Cross-Modality Masked Learning for Survival Prediction in ICI Treated NSCLC Patients

Qilong Xing[1], Zikai Song[1] *, Bingxin Gong[2], Lian Yang[2], Junqing Yu[1], and Wei Yang[1]

[1] School of Computer Science and Technology
[2] Department of Radiology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China
{qlxing, skyesong}@hust.edu.cn

Abstract. Accurate prognosis of non-small cell lung cancer (NSCLC) patients undergoing immunotherapy is essential for personalized treatment planning, enabling informed patient decisions, and improving both treatment outcomes and quality of life. However, the lack of large, relevant datasets and effective multi-modal feature fusion strategies pose significant challenges in this domain. To address these challenges, we present a large-scale dataset and introduce a novel framework for multi-modal feature fusion aimed at enhancing the accuracy of survival prediction. The dataset comprises 3D CT images and corresponding clinical records from NSCLC patients treated with immune checkpoint inhibitors (ICI), along with progression-free survival (PFS) and overall survival (OS) data. We further propose a cross-modality masked learning approach for medical feature fusion, consisting of two distinct branches, each tailored to its respective modality: a Slice-Depth Transformer for extracting 3D features from CT images and a graph-based Transformer for learning node features and relationships among clinical variables in tabular data. The fusion process is guided by a masked modality learning strategy, wherein the model utilizes the intact modality to reconstruct missing components. This mechanism improves the integration of modality-specific features, fostering more effective inter-modality relationships and feature interactions. Our approach demonstrates superior performance in multi-modal integration for NSCLC survival prediction, surpassing existing methods and setting a new benchmark for prognostic models in this context.

Keywords: Survival analysis · Multimodal · Masked learning.

1 Introduction

Lung cancer remains one of the leading causes of death worldwide, with non-small cell lung cancer (NSCLC) accounting for approximately 85% [21]. Immunotherapy, represented by immune checkpoint inhibitors (ICIs), targets the anti-programmed cell death protein 1 (PD-1), anti-programmed death-ligand 1

* Corresponding author: skyesong@hust.edu.cn

(PD-L1), and anti-cytotoxic T-lymphocyte-associated protein 4 (CTLA-4) pathways to enhance the body's anti-tumor immune response, significantly improving the treatment of advanced or metastatic NSCLC [19]. However, many patients do not benefit from ICIs [15]. It has been reported that only around 20% of patients respond to ICIs, while an excessive immune response can disrupt immune tolerance, resulting in immune-related adverse events (irAEs) [14]. Accurate prognosis in immunotherapy is crucial, as it enables healthcare providers to predict disease progression and tailor treatment plans accordingly. However, the lack of comprehensive datasets poses a significant challenge.

Regarding techniques used in survival prediction, recent advancements have shifted from traditional methods, such as the Kaplan-Meier estimator [11] and Cox proportional hazards regression [3], to neural network-based approaches [17,13,2]. While single-modality methods [23,5] have shown promise, recent research [6,7,25,10,18] increasingly adopts multimodal strategies to leverage complementary information from diverse data sources. Compared to omics data, clinical data and medical images are generally more accessible. DAFT [24] introduces a dynamic affine module to fuse clinical data with 3D image features, while DOF [1] proposes an orthogonal loss function to encourage distinct modality representations. Additionally, Multisurv [20] combines clinical and whole-slide image features using an attention-based summation approach. However, these methods often rely on simplistic feature fusion techniques, which do not fully capture the complex inter-relationships between modalities, thus limiting the potential synergy of multimodal features.

To address these challenges, we propose two main contributions. First, we curate a large dataset comprising 3D CT images and corresponding clinical records of immunotherapy patients. This dataset includes progression-free survival (PFS) and overall survival (OS) data, which are essential for developing a robust survival prediction model for immunotherapy prognosis. Second, we introduce a cross-modality masked learning approach to enhance feature fusion, inspired by recent advancements in masked learning methods [8]. Specifically, we use a 3D visual transformer to encode the 3D CT images and employ a graph transformer to extract clinical features from the tabular data. The feature interaction and fusion are further strengthened by the cross-modality masked learning approach, which requires the model to complete missing information from one modality using the intact features from the other modality. Results confirm our method's effectiveness in multi-modal fusion and survival analysis.

2 Method

The proposed framework consists of two distinct branches: one dedicated to processing tabular clinical data and the other focused on 3D CT images. Our method follows a two-stage training procedure. The first stage involves pretraining the modality-specific encoders, while the second stage fine-tunes additional multilayer perceptron (MLP) module for the survival prediction task. The overall architecture of the pretraining process is illustrated in Fig.1.

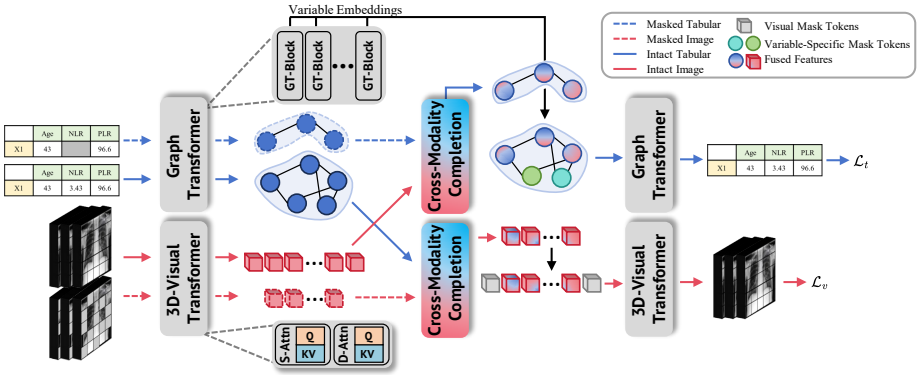

Fig. 1. During pretraining, both intact and masked versions of each modality are input into their respective branches. In the multi-modal completion process, the masked modality integrates features from the intact version of the other modality, which are then passed into the decoder for reconstruction. In the visual decoder, a learnable mask token is used, while in the graph transformer decoder, clinical variable-specific features from the encoder serve as masked tokens.

2.1 Visual Branch

We employ a 3D visual transformer to extract image features from CT scans, drawing inspiration from recent advancements in video transformer models [22]. Given an input 3D CT image $\mathcal{I} \in \mathbb{R}^{H \times W \times D}$ and predefined patch sizes $p_1 \times p_2 \times p_3$, the model first reshapes the image into non-overlapping patches and transform to patch embeddings $\mathcal{V} \in \mathbb{R}^{hw \times d \times C}$, where $h = \lfloor \frac{H}{p_1} \rfloor$, $w = \lfloor \frac{W}{p_2} \rfloor$, $d = \lfloor \frac{D}{p_3} \rfloor$, and $C = p_1 \times p_2 \times p_3$. The transformer encoder consists of a slice-based transformer (ST) block and a depth-based (DT) transformer block. In ST block, the slice-based attention (S-Attn) operates on the transposed patch embeddings $\mathcal{V}^T \in \mathbb{R}^{d \times hw \times C}$, using self-attention to capture global context across patches along the second dimension. The processed features are then transposed back and passed to the DT blocks, where the depth-based attention (D-Attn) facilitates information interaction along the depth dimension. The feature is then flattened for further computation:

$$\mathcal{V}^{'} = \text{D-Attn}(\text{S-Attn}(\mathcal{V}_p^T)^T) \in \mathbb{R}^{hwd \times c}, \tag{1}$$

During the masked learning procedure, image patches are randomly masked with a masking ratio k. Only the unmasked patches are fed into the model encoder for feature extraction. The decoder shares the same architecture as the encoder. To reconstruct the masked patches, following MAE [8], the learnable masked embedding is combined with the visible patches and fed into the decoder. The reconstruction loss $\mathcal{L}_v$ is computed as the mean squared error (MSE) between the predicted and original patches.

2.2 Tabular Branch

We adapt the graph-based transformer T2G [26] for our masked learning procedure. Categorical variables are encoded using learnable embeddings, while numerical variables are mapped via a linear module. All features are combined as input to the graph transformer. We treat each clinical variable as a node in the graph, and construct the relationships between variables as edges. Following the T2G framework, the graph edges built in each Graph Transformer block (GT-Block) consist of two components: one encodes the relationships based on specific sample features using a learnable relation diagonal matrix R, and the other encodes the global relationships between variables using learnable embeddings $\mathcal{Z} = \{z_1; ...; z_d\} \in \mathbb{R}^{d \times c}$ for each clinical variable, where d denotes the number of clinical variables and c represents the feature channel size. The final graph is formed by combining these two types of edges.:

$$G_s = \mathcal{X} R \mathcal{X}^T, \quad G_z = \mathcal{Z}\mathcal{Z}^T, \quad G = \text{Softmax}(G_s + G_z) \in \mathbb{R}^{d \times d}, \tag{2}$$

where $\mathcal{X} = \{x_1; ...; x_d\} \in \mathbb{R}^{d \times c}$ represents the node features. The graph is used to update the node features, resulting in the processed features $\mathcal{X}^{'}$ obtained by the encoder.

In the masked learning procedure, $\lfloor dk \rfloor$ clinical variables are randomly masked per sample. Only the unmasked variables $\mathcal{X}_{[U]}$ are processed by the encoder, and edges associated with the masked nodes are removed, ensuring interaction occurs only between the remaining nodes.

Unlike visual features, graph features are less influenced by positional information. Therefore, simply combining position embeddings with a learnable masked embedding, as done in MAE [8], is inadequate for providing the necessary information to facilitate effective reconstruction of masked clinical variables. To address this, we propose collecting the embeddings for each clinical variable, $\mathcal{Z}$, from all n blocks in the encoder and fusing them using a learnable matrix $W_z \in \mathbb{R}^{nc \times c}$ to construct a specific masked embedding for each clinical variable:

$$\mathcal{Z}_a = \text{Concat}([\mathcal{Z}_1, ..., \mathcal{Z}_n]) \in \mathbb{R}^{d \times nc}, \quad \mathcal{X}_{[M]} = \mathcal{Z}_a W_z \in \mathbb{R}^{d \times c}, \tag{3}$$

During reconstruction, the embedding for each masked variable is retrieved and combined with unmasked features for processing in the decoder, which shares the same architecture as the encoder. For the tabular reconstruction loss $\mathcal{L}_x$, we propose using different loss functions for numerical and categorical variables. In the final output layer of the decoder, numerical features are mapped to a single value, and the MSE loss is applied. For categorical variables, the features are mapped to logits. If the variable has only two categories, binary cross-entropy (BCE) loss is used; otherwise, cross-entropy (CE) loss is employed.

2.3 Cross-Modality Completion

To enhance multimodal fusion, we propose a cross-modality completion (CMC) process based on masked learning. The CMC process requires the masked modal-

ity to complete itself by extracting and combining features from the other modality, strengthening the relationship between them and improving the feature fusion process. The CMC process is illustrated in Fig.1.

Specifically, we employ m transformer layers to progressively integrate cross-modality features within each branch. For instance, given the tabular features after random variable masking, $\mathcal{X}_{[U]}^{'}$, and the intact visual features, the self-attention mechanism is first applied to the tabular data to enhance its features:

$$\begin{aligned} Q_s = \mathcal{X}_{[U]}^{'(m-1)} W_{s,q}^{x,m}, \quad K_s = \mathcal{X}_{[U]}^{'(m-1)} W_{s,k}^{x,m}, \quad V_s = \mathcal{X}_{[U]}^{'(m-1)} W_{s,v}^{x,m}, \\ \mathcal{X}_{[S,U]}^{'(m)} = \mathrm{LN}(\text{Self-Attn}(Q_s, K_s, V_s) + \mathcal{X}_{[U]}^{'(m-1)}), \end{aligned} \tag{4}$$

Here, $W_{s,q}^{x,m}$, $W_{s,k}^{x,m}$ and $W_{s,v}^{x,m}$ are learnable matrices in the self-attention mechanism at the m-th layer of the tabular branch, and LN denotes layer normalization. To explore and fuse visual features for the reconstruction of masked clinical variables, we use the tabular features as the Query, while the visual features act as the Key and Value in the cross-modality attention mechanism:

$$\begin{aligned} Q_c = \mathcal{X}_{[S,U]}^{'(m)} W_{c,q}^{x,m}, \quad K_c = \mathcal{V}^{'} W_{c,k}^{x,m}, \quad V_c = \mathcal{V}^{'} W_{c,v}^{x,m}, \\ \mathcal{X}_{[C,U]}^{''(m)} = \mathrm{LN}(\text{Cross-Attn}(Q_c, K_c, V_c) + \mathcal{X}_{[S,U]}^{'(m)}), \end{aligned} \tag{5}$$

where $W_{c,q}^{x,m}$, $W_{c,k}^{x,m}$ and $W_{c,v}^{x,m}$ are learnable matrices in the cross-attention process. Subsequently, a Multi-Layer Perceptron (MLP) with a residual connection is applied to produce the final feature at layer m:

$$\mathcal{X}_{[U]}^{'(m)} = \mathrm{MLP}(\mathcal{X}_{[C,U]}^{''(m)}) + \mathcal{X}_{[C,U]}^{''(m)}, \tag{6}$$

The output of the CMC is fed into the graph transformer decoder, where it is combined with the masked embeddings. The operation of the visual branch is symmetric and therefore omitted for simplicity.

2.4 Survival Analysis

To fine-tune for the survival analysis task, we remove the decoder parts of the two branches and keep the remaining modules frozen. The intact clinical data and 3D CT images are input for feature extraction. After feature fusion through the transformers in the CMC process, the visual feature $\mathcal{V}_{[U]}^{'(m)}$ is pooled and concatenated with the $[CLS]$ token of the tabular feature $\mathcal{X}_{[U]}^{'(m)}$. These combined features are then processed by a MLP module to generate the survival prediction. For fine-tuning, we adopt the widely used Cox loss [12].

3 Experiments

3.1 Dataset

We curated a multimodal dataset for NSCLC immunotherapy patients, consisting of 3D CT images and clinical data from 2,128 samples. The CT images have

Table 1. Sample distribution of categorical variables. Sample numbers are provided in parentheses. Abbreviations: AD (Adenocarcinoma), SCC (Squamous Cell Carcinoma), SD (Stable Disease), PD (Progressive Disease), PR (Partial Response). Abbreviations: Adenocarcinoma (AD), Squamous Cell Carcinoma (SCC), Stable Disease (SD), Progressive Disease (PD), Partial Response (PR). The Response Evaluation variable is excluded from training, while all other variables are used as input clinical features.

Categorical Variables	Distributions
Sex	Male (1796) vs Female (332)
Lung cancer stage	Stage 3 (780) vs Stage 4 (1332)
Pathological types	AD (897) vs SCC (1040) vs Others (190)
Diabetes	With diabetes (208) vs Without diabetes (1920)
Hypertension	With Hypertension (527) vs Without Hypertension (1599)
Smoking	Smoking (1134) vs Not smoking (842)
Drinking	Drinking (512) vs Not drinking (1388)
Hyperlipidemia	With hyperlipidemia (437) vs Without Hyperlipidemia (1568)
Response Evaluation	SD (1037) vs PD (392) vs PR (699)

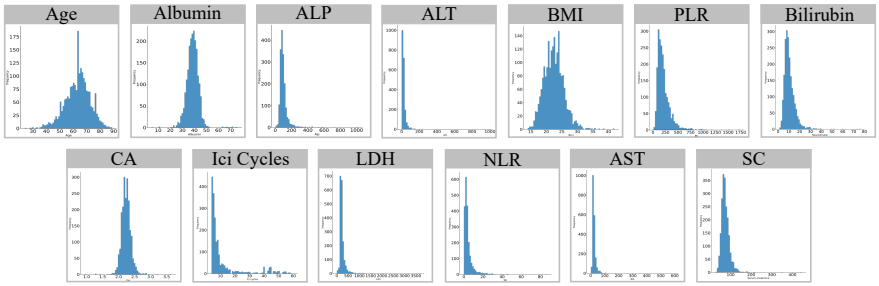

Fig. 2. Sample distribution of numerical variables. Abbreviations: alkaline phosphatase (ALP), alanine aminotransferase (ALT), body mass index (BMI), platelet-to-lymphocyte ratio (PLR), calcium (CA), immune checkpoint inhibitor cycles (ICI cycles), lactate dehydrogenase (LDH), neutrophil-to-lymphocyte ratio (NLR), aspartate aminotransferase (AST), Serum creatinine (SC).

a median resolution of $512 \times 512 \times 268$ voxels and a median voxel spacing of $0.73 \times 0.73 \times 1.25$ mm. The clinical data comprises 13 numerical (see Fig. 2) and 9 categorical variables (see Table 1). The PFS and OS times are measured in months.

3.2 Implementation Details

We preprocess CT images by resampling them to a uniform voxel spacing of $0.75 \times 0.75 \times 1.5$ and center-cropping them to a size of $480 \times 480 \times 240$. For clinical data, missing numerical values are imputed with the mean, and categorical values with the most frequent class. We use five-fold cross-validation to evaluate PFS and OS predictions, with the Concordance Index (CI) as the metric [9]. The training consists of two stages. In pretraining, the default masking ratio is set to

Table 2. The evaluation is based on two tasks: progression-free survival (PFS) prediction and overall survival (OS) prediction. * indicates statistical significance at the 0.05 level compared to the second-best performing model, the Interactive-Model.

Methods	PFS CI	OS CI
Cox [3]	0.653 ± 0.015	0.646 ± 0.016
Interactive-Model [4]	0.690 ± 0.018	0.687 ± 0.017
FiLM [16]	0.681 ± 0.012	0.665 ± 0.020
DAFT [24]	0.686 ± 0.015	0.683 ± 0.016
Ours	**0.701 ± 0.018***	**0.705 ± 0.021***

Table 3. The ablation results. * indicates statistical significance at the 0.05 level compared to the Interactive-Model.

Methods	Image	Tabular	PFS CI	OS CI
VS	✔	✗	0.583 ± 0.011	0.591 ± 0.019
VM	✔	✗	0.600 ± 0.015	0.604 ± 0.013
TS	✗	✔	0.672 ± 0.010	0.643 ± 0.017
TM	✗	✔	0.670 ± 0.013	0.645 ± 0.014
CS	✔	✔	0.685 ± 0.010	0.682 ± 0.019
CM	✔	✔	0.689 ± 0.015	0.685 ± 0.010
Ours w/o CVS	✔	✔	0.695 ± 0.017	0.696 ± 0.011
Ours	✔	✔	**0.701 ± 0.018***	**0.705 ± 0.021***

50%, the learning rate is $1e^{-4}$, and training lasts 80 epochs. In fine-tuning, the model is trained for 100 epochs with a learning rate of $1e^{-2}$. For comparison, we rerun competitive methods using their official implementations. All experiments are conducted on a single NVIDIA V100 GPU [1].

3.3 Main Results

We compare our method with the baseline Cox proportional hazards model [3], which uses only tabular data, and other competitive methods that integrate both tabular and imaging data. The main results are shown in Table 2. Our method outperforms the baseline Cox model and other multimodal approaches on both PFS and OS prediction tasks, demonstrating the effectiveness of the CMC procedure. Notably, unlike methods that require retraining the entire model for each task (which takes hours), our method only fine-tunes a lightweight MLP module, taking less than 1 minute after pre-storing the features. This highlights the generalizability of our fused features across tasks.

3.4 Model Analysis

We conduct ablation studies to further evaluate the proposed method, with results shown in Table 3. First, we assess models based on individual modalities.

[1] The code will be available at `https://github.com/qlxing/CMC_Surv`.

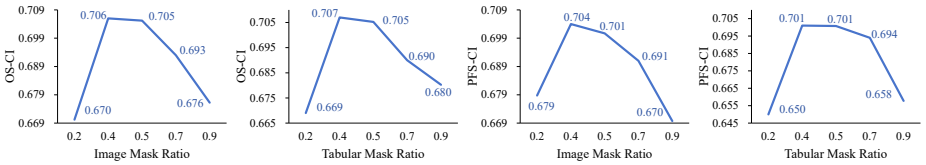

Fig. 3. Effect of mask ratio on OS and PFS when varying the mask ratio for a single modality.

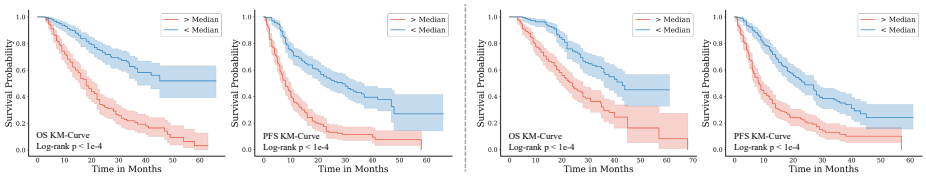

Fig. 4. Kaplan-Meier curves and Log-rank tests of different CV folds.

Using only image data with the Cox loss (VS model) results in limited performance, while pretraining with the masked learning approach (VM model) improves performance. For tabular data, directly training with the Cox loss (TS model) yields similar performance to the masked learning-pretrained model, indicating that masked learning enhances the visual features more effectively than the tabular features.

For multimodal fusion, directly concatenating features from different branches and training with the Cox loss (CS model) outperforms single-modality models but performs worse than concatenating features enhanced by masked training for each branch separately (CM model). Our method achieves the highest accuracy. Additionally, we explore clinical-variable-specific (CVS) masked embeddings in the graph transformer, which prove more effective than using a single learnable mask token for different clinical variables.

We also explore the effect of the mask ratio for different modalities, fixing one at 50%. As shown in Fig. 3, a moderate mask ratio for both modalities yields the best performance. Too low or too high a mask ratio for either modality prevents effective feature fusion, making the branch ineffective.

In Fig. 4, we present the Kaplan-Meier curve, splitting the test data using the median value of the predicted logits. The results show a very small p-value from the log-rank test, indicating that our method is effective for survival prediction.

4 Conclusion

Accurately predicting survival in ICI-treated NSCLC patients is essential for guiding treatment decisions. It allows for adaptive adjustments to treatment plans, ensuring personalized and effective care. To develop a robust model for survival prediction in NSCLC patients undergoing immunotherapy, we compile a

large, relevant dataset and propose a cross-modality fusion method to optimize the use of multimodal data. Our experiments demonstrate that the proposed model effectively integrates features from multiple modalities and outperforms existing methods in survival prediction for NSCLC patients treated with immunotherapy.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (NSFC No. 62272184 and No. 62402189), the China Postdoctoral Science Foundation under Grant Number GZC20230894, the China Postdoctoral Science Foundation (Certificate Number: 2024M751012), the Postdoctor Project of Hubei Province under Grant Number 2024HBBHCXB014, the Noncommunicable Chronic Diseases-National Science and Technology Major Project (2024ZD0522800 / 2024ZD0522806). The computation is completed in the HPC Platform of Huazhong University of Science and Technology.

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

References

1. Braman, N., Gordon, J.W., Goossens, E.T., Willis, C., Stumpe, M.C., Venkataraman, J.: Deep orthogonal fusion: multimodal prognostic biomarker discovery integrating radiology, pathology, genomic, and clinical data. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24. pp. 667–677. Springer (2021)
2. Cai, S., Huang, W., Yi, W., Zhang, B., Liao, Y., Wang, Q., Cai, H., Chen, L., Su, W.: Survival analysis of histopathological image based on a pretrained hypergraph model of spatial transcriptomics data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 455–466. Springer (2024)
3. Cox, D.R.: Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) **34**(2), 187–202 (1972)
4. Duanmu, H., Huang, P.B., Brahmavar, S., Lin, S., Ren, T., Kong, J., Wang, F., Duong, T.Q.: Prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer using deep learning with integrative imaging, molecular and demographic data. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23. pp. 242–252. Springer (2020)
5. Haarburger, C., Weitz, P., Rippel, O., Merhof, D.: Image-based survival prediction for lung cancer patients using cnns. In: 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019). pp. 1197–1201. IEEE (2019)
6. Hao, J., Kim, Y., Mallavarapu, T., Oh, J.H., Kang, M.: Cox-pasnet: pathway-based sparse deep neural network for survival analysis. In: 2018 IEEE international conference on bioinformatics and biomedicine (BIBM). pp. 381–386. IEEE (2018)
7. Hao, J., Kosaraju, S.C., Tsaku, N.Z., Song, D.H., Kang, M.: Page-net: interpretable and integrative deep learning for survival analysis using histopathological images and genomic data. In: Pacific Symposium on Biocomputing 2020. pp. 355–366. World Scientific (2019)

8. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
9. Jiang, L., Xu, C., Bai, Y., Liu, A., Gong, Y., Wang, Y.P., Deng, H.W.: Autosurv: interpretable deep learning framework for cancer survival analysis incorporating clinical and multi-omics data. NPJ precision oncology **8**(1), 4 (2024)
10. Jiang, S., Gan, Z., Cai, L., Wang, Y., Zhang, Y.: Multimodal cross-task interaction for survival analysis in whole slide pathological images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 329–339. Springer (2024)
11. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. Journal of the American statistical association **53**(282), 457–481 (1958)
12. Katzman, J.L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y.: Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. BMC medical research methodology **18**, 1–12 (2018)
13. Kim, K., Lee, Y., Park, D., Eo, T., Youn, D., Lee, H., Hwang, D.: Llm-guided multi-modal multiple instance learning for 5-year overall survival prediction of lung cancer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 239–249. Springer (2024)
14. Okiyama, N., Tanaka, R.: Immune-related adverse events in various organs caused by immune checkpoint inhibitors. Allergology International **71**(2), 169–178 (2022)
15. Passaro, A., Brahmer, J., Antonia, S., Mok, T., Peters, S.: Managing resistance to immune checkpoint inhibitors in lung cancer: treatment and novel strategies. Journal of Clinical Oncology **40**(6), 598–610 (2022)
16. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
17. Qu, L., Huang, D., Zhang, S., Wang, X.: Multi-modal data binding for survival analysis modeling with incomplete data and annotations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 501–510. Springer (2024)
18. Saeed, N., Ridzuan, M., Maani, F.A., Alasmawi, H., Nandakumar, K., Yaqub, M.: Survrnc: Learning ordered representations for survival prediction using rank-n-contrast. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 659–669. Springer (2024)
19. Sharma, P., Goswami, S., Raychaudhuri, D., Siddiqui, B.A., Singh, P., Nagarajan, A., Liu, J., Subudhi, S.K., Poon, C., Gant, K.L., et al.: Immune checkpoint therapy—current perspectives and future directions. Cell **186**(8), 1652–1669 (2023)
20. Silva, L.A.V., Rohr, K.: Pan-cancer prognosis prediction using multimodal deep learning. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). pp. 568–571. IEEE (2020)
21. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **71**(3), 209–249 (2021)
22. Villegas, R., Babaeizadeh, M., Kindermans, P.J., Moraldo, H., Zhang, H., Saffar, M.T., Castro, S., Kunze, J., Erhan, D.: Phenaki: Variable length video generation from open domain textual description. arXiv preprint arXiv:2210.02399 (2022)
23. Wang, J., Chen, N., Guo, J., Xu, X., Liu, L., Yi, Z.: Survnet: a novel deep neural network for lung cancer survival analysis with missing values. Frontiers in Oncology **10**, 588990 (2021)

24. Wolf, T.N., Pölsterl, S., Wachinger, C., Initiative, A.D.N., et al.: Daft: A universal module to interweave tabular data and 3d images in cnns. NeuroImage **260**, 119505 (2022)
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. Yan, J., Chen, J., Wu, Y., Chen, D.Z., Wu, J.: T2g-former: organizing tabular features into relation graphs promotes heterogeneous feature interaction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 10720–10728 (2023)