arXiv:2401.01646* (2024)
30. Zhou, F., Chen, H.: Cross-modal translation and alignment for survival analysis. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 21485–21494 (2023)
31. Zhou, H., Zhou, F., Chen, H.: Cohort-individual cooperative learning for multimodal cancer survival analysis. *IEEE Transactions on Medical Imaging* (2024)