MuGI: Multi-Granularity Interactions of Heterogeneous Biomedical Data for Survival Prediction

Lifan Long[1], Jiaqi Cui[1], Pinxian Zeng[1], Yilun Li[1], Yuanjun Liu[2] and Yan Wang[1, ✉]

[1] School of Computer Science, Sichuan University, China
wangyanscu@hotmail.com
[2] Department of Hepatobiliary Surgery, Suining Central Hospital, China

Abstract. Multimodal learning significantly benefits survival analysis for cancer, particularly through the integration of pathological images and genomic data. However, this presents new challenges on how to effectively integrate multi-modal biomedical data. Existing multi-modal survival prediction methods focus on mining the consistency or modality-specific information, failing to capture cross-modal interactions. To address this limitation, attention-based methods are proposed to enhance both the consistency and interactions. However, these methods inevitably introduce redundancy due to the overlapped information of multi-modal data. In this paper, we propose a **Mu**lti-**G**ranularity **I**nteractions of heterogeneous biomedical data framework (MuGI) for precise survival prediction. MuGI consists of: a) unimodal extractor for exploring preliminary modality-specific information, b) multimodal optimal features capture (MOFC) for extracting ideal multi-modal representations, eliminating redundancy through decomposed multi-granularity information, as well as capturing consistency in a common space and enhancing modality-specific features in a private space, and c) multi-modal hierarchical interaction for sufficient acquisition of cross-modal correlations and interactions through the cooperation of two Bilateral Cross Attention (BCA) modules. We conduct extensive experiments on three cancer cohorts from the Cancer Genome Atlas (TCGA) database. The experimental results demonstrate that our MuGI achieves the state-of-the-art performance, outperforming both unimodal and multi-modal survival prediction methods.

Keywords: Survival Prediction, Multimodal Learning, Heterogeneous Biomedical Data, Transformer.

1 Introduction

Cancer survival analysis aims to predict the relative risk of death in prognosis. Precision survival prediction is of great significance for cancer prevention and treatment.

In clinical settings, a substantial quantity of biomarkers has been developed and employed in cancer survival analysis, including imaging and genomic biomarkers. Of all these types of biomarkers, histopathology images, also known as whole slide images (WSIs), are generally considered to be the gold standard for cancer survival prediction since they can provide morphological attributes of cells that are highly related to the degree of cancer aggressiveness [1]. Numerous unimodal survival prediction methods [1–4] based on pathological images have been proposed to automate survival prediction for cancer. Although these unimodal survival prediction methods have demonstrated promising results, they fail to utilize the complementary information included in the multi-modal biomedical data. For example, genomic data which provides profound molecular characteristics of tumors has the potential to facilitate precision survival prediction when integrated with histopathology images [5, 6]. Along the multimodal survival prediction research direction, the core problem lies in how to effectively integrate the heterogeneous pathological images and genomic profiles.

The multimodal feature integration strategies in current multimodal survival prediction methods can be roughly categorized into intra-modal representation and inter-modal fusion. Intra-modal representation primarily extracts complementarity and consistency of multiple modalities. Existing approaches often project each modality into a common subspace [5, 7, 8], striving to eliminate redundancy. However, they may overlook that diverse modalities unveil distinctive characteristic of survival outcomes from various perspectives, leading to suboptimal results. For inter-modal fusion, some tensor-based methods use simple operation such as concatenation [9, 10], or taking the Kroncecker product [11, 12]. Moreover, these approaches are typically employed in either early or late fusion stages to integrate multiple modalities. However, early fusion techniques may overlook the tremendous gap between multiple heterogenous modalities, whereas late fusion techniques hinder the mining of potential multimodal correlations and interactions [13]. Therefore, intermediate fusion has been proposed to overcome this trade-off. This strategy can capture cross-modality information while leveraging the inherent information of each modality. Among these intermediate fusion approaches, cross-attention-based algorithms [14, 15] rescale features through complementary information from another modality to enhance correlations and interactions. However, due to the duplicated information in multimodal data, it may lead to redundancy. Decoupling information into the modality-common information and modality-specific information [16], on the other hand, can improve the effective utilization of features by minimizing redundant information. Furthermore, to extract valuable insights from multimodal redundancy, alignment-based methods [17, 18] focus on latent cross-modal adaptation, integrating modality-common information, and emphasizing the inherent consistency. However, alignment and integration of multimodal information frequently prioritize modality-common features, potentially overlooking modality-specific information, thus neglecting the richness of distinctive perspectives.

To address the above limitations and encourage the effective exploitation of multimodal data, in this paper, we propose a Multi-Granularity Interactions of heterogeneous biomedical data (MuGI) framework for survival prediction, which bridges the heterogeneity gap between pathological and genomic data, thereby sufficiently integrating the complementary information from each individual modality while removing redundancy. Specifically, our MuGI contains three stages, including unimodal extractor, multimodal optimal features capture (MOFC) and multimodal hierarchical interaction. The unimodal extractor obtains unimodal feature representation (i.e., modality-specific information) through two transformer encoders. Then, in MOFC, supported by the adversarial learning, we construct multi-granularity spaces to decouple multimodal information, i.e., a private space to enhance the modality-specific information and a shared latent space to mine the modality-common information. Subsequently, multimodal hierarchical interaction narrows the heterogeneity gap through two Bilateral Cross Attention (BCA) modules to achieve multi-granularity interactions. The main contributions of this paper can be summarized as follows:

- We present a novel multimodal survival prediction method, which effectively integrates complementary information from genomic and pathological data, preserving complementarity while eliminating redundancy, with the aim of enhancing the representation capability for precise survival prediction.
- A multimodal optimal features capture (MOFC) stage is designed to decouple information from pathological images and genomic profiles into modality-common information and enhanced modality-specific information, thereby minimizing redundancy.
- A multimodal hierarchical interaction stage is devised to obtain the final fusion features. In this stage, two Bilateral Cross Attention (BCA) modules are introduced to explore correlations and interactions between multi-granularity features.
- We conduct extensive experiments on three public TCGA datasets. The results demonstrate that our method consistently outperforms current state-of-the-art methods.

2 Methodology

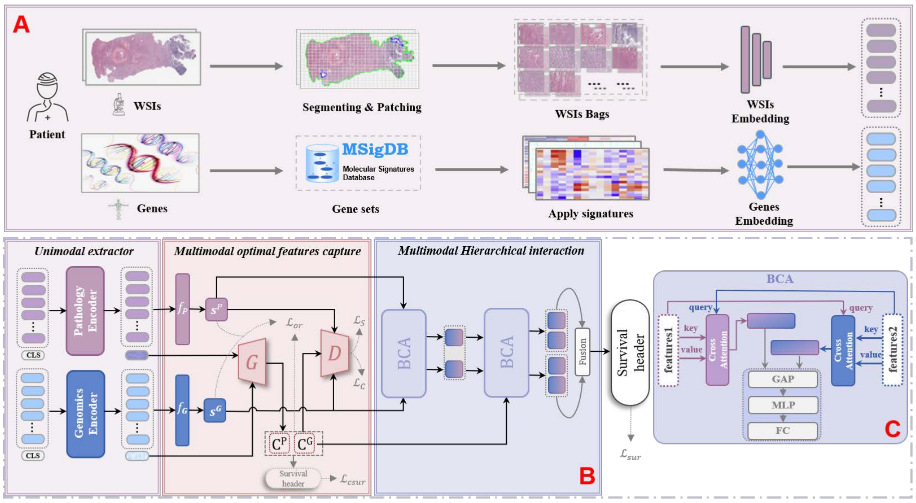

Fig. 1. Illustration of the proposed method, which includes: A) flowchart for pathology-genomic embedding; B) the three stages of our MuGI framework; and C) the structure of Bilateral Cross Attention (BCA).

The overview our proposed MuGI framework is illustrated in Fig. 1. Specifically, Fig. 1A illustrates the flowchart for pathology-genomic embedding process. Fig. 1B presents the three stages of our MuGI framework, while Fig. 1C depicts a more detailed diagram of the BCA module.

2.1 Preliminaries

In this multimodal survival prediction task, our objective is learning a robust multimodal representation by utilizing paired pathology data P and genomic data G, and developing a survival model $F(\cdot)$ to estimate the hazard function $h(\cdot)$. Let $X = \{x_1, x_2, \cdots, x_N\}$ be all patients samples on the dataset, each patient data can be represented as a quadruple $x_i = \{P_i, G_i, c_i, t_i\}_{i=1}^{N}$, where N indicates the number of patients, $c_i \in \{0,1\}$ is right uncensorship status, and $t_i \in \mathbb{R}^{+}$ is either survival time or censored time. Notably, a patient may have multiple WSIs. In mathematical terms, $h(\cdot)$ can be defined as:

$$h(t_i|P_i, G_i) = F(P_i, G_i; \theta) \tag{1}$$

where θ denotes the set of learnable parameters. Instead of predicting t_i directly, the survival model aims to estimate the ordinal risk of an event (death occurrence) via the survival function, which can be derived by the following cumulative hazard function:

$$S(t_i|P_i, G_i) = \prod_{u=1}^{t}(1 - h(u_i|P_i, G_i)) \tag{2}$$

2.2 Pathology-Genomic Embedding

For pathology, we construct each WSI as a “bag” structure within a conventional multiple instance learning (MIL) setting. Similarly, genomic data is structured as a “bag” taking into account the biological functional impact of genes. The flowchart for data processing is illustrated in Fig. 1A.

Histology Feature Embedding. Since a WSI is of large size (e.g., 100,000-by-100,000 pixels), we adopt CLAM [19] to automatically cut out each WSI to no-overlapping patches (constitute a bag). Then, a pre-trained CNN is used to embed each patch in an instance-level feature vector. Finally, each instance is converted into a d-dimensional embedding $p_k \in \mathbb{R}^{d}$ through a fully-connected layer. Hence, the histology images of each patient can be represented as $P_i = \{p_{1j}, p_{2j}, \cdots, p_{kj}\}_{j=1}^{m} \in \mathbb{R}^{m\times k\times d}$, where m is the number of WSIs and k denotes the number of patches.

Genomic Feature Embedding. For genomic features such as RNA-Seq, copy number variation (CNV), and gene mutation status are typically denoted as 1×1 attributes. The data exhibits a high-dimensional low-sample size characteristic. To address this issue, following [18], we using six functional categories obtained from [20], and leveraging SNN [21] to obtain d-dimension genomic feature embeddings. Thus, each patient genomic bag can be constructed as $G_i = \{g_1, g_2, \cdots, g_k\} \in \mathbb{R}^{k\times d}$, where k = 6.

2.3 Unimodal Extractor and Multimodal Optimal Features Capture

Unimodal Extractor. To preserve the inherent structural information of single modality, we use dual transformer encoders (i.e., a pathology encoder and a gene encoder) to enrich the features in both modalities. Due to the significant difference in bag count, a class token (CLS) is utilized to attain information balance between genes and WSIs in the shared feature space, ensuring equilibrium between common and specific features during subsequent BCA module. Also, we utilize the Nystrom attention [21] in the two encoders to tackle the bags with extremely large size. Similar to [14], we adopted Pyramid Position Encoding Generator (PPEG) module [22] in the pathology encoder. Therefore, we can obtain inherent gene and pathology representation, denoted as $P_i^u = \{p^u, p_{1j}^u, p_{2j}^u, \cdots, p_{kj}^u\}_{j=1}^m \in \mathbb{R}^{(1+m\times k)\times d}$ and $G_i^u = \{g^u, g_1^u, g_2^u, \cdots, g_k^u\} \in \mathbb{R}^{(1+k)\times d}$, where p^u and g^u are class tokens.

Multimodal Optimal Features Capture. Due to the overlapped information in multimodal data [14, 15], integrating information from diverse modalities in a simple intermediate inter-modal fusion manner inevitably results in redundancy. To mitigate such redundancy and obtain more complementary features, we design a Multimodal Optimal Features Capture (MOFC) module for the subsequent multimodal feature fusion. It decouples genes and pathology into modality-common information and modality-specific information to represent consistency and the specificity of heterogeneous modalities, respectively. In this case, inter-modal representations can be characterized in two disjoint parts, thus reducing the redundancy between the modalities. Specifically, a generative adversarial network [23] is employed. *On the one hand*, we develop a generator $G(\cdot)$ to capture modality-common information, which projects gene and histology global features (CLS) into a common latent subspace, aiming to maintain the consistency. *On the other hand*, two fully connected deep neural networks $f_P(\cdot)$ and $f_G(\cdot)$ are developed to capture their respective modality-specific information (enhanced specificity). The common representation $C^{\{P,G\}} \in \mathbb{R}^d$ and the specific representation $S^{\{P,G\}} \in \mathbb{R}^{m\times k\times d}$ can be formulated as follows:

$$\begin{aligned} C^P &= G(p^u), & C^G &= G(g^u) \\ S^p &= f_P\left(p_{1j}^u, p_{2j}^u, \cdots, p_{kj}^u\right), & S^G &= f_G(g_1^u, g_2^u, \cdots, g_k^u) \end{aligned} \tag{3}$$

Meanwhile, we design a discriminator $D(\cdot)$ during training to estimate which modality (modality-common or -specific) the representation comes from, providing feedback to the generator. The above process can be described as follows:

$$\begin{aligned} D\left(C^{\{P,G\}}\right) &= Sigmod\left(C^{\{P,G\}}\boldsymbol{W}^C + \boldsymbol{b}^C\right), \\ D\left(S^{\{P,G\}}\right) &= Sigmod(S^{\{P,G\}}\boldsymbol{W}^S + \boldsymbol{b}^S) \end{aligned} \tag{4}$$

where $\boldsymbol{W}^C \in \mathbb{R}^d$ and $\boldsymbol{W}^S \in \mathbb{R}^{d\times m}$ are the weight matrices; $b^C \in \mathbb{R}^d$ and $b^S \in \mathbb{R}^{d\times m}$ are the bias matrices.

2.4 Multimodal Hierarchical interaction

Through the above modules, the common and enhanced specific representations of modalities contain consistency and complementarity, as well as minimizing redundancy.

However, there remains a deficiency in the interactions between modalities. To capture interactions between genes and tumor microenvironment in WSIs, we propose two BCA modules, as shown in Fig. 1C.

BCA adopts two cross-attention mechanisms to comprehensively integrate multi-modal features, exploring cross-modal correlations and interactions. An optional block, consisting a Gated Attention Pooling (GAP) [24], a Multi-Layer Perceptron (MLP), and a fully connected layer, are followed to adaptively refine the integrated features. Specifically, for cross-attention, Q (query) is obtained from one modality (features1) while K (key) and V (value) are derived from another modality (features2). Then, the two integrated features, each guided by features1 and features2 respectively, are fed into the into the subsequent optional block. The first BCA integrates gene specific representation S^P and pathology specific representation S^G, obtaining an integrated specific representation S, and the second integrates concatenated common representation $C^{\{P,G\}}$ and S, deriving the final promising representation $\mathcal{H}$. Finally, $\mathcal{H}$ fed into the survival header for survival prediction. In this manner, hierarchical interactions are achieved through the close collaboration of the two BCAs.

2.5 Loss Function

The final objective function contains five parts, as follows. (1) The overall survival loss $\mathcal{L}_{sur}$ is calculated based on the risk values output by final survival header, that adopts the Negative Log-Likelihood (NLL) function [18]. (2) The common NLL loss $\mathcal{L}_{csur}$ is calculated based on the risk values output by an additional survival header (a linear layer), the header accepts the average of C^P and C^G, exhibiting the same semantics between common and private information. (3) The common adversarial loss $\mathcal{L}_C$ and (4) the specific adversarial loss $\mathcal{L}_S$. We employ ground truth modality labels represented by one-hot encoding and use Binary Cross Entropy Loss to compute them in adversarial learning for modality binary classification (WSI or gene), ensuring the purity of common and specific features. (5) To further eliminant redundancy, we penalize redundancy in $C^{\{P,G\}}$ and $S^{\{P,G\}}$ with orthogonal loss as follows:

$$\mathcal{L}_{or} = -\sum_{o\in\{P,G\}}\sum_{i=1}^{N} \| (S_i^o)^T(C_i^o) \|_2 \tag{5}$$

where $\|\cdot\|_2$ is the Euclidean norm.

Above all, the final objective function is computed as:

$$\mathcal{L}_{final} = \mathcal{L}_{sur} + \gamma\mathcal{L}_{csur} + \alpha_1\mathcal{L}_C + \alpha_2\mathcal{L}_S + \beta\mathcal{L}_{or} \tag{6}$$

where α_1, α_2, β, and γ are the trade-off parameters.

3 Experiment

3.1 Datasets and Evaluation Metrics

To validate the performance of the proposed MuGI, we apply our method on three cancer survival datasets from TCGA database, including Uterine Corpus Endometrial Carcinoma (UCEC) (n = 480), Lung Adenocarcinoma (LUAD) (n = 452) and Bladder Urothelial Carcinoma (BLCA) (n = 373). Each dataset comprises diagnostic WSIs and genomic data with labeled survival times and censorship statuses. We evaluate the model performance by the concordance index (C-Index). It evaluates a model's capacity to accurately rank pairs of individuals based on their predicted survival times.

3.2 Implementation details

For each experiment, we perform 5-fold cross-validation with a training-validation split ratio of 4:1. The number of survival bins is set as 4. For a fair comparison during training, we follow the setting in [14] and accordingly utilize the Adam optimizer with an initial learning rate of 2e-4. In our experiment, the batch size set as 1, and the model trained for 20 epochs. We set α_1 as 0.1, α_2 as 0.01 and γ as 5e-6 in all datasets, β as 0.7 in BLCA and 0.5 in the other two datasets. Experiments are conducted on a single NVIDIA GeForce RTX 4060Ti GPU with 16 GB memory.

3.3 Experimental Results

In all experiments, we compare MuGI with state-of-the-art unimodal models (leveraging only gene or WSI) and multimodal models (leveraging the both data). Table 1 shows the experimental results across three datasets. Besides, we stratify patients into low (blue) and high-risk (orange) groups based on predicted risk scores to validate the stratification capability of MuGI. The Kaplan-Meier is adopted to visualize survival events, and Logrank test is employed to assesses statistical significance (A p-value ≤ 0.05 is generally considered statistically significant). The results are demonstrated in Fig. 2.

Table 1. The performance of different SOTA survival analysis methods on the three datasets. The results are presented as "averages ± standard" deviations from 5-fold cross-validation. The methods marked * are re-implemented. G: use gene, W: use WSI, and M: use gene and WSI.

	Methods	UCEC	LUAD	BLCA	Overall
G	SNN* [21]	0.627 ± 0.031	0.553 ± 0.051	0.587 ± 0.019	0.589 ± 0.034
	DeepSurv [25]	0.634 ± 0.049	0.608 ± 0.026	0.567 ± 0.049	0.603 ± 0.041
W	TransMIL* [22]	0.632 ± 0.051	0.623 ± 0.031	0.590 ± 0.037	0.615 ± 0.040
	CLAM* [19]	0.651 ± 0.031	0.595 ± 0.033	0.587 ± 0.033	0.611 ± 0.032
M	MCAT* [18]	0.598 ± 0.030	0.616 ± 0.027	0.630 ± 0.025	0.615 ± 0.027
	MGCT [15]	0.645 ± 0.039	0.596 ± 0.078	0.640 ± 0.039	0.627 ± 0.052
	CMTA* [14]	0.665 ± 0.053	0.675 ± 0.047	0.678 ± 0.012	0.672 ± 0.037
	MuGI(ours)	**0.682 ± 0.057**	**0.684 ± 0.045**	**0.681 ± 0.056**	**0.682 ± 0.052**

Comparison with Unimodal Models: As shown in Table 1, MuGI achieved the highest performance on all TCGA datasets compared with all unimodal methods. Concretely, our model obtains C-Index of 68.2% on UCEC, 68.4% on LUAD, and 68.1% on BLCA, improving the second-best unimodal methods by 3.1%, 6.1% and 9.1% respectively. Compared to the best gene model DeepSurv [25], the performance increases by 7.9%. Compared to the best pathology model TransMIL [22] , the overall C-Index performance increasing 6.7%.

Comparison with Multimodal Models: MuGI outperforms all multimodal methods, with an overall C-Index performance increase ranging from 1% to 6.7%. Against CMTA which uses the same encoder and directly utilizes cross-attention between the two embeddings, our model achieves superior performance. This indicates that decoupling information into common and specific information aids in survival prediction to some extent. The superior prediction accuracy achieved by our method demonstrates the effectiveness of integrating multimodal information.

Table 2. Ablation results of our method. Specific: $f(\cdot)$, GAN: $G(\cdot)$ and $D(\cdot)$.

Modules	LUAD	BLCA
w/o $\mathcal{L}_{or}$	0.672 ± 0.024	0.657 ± 0.044
w/o Specific	0.632 ± 0.044	0.592 ± 0.030
w/o GAN	0.664 ± 0.047	0.666 ± 0.046
w/o BCA	0.631 ± 0.036	0.665 ± 0.052
MuGI (all)	**0.684 ± 0.045**	**0.681 ± 0.056**

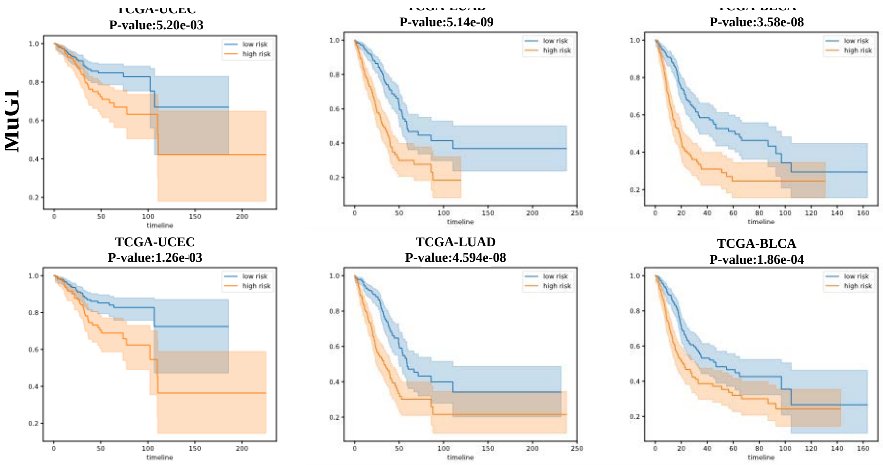

Fig. 2. The Kaplan-Meier curves on the three datasets of MuGI and CMTA.

Ablation Studies. We conducted ablation studies on two datasets, i.e., LUAD and BLCA, to validate the effectiveness of the proposed modules. We construct the ablation variants as follows. (1) w/o $\mathcal{L}_{or}$: To verify the redundancy elimination ability of $\mathcal{L}_{or}$; (2) w/o Specific: Removing the two FC layers of MOFC to verify its ability to enhance the specificity of the layers; (3) w/o GAN: To confirm whether both consistent and reduced redundant information improve the performance; (4) w/o BCA: To verify multi-modalities, multi-granularity integration and interaction ability of BCAs, including i) the integration and interaction between specific genetic and specific pathological features; ii) the integration and interaction between common and specific features. The results, presented in Table 2, demonstrate the effectiveness of each module. As can been seen, (a) according to the results of the second ablation variant, confirming that decoupling of information is prominently effective for survival prediction; (b) modality-specific information is most important for survival analysis. The removal of this module (i.e., w/o Specific) results in a 5.2% performance drop on LUAD, and an 8.9% performance drop on BLCA.

3.4 Conclusion

We propose MuGI for survival analysis using WSIs and genomic profiles. MuGI adopts a unimodal extractor to explore inherent information of each modality. Then, supported by adversarial learning, the MOFC decompose the extracted information into multi-granularity details, including modality-common and modality-specific information. Finally, by collaborating with the BCA module, heterogeneous information is effectively integrated for precise survival prediction. Extensive experiments on three TCGA cancer datasets indicate the effectiveness of our MuGI.

Acknowledgement. This work is supported by National Natural Science Foundation of China (NSFC 62371325, 62071314), Sichuan Science and Technology Program 2023YFG0025, 2023YFG0101, and 2023 Science and Technology Project of Sichuan Health Commission 23LCYJ002.

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

References

1. Shao, W., Wang, T., Huang, Z., et al.: Weakly Supervised Deep Ordinal Cox Model for Survival Prediction From Whole-Slide Pathological Images. IEEE Trans. Med. Imaging. 40, 3739–3747 (2021).
2. Chen, R.J., Lu, M.Y., Shaban, M., et al.: Whole Slide Images are 2D Point Clouds: Context-Aware Survival Prediction Using Patch-Based Graph Convolutional Networks. In: De Bruijne, M., Cattin, P.C., Cotin, S., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. pp. 339–349. Springer International Publishing, Cham (2021).
3. Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., Huang, J.: Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. Medical Image Analysis. 65, 101789 (2020).
4. Wang, Z., Li, J., Pan, Z., et al.: Hierarchical Graph Pathomic Network for Progression Free Survival Prediction. In: De Bruijne, M., Cattin, P.C., Cotin, S., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. pp. 227–237. Springer International Publishing, Cham (2021).
5. Ding, K., Zhou, M., Metaxas, D.N., Zhang, S.: Pathology-and-Genomics Multimodal Transformer for Survival Outcome Prediction. In: Greenspan, H., Madabhushi, A., Mousavi, P., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 622–631. Springer Nature Switzerland, Cham (2023).

6. Xing, X., Chen, Z., Zhu, M., et al.: Discrepancy and Gradient-Guided Multi-modal Knowledge Distillation for Pathological Glioma Grading. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., and Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. pp. 636–646. Springer Nature Switzerland, Cham (2022).
7. Ning, Z., Lin, Z., Xiao, Q., et al.: Multi-Constraint Latent Representation Learning for Prognosis Analysis Using Multi-Modal Data. IEEE Trans. Neural Netw. Learning Syst. 34, 3737–3750 (2023).
8. Ning, Z., Du, D., Tu, C., Feng, Q., Zhang, Y.: Relation-Aware Shared Representation Learning for Cancer Prognosis Analysis With Auxiliary Clinical Variables and Incomplete Multi-Modality Data. IEEE Trans. Med. Imaging. 41, 186–198 (2022).
9. Mobadersany, P., Yousefi, S., Amgad, M., et al.: Predicting cancer outcomes from histology and genomics using convolutional networks. Proc. Natl. Acad. Sci. U.S.A. 115, (2018).
10. Zheng, H., Lin, Z., Zhou, Q., et al.: Multi-transSP: Multimodal Transformer for Survival Prediction of Nasopharyngeal Carcinoma Patients. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., and Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. pp. 234–243. Springer Nature Switzerland, Cham (2022).
11. Chen, R.J., Lu, M.Y., Wang, J., et al.: Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis. IEEE Trans. Med. Imaging. 41, 757–770 (2022).
12. Chen, R.J., Lu, M.Y., Williamson, D.F.K., et al.: Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell. 40, 865-878.e6 (2022).
13. Liang, P.P., Zadeh, A., Morency, L.-P.: Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions. Presented at the (2022).
14. Zhou, F., Chen, H.: Cross-Modal Translation and Alignment for Survival Analysis. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 21428–21437. IEEE, Paris, France (2023).
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, Istanbul, Turkiye (2023).
16. Liang, P.P., Cheng, Y., Fan, X., et al.: Quantifying & Modeling Feature Interactions: An Information Decomposition Framework. ArXiv. abs/2302.12247, (2023).
17. Xu, Y., Chen, H.: Multimodal Optimal Transport-based Co-Attention Transformer with Global Structure Consistency for Survival Prediction. 2023 IEEE/CVF International Conference on Computer Vision (ICCV). 21184–21194 (2023).
18. Chen, R.J., Lu, M.Y., Weng, W.-H., et al.: Multimodal Co-Attention Transformer for Survival Prediction in Gigapixel Whole Slide Images. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3995–4005. IEEE, Montreal, QC, Canada (2021).
19. Lu, M.Y., Williamson, D.F.K., Chen, T.Y., et al.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng. 5, 555–570 (2021).
20. Liberzon, A., Birger, C., Thorvaldsdóttir, H., et al.: The Molecular Signatures Database Hallmark Gene Set Collection. Cell Systems. 1, 417–425 (2015).
21. Xiong, Y., Zeng, Z., Chakraborty, R., et al.: Nyströmformer: A nyström-based algorithm for approximating self-attention. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 14138–14148 (2021).

22. Shao, Z., Bian, H., Chen, Y., et al.: TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classication. In: Neural Information Processing Systems (2021).
23. Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al.: Generative Adversarial Nets. In: Neural Information Processing Systems (2014).
24. Ilse, M., Tomczak, J.M., Welling, M.: Attention-based Deep Multiple Instance Learning. In: International Conference on Machine Learning (2018).
25. Katzman, J., Shaham, U., Cloninger, A., et al.: DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Medical Research Methodology. 18, (2016).