TMSE: Tri-Modal Survival Estimation with Context-aware Tissue Prototype and Attention-Entropy Interaction

Ruofan Zhang[1,2], Mengjie Fang[2], Shengyuan Liu[3], Zipei Wang[1,2], Jie Tian✉[2,4], and Di Dong✉[1,2]

[1] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[2] CAS Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3] The Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, China
[4] Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beihang University, Beijing, China
tian@ieee.org, di.dong@ia.ac.cn

Abstract. Survival prediction plays a crucial role in clinical decision-making, enabling personalized treatments by integrating multi-modal medical data, such as histopathology images, pathology reports, and genomic profiles. However, the heterogeneity across these modalities and the high dimensionality of Whole Slide Images (WSI) make it challenging to capture survival-relevant features and model their interactions. Existing methods, typically focused on single-modal WSI, fail to leverage multimodal information, such as expert-driven pathology reports, and struggle with the computational complexity of WSI. To address these issues, we propose a novel Tri-Modal Survival Estimation framework (TMSE), which includes three components: (1) Pathology report processing pipeline, curated with expert knowledge, with both the pipeline and the processed structured report being publicly available; (2) Context-aware Tissue Prototype (CTP) module, which uses Mamba and Gaussian mixture models to extract compact, survival-relevant features from WSI, reducing redundancy while preserving histological details; (3) Attention-Entropy Interaction (AEI) module, a attention mechanism enhanced with entropy-based optimization to align and fuse three modalities: WSI, pathology reports, and genomic data. Extensive evaluation on three TCGA datasets (BLCA, BRCA, LUAD) shows that our approach achieves superior performance in survival prediction. Data and code are available: https://github.com/RuofanZhang8/TMSE

Keywords: Multi-modal learning · Survival Prediction · Heterogeneous Biomedical Data.

1 Introduction

Survival prediction is vital in clinical decision-making, offering prognostic insights for personalized treatment and advancing precision medicine by integrating machine learning with multi-modal medical data [23]. As precision medicine evolves, cancer tissue pathology slides are digitized into Whole Slide Images (WSI) with expert reports, while patient genomic profiles are increasingly collected, enhancing survival analysis [1]. Integrating this multi-modal data is crucial to improving the accuracy and robustness of prognostic predictions despite complex clinical relationships [15, 27].

Analyzing multi-modal information presents significant challenges due to the substantial heterogeneity [4, 5, 16] across modalities such as histopathology images, pathology reports, and genomic data. A common difficulties is capturing the complex relationships between these modalities. Although each modality contains rich information, identifying survival-relevant features from the vast amounts of data they contain is difficult for effective integrated survival prediction [24]. Previous studies on multi-modal fusion largely focused on pairing histopathology images with genomic data, using techniques like cross-attention [3] or optimal transport [29]. However, these methods often face challenges when models simulate the interactions of thousands of WSI patch tokens and other modality information to make patient-level predictions. It is difficult for the model to extract critical interaction information from the large number of WSI tokens, and the overwhelming amount of WSI data can easily obscure information from other modalities, posing an extremely difficult challenge. Furthermore, pathology reports, which offer expert-driven descriptive insights [14], represent a valuable complement to image data and are readily accessible. However, earlier research often neglects this expert-driven information or lacks efficient methods to integrate pathology reports with clinical knowledge, limiting their potential utility in multi-modal analysis.

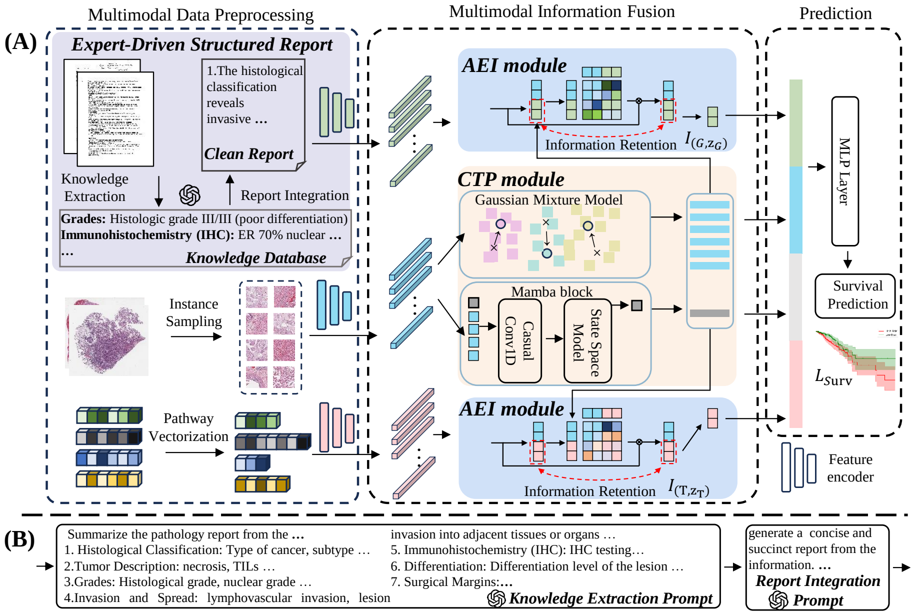
To address these challenges, we propose the Tri-Modal Survival Estimation framework (TMSE). First, we design a pathology report processing pipeline that integrates expert knowledge to efficiently and uniformly extract information from pathology reports for survival prediction assistance. Second, we propose a Context-aware Tissue Prototype (CTP) module that integrates Mamba architecture and Gaussian Mixture Model (GMM) for efficient tissue representation learning. The GMM component compresses similar tissue patches into a compact set of prototype representations, effectively reducing the redundant information caused by high-resolution WSI sampling. Meanwhile, the Mamba architecture models contextual relationships among thousands of WSI patch tokens, capturing survival-relevant tumor microenvironment (TME) features that emerge from large-scale tissue interactions. This information cannot be adequately represented by GMM's tissue-level prototypes alone. This dual approach efficiently condenses extensive patch-level data into concise yet comprehensive feature representations, effectively addressing the "needle in a haystack" challenge in pathological image analysis while facilitating robust multi-modal feature alignment and fusion. Third, we propose a Attention-Entropy Interaction (AEI) module to

facilitate interactions and fusion between tri-modal information. We decouple the information interaction in AEI into an information interaction mechanism and an information retention mechanism, helping the model avoid being overwhelmed by the large volume of WSI data. This enhancing multi-modal information fusion efficiency and improving the model's capacity to extract meaningful insights from diverse modalities. Our contributions can be summarized as follows:
1. We utilize an expert-knowledge-based prompt chain pipeline to organize approximately ten thousand TCGA pathology reports using large language model (LLM). The refined reports will enhance survival analysis and be made publicly available to facilitate further research.
2. We propose the CTP module, which addresses WSI redundancy through prototype-based representation learning, enabling efficient extraction and fusion of critical histological features.
3. We develop the AEI module, which mitigates multimodal alignment challenges through entropy-based fusion, achieving effective integration of WSI, pathology reports, and genomic data.
4. We validate the effectiveness of our model on three TCGA datasets: BLCA, BRCA, and LUAD, demonstrating the superior performance of our method.

2 Methods

2.1 Overview and MultiModal Data Preprocessing

We present TMSE in Fig. 1(A), a Tri-Modal Survival Estimation framework for survival prediction. It consists of two main modules: the first efficiently represents histological features using the CTP (Section 2.1), and the second aligns and fuses information across modalities with the AEI (Section 2.2). Our model innovatively leverages information from three modalities. Below, we explain how patient-level tri-modal feature representations are constructed from WSI, expert-driven pathology reports, and genomic data.
WSI Given a WSI of one patient, we crop it into N_p patches at 20× magnification with a size of 256×256 by CLAM [18], denoted as $\{x_i\}_{i=1}^{N_w}$. Each patch image x_i is then encoded into a feature representation w_i using a pre-trained pathology foundation model PLIP [9]. The resulting representation is $W = \{w_i\}_{i=1}^{N_w}$.
Genomics The genomic data consists of transcriptomic profiles from bulk RNA sequencing, tokenized into biological pathway vectors [10]. The resulting gene expression vectors are organized as $G = \{g_i\}_{i=1}^{N_g}$, where $g_i \in \mathbb{R}^{N_{c,g}}$ represents the expression vector for the g-th pathway, consisting of $N_{c,g}$ genes.
Pathology Report Pathology reports, often encumbered by extraneous details and a lack of structure, pose challenges to unified feature encoding. To address this, we organize these reports using LLM (DeepSeek-V3 [17]), employing a expert-knowledge-based, chain-of-thought pipeline to extract critical information and refine it into concise diagnostic reports. Our preprocessing pipeline consists of two key steps in Fig. 1(B). First, guided by clinical expertise, we devise a expert-base question prompt, which leverages seven predefined, domain-specific



1.The histological classification reveals invasive ...

Grades: Histologic grade III/III (poor differentiation)
Immunohistochemistry (IHC): ER 70% nuclear ...
...

Summarize the pathology report from the ...
1. Histological Classification: Type of cancer, subtype ...
2.Tumor Description: necrosis, TILs ...
3.Grades: Histological grade, nuclear grade ...
4.Invasion and Spread: lymphovascular invasion, lesion invasion into adjacent tissues or organs ...
5. Immunohistochemistry (IHC): IHC testing...
6. Differentiation: Differentiation level of the lesion ...
7. Surgical Margins:...

generate a concise and succinct report from the information. ...

Fig. 1. (A) TMSE Overview: A tri-modal alignment pipeline featuring report preprocessing, WSI prototype extraction with the CTP module, and coupled AEI modules for information interaction. (B) Report Preprocessing Prompts: Extract report details using expert knowledge and consolidate them into a unified report representation.

questions. This approach directs the LLM to focus on and extract multifaceted information from the reports, constructing a patient-level knowledge database. Second, we aggregate this patient-level knowledge database into a standardized, unified report, mitigating the risk of overlooking essential details that might occur when models directly process raw reports. We have applied this workflow to process all TCGA reports, openly sharing our pipeline and results to facilitate further research advancements. The refined report is then embedded using BiomedBERT [8], resulting in a feature $T = \{t_i\}, t_i \in \mathbb{R}^{d_T}$.

Dimension Matching To merge features from different modalities, we align their token dimensions to a uniform length d. A linear projection layer adjusts the WSI token representations $W \in \mathbb{R}^{N_w \times d}$ and pathology reports $T \in \mathbb{R}^{1 \times d}$, while self-normalizing neural networks (SNN) [13] map variable-length genomics data via pathway-specific functions $f_{c,g}$ to $G \in \mathbb{R}^{N_g \times d}$.

2.2 Context-aware Tissue Prototype

In this section, we introduce a Context-aware Tissue Prototype (CTP) for extracting comprehensive WSI representations from local to global scales. The framework employs a GMM to derive representative local histological feature embeddings, coupled with a Mamba block to capture survival-relevant patterns across extensive WSI patches. We ultimately extract the prototype as the result of combining two methods: $\text{Proto}_{\text{CTP}} = \text{Cat}(\text{Proto}_{\text{GMM}}, \text{Proto}_{\text{Mam}})$.

GMM-Based Prototype Aggregation We compress N_w WSI patches into C_w prototypes ($C_w \ll N_w$), preserving patch distribution by aggregating similar regions and emphasizing distinct tissue types using GMM [12, 21]. For a WSI, we model patch feature w_i distribution with a probabilistic framework:

$$p(\mathbf{w}_i;\theta) = \sum_{c=1}^{C_w} \pi_c \cdot \mathcal{N}(\mathbf{w}_i;\mu_c,\varSigma_c), \quad \text{s.t.} \sum_{c=1}^{C_w} \pi_c = 1, \tag{1}$$

where $\theta = \{\pi_c, \mu_c, \varSigma_c\}$ defines mixture probability, mean, and covariance. We interpret these as tissue morphology descriptors—similar regions share mean and variance, while π_c reflects tissue type proportions. Prototypes are formed as $\text{Proto}_{\text{GMM}} = [\pi_c, \mu_c, \varSigma_c]$. We optimize θ by initializing μ_c with k-means cluster centers, setting $\pi_c = 1/C_p$ and $\varSigma_c = I$, then maximizing $\sum_{i=1}^{N_w} \log p(\mathbf{w}_i;\theta)$ via expectation maximization.

Contextual Mamba Prototype To model the TME information within tissue, it is essential to capture the contextual information across a large number of tissue blocks. For the feature sequence of WSI $W \in \mathbb{R}^{N_w \times d}$, we prepend a class token $t_{\text{cls}} \in \mathbb{R}^d$ to represent the entire slide patch sequence, forming $W' = [t_{\text{cls}}, W]$. After normalizing, we input the concatenated sequence into the Mamba [7] module as follows:

$$Y = \text{SSM}\left(\text{SiLU}\left(\text{Conv1D}\left(\text{Linear}\left(W'\right)\right)\right)\right) \tag{2}$$

Finally, we extract the class token from the output sequence Y as Mamba's prototype, $\text{Proto}_{\text{Mam}} = Y^{cls}$, which represents the global sequence information derived from the long feature sequence.

2.3 Attention-Entropy Interaction

In this section, we introduce the Attention-Entropy Interaction (AEI) module, which integrates attention-based interaction for bimodal alignment and fusion with entropy-based information retention. This takes into account both the fusion and preservation of multimodal information. Since WSI contain extensive information (from macro to micro details), we separately align report data (descriptive information) and genomic profiles (micro-level information) with WSI.

Cross-Modal Attention Alignment In our approach, we replace the whole WSI patch token used in traditional attention mechanisms with CTP prototypes to compute interactive attention, improving efficiency and reducing computational complexity. For queries, we introduce learnable matrices $\mathbf{W}_Q$ and define $\mathbf{Q} = \left(\mathbf{Q}_W^{\mathrm{T}}, \mathbf{Q}_{\mathrm{G}}^{\mathrm{T}}\right)^{\mathrm{T}} = \left(\mathrm{Proto}_{\mathrm{CTP}}, G\right)^{\mathrm{T}} \mathbf{W}_Q$, and similarly for $\mathbf{K}$ and $\mathbf{V}$. The interactive attention is expressed as [28]:

$$\begin{pmatrix} \mathbf{Z}_W \\ \mathbf{Z}_G \end{pmatrix} = \sigma\left(\frac{1}{\sqrt{d}}\begin{pmatrix} \mathbf{Q}_W\mathbf{K}_W^{\mathrm{T}} & \mathbf{Q}_W\mathbf{K}_G^{\mathrm{T}} \\ \mathbf{Q}_G\mathbf{K}_W^{\mathrm{T}} & \mathbf{Q}_G\mathbf{K}_G^{\mathrm{T}} \end{pmatrix}\right)\begin{pmatrix} \mathbf{V}_W \\ \mathbf{V}_G \end{pmatrix} \tag{3}$$

where $\sigma(\cdot)$ denotes the softmax operation. We treat $\mathbf{Z}_G$ as the fused pathway information features after the interaction. Using the same approach, we can obtain the fused report information $\mathbf{Z}_T$.

Entropy-based Information Retention In this section, we propose using mutual information entropy to preserve information during the cross-modal attention alignment process. Specifically, we exclude WSI information when computing the entropy, aiming to preserve the information from the report data and genomic profiles. We apply matrix-based Rényi's α-order entropy functional [6,20] to estimate information-theoretic measures. For gene pathway data, we use pre-AEI G and post-AEI Z_G, selecting N sample pairs $\{G^m, Z_G^m\}_{m=1}^N$ from a mini-batch. Self-information and mutual Information are computed using the normalized eigenspectrum of a Gram matrix $A_G = K_G / \operatorname{tr}(K_G)$, where $K_G(m,n) = k\left(G^m, G^n\right)$ and k is a Gaussian kernel.

$$H_\alpha\left(A_G\right) = \frac{1}{1-\alpha}\log_2\left(\operatorname{tr}\left(A_G^\alpha\right)\right) \quad H_\alpha\left(A_G, A_{Z_G}\right) = H_\alpha\left(\frac{A_G \circ A_{Z_G}}{\operatorname{tr}\left(A_G \circ A_{Z_G}\right)}\right) \tag{4}$$

We set $\alpha = 1.01$ to approximate Shannon entropy, and $\circ$ denotes the element-wise product. Finally, the matrix-based Rényi's α-order mutual information entropy $I_\alpha\left(G; Z_G\right)$ used for information retention, is defined as:

$$I_\alpha\left(G; Z_G\right) = H_\alpha\left(A_G\right) + H_\alpha\left(A_{Z_G}\right) - H_\alpha\left(A_G, A_{Z_G}\right) \tag{5}$$

For textual information, $I\left(T; Z_T\right)$ can be calculated in the same manner.

2.4 Training Strategy for Survival Prediction

For a patient, the final multimodal feature is derived by applying layer normalization and mean pooling to the CTP prototypes, post-AEI text feature, and post-AEI genomic feature. For survival prediction, we employ the Negative Log-Likelihood (NLL) loss function [30], as $\mathcal{L}_{\mathrm{surv}}$. The total training loss for end-to-end model training is then calculated with weight factors α:

$$\mathcal{L} = \mathcal{L}_{\mathrm{surv}} + \alpha_1\mathcal{L}_{I(G;Z_G)} + \alpha_2\mathcal{L}_{I(T;Z_T)}. \tag{6}$$

Table 1. C-Index (mean ± std) performance over three cancer datasets. The best results are shown in **bold**, and the second best ones are underlined. G: use gene, W: use WSI, and T: use pathology report.

Model	Mod.	BLCA	BRCA	LUAD	Avg.↑
SNN [13]	G	0.649±0.061	0.694±0.094	0.623±0.035	0.655
CLAM-MB [18]	W	0.655±0.026	0.604±0.052	0.594±0.046	0.618
TransMIL [19]	W	0.647±0.041	0.606±0.054	0.627±0.420	0.627
MHIM [25]	W	0.620±0.031	0.572±0.085	0.593±0.054	0.595
R^2T-MIL [26]	W	0.651±0.036	0.605±0.063	0.611±0.026	0.622
PANTHER [21]	W	0.612±0.039	0.616±0.071	0.581±0.072	0.603
MCAT [3]	G+W	0.668±0.065	0.723±0.080	0.627±0.055	0.673
MOTCAT [29]	G+W	0.673±0.064	0.715±0.052	0.617±0.041	0.668
CMTA [31]	G+W	0.665±0.044	0.710±0.081	0.661±0.035	0.679
MMP [22]	G+W	0.674±0.052	0.716±0.066	0.607±0.026	0.666
Pathomic [2]	G+W+T	0.657±0.043	0.727±0.028	0.614±0.063	0.666
TMSE(Ours)	G+W+T	**0.687±0.057**	**0.738±0.067**	**0.682±0.057**	**0.702**

3 Experiments

Datasets and Evaluation Metrics We apply our method to three cancer survival datasets from the TCGA database, including Breast invasive carcinoma (BRCA) (n= 875), Bladder Urothelial Carcinoma (BLCA) (n= 328), and Lung Adenocarcinoma (LUAD) (n= 401). These datasets provide WSI, genomic data, patient pathology reports [11], and overall survival (OS) times. We assess the model performance using the concordance index (C-Index).

Implementation Details In each experiment, we perform 5-fold cross validation. All models are trained for 40 epochs with a learning rate of 10^{-4}, using a cosine decay scheduler, AdamW optimizer with weight decay of 10^{-5}. We set the batch size to 1 and the number of bins to 4 in the NLL loss setting. The Gaussian prototype number is set to $C_{\mathrm{w}} = 16$.

3.1 Comparison Results

We compare our model with the current state-of-the-art models in Tab. 1. All multimodal fusion models utilize both WSI and gene data, and we maintain this setup. Due to the scarcity of tri-modal survival prediction methods, we adapt an existing tri-modal fusion approach Pathomic [2] for comparison.

Our TMSE model outperforms all baseline methods, achieving best performance across three datasets, demonstrating its effectiveness and generalizability to various cancer datasets. By integrating three modalities, TMSE significantly outperforms unimodal models. Furthermore, compared to the contrasted tri-modal model, our model greatly enhances the efficiency of interactions. Kaplan-Meier analysis in Fig. 2 further confirms TMSE's superior patient risk stratification, enhancing its p-value over the second-best model.

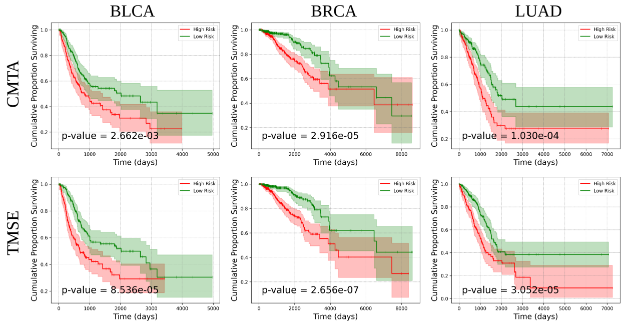


Fig. 2. Kaplan-Meier analysis for TMSE and CMTA on the three datasets.

Table 2. Ablation results of our method.

Module	Mod.	BLCA	BRCA	LUAD	Avg.↑
w/o Proto.	W+G+T	0.663±0.068	0.690±0.076	0.671±0.043	0.675
w/o Mamba	W+G+T	0.684±0.060	0.728±0.039	0.678±0.060	0.697
w/o IR	W+G+T	0.668±0.053	0.713±0.044	0.671±0.043	0.684
w/o AEI	W+G+T	0.681±0.062	0.677±0.053	0.682±0.051	0.680
TMSE	W	0.605±0.030	0.580±0.066	0.600±0.074	0.595
TMSE	W+T	0.637±0.052	0.608±0.041	0.668±0.051	0.638
TMSE	W+G	0.675±0.052	0.716±0.044	0.648±0.055	0.680
TMSE(Ours)	W+G+T	**0.687±0.057**	**0.738±0.067**	**0.682±0.057**	**0.702**

3.2 Ablation Studies

Effect of Prototypes We replace prototypes in the AEI module with all patch tokens, while retaining prototypes for the final survival prediction to preserve the integrity of the architecture. A C-index drop across three datasets shows that compressed prototypes boost information interaction efficiency. We drop the Mamba module, keeping only GMM-based prototypes. The C-index decreases, hinting at limitations in GMM-based feature extraction. Mamba-derived tokens, capturing broader context, prove vital for the task.

Effect of Different Modules in AEI We explore the tradeoff in AEI between information retention and fusion by removing information retention (IR) while preserving the attention mechanism, using only the fused information, or completely removing the AEI module and relying on the original features. For BLCA and LUAD, dropping information retention significantly lowers performance, while using the original feature causes a smaller decline, indicating the value of retaining multimodal information for survival prediction. Conversely, BRCA shows an opposite pattern. Our AEI module optimally balances information fusion and information retention, delivering the best performance overall.

Effect of Multimodal Fusion We test our model by excluding certain modalities. Using only WSI data reduces performance, while adding text and gene profiles enhances it, showing the benefit of multi-modal integration. The highest performance occurs when all three modalities are combined, validating our three-modal fusion method.

4 Conclusion

We propose TMSE, a tri-modal survival analysis method integrating WSI, genomic profiles, and pathology reports. It uses expert-driven report extraction, a GMM-Mamba hybrid for WSI features, and entropy-combined attention to fusion modality information. Extensive tests on three cancer datasets show that our method outperforms other methods in OS survival prediction.

Acknowledgments. This work was supported by the National Key R&D Program of China (2023YFC2415200), National Natural Science Foundation of China (82441018, U24A20759, 82372053, 82361168664, 82302296), Beijing Natural Science Foundation (JQ24048), Scientific and Technological Innovation Project of China Academy of Chinese Medical Sciences (CI2023C008YG), Science and Technology Development Fund of Macao Special Administrative Region (0006/2023/AFJ).

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

References

1. Acosta, J.N., Falcone, G.J., Rajpurkar, P., Topol, E.J.: Multimodal biomedical ai. Nature medicine **28**(9), 1773–1784 (2022)
2. 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)
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. Dong, D., Fang, M.J., Tang, L., Shan, X.H., Gao, J.B., Giganti, F., Wang, R.P., Chen, X., Wang, X.X., Palumbo, D., et al.: Deep learning radiomic nomogram can predict the number of lymph node metastasis in locally advanced gastric cancer: an international multicenter study. Annals of oncology **31**(7), 912–920 (2020)
5. Dong, D., Tang, L., Li, Z.Y., Fang, M.J., Gao, J.B., Shan, X.H., Ying, X.J., Sun, Y.S., Fu, J., Wang, X.X., et al.: Development and validation of an individualized nomogram to identify occult peritoneal metastasis in patients with advanced gastric cancer. Annals of Oncology **30**(3), 431–438 (2019)
6. Giraldo, L.G.S., Rao, M., Principe, J.C.: Measures of entropy from data using infinitely divisible kernels. IEEE Transactions on Information Theory **61**(1), 535–548 (2014)

7. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
8. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH) **3**(1), 1–23 (2021)
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. 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)
11. Kefeli, J., Tatonetti, N.: Tcga-reports: A machine-readable pathology report resource for benchmarking text-based ai models. Patterns **5**(3) (2024)
12. Kim, M.: Differentiable expectation-maximization for set representation learning. In: International Conference on Learning Representations (2022)
13. Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S.: Self-normalizing neural networks. Advances in neural information processing systems **30** (2017)
14. Li, H., Chen, Y., Chen, Y., Yu, R., Yang, W., Wang, L., Ding, B., Han, Y.: Generalizable whole slide image classification with fine-grained visual-semantic interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11398–11407 (2024)
15. Lin, D.F., Li, H.L., Liu, T., Lv, X.F., Xie, C.M., Ou, X.M., Guan, J., Zhang, Y., Yan, W.B., He, M.L., et al.: Radiomic signatures associated with tumor immune heterogeneity predict survival in locally recurrent nasopharyngeal carcinoma. JNCI: Journal of the National Cancer Institute **116**(8), 1294–1302 (2024)
16. Lipkova, J., Chen, R.J., Chen, B., Lu, M.Y., Barbieri, M., Shao, D., Vaidya, A.J., Chen, C., Zhuang, L., Williamson, D.F., et al.: Artificial intelligence for multimodal data integration in oncology. Cancer cell **40**(10), 1095–1110 (2022)
17. Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024)
18. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)
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. Shi, J., Tang, L., Gao, Z., Li, Y., Wang, C., Gong, T., Li, C., Fu, H.: Mg-trans: Multi-scale graph transformer with information bottleneck for whole slide image classification. IEEE Transactions on Medical Imaging **42**(12), 3871–3883 (2023)
21. Song, A.H., Chen, R.J., Ding, T., Williamson, D.F., Jaume, G., Mahmood, F.: Morphological prototyping for unsupervised slide representation learning in computational pathology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11566–11578 (2024)
22. Song, A.H., Chen, R.J., Jaume, G., Vaidya, A.J., Baras, A., Mahmood, F.: Multimodal prototyping for cancer survival prediction. In: Forty-first International Conference on Machine Learning (2024)

23. Song, A.H., Jaume, G., Williamson, D.F., Lu, M.Y., Vaidya, A., Miller, T.R., Mahmood, F.: Artificial intelligence for digital and computational pathology. Nature Reviews Bioengineering **1**(12), 930–949 (2023)
24. Steyaert, S., Pizurica, M., Nagaraj, D., Khandelwal, P., Hernandez-Boussard, T., Gentles, A.J., Gevaert, O.: Multimodal data fusion for cancer biomarker discovery with deep learning. Nature machine intelligence **5**(4), 351–362 (2023)
25. Tang, W., Huang, S., Zhang, X., Zhou, F., Zhang, Y., Liu, B.: Multiple instance learning framework with masked hard instance mining for whole slide image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4078–4087 (2023)
26. Tang, W., Zhou, F., Huang, S., Zhu, X., Zhang, Y., Liu, B.: Feature re-embedding: Towards foundation model-level performance in computational pathology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11343–11352 (2024)
27. Wang, Z., Fang, M., Zhang, J., Tang, L., Zhong, L., Li, H., Cao, R., Zhao, X., Liu, S., Zhang, R., et al.: Radiomics and deep learning in nasopharyngeal carcinoma: a review. IEEE Reviews in Biomedical Engineering **17**, 118–135 (2023)
28. Xu, P., Zhu, X., Clifton, D.A.: Multimodal learning with transformers: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(10), 12113–12132 (2023)
29. Xu, Y., Chen, H.: Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 21241–21251 (2023)
30. Zadeh, S.G., Schmid, M.: Bias in cross-entropy-based training of deep survival networks. IEEE transactions on pattern analysis and machine intelligence **43**(9), 3126–3137 (2020)
31. 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)