Enhancing WSI-Based Survival Analysis with Report-Auxiliary Self-Distillation

Zheng Wang[1], Hong Liu[1], Zheng Wang[1,2], Danyi Li[3], Min Cen[4], Baptiste Magnier[5,6], Li Liang[3], and Liansheng Wang[1(✉)]

[1] School of Informatics, Xiamen University, Xiamen, China
{zhengwang, liuhong, zwang}@stu.xmu.edu.cn, lswang@xmu.edu.cn
[2] Shanghai Innovation Institution, Shanghai, China
[3] Nanfang Hospital, Southern Medical University, Guangzhou, China
lidanyi26@163.com,lli@smu.edu.cn
[4] School of Artificial Intelligence and Data Science, University of Science and Technology of China, Hefei, China
cenmin0127@mail.ustc.edu.cn
[5] EuroMov Digital Health in Motion, Univ Montpellier, IMT Mines Ales, Ales, France
baptiste.magnier@mines-ales.fr
[6] Service de Médecine Nucléaire, Centre Hospitalier Universitaire de Nîmes, Université de Montpellier, Nîmes, France

Abstract. Survival analysis based on Whole Slide Images (WSIs) is crucial for evaluating cancer prognosis, as they offer detailed microscopic information essential for predicting patient outcomes. However, traditional WSI-based survival analysis usually faces noisy features and limited data accessibility, hindering their ability to capture critical prognostic features effectively. Although pathology reports provide rich patient-specific information that could assist analysis, their potential to enhance WSI-based survival analysis remains largely unexplored. To this end, this paper proposes a novel **R**eport-**a**uxiliary **s**elf-distill**a**tion (**Rasa**) framework for WSI-based survival analysis. First, advanced large language models (LLMs) are utilized to extract fine-grained, WSI-relevant textual descriptions from original noisy pathology reports via a carefully designed task prompt. Next, a self-distillation-based pipeline is designed to filter out irrelevant or redundant WSI features for the student model under the guidance of the teacher model's textual knowledge. Finally, a risk-aware mix-up strategy is incorporated during the training of the student model to enhance both the quantity and diversity of the training data. Extensive experiments carried out on our collected data (CRC) and public data (TCGA-BRCA) demonstrate the superior effectiveness of Rasa against state-of-the-art methods. Our code is available at https://github.com/zhengwang9/Rasa.

Keywords: Survival prediction · Multimodal learning · Whole slide image · Self-distillation · Mix-up augmentation.

1 Introduction

Survival analysis based on Whole Slide Images (WSIs) is crucial for evaluating cancer prognosis, as they offer detailed microscopic information essential for predicting patient outcomes. As the gold standard for cancer diagnosis and prognosis [18], WSIs capture critical features such as cellular structures, tumor microenvironment, and tissue phenotypes. However, the effectiveness of traditional WSI-based survival analysis is often limited by two major challenges. First, the ultra-high resolution of WSIs introduces a vast number of irrelevant and redundant features, compromising analysis accuracy. Second, the acquisition of large-scale, high-quality data faces significant obstacles, including the need to meticulously label samples, privacy concerns, and extended follow-up periods.

Previous studies have attempted to tackle the two challenges by employing data augmentation [15,21,25] or de-noising directly on WSIs and labels [4,18], while others have incorporated rich multimodal data to facilitate more sophisticated survival analysis [3,11,24]. More recently, descriptions generated by advanced large language models (LLMs) have been introduced to enhance WSI-related tasks [7,17]. Compared with them, pathology reports offer richer patient-specific information, including high-level semantic descriptions of key findings, which could significantly aid analysis. However, the potential of leveraging these reports to enhance WSI-based survival analysis remains largely unexplored. This motivates our investigation into utilizing pathology reports to help overcome the two distinct limitations (*i.e.*, noisy features and limited data accessibility).

To this end, we proposes a novel **R**eport-**a**uxiliary **s**elf-distill**a**tion (**Rasa**) framework for WSI-based survival analysis. First, to tackle the issue of noise (*e.g.*, unmatched content with WSIs) in pathology reports, we employ advanced LLMs to transfer the noisy raw texts into detailed, WSI-aligned textual descriptions with a carefully crafted task-specific prompt. Next, to facilitate the student model's focus on prognostically relevant information, we design a self-distillation pipeline that filters out irrelevant and redundant WSI features, guided by the teacher model's textual knowledge. Finally, we introduce a risk-aware mix-up strategy to enhance both data quantity and diversity during the student model training. Our contributions are summarized as follows:

1. We are the first to leverage pathology reports to improve WSI-based survival analysis, demonstrating the significant potential of integrating pathology reports with WSI analysis to advance computational pathology.
2. We develop a novel **R**eport-**a**uxiliary **s**elf-distill**a**tion (**Rasa**) framework to enhance WSI-based survival analysis by addressing two core challenges, *i.e.*, noisy features and limited data accessibility in Sec. 2.
3. We extensively evaluate our method on our collected data (CRC) and public data (TCGA-BRCA) in Sec. 3. The results demonstrate the superior performance of our method against the state-of-the-art (SOTA) methods.

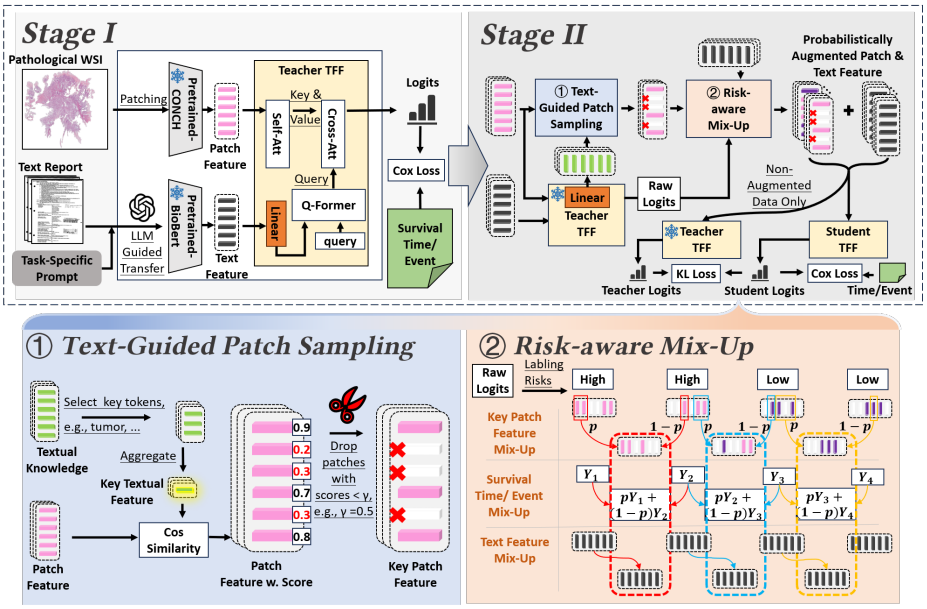

Fig. 1. The overview of Rasa framework.

2 Method

An overview of Rasa framework is available in the workflow presented in Fig. 1. It mainly includes two stages: Text-Fused Teacher Model Training and Tumor-Focused Student Model Training. In the first stage, we pre-process the raw data (*i.e.*, WSIs and pathology reports) into representative features and use them to train a teacher model. In the second stage, we employ the text-guided sampling module to filter noisy features and the risk-aware mix-up module to augment data. Subsequently, we train the student model under teacher model's guidance.

2.1 Data Processing

For each WSI $\mathcal{X}_i$, we use a pre-trained vision encoder of CONCH [16] to extract patch features $X_i = \{X_{ij}\}_{j=0}^{|X_i|}$. For each pathology report $\mathcal{T}_i$, we first use GPT-4 [1] to transfer it into detailed, WSI-aligned textual descriptions and then encode them into token features $T_i = \{T_{ij}\}_{j=0}^{|T_i|}$ by a pre-trained BioClinicalBert [23]. Specifically, we carefully designed a text prompt for GPT-4 [1] to emphasize the details of microscopic visual characteristics in WSIs and eliminate WSI-irrelevant information (*e.g.*, lymph node information, immunohistochemistry results, and certain genetic data). This facilitates the interaction between textual information and WSIs, as empirically evidenced by Table 2 and Fig. 4.

2.2 Text-Fused Teacher Model Training

After processing data, we use all extracted features (*i.e.*, patch feature X_i and text feature T_i) to pre-train a Text-Fused Former (TFF) model as the teacher (*i.e.*, Stage I in Fig. 1). Concretely, T_i is converted into the teacher's textual knowledge $T_i^{(proj)}$ via a projector (*i.e.*, the orange Linear in Fig. 1) and refined with a Q-Former [14]. For X_i, we directly encode it via a self-attention module. By using T_i' as the query and X_i' as the key and value, we compute cross-attention between T_i' and X_i' to summarize the task-essential information, which is pooled and passed through the head for predictions. The training objective is presented in Eq. (4), and the forward process of the TFF model is outlined as follows:

$$\begin{aligned} T_i' = \text{Q-Former}(T_i^{(proj)}), \quad T_i^{(proj)} = \text{Linear}(T_i), \quad X_i' = \text{self-attn}(X_i), \\ y_i = \text{head}(\text{Pool}(Z_i)), \quad Z_i = \text{cross-attn}(T_i', X_i', X_i'). \end{aligned} \tag{1}$$

2.3 Tumor-Focused Student Model Training

Text-Guided Patch Sampling: While the extremely high resolution of WSIs benefits survival analysis by offering rich information, it also introduces numerous irrelevant and redundant patches that can distract the model and compromise its performance. Since pathologists typically draw conclusions based on lesion regions (*i.e.*, cancerous areas) in clinical practice [6], we propose to filter out noise in WSIs by leveraging the teacher model's textual knowledge (*e.g.*, Block ① in Fig. 1). First, we select token features $\mathcal{S}_i = \{T_{ij}^{(proj)} | j \text{ is selected}\}$ from the teacher's textual knowledge $T_i^{(proj)}$ corresponding to manually specified keywords (*e.g.*, "tumor" or "cancer") in tokenized input texts. Next, we average selected features into the key textual feature $T_i^{(key)} = \frac{1}{|\mathcal{S}_i|} \sum_{j \in \mathcal{S}_i} T_{ij}^{(proj)}$. Finally, we filter out patches dissimilar to the key textual feature using cosine similarity:

$$X_i^{(key)} = \left\{ X_{ij} \mid \frac{X_{ij} \cdot T_i^{(key)}}{\|X_{ij}\|_2 \|T_i^{(key)}\|_2} \geq \gamma, \quad X_{ij} \in X_i \right\}, \tag{2}$$

where γ is a pre-defined threshold. This strategy effectively filters out a large number of noisy patches by retaining only key patches strongly associated with tumor regions, as shown in Fig. 4, and it also enhances efficiency by eliminating time-consuming, expertise-dependent manual labeling of cancerous areas.

Risk-aware Mix-up: To tackle the challenge of limited data accessibility, we consider employing mix-up-based data augmentation to improve the quantity and diversity of the training data. However, directly mixing up pairs of samples (*e.g.*, WSIs, labels and reports) without recognizing their risks could potentially yield misleading fusion. For example, it would be improper to associate a mixed WSI that contains high-risk patches with a low-risk text report. To this end, we first use the pre-trained teacher model to label the samples' risk $r_i = \mathbb{I}(s_i \geq s_{\text{medium}})$, where $s_i = \text{sigmoid}(y_i)$ and s_{medium} is the medium number of $\{s_i\}_{i=1}^N$

over the training set. Then, we respectively mix each sample's text features T_a, key patch features $X_a^{(key)}$, and labels Y_a with another one's (*e.g.*, $(T_b, X_b^{(key)}, Y_b)$) into a new sample $(T_{\text{mix}}, X_{\text{mix}}, Y_{\text{mix}})$ with a probability p_{aug}. For the text feature, we directly set $T_{\text{mix}} = T_a$ for pairs of samples with homogeneous risks (*e.g.*, high-high and low-low risks) while using $T_{\text{mix}} = T_{\text{argmax}(r_a, r_b)}$ for pairs of samples with heterogeneous risks (*e.g.*, high-low risks). This is because high-risk samples often provide more critical characteristics than low-risk ones in survival analysis. For patch features, we randomly select $100 * p_{\text{mix}}\%$ from $X_a^{(key)}$ and $100 * (1 - p_{\text{mix}})\%$ from $X_b^{(key)}$ and combine them into X_{mix}. Compared with previous bag mix-up strategies [4,15], we only mix the key patches selected by the textual knowledge, as mixing low-information patches causes low efficiency in increasing data diversity. For the labels $Y = (c, t)$ –*e.g.*, the censoring status c and survival time t– we adopt a soft-mixing strategy to ensure that the mixed label accurately reflects the contribution of both participants, as is computed below:

$$c_{mix} = (1 - p_{mix})c_a + p_{mix}c_b, \quad t_{mix} = (1 - p_{mix})t_a + p_{mix}t_b. \tag{3}$$

2.4 Training Procedure

The task objective is to minimize the Cox loss [27] as defined below:

$$\mathcal{L}_{cox} = -\sum_{i=1}^{N} \delta_i \left(y_i - \log \sum_{j \in R(t_i)} e^{y_j} \right), \tag{4}$$

where y_i represents the output of the model, and $R(t_i)$ is the risk set at time t_i. Besides, we introduce Kullback-Leibler (KL) divergence [20] to leverage the teacher model to guide the student model on non-augmented samples as:

$$\mathcal{L}_{KL} = D_{KL}(y_{i,student} \| y_{i,teacher}), \tag{5}$$

as the teacher model might yield unreliable results on unseen augmented samples. The student's objective $\mathcal{L}_{cox} + \lambda \mathcal{L}_{KL}^{(non-aug)}$ enables it to additionally learn from the teacher's refined knowledge, where λ is set to balance the two objectives.

3 Experiment

3.1 Experimental Settings

Datasets: The experiments were conducted on a Colorectal Cancer (CRC) cohort comprising 302 cases collected from a collaborating hospital, and a publicly available Breast Invasive Carcinoma cohort from The Cancer Genome Atlas (TCGA-BRCA) [9], consisting of 331 cases. We employed a 5-fold cross-validation where the train/validation/test ratio is 0.6/0.2/0.2 within each trial.

Table 1. The performance of our model compared with SOTA methods.

Type	Method	CRC	TCGA-BRCA
Vision-only	ABMIL [10]	$0.5132_{\pm 0.0982}$	$0.6368_{\pm 0.0437}$
	PatchGCN [2]	$0.5474_{\pm 0.1144}$	$0.6372_{\pm 0.0611}$
	TransMIL [19]	$0.5348_{\pm 0.0787}$	$0.5934_{\pm 0.0232}$
	DSMIL [13]	$0.5234_{\pm 0.1256}$	$0.6284_{\pm 0.0509}$
	MambaMIL [26]	$0.5416_{\pm 0.0954}$	$0.6366_{\pm 0.0149}$
Vision & Text	QPMIL-VL [7]	$0.5748_{\pm 0.0733}$	$0.5826_{\pm 0.0517}$
	TOP [17]	$0.5488_{\pm 0.0647}$	$0.5434_{\pm 0.0291}$
	MCAT [8]	$0.5592_{\pm 0.1020}$	$0.6198_{\pm 0.0520}$
Bag Mix-up	PseMix [15]	$0.5824_{\pm 0.1018}$	$0.6500_{\pm 0.0432}$
	RankMix [4]	$0.5262_{\pm 0.1349}$	$0.6120_{\pm 0.0803}$
	Rasa (Ours)	$\mathbf{0.6834}_{\pm 0.1331}$	$\mathbf{0.6972}_{\pm 0.0500}$

Metric: We use the Concordance Index (CI) [22] as the metric to measure the performance in predicting survival outcomes. We fairly report the averaged testing results of the models that optimally perform on the validation set.

Implementation: We use Adam optimizer to train the model [12] for 60 epochs with a fixed batch size of 8 and a learning rate of 1×10^{-5}. λ is optimally tuned to 1×10^{-2} and 1×10^{-5}, respectively for CRC and TCGA-BRCA datasets.

3.2 Comparison with SOTA Methods

We compare our methods with three types of baselines: *i)* vision-only models [10,2,19,13,26], *ii)* vision-language (VL) models [7,17,3], and *iii)* bag mix-up strategies [4,15]. The results in Table 1 suggest the superiority of our approach on both datasets in survival prediction tasks. First, vision-only models achieve moderate performance, revealing the limitations of relying solely on noisy visual information for capturing nuanced details essential for accurate survival prediction. Second, vision-language (VL) models show only marginal improvements on CRC and a slight decline on TCGA-BRCA against vision-only models. We attribute this to the simplistic textual information (*e.g.*, generic class names and GPT-generated descriptions) used by these VL models, lacking sufficient slide-specific details for further improvement. Third, among the bag mix-up augmentation strategies, PseMix (Pseudo-bag mixup augmentation [15]) stands out as the best-performing baseline, demonstrating the effectiveness of mix-up techniques in improving model performance. The poor performance of RankMix [4] may be due to its heavy reliance on the ability of the teacher model. Finally, our Rasa achieves the highest performance on both datasets, highlighting the effectiveness of generating slide-specific textual descriptions and integrating them into the decision process.

Table 2. Results of textual variates

Config	CRC	TCGA-BRCA
w/o Text	$0.5566_{\pm 0.1137}$	$0.6362_{\pm 0.0335}$
'Tumor cells'	$0.6020_{\pm 0.0988}$	$0.6618_{\pm 0.0735}$
GPT Text	$0.6450_{\pm 0.0823}$	$0.6636_{\pm 0.0407}$
Original Report	$0.6394_{\pm 0.0973}$	$0.6564_{\pm 0.0224}$
CONCH Text-Encoder	$0.6178_{\pm 0.1529}$	$0.6354_{\pm 0.0532}$
Ours	$\mathbf{0.6834}_{\pm 0.1331}$	$\mathbf{0.6972}_{\pm 0.0500}$

Table 3. Ablation study on sub-modules

Config	CRC	TCGA-BRCA
Teacher Model	$0.6196_{\pm 0.1336}$	$0.6494_{\pm 0.0269}$
w/o Loss KL & Mix-up	$0.6678_{\pm 0.1663}$	$0.6618_{\pm 0.0620}$
w/o Mix-up	$0.6746_{\pm 0.1465}$	$0.6750_{\pm 0.0510}$
w/o Loss KL	$0.6726_{\pm 0.1316}$	$0.6908_{\pm 0.0430}$
Ours	$\mathbf{0.6834}_{\pm 0.1331}$	$\mathbf{0.6972}_{\pm 0.0500}$

3.3 Ablation Studies

Impact of Text: To validate our designed textual modality, we tested various alternatives in Table 2: no text (w/o Text), "Tumor cells", GPT-generated descriptions (GPT Text), original pathology reports (Original Report), and CONCH text embeddings (CONCH Text-Encoder). Removing text (*i.e.*, w/o text) yields the worst results while using simple "Tumor cells" shows non-trivial improvement, highlighting the indispensable importance of texts. GPT Text achieves suboptimal performance by offering rich textual context. Although Original Report introduces more patient-specific information than GPT Text while similarly offering context, its effectiveness is limited by noise. CONCH text embeddings also underperform our method that uses BioClinicalBert. Our approach, leveraging LLMs for precise, contextually aligned text descriptions, achieves the best performance, demonstrating the effectiveness of advanced text processing in pathological image analysis. These findings underscore the importance of sophisticated text integration for accurate and robust survival prediction.

Effect of Sub-Module: We evaluate the impact of each sub-module in Rasa through ablation studies in Table 3. The results show that directly using the teacher model yields the poorest performance. Removing either the Mix-up module or the teacher guidance from $\mathcal{L}_{KL}$ leads to a performance drop, while removing both results in a more significant decline. This indicates the effectiveness of each sub-module and the importance of their collaborative interaction.

Effect of Hyper-Parameters: We investigate the impact of p_{aug} (*i.e.*, the augmentation probability), λ (*i.e.*, the coefficient of $\mathcal{L}_{KL}$) in Fig. 2. For p_{aug}, a lower value may lead to insufficient diversity of data, while a higher one could

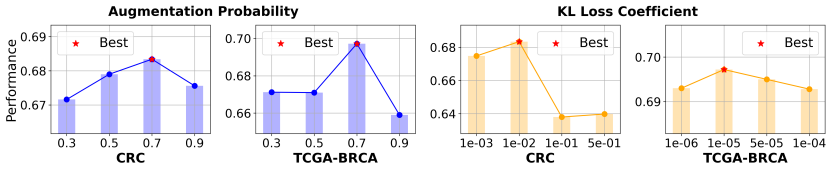

Fig. 2. Impact of the augmentation probability p_{aug} and the coefficient λ

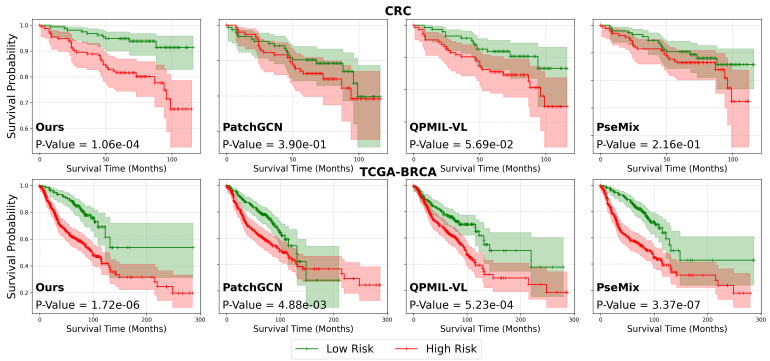

Fig. 3. Kaplan-Meier curves for predicted high-risk (red) and low-risk (green) groups on the test sets of the CRC (top) and TCGA-BRCA (bottom) datasets

introduce excessive noise, potentially degrading the model's performance. The optimal value of p_{aug} is 0.7 for both the CRC and TCGA-BRCA datasets. For λ, optimal performance was achieved with values of 1×10^{-2} and 1×10^{-5} for the CRC and TCGA-BRCA datasets, respectively.

3.4 Visualization

Kaplan–Meier Analysis: We further validate the effectiveness of our method via Kaplan-Meier (KM) curves in Fig. 3. We follow [5] to divide patients in the test set into low-risk and high-risk groups based on the median risk score from the training set. The statistical significance of the survival time differences between these groups was evaluated using the log-rank test, with a p-value below 0.01 indicating statistical significance. Compared with other SOTA methods, our method achieved remarkably low p-values on both datasets while demonstrating a clear and robust separation between low-risk and high-risk groups.

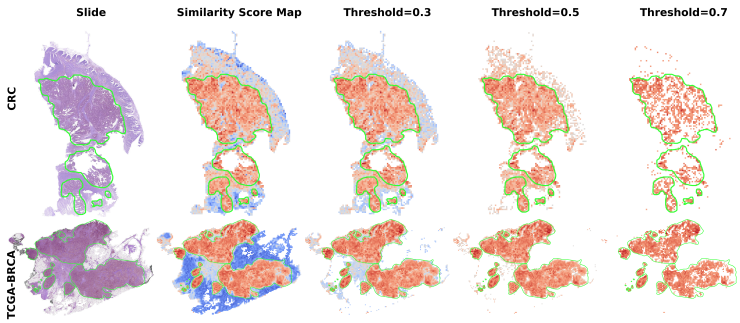

Fig. 4. Text-patch similarity score maps and preserved patches with different γ.

Effect of Threshold: The text-patch similarity score map with different values of γ (the patch sampling filter threshold) is shown in Fig. 4. Regions with high text-patch similarity predominantly overlap with cancerous areas (*i.e.*, green-lined regions), demonstrating that the teacher model's refined text features effectively align the embeddings of text keywords with patch embeddings. For another thing, the impact of varying thresholds on patch filtering is evident, as increasing the threshold retains more cancerous patches but introduces noisy patches. We selected a threshold of 0.5 in experiments, as it optimally balances the retention of cancerous regions with minimal inclusion of non-cancerous patches.

4 Conclusion

In this paper, we propose a **R**eport-**a**uxiliary **s**elf-distill**a**tion (**Rasa**) framework to address two core challenges in WSI-based survival analysis: noisy features and limited data accessibility. By leveraging advanced LLMs and carefully designing modules, we successfully enhanced WSI-based survival analysis with the assistance of reports. Extensive experiments on CRC and TCGA-BRCA datasets confirm the superiority of Rasa. We plan to fully release the power of the report-auxiliary data enhancement technique in more WSI analysis tasks in the future.

Acknowledgments. This work was supported by National Natural Science Foundation of China (Grant No. 62371409) and Fujian Provincial Natural Science Foundation of China (Grant No. 2023J01005).

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

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Chen, R.J., Lu, M.Y., Shaban, M., Chen, C., Chen, T.Y., Williamson, D.F., Mahmood, F.: Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. 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. 339–349. Springer (2021)
3. 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)
4. Chen, Y.C., Lu, C.S.: Rankmix: Data augmentation for weakly supervised learning of classifying whole slide images with diverse sizes and imbalanced categories. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23936–23945 (2023)

5. Cheng, J., Mo, X., Wang, X., Parwani, A., Feng, Q., Huang, K.: Identification of topological features in renal tumor microenvironment associated with patient survival. Bioinformatics **34**(6), 1024–1030 (2018)
6. Connolly, J.L., Schnitt, S.J., Wang, H.H., Longtine, J.A., Dvorak, A., Dvorak, H.F.: Role of the surgical pathologist in the diagnosis and management of the cancer patient. In: Holland-Frei Cancer Medicine. 6th edition. BC Decker (2003)
7. Gou, J., Ji, L., Liu, P., Ye, M.: Queryable prototype multiple instance learning with vision-language models for incremental whole slide image classification. In: Proceedings of the AAAI conference on artificial intelligence (2025), https://arxiv.org/abs/2410.10573
8. Han, M., Qu, L., Yang, D., Zhang, X., Wang, X., Zhang, L.: Mscpt: Few-shot whole slide image classification with multi-scale and context-focused prompt tuning. arXiv preprint arXiv:2408.11505 (2024)
9. Hutter, C., Zenklusen, J.C.: The cancer genome atlas: creating lasting value beyond its data. Cell **173**(2), 283–285 (2018)
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. 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)
12. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
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, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
15. Liu, P., Ji, L., Zhang, X., Ye, F.: Pseudo-bag mixup augmentation for multiple instance learning-based whole slide image classification. IEEE Transactions on Medical Imaging (2024)
16. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature Medicine **30**(3), 863–874 (2024)
17. Qu, L., Fu, K., Wang, M., Song, Z., et al.: The rise of ai language pathologists: Exploring two-level prompt learning for few-shot weakly-supervised whole slide image classification. Advances in Neural Information Processing Systems **36** (2024)
18. Shao, W., Wang, T., Huang, Z., Han, Z., Zhang, J., Huang, K.: Weakly supervised deep ordinal cox model for survival prediction from whole-slide pathological images. IEEE Transactions on Medical Imaging **40**(12), 3739–3747 (2021)
19. 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)
20. Van Erven, T., Harremos, P.: Rényi divergence and kullback-leibler divergence. IEEE Transactions on Information Theory **60**(7), 3797–3820 (2014)
21. Wang, H., Luo, L., Wang, F., Tong, R., Chen, Y.W., Hu, H., Lin, L., Chen, H.: Rethinking multiple instance learning for whole slide image classification: A bag-level classifier is a good instance-level teacher. IEEE Transactions on Medical Imaging (2024)

22. Wang, P., Li, Y., Reddy, C.K.: Machine learning for survival analysis: A survey. ACM Computing Surveys (CSUR) **51**(6), 1–36 (2019)
23. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Huggingface's transformers: State-of-the-art natural language processing (2020), https://arxiv.org/abs/1910.03771
24. 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)
25. Xiong, C., Lin, Y., Chen, H., Zheng, H., Wei, D., Zheng, Y., Sung, J.J., King, I.: Takt: Target-aware knowledge transfer for whole slide image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 503–513. Springer (2024)
26. Yang, S., Wang, Y., Chen, H.: Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 296–306. Springer (2024)
27. 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)