Multi-modal Data Binding for Survival Analysis Modeling with Incomplete Data and Annotations

Linhao Qu*[1], Dan Huang*[2], Shaoting Zhang[1], and Xiaosong Wang(✉)[1]

[1] Shanghai Artificial Intelligence Laboratory, Shanghai, China
[2] Department of Pathology, Fudan University Shanghai Cancer Center, Shanghai, P.R. China. 270 Dong An Road, Shanghai 200032, China.
wangxiaosong@pjlab.org.cn

Abstract. Survival analysis stands as a pivotal process in cancer treatment research, crucial for predicting patient survival rates accurately. Recent advancements in data collection techniques have paved the way for enhancing survival predictions by integrating information from multiple modalities. However, real-world scenarios often present challenges with incomplete data, particularly when dealing with censored survival labels. Prior works have addressed missing modalities but have overlooked incomplete labels, which can introduce bias and limit model efficacy. To bridge this gap, we introduce a novel framework that simultaneously handles incomplete data across modalities and censored survival labels. Our approach employs advanced foundation models to encode individual modalities and align them into a universal representation space for seamless fusion. By generating pseudo labels and incorporating uncertainty, we significantly enhance predictive accuracy. The proposed method demonstrates outstanding prediction accuracy in two survival analysis tasks on both employed datasets. This innovative approach overcomes limitations associated with disparate modalities and improves the feasibility of comprehensive survival analysis using multiple large foundation models.

Keywords: Multi-modality · Survival Analysis · Missing Data.

1 Introduction

Survival analysis is a vital process in cancer treatment research, enabling the prediction of important outcomes such as patient survival rates [21,1,3,4,18]. Recent advancements [16,11] in data collection techniques have opened new avenues for improving the accuracy of survival predictions by leveraging information from multiple modalities. An accurate prediction of the survival rate and time could primarily facilitate the precise composition of treatment planning.

The fusion of information from multiple modalities has been the recent trend of research in survival analysis. By leveraging cutting-edge techniques (e.g., cross-attention [20] and co-attention [2]) from the field of vision and language, it is

*Linhao Qu and Dan Huang contributed equally.

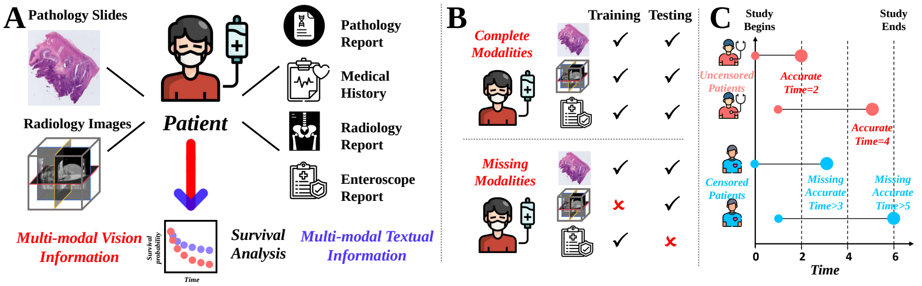

Fig. 1. Overview of the problem definition. (A) Employing multi-modal vision and textual data for patient-wise survival analysis. (B) Missing modality issues during training and testing. (C) Missing accurate label issues for censored patients.

popular to jointly train a model with two data modalities, e.g., pathology images and genomics. Nonetheless, these methods often require high data integrity and are constrained by limitations involving more than two modalities, as shown in Fig. 1 A and B. Moreover, in real-world survival analysis, it is typical to encounter a significant proportion of right-censored data, where the event of interest has not occurred or remains unknown by the end of the follow-up period, as is defined as missing accurate label issues shown in Fig. 1 C.

To address this challenge, previous works first tackle the incompleteness of input data in multi-modal survival analysis. These methods [5,8] focus on synthesizing the missing modalities, leveraging statistical techniques and generative model-based reconstruction algorithms to estimate the missing values. By filling in the missing information, these approaches strive to ensure that the analysis is conducted on as complete information as possible. Nonetheless, they still overlook the equally important issue of incomplete labels, which may introduce bias and limit the training efficacy of survival models. There is an urgent need for innovative approaches that account for both incomplete modalities and labels.

In this work, we demonstrate a joint framework that revolutionizes survival analysis by handling incomplete data across modalities and censored survival labels together. First, we take advantage of the current availability of advanced foundation models capable of encoding individual modalities. Then, the computed multi-modal embeddings are bound into a universal representation space via multi-modal feature alignment, paving the way for a seamless fusion of diverse modalities in a missing modality setting. Unlike previous methods, our method goes beyond merely handling missing modalities, instead addressing the formidable challenge of incomplete survival labels. By generating pseudo labels and incorporating uncertainty in the training of censored data, we significantly improve the predictive accuracy of survival prediction. This innovative approach eliminates the limitations associated with disparate modalities and enhances the feasibility of conducting comprehensive survival analysis. Crucially, our method also allows for clear interpretability of each modality's importance via an explicit attention mechanism.

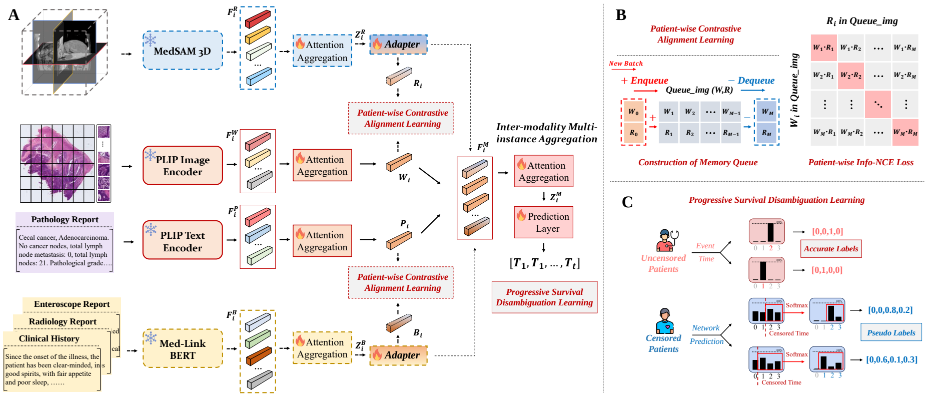
A
MedSAM 3D
F_i^R
Attention Aggregation
Z_i^R
Adapter
R_i
Patient-wise Contrastive Alignment Learning
PLIP Image Encoder
F_i^W
Attention Aggregation
W_i
Pathology Report
Cecal cancer, Adenocarcinoma. No cancer nodes, total lymph node metastasis: 0, total lymph nodes: 21. Pathological grade….
PLIP Text Encoder
F_i^P
Attention Aggregation
P_i
Patient-wise Contrastive Alignment Learning
Enteroscope Report
Radiology Report
Clinical History
Since the onset of the illness, the patient has been clear-minded, in good spirits, with fair appetite and poor sleep, ……
Med-Link BERT
F_i^B
Attention Aggregation
Z_i^B
Adapter
B_i
Inter-modality Multi-instance Aggregation
F_i^M
Attention Aggregation
Z_i^M
Prediction Layer
$[T_1, T_1, \dots, T_t]$
Progressive Survival Disambiguation Learning
B
Patient-wise Contrastive Alignment Learning
New Batch
+ Enqueue
− Dequeue
Queue_img (W,R)
W_0 R_0
W_1 W_2 … W_{M-1} W_M
R_1 R_2 … R_{M-1} R_M
Construction of Memory Queue
R_i in Queue_img
W_i in Queue_img
$W_1 \cdot R_1$ $W_1 \cdot R_2$ … $W_1 \cdot R_M$
$W_2 \cdot R_1$ $W_2 \cdot R_2$ … $W_2 \cdot R_M$
$W_M \cdot R_1$ $W_M \cdot R_2$ … $W_M \cdot R_M$
Patient-wise Info-NCE Loss
C
Progressive Survival Disambiguation Learning
Uncensored Patients
Event Time
0 1 2 3
[0,0,1,0]
Accurate Labels
[0,1,0,0]
Censored Patients
Network Prediction
Softmax
Censored Time
[0,0,0.8,0.2]
Pseudo Labels
[0,0.6,0.1,0.3]

Fig. 2. (A) Overview of the proposed framework. Solid lines: constant modalities; dashed lines: potentially missing modalities. (B) Diagram of Patient-wise Contrastive Alignment Learning. (C) Diagram of Progressive Survival Disambiguation Learning.

The contribution of the proposed multi-modal survival analysis framework is three-fold: 1) We proposed a multi-modal survival analysis framework by considering the incompleteness in both the input data and label; 2) we utilize a variety of foundation models to encode each modality and bind them into aligned representations for a more generalizable means of multi-modal data fusion; 3) We demonstrate outstanding prediction accuracy in two survival tasks across two real clinical datasets.

2 Method

The overall training framework is shown in Fig. 2. A. We first categorize all modalities into four types: radiology images, pathology Whole Slide Images (WSIs), pathology reports, and other clinical notes. Then, we utilize pre-trained modality-specific Foundation Models (FMs) to extract features from each modality separately (see Sec. 2.1); next, we introduce attention-based multi-instance feature aggregation for both intra-modality and inter-modality (see Sec. 2.2). To learn a single joint embedding space that encompasses patient-specific modalities, we propose a Patient-wise Contrastive Alignment Learning based on the Adapter and Contrastive Learning techniques (see Sec. 2.3). Finally, we introduce Progressive Survival Disambiguation Learning to address the issue of estimating unknown interval risks for censored patients (see Sec. 2.4). The overall loss function of the framework is defined as $L = \lambda L_{\text{con}} + L_{\text{surv}}$, where L_{con} denotes the contrastive loss (see Sec. 2.3), L_{surv} denotes the survival loss (see Sec. 2.4), and λ denotes the balancing weight coefficient.

2.1 Encoding Modalities with Modality-specific Foundation Models

Multi-modal Input Data: This paper includes a variety of important modalities for survival analysis of colorectal cancer patients, which are: (1) pathology WSIs; (2) radiology images such as CT or MRI; (3) pathology reports: descriptions of tumors regarding the patient's pathology slides, including tumor location, shape and size, grading, infiltration, invasion, and the status of major genetic targets; (4) radiology reports: descriptions of tumors corresponding to the patient's radiology images; (5) clinical history; (6) colonoscopy reports. All patients have at least one WSI and a corresponding pathology report, while the presence of other modalities can vary with certain incompleteness.

A single general FM may not be adequate for comprehensive encoding all modalities due to the diversity in format and content. For instances, FMs trained predominantly on natural image datasets fall short in accurately interpreting radiologic anatomical structures. Models trained on pathological images may not grasp the nuances of radiological anatomy. To overcome these limitations, we leverage modality-specific FMs, enhancing encoding precision.

Pathology Images and Reports: PLIP [9] is applied to extract visual features from WSIs and textual features from pathology reports. Each WSI, after background removal, is divided into n_W patches of 224×224 at 10x magnification. PLIP's Image Encoder outputs a 1×512 visual feature vector for each patch and together forms a set sized $n_W \times 512$. The patch count per WSI can vary, and for individuals with multiple WSIs, the vectors are concatenated to form an extensive WSI feature set $F_i^W \in \mathbb{R}^{N_i^W \times 512}$, where N_i^W is the total patch count from all WSIs of patient i. Pathology reports are segmented into sections by keywords (avoiding length limit of FM) and each section is encoded into a 1×512 vector by PLIP's Text Encoder, creating a feature set $F_i^P \in \mathbb{R}^{N_i^P \times 512}$, with N_i^P representing the section count, which varies by patient. These keywords used to segment the reports were provided by pathologists through structured text report data, including tumor location, size, grading, etc.

Radiology Images: For 3D radiology images (CT or MRI), the pre-trained MedSAM-3D [12] is utilized. Initially, each image is segmented into n_R 3D patches of size 128×128×128 via sliding window. Subsequently, MedSAM-3D maps these patches to 1×512 feature vectors, as an array with dimensions of $n_R \times 512$. For individuals with variable number of images, feature vectors are concatenated into an array $F_i^R \in \mathbb{R}^{N_i^R \times 512}$, where N_i^R denotes the aggregate number of 3D patches from all radiographic images for patient i.

Other Clinical Notes: The pre-trained BioLinkBERT-large [19] is employed to tokenize and encode other clinical notes, including radiology reports, medical history, and colonoscopy reports. These textual data are segmented by keywords and encoded into feature vectors. Embeddings from multiple reports are concatenated to form a comprehensive feature set $F_i^B \in \mathbb{R}^{N_i^B \times 1024}$, with N_i^B indicating the total segment count across all clinical notes for a patient.

2.2 Attention-based Intra-modality and Inter-modality Multi-instance Aggregation

In multi-modal analysis, the challenges include (1) the potential absence of multiple modalities during both training and testing, (2) the variable number of feature vectors within each modality, and (3) the necessity to quantify the importance of intra-modal and inter-modal factors on outcomes. Innovatively, we address these issues by unifying the aggregation of intra-modal and inter-modal features into a problem of multi-instance aggregation based on the attention mechanism, offering a flexible and efficient solution.

Intra-modality Aggregation: We aggregate the patient-wise feature vector sets extracted from radiology, pathology, and other medical notes into a uniform-dimensional feature vector, respectively. Specifically, we take the feature vector set of radiology data $F_{i,j}^{R}$ as an example, where the instances being aggregated are the feature vectors extracted from each 3D patch.

$$Z_i^R = \sum_{j=1}^{N_i^R} a_{i,j}^R F_{i,j}^R \ , \qquad a_{i,j}^R = \frac{\exp\left\{w^\top \tanh\left(V F_{i,j}^{R^\top}\right)\right\}}{\sum_{j=1}^{N_i^R} \exp\left\{w^\top \tanh\left(V F_{i,j}^{R^\top}\right)\right\}} \tag{1}$$

where $a_{i,j}^R$ is the attention score predicted by the self-attention network with learnable parameters w and V, reflecting each instance's contribution during aggregation [10]. Its flexibility is demonstrated by the fact that the number of instances per patient input does not need to be strictly equal. Ultimately, we obtain the attention aggregated imaging features Z_i^R, pathology WSI features W_i, pathology report features P_i, and other medical report features Z_i^B, respectively.

Inter-modality Aggregation: All modal features of a patient are aggregated into a single-dimensional feature vector. Given that a patient's modal quantity varies, we first concatenate all modal features together to form the modal feature vector set F_i^M. Then, we aggregate F_i^M in the same manner as described in Eq. 1, resulting in the final patient-level feature vector $Z_i^M \in \mathbb{R}^{1\times 256}$.

2.3 Patient-wise Contrastive Alignment Learning

We aim to create a unified joint embedding space for all modalities, facilitating more distinctive representations and the alignment of features from different modalities. It can also suppress the challenge of missing modalities when performing inter-modality aggregation. Nonetheless, we face the challenge of enforcing the alignment of multiple modalities.

This work utilizes pathology images and corresponding reports (encoded and aligned using the pre-trained PLIP model) as the hub. The strategy segregates the contrastive learning process into separate image and text sides, i.e., aligning embeddings of other visual modalities with pathology images and other textual modalities with pathology reports. The InfoNCE Loss [14] which compares the similarity of samples and encourages the model to identify positive samples among the negatives is applied for each alignment. Due to the issue of missing modalities for each patient and the varying number of instances within each modality, it results in a batch size of one (patient) during specific training sessions. Inspired by MOCO [7], Memory Queues have been constructed on both the image and text sides to provide a substantial number of negative samples, as shown in Fig. 2.B.

Specifically, we first import the paired radiology image and pathology WSI of M patients as the initial memory queue, denoted as $Queue_{img}$; and the paired pathological report data and other medical report data as the initial $Queue_{text}$. Samples in the memory queue will be popped out as the inter-patient negative samples for contrastive learning and refilled whenever new patient data are processed for training.

The process of contrastive learning on both the image and text sides is similar. Here, we take the image side as an example. First, we use an Adapter (two-layer Fully-connected layer [6]) to map the image features Z_i^R after attention aggregation, obtaining features R_i. Then, the current patient's WSI features W_i and R_i (as a pair) are added to $Queue_{img}$, and one patient's feature pair in the queue is discarded. Next, for all feature pairs (W, R) in $Queue_{img}$, with the current patient's W_i and R_i as the positives and other combinations as negatives, we can compute the InfoNCE Loss $L_{W,R}$. Similarly, we use an Adapter to map the features Z_i^B of other medical reports after attended aggregation, obtaining features B_i. Then, based on (P, B), we construct $L_{P,B}$.

The overall form of the contrastive loss is defined as $L_{con} = L_{W,R} + \lambda_{con} L_{P,B}$, where λ_{con} is a balancing coefficient determined by the ratio of the number of complete feature pairs (W, R) and (P, B) among all patients.

2.4 Progressive Survival Disambiguation Learning

Survival analysis is challenging because it involves ordinal regression to model time-to-event (e.g., death) data, with some events potentially not observed (right-censored). Following [2,20], we divided survival times of uncensored patients into set intervals (e.g., {0,1,2,3}) as discrete labels for all patients, forming a maximum likelihood loss from these labels. A two-layer Prediction Layer is added to regress the death hazard and survival probability for each interval. For uncensored patients with accurate labels, we maximize their risk of death and minimize their survival probability in the actual discrete interval of death. For censored patients without accurate labels, previous works [2,20,8] could only maximize survival probability in the current interval while neglecting their hazard loss.

Therefore, we propose Progressive Survival Disambiguation Learning to address this challenge by estimating unknown interval hazards for censored patients during training. Specifically, we first use the Prediction Layer to predict risks for each interval and then use censored time to retain the hazards for subsequent intervals while setting the hazards for earlier intervals to zero. Softmax is applied to normalize these predictions, and they are then used to weight the ground-truth label to form the soft labels for training. See Fig. 2.C. for an example. Considering the network's limited predictive capability at the beginning of training, we further employ a time-dependent Gaussian warming up function [13], $\lambda_{\text{pro}}(t) = 0.1 \cdot e^{\left(-5\left(1-\frac{t_i}{t_{\text{total}}}\right)^2\right)}$, to weight on these pseudo labels increasingly. Here, t_i and t_{total} denote the current and total iterations.

The survival loss is defined as $L_{surv} = L_{uncen} + \lambda_{cen}\left(L_{cen} + \lambda_{pro}(t) L_{cen\_p}\right)$, where L_{uncen}, L_{cen}, and L_{cen_p}, which represent the loss for uncensored patients, censored patients, and the proposed risk probability estimation loss for censored patients, respectively. λ_{cen} is the weight coefficient for the overall loss of censored patients.

3 Experiments

3.1 Datasets and Tasks

We assessed our algorithm's performance through two real-world in-house datasets, each partitioned into training and test sets on a per-patient basis for five-fold cross-validation, with mean outcomes reported. Notably, all patients have WSIs and pathology reports, while radiology data and additional clinical notes are intermittently absent. Dataset 1 contains 367 patients, of which 180 are with radiology images, 303 medical history reports, 205 colonoscopy reports, and 89 radiology reports. Dataset 2 includes 193 patients, with 133 having radiology images, 181 with medical history reports, 154 with colonoscopy reports, and 129 with radiology reports. In both, we assessed two important prognosis tasks: overall survival (OS) prediction and disease-free survival (DFS) prediction.

3.2 Evaluation Metrics and Compared Methods

The evaluation of our model employed three metrics. Firstly, the Concordance Index (CI) served as the primary metric, quantifying the proportion of patient pairs whose survival risks are accurately ranked. CI values span from 0 to 1, with higher values indicating superior model performance. Secondly, the Brier Score (BS) assessed prediction accuracy, calculating the mean squared difference between observed survival statuses and predicted probabilities, where a BS of 0 represents optimal accuracy. Lastly, Kaplan-Meier (KM) analysis determined patient stratification efficacy by dividing patients into high-risk and low-risk groups based on the median prediction model scores for each cohort. Superior stratification is reflected by lower p-values in the Logrank test.

Given that our data are real clinical data with multimodal missing situations in both training and testing, we primarily compared against the current SOTA multimodal algorithms for missing modalities, including MMD [5], HGCN [8], and ShaSpec [17]. To ensure a fair comparison, we used the same FMs to extract features for different modalities and maintained exactly the same data division.

3.3 Results

As shown in Table 1, the performance of our algorithm significantly surpasses all comparison algorithms across all tasks in all datasets, indicating the high efficiency of our algorithm.

Table 1. Performance comparison on the two tasks across the two datasets.

Dataset	Dataset 1				Dataset 2			
Task	OS		DFS		OS		DFS	
Method	CI ↑	BS ↓	CI ↑	BS ↓	CI ↑	BS ↓	CI ↑	BS ↓
MMD (MICCAI'22)	0.877	0.108	0.888	0.104	0.893	0.102	0.884	0.105
HGCN (TMI'23)	0.882	0.105	0.891	0.101	0.899	0.101	0.891	0.102
ShaSpec (CVPR'23)	0.889	0.104	0.896	0.100	0.905	0.098	0.899	0.097
Ours	**0.897**	**0.100**	**0.905**	**0.095**	**0.914**	**0.083**	**0.907**	**0.091**

Table 2. Results of ablation studies.

Method	CI ↑	BS ↓
Baseline	0.871	0.109
w/o L_{con}	0.884	0.105
w/o L_{cen_p}	0.887	0.104
Ours	**0.897**	**0.100**

Method	CI ↑	BS ↓
WSI	0.856	1.116
P_{rep}	0.842	1.119
WSI&P_{rep}	0.865	1.111
Ours	**0.897**	**0.100**

Method	CI ↑	BS ↓
PLIP	0.889	0.104
CLIP	0.885	0.105
Ours	**0.897**	**0.100**

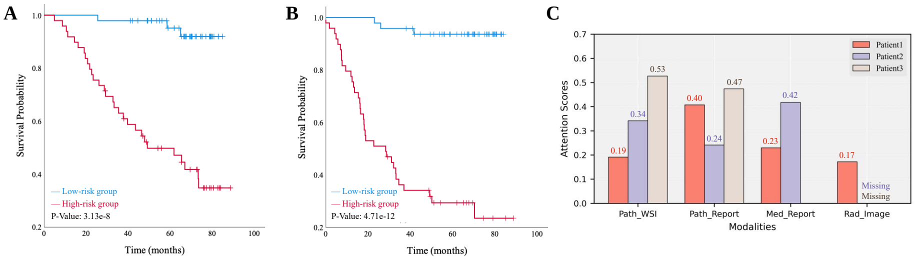

Fig. 3. The KM analysis curves for (A) OS prediction task and (B) DFS prediction task. (C) the visualization of modal attention scores for three patients.

We further divided the test cohort of Dataset 1 into high-risk and low-risk groups based on the median risk score predicted by our model. If our model's predictions are efficient, then there should be a significant difference between the KM curves of these two groups. The results, as shown in Fig. 3.A (OS prediction task) and Fig. 3.B (DFS prediction task), reveal that the p-values for both groups are less than 1e-7, indicating the significant efficacy of our method.

Our method also possesses strong clinical interpretability, allowing for the flexible quantification of the importance of each modality engaged by the test patients toward the outcomes. Fig. 3.C illustrates the visualization of modal attention scores for three typical patients, demonstrating our method's robust and flexible interpretative advantage in clinical applications.

Ablation Study: In our detailed ablation studies for the Overall Survival (OS) prediction task using Dataset 1, we explored three different aspects, with results presented in Table 2. First, we conducted ablation experiments on the proposed contrastive learning loss L_{con}, and the loss L_{cen_p} tailored for censored data. The baseline represents scenarios without any contrastive learning or disambiguation learning, highlighting the effectiveness of these two key components we introduced. Second, we performed ablation experiments with pathology WSIs

(WSI) only, pathology reports (P_{rep}) only, and both, demonstrating that the joint learning of multiple modalities can effectively enhance performance. Third, we compared the performance of using a single vision-language large model, either PLIP [9] or CLIP [15], for all tasks. The results indicated that using a specialized medical large model for each modality can improve performance.

4 Conclusion

Our paper presents a novel multi-modal survival analysis framework tailored to address critical challenges in cancer treatment research, including incomplete data and censored survival labels. It represents a significant advancement in leveraging multi-modal data and overcoming critical challenges for precise survival predictions in cancer treatment research. Future research avenues include exploring application and validation in large-scale multi-center studies.

Acknowledgement

This work is funded by the National Key R&D Program of China (2022ZD0160700) and Shanghai AI Laboratory.

Disclosure of Interests

We declare no competing interests.

References

1. Chen, R.J., Lu, M.Y., Wang, J., Williamson, D.F., Rodig, S.J., Lindeman, N.I., Mahmood, F.: Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. IEEE Transactions on Medical Imaging **41**(4), 757–770 (2020)
2. 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)
3. Chen, R.J., Lu, M.Y., Williamson, D.F., Chen, T.Y., Lipkova, J., Shaban, M., Shady, M., Williams, M., Joo, B., Noor, Z., et al.: Pan-cancer integrative histology-genomic analysis via interpretable multimodal deep learning. arXiv preprint arXiv:2108.02278 (2021)
4. Courtiol, P., Maussion, C., Moarii, M., Pronier, E., Pilcer, S., Sefta, M., Manceron, P., Toldo, S., Zaslavskiy, M., Le Stang, N., et al.: Deep learning-based classification of mesothelioma improves prediction of patient outcome. Nature Medicine **25**(10), 1519–1525 (2019)
5. Cui, C., Liu, H., Liu, Q., Deng, R., Asad, Z., Wang, Y., Zhao, S., Yang, H., Landman, B.A., Huo, Y.: Survival prediction of brain cancer with incomplete radiology, pathology, genomic, and demographic data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 626–635. Springer (2022)

6. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision **132**(2), 581–595 (2024)
7. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)
8. Hou, W., Lin, C., Yu, L., Qin, J., Yu, R., Wang, L.: Hybrid graph convolutional network with online masked autoencoder for robust multimodal cancer survival prediction. IEEE Transactions on Medical Imaging (2023)
9. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature Medicine **29**(9), 2307–2316 (2023)
10. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning. pp. 2127–2136. PMLR (2018)
11. Kashyap, A., Fomitcheva Khartchenko, A., Pati, P., Gabrani, M., Schraml, P., Kaigala, G.V.: Quantitative microimmunohistochemistry for the grading of immunostains on tumour tissues. Nature Biomedical Engineering **3**(6), 478–490 (2019)
12. Lei, W., Wei, X., Zhang, X., Li, K., Zhang, S.: Medlsam: Localize and segment anything model for 3d medical images. arXiv preprint arXiv:2306.14752 (2023)
13. Luo, X., Hu, M., Song, T., Wang, G., Zhang, S.: Semi-supervised medical image segmentation via cross teaching between cnn and transformer. In: International Conference on Medical Imaging with Deep Learning. pp. 820–833. PMLR (2022)
14. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
16. Shen, S., Wang, Y., Wang, C., Wu, Y.N., Xing, Y.: Surviv for survival analysis of mrna isoform variation. Nature Communications **7**(1), 11548 (2016)
17. Wang, H., Chen, Y., Ma, C., Avery, J., Hull, L., Carneiro, G.: Multi-modal learning with missing modality via shared-specific feature modelling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15878–15887 (2023)
18. Yao, J., Zhu, X., Huang, J.: Deep multi-instance learning for survival prediction from whole slide images. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22. pp. 496–504. Springer (2019)
19. Yasunaga, M., Leskovec, J., Liang, P.: Linkbert: Pretraining language models with document links. arXiv preprint arXiv:2203.15827 (2022)
20. 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)
21. Zhu, X., Yao, J., Zhu, F., Huang, J.: Wsisa: Making survival prediction from whole slide histopathological images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7234–7242 (2017)