Knowledge-Enhanced Representation Learning with Retrieval-Augmented Multimodal Fusion for Survival Prediction

Zeyu Zhang[1], Puria Azadi[1], Behnam Maneshgar[1], Ali Khajegili Mirabadi[1], Nikolay Alabi[1], Hossein Farahani[1,2], and Ali Bashashati[1,2(✉)]

[1] University of British Columbia, Vancouver, BC, Canada
[2] Jointly supervised the study
Correspondence: ali.bashashati@ubc.ca

Abstract. Multimodal survival prediction using whole-slide images (WSIs) and transcriptomics has shown promise in cancer prognostication, yet remains fundamentally limited by two challenges: (1) substantial disease-irrelevant redundancy within each modality, and (2) unstable discriminative signals caused by intratumoral heterogeneity. Existing multimodal methods primarily focus on cross-modal interaction modeling but lack mechanisms to explicitly suppress redundancy and stabilize clinically meaningful representations. To address these challenges, we propose **KERA**, a **K**nowledge-**E**nhanced **R**etrieval-**A**ugmented multimodal framework for survival prediction. KERA introduces a two-stage training strategy. First, a pan-cancer pretraining paradigm (ARPnet) performs modality alignment while explicitly retaining disease-relevant information under cancer-type supervision, enhancing informative representations while alleviating redundancy. Second, a retrieval-augmented fusion module (RAFnet) enriches patient representations by incorporating contextually similar intra- and inter-modality features, mitigating representation instability caused by sparse malignant patterns. Across five independent TCGA cohorts, KERA achieves a mean C-index of 0.716, yielding a 3.17% relative improvement over the strongest multimodal baseline, with statistically significant gains (Wilcoxon signed-rank test, $p < 0.05$). These results demonstrate that structured knowledge-enhanced pretraining combined with retrieval augmentation provides a principled solution to redundancy and heterogeneity in multimodal survival modeling. The code is available at `https://github.com/AIMLab-UBC/KERA`.

Keywords: Multimodality · Retrieval augmentation · Survival analysis.

1 Introduction

Cancer survival analysis, aiming to estimate the relative risk of patient death, has been widely investigated due to its significant clinical value in prognostic assessment and treatment planning [2]. Over the past decade, advances in digital pathology and sequencing techniques have enabled multimodal integration

for cancer prognostication, providing complementary information across distinct perspectives. However, survival prediction is a multifactorial task influenced by heterogeneous biological determinants, while multimodal clinical data often exhibit limited discriminative signals and substantial redundancy. For instance, gigapixel WSIs typically have extremely high resolutions (up to $100{,}000 \times 100{,}000$ pixels) but only a small fraction of regions are associated with tumor progression [21]. Likewise, transcriptomics data are inherently high-dimensional and contain substantial redundancy, impeding the extraction of prognostically relevant information and resulting in suboptimal performance [3].

Previous studies primarily focused on modeling complex associations between pathological morphological features and gene expression patterns [1,4]. However, these methods lack dedicated constraints to mitigate the redundant information irrelevant to the disease, such as normal tissue in pathological WSIs or molecular alterations unrelated to tumor progression, which may dominate the learned representation and obscure clinically meaningful information. Recently, several studies have attempted to decouple the multimodal knowledge into modality-shared and modality-specific components to better characterize cross-modal interactions while reducing redundancy [5,6,7,8,9]. For instance, MurreNet [5] employs feature alignment to capture modality-common knowledge and Kullback-Leibler divergence to enforce feature disentanglement. DRIM [6] proposes a disentangled multimodal framework based on mutual information maximization and adversarial decoupling for cancer survival prediction. Nevertheless, the contextual information within WSIs and transcriptomics data is inherently unstable due to the spatial intratumoral heterogeneity. Moreover, available medical datasets are often limited in scale, further hindering the robustness of multimodal representations [2]. Importantly, we hypothesize that the limitation of current multimodal survival prediction methods may arise not only from insufficient cross-modal interaction modeling, but also from unstable and overly redundant underlying representations. Without explicitly modeling shared biological signals and disease-discriminative patterns, multimodal fusion may amplify noise rather than enhance prognosis. As a result, these models may struggle to capture the subtle yet clinically meaningful attributes, such as scattered malignant foci in WSIs or rare genetic abnormalities in genomic profiles.

In this paper, we propose a Knowledge-Enhanced multimodal Retrieval-Augmented framework (KERA) for cancer survival prediction, consisting of modality Alignment and Retention-based Pre-training network (ARPnet) and Retrieval-Augmented multimodal Fusion network (RAFnet). Our contributions are threefold: (1) We introduce ARPnet, a knowledge-enhanced pan-cancer pretraining paradigm that facilitates modality alignment while preserving disease-relevant information, thereby enhancing the clinically meaningful knowledge and reducing feature redundancy. (2) We propose RAFnet, a retrieval-augmented multimodal fusion module that performs structured intra- and inter-modality retrieval in a computationally efficient manner, contributing to more robust and informative representations. (3) The proposed framework provides a scalable pipeline for multimodal analysis that can be extended to other data modalities

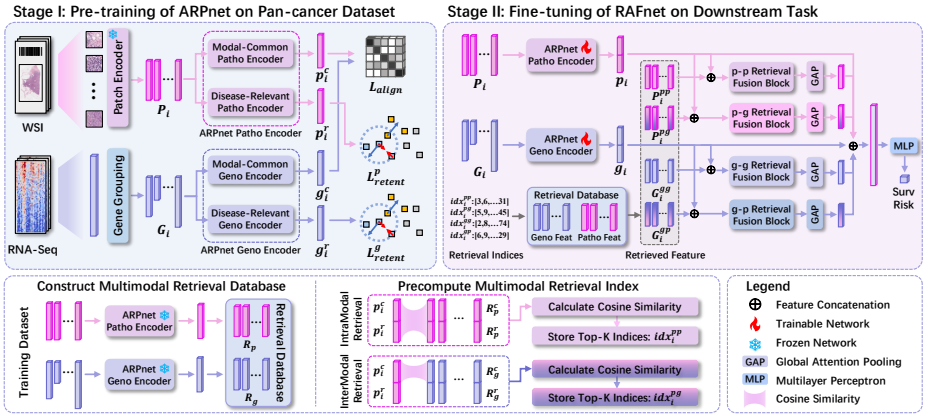

Fig. 1. Overview of the proposed KERA framework, including two stages: (I) Pre-training of ARPnet on pan-cancer dataset and (II) Fine-tuning of RAFnet on downstream task. After Stage I, the pretrained ARPnet is utilized as a feature extractor to construct a multimodal retrieval database, and the retrieval indices are precomputed prior to model fine-tuning in Stage II, as illustrated in the lower-left panel.

and clinical tasks. Extensive experiments demonstrate its superior performance in survival prediction, further validating the effectiveness of knowledge-enhanced representation learning and retrieval-augmented fusion strategies.

2 Method

KERA adopts a two-stage training paradigm, as illustrated in Fig. 1. In Stage I, ARPnet is pretrained on a pan-cancer dataset to jointly perform multimodal alignment and disease-relevant information retention across diverse cancer types. The pretrained ARPnet is then frozen and used to extract modality-common and disease-relevant embeddings for constructing a multimodal retrieval database with precomputed similarity-based retrieval indices. In Stage II, the pretrained ARPnet is integrated with RAFnet, a retrieval-augmented fusion module that incorporates similar multimodal features to enhance the robustness of representation. The unified framework is then fine-tuned for cancer survival prediction.

2.1 Feature Alignment and Retention based Pre-training

In Stage I, each pathological WSI is partitioned into non-overlapping 224×224 image patches at $20\times$ magnification. Each patch is encoded using UNI [10], a pathology foundation model, to obtain a one-dimensional feature vector. For the i-th patient, the pathological feature is denoted as $P_i \in \mathbb{R}^{M\times d}$, where M is the number of patches and d is the feature dimension. A learnable [CLS] token is then concatenated with the pathological features and fed into two parallel

TransMIL-based encoders: a modality-common pathology encoder and a disease-relevant pathology encoder. The output [CLS] tokens are denoted as $p_i^c \in \mathbb{R}^{1\times d}$ and $p_i^r \in \mathbb{R}^{1\times d}$, respectively. For the gene expression data (RNA-seq), we follow previous study [11,12] and group genes into six functional categories: (1) Tumor Suppression, (2) Oncogenesis, (3) Protein Kinases, (4) Cellular Differentiation, (5) Transcription, (6) Cytokines and Growth. For the i-th patient, the RNA sequencing data are represented as $G_i = \{g_i^0, g_i^1, ...g_i^5\}$. Similar to the pathology branch, a modality-common genomic encoder and a disease-relevant genomic encoder are adopted. Each encoder is implemented using MLP layers with global attention pooling to unify the dimensionality across functional groups and extract global genomic representations, denoted as $g_i^c \in \mathbb{R}^{1\times d}$ and $g_i^r \in \mathbb{R}^{1\times d}$.

In order to mitigate the modality discrepancies and capture shared information between morphological features and genomic characteristics, we employ a CLIP-based contrastive loss [13] to align the modality-common pathological feature p_i^c and genomic feature g_i^c, which can be formulated as:

$$\mathcal{L}_{\text{align}} = -\frac{1}{2B}\sum_{i=1}^{B}\log\frac{\exp\left(\frac{p_i^{c\top} g_i^c}{\tau_a}\right)}{\sum_{j=1}^{B}\exp\left(\frac{p_i^{c\top} g_j^c}{\tau_a}\right)} - \frac{1}{2B}\sum_{i=1}^{B}\log\frac{\exp\left(\frac{g_i^{c\top} p_i^c}{\tau_a}\right)}{\sum_{j=1}^{B}\exp\left(\frac{g_i^{c\top} p_j^c}{\tau_a}\right)} \quad (1)$$

where B is the batch size, and τ_a is a temperature parameter. However, overemphasizing modality-common features may obscure disease-specific discriminative signals, while the information redundancy remains unsolved. To further address these limitations, we introduce a metric learning-based disease-relevant information retention module that employs adaptive sparse pairwise (AdaSP) loss [14,15] with cancer-type supervision. Specifically, the AdaSP loss pulls together features from the same cancer type while separating features from different cancer types in an adaptive sparse manner, thereby enhancing discriminative disease-relevant representations while suppressing redundant information. Notably, a pan-cancer dataset is curated for model pre-training, where each batch contains K random cancer types with N patients per type. Taking the pathological disease-relevant information retention loss $\mathcal{L}_{\text{retent}}^p$ as an example, it can be formulated as:

$$\mathcal{L}_{\text{retent}}^p = \frac{1}{K}\sum_{i=1}^{K}\log\left(1+\exp\left(\frac{S_{pi}^- - S_{pi}^+}{\tau_r}\right)\right) \quad (2)$$

$$S_{pi}^+ = \max_m \min_n \cos(p_m^{ri}, p_n^{ri}) \approx \tau_r \log\left(\sum_{m=1}^{N}\frac{1}{\sum_{n=1}^{N}\exp\left(-\frac{\cos(p_m^{ri}, p_n^{ri})}{\tau_r}\right)}\right) \quad (3)$$

$$S_{pi}^- = \max_{j,m,n} \cos(p_m^{ri}, p_n^{rj}) \approx \tau_r \log\left(\sum_{m=1}^{N}\sum_{j=1,j\neq i}^{K}\sum_{n=1}^{N}\exp\left(\frac{\cos(p_m^{ri}, p_n^{rj})}{\tau_r}\right)\right) \quad (4)$$

where $\cos(\cdot)$ denotes cosine similarity and τ_r is a temperature parameter. p_m^{ri} represents the m-th disease-relevant pathological feature of cancer type i. S_{pi}^+ and

S_{pi}^{-} denote the similarities of the least-hard positive pair and the hardest negative pair, respectively. Intuitively, S_{pi}^{-} captures the largest inter-class similarity (i.e., the smallest separation between different cancer types), while S_{pi}^{+} indicates the least-hard intra-class similarity, preventing dominance of large intra-class variations induced by redundant information and preserving discriminative features relevant to cancer types. The overall ARPnet pretraining loss is defined as:

$$\mathcal{L}_{\text{ARPnet}} = \alpha\mathcal{L}_{\text{align}} + \beta_1\mathcal{L}_{\text{retent}}^{p} + \beta_2\mathcal{L}_{\text{retent}}^{g} \tag{5}$$

where α, β_1 and β_2 are hyperparameters.

2.2 Retrieval-Augmented Multimodal Fusion

Prognostically relevant signals in pathological WSIs and transcriptomic data are inherently unstable and limited due to severe intratumoral heterogeneity. To address this issue, we introduce RAFnet, a retrieval-augmented multimodal fusion module that enhances input representations by retrieving similar contextual information through both intra-modality and inter-modality retrieval. Specifically, we first employ the pretrained ARPnet to encode multimodal features and construct a retrieval database composed of pathology features $R_p = Concat(R_p^c; R_p^r)$ and genomic features $R_g = Concat(R_g^c; R_g^r)$, where $R_p, R_g \in \mathbb{R}^{H\times 2d}$, H denotes the number of training samples. To prevent data leakage, only samples from the training split are utilized for retrieval database construction.

To facilitate efficient feature retrieval, patient-specific top-k retrieval indices based on cosine similarity are precomputed and directly accessed during model fine-tuning. Notably, intra-modality retrieval is performed using concatenated modality-common and disease-relevant features, whereas inter-modality retrieval relies solely on modality-common features to maintain cross-modality consistency. For instance, given a pathological feature $p_i = Concat(p_i^c, p_i^r) \in \mathbb{R}^{1\times 2d}$, the corresponding intra- and inter-modality retrieval indices are computed by:

$$idx_i^{pp} = TopK(\cos(p_i, R_p), k) \tag{6}$$

$$idx_i^{pg} = TopK(\cos(p_i^c, R_g^c), k) \tag{7}$$

where k denotes the number of retrieved features, idx_i^{pp}, $idx_i^{pg} \in \mathbb{N}^k$ denote the indices of the top-k retrieved pathological and genomic features, respectively.

In Stage II, we combine the pretrained ARPnet with RAFnet for survival prediction. Specifically, for the i-th patient, the patch-level features P_i and genomic expression data G_i are first fed into ARPnet to obtain the global pathological representation p_i and genomic representation g_i. Meanwhile, the corresponding retrieval indices $\{idx_i^{pp}, idx_i^{pg}, idx_i^{gg}, idx_i^{gp}\}$ are accessed to retrieve the top-k most similar features from the retrieval database, which includes pathological features $P_i^{pp}, P_i^{gp} \in \mathbb{R}^{k\times 2d}$ retrieved from R_p and genomic features $G_i^{gg}, G_i^{pg} \in \mathbb{R}^{k\times 2d}$ retrieved from R_g. We adopt an attention-based retrieval fusion block to incorporate complementary information from the retrieved features.

Taking the pathological feature p_i and its intra-modality retrieved feature P_i^{pp} as an example, the attention-based fusion operation can be formulated as:

$$F_i^{pp} = \text{Softmax}\left(\frac{([p_i; P_i^{pp}] \cdot W_Q)([p_i; P_i^{pp}] \cdot W_K)^\top}{\sqrt{2d}}\right)([p_i; P_i^{pp}] \cdot W_V) \quad (8)$$

where Softmax$(\cdot)$ denotes the row-wise softmax function. W_Q, W_K, and W_V are learnable parameters. $F_i^{pp} \in \mathbb{R}^{(1+k)\times 2d}$ denotes the fused query–retrieval representation. The similar attention-based fusion is applied to obtain F_i^{pg}, F_i^{gg} and F_i^{gp}. Subsequently, the retrieval components (e.g., $F_i^{pp}[1{:},:]$) are aggregated via global attention pooling and concatenated with p_i and g_i for survival risk estimation. Negative log-likelihood (NLL) loss [16] is used for model fine-tuning.

3 Experiments

3.1 Experimental settings

Dataset We collected 5,369 samples across 24 cancer types with paired pathological WSIs and RNA-seq data from The Cancer Genome Atlas (TCGA) for pretraining. To prevent data leakage, five additional cancer types were strictly excluded from pan-cancer pretraining and reserved for downstream fine-tuning, including Glioblastoma Multiforme & Lower Grade Glioma (GBMLGG, $n = 654$), Bladder Urothelial Carcinoma (BLCA, $n = 353$), Breast Invasive Carcinoma (BRCA, $n = 983$), Uterine Corpus Endometrial Carcinoma (UCEC, $n = 486$), and Lung Adenocarcinoma (LUAD, $n = 401$). To obtain a more accurate assessment of disease-related outcomes, we use disease-specific survival (DSS) data from the UCSC Xena database [25] for survival prediction.

Implementation Details ARPnet was pretrained on the pan-cancer dataset for 10 epochs using AdamW (weight decay 1×10^{-5}) with a learning rate of 1×10^{-4} and a batch size of 32. The hyperparameters α, β_1, and β_2 were set to 0.2, 0.3, and 0.1, respectively. The pretrained ARPnet was then integrated with RAFnet and fine-tuned on cancer-specific cohorts for 5 epochs with a learning rate of 1×10^{-4} and a batch size of 1 with 32 gradient accumulation steps. The retrieval number k was set to 10. We adopted 5-fold cross-validation and used concordance index (C-index) for model evaluation. Kaplan–Meier analysis with log-rank test was used to assess differences between high- and low-risk groups.

3.2 Comparisons with State-of-the-art Methods

We compared our method with multiple baselines, including genomic uni-modal baselines: MLP, SNN [17], and SMLP [18]; pathology uni-modal baselines: ABMIL [19], TMIL [20], DeepMISL [22], and CLAM-SB [21]; and multimodal baselines: Porpoise [23], MCAT [24], SurvPath [1], MOTcat [4], and PathOmics [12].

Table 1. Comparison of C-index (mean ± std) with other baselines. The best performance is shown in red and the second-best in blue. * indicates $p < 0.05$ (Wilcoxon).

	Method	GBMLGG	BLCA	BRCA	UCEC	LUAD	Mean
Geno	SNN	0.826±0.024	0.526±0.068	0.585±0.059	0.665±0.063	0.533±0.014	0.627*
	MLP	0.702±0.155	0.527±0.080	0.585±0.041	0.478±0.085	0.537±0.084	0.566*
	SMLP	0.820±0.025	0.633±0.036	0.609±0.081	0.695±0.093	0.573±0.038	0.666*
Patho	ABMIL	0.808±0.036	0.576±0.076	0.673±0.062	0.696±0.070	0.581±0.046	0.667*
	TMIL	0.809±0.035	0.597±0.080	0.693±0.059	0.709±0.060	0.601±0.026	0.682*
	DeepMISL	0.810±0.032	0.603±0.077	0.674±0.071	0.750±0.069	0.588±0.038	0.685*
	CLAM-SB	0.807±0.037	0.594±0.063	0.636±0.050	0.698±0.106	0.603±0.059	0.668*
Multimodal	Porpoise	0.830±0.027	0.548±0.057	0.641±0.041	0.717±0.067	0.566±0.038	0.660*
	MCAT	0.806±0.024	0.608±0.044	0.621±0.042	0.687±0.131	0.611±0.049	0.667*
	SurvPath	0.819±0.030	0.583±0.035	0.634±0.061	0.714±0.111	0.554±0.060	0.661*
	MOTcat	0.817±0.032	0.637±0.044	0.675±0.054	0.740±0.098	0.601±0.024	0.694*
	PathOmics	0.816±0.027	0.594±0.040	0.667±0.035	0.731±0.062	0.624±0.033	0.686*
	KERA	0.832±0.023	0.653±0.043	0.701±0.043	0.759±0.077	0.635±0.060	0.716

Table 2. Ablation study of KERA across five TCGA cancer cohorts.

Config	GBMLGG	BLCA	BRCA	UCEC	LUAD	Mean
w/o L_{retent}	0.814±0.028	0.541±0.061	0.686±0.059	0.649±0.082	0.546±0.036	0.647
w/o L_{align}	0.807±0.033	0.578±0.039	0.607±0.080	0.612±0.102	0.564±0.067	0.634
w/o RA	0.824±0.034	0.578±0.054	0.688±0.074	0.727±0.063	0.566±0.066	0.677
w/o Pre-T	0.827±0.038	0.590±0.049	0.723±0.064	0.724±0.047	0.591±0.033	0.691
Ours	0.832±0.023	0.653±0.043	0.701±0.043	0.759±0.077	0.635±0.060	0.716

As shown in Table 1, our model achieves superior performance across all cohorts, with an average C-index of 0.716, surpassing the strongest multimodal (MOTcat) and uni-modal (DeepMISL) baselines by 3.17% and 4.53%, respectively. Notably, some uni-modal baselines (e.g., TMIL and DeepMISL) achieve slightly higher C-index values than certain multimodal approaches (e.g., Porpoise and MCAT), which suggests that integrating additional modalities does not necessarily improve the performance. Instead, the redundant features introduced from other modalities may even negatively impact prognostic prediction. To further assess the robustness of our results, the 5-fold C-index values were aggregated across all cohorts for each method, and the statistical significance of the differences between KERA and each baseline was evaluated using the Wilcoxon signed-rank test. All pairwise comparisons yielded P values < 0.05, indicating significant differences between KERA and the baselines. Overall, our model consistently outperforms other baselines, demonstrating the strength of KERA and its potential clinical utility for prognostic evaluation.

3.3 Ablation Study

We conducted ablation studies to systematically validate the impact of key components in KERA. We first evaluated the effectiveness of multimodal alignment and disease information retention losses in ARPnet pretraining. The re-

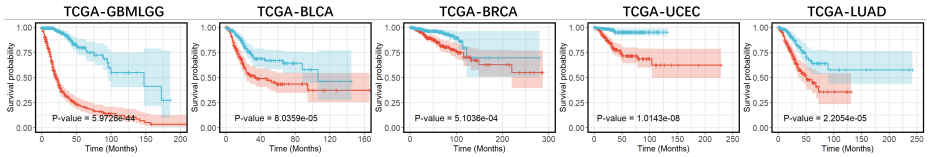

Fig. 2. Kaplan–Meier analysis and log-rank test of KERA on TCGA cohorts.

Table 3. Impact of different Retrieval Numbers (RN) in RAFnet.

RN	BLCA	LUAD
5	0.662±0.042	0.626±0.047
10	0.653±0.043	0.635±0.060
15	0.641±0.056	0.627±0.047
20	0.637±0.051	0.629±0.052
30	0.636±0.053	0.633±0.052

Table 4. Ablation study on feature Retrieval (R) across data modalities.

Modal	R	BLCA	LUAD
Geno	wo	0.579±0.047	0.542±0.081
	w	0.647±0.045	0.591±0.028
Patho	wo	0.540±0.058	0.579±0.052
	w	0.572±0.088	0.602±0.042
Ours	w	0.653±0.043	0.635±0.060

trieval module was excluded in these experiments since the pretrained modality-common or disease-relevant features required for retrieval were unavailable. As shown in Table 2, removing L_{retent} and L_{align} leads to 9.64% and 11.45% performance drops compared to our model, indicating that both modality-common and disease-relevant information are crucial for prognostic prediction. If both constraints are retained while the retrieval module is excluded (w/o RA), the average C-index reaches 0.677, which remains lower than the full model.

To assess the impact of pan-cancer pretraining, we retrained KERA from scratch without loading the pretrained ARPnet weights (w/o Pre-T). A clear performance drop is observed across most cohorts, with an average reduction of 3.49%. Notably, the w/o Pre-T baseline achieves slightly higher performance on the BRCA cohort. One possible reason is that BRCA is the largest fine-tuning cohort but was strictly excluded from pan-cancer pretraining to prevent data leakage, and the representations learned from other cancer types may not fully generalize to BRCA-specific characteristics. Overall, the ablation studies demonstrate the contributions of each loss term in ARPnet and the retrieval-augmented fusion module. Figure 2 shows the Kaplan–Meier curves across all cohorts. All log-rank p-values are below 0.05, indicating that KERA can effectively stratify patients into high- and low-risk groups with statistical significance.

We further evaluated the effect of the retrieval number. As shown in Table 3, the impact of the retrieval number varies across cancer types. For instance, in the BLCA cohort, increasing the retrieval number may negatively affect model performance, potentially due to the inclusion of unrelated information. In contrast, the LUAD cohort is less sensitive to the retrieval number. The ablation study on data modalities is presented in Table 4. We compared our method with the uni-modal baselines, both with and without feature retrieval. The results indicate that feature retrieval consistently improves the C-index in the uni-modal setting and the integration of both modalities further boosts performance.

4 Conclusion

In summary, we present KERA, a knowledge-enhanced retrieval-augmented framework that addresses two fundamental challenges in multimodal survival prediction: feature redundancy and representation instability. By explicitly modeling modality-common and disease-relevant signals during pan-cancer pretraining and stabilizing patient representations through structured retrieval augmentation, KERA moves beyond conventional cross-modal fusion strategies. Our findings suggest that robust survival modeling requires not only multimodal interaction learning, but principled representation structuring and contextual enrichment. This paradigm opens new directions for scalable multimodal modeling in oncology and other heterogeneous clinical domains.

Acknowledgments. This work is supported by funding from NSERC, CIHR, and MSFHR.

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

References

1. Jaume, G., Vaidya, A., Chen, R.J., Williamson, D.F.K., Liang, P.P., Mahmood, F.: Modeling Dense Multimodal Interactions Between Biological Pathways and Histology for Survival Prediction. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11579–11590. IEEE, Seattle, WA, USA (2024). https://doi.org/10.1109/CVPR52733.2024.01100
2. Song, A.H., Jaume, G., Williamson, D.F.K., Lu, M.Y., Vaidya, A., Miller, T.R., Mahmood, F.: Artificial intelligence for digital and computational pathology. Nat Rev Bioeng. 1, 930–949 (2023).
3. Swain, M.K., Kamila, N.K., Jena, L., Chithaluru, P., Albeshri, M.Y., Alsekait, D.M.: Hybrid deep learning framework for accurate classification of high dimensional genomic data. Sci Rep. 16, 5919 (2026).
4. Xu, Y., Chen, H.: Multimodal Optimal Transport-based Co-Attention Transformer with Global Structure Consistency for Survival Prediction. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 21184–21194. IEEE, Paris, France (2023). https://doi.org/10.1109/ICCV51070.2023.01942
5. Liu, M., Cai, C., Li, J., Xu, P., Li, J., Ma, J., Xu, J.: MurreNet: Modeling Holistic Multimodal Interactions Between Histopathology and Genomic Profiles for Survival Prediction. In: Gee, J.C., Alexander, D.C., Hong, J., Iglesias, J.E., Sudre, C.H., Venkataraman, A., Golland, P., Kim, J.H., and Park, J. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. pp. 396–406. Springer Nature Switzerland, Cham (2026). https://doi.org/10.1007/978-3-032-05182-0_39
6. Robinet, L., Berjaoui, A., Kheil, Z., Cohen-Jonathan Moyal, E.: DRIM: Learning Disentangled Representations from Incomplete Multimodal Healthcare Data. In: Linguraru, M.G., Dou, Q., Feragen, A., Giannarou, S., Glocker, B., Lekadir, K., and Schnabel, J.A. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 163–173. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-72384-1_16

7. Zhang, Y., Xu, Y., Chen, J., Xie, F., Chen, H.: Prototypical Information Bottlenecking and Disentangling for Multimodal Cancer Survival Prediction. arXiv preprint arXiv:2401.01646 (2024)
8. Wang, T., Fan, J., Zhang, D., Liu, D., Xia, Y., Huang, H., Cai, W.: MIRROR: Multi-Modal Pathological Self-Supervised Representation Learning via Modality Alignment and Retention. IEEE Transactions on Medical Imaging. 1–1 (2025).
9. Liu, H., Shi, Y., Xu, Y., Li, A., Wang, M.: Agnostic-Specific Modality Learning for Cancer Survival Prediction from Multiple Data. IEEE Journal of Biomedical and Health Informatics. 1–12 (2024).
10. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F.K., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., Williams, M., Oldenburg, L., Weishaupt, L.L., Wang, J.J., Vaidya, A., Le, L.P., Gerber, G., Sahai, S., Williams, W., Mahmood, F.: Towards a general-purpose foundation model for computational pathology. Nat Med. 30, 850–862 (2024).
11. Zhou, F., Chen, H.: Cross-Modal Translation and Alignment for Survival Analysis. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 21428–21437. IEEE, Paris, France (2023). https://doi.org/10.1109/ICCV51070.2023.01964
12. Ding, K., Zhou, M., Metaxas, D.N., Zhang, S. (2023). Pathology-and-Genomics Multimodal Transformer for Survival Outcome Prediction. In: Greenspan, H., et al. Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. MICCAI 2023. Lecture Notes in Computer Science, vol 14225. Springer, Cham. https://doi.org/10.1007/978-3-031-43987-2_60
13. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., et al.: Learning Transferable Visual Models From Natural Language Supervision. In: Proceedings of the 38th International Conference on Machine Learning. pp. 8748–8763. PMLR (2021).
14. Zhou, X., Zhong, Y., Cheng, Z., Liang, F., Ma, L.: Adaptive Sparse Pairwise Loss for Object Re-Identification. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19691–19701. IEEE, Vancouver, BC, Canada (2023). https://doi.org/10.1109/CVPR52729.2023.01886
15. Zhou, X., Zhang, X., Wu, C., Zhang, Y., Xie, W., Wang, Y.: Knowledge-Enhanced Visual-Language Pretraining for Computational Pathology. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., and Varol, G. (eds.) Computer Vision – ECCV 2024. pp. 345–362. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-72943-0_20
16. Zadeh, S.G., Schmid, M.: Bias in Cross-Entropy-Based Training of Deep Survival Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. 43, 3126–3137 (2021).
17. Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S.: Self-Normalizing Neural Networks. In: Advances in Neural Information Processing Systems. Curran Associates, Inc. (2017).
18. Elmarakeby, H.A., Hwang, J., Arafeh, R., Crowdis, J., Gang, S., Liu, D., AlDubayan, et al.: Biologically informed deep neural network for prostate cancer discovery. Nature. 598, 348–352 (2021).
19. Ilse, M., Tomczak, J., Welling, M.: Attention-based Deep Multiple Instance Learning. In: Proceedings of the 35th International Conference on Machine Learning. pp. 2127–2136. PMLR (2018).
20. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., Zhang, Y.: TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification. In: Advances in Neural Information Processing Systems. pp. 2136–2147. Curran Associates, Inc. (2021).

21. Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng. 5, 555–570 (2021).
22. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream Multiple Instance Learning Network for Whole Slide Image Classification with Self-supervised Contrastive Learning. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14313–14323. IEEE, Nashville, TN, USA (2021). https://doi.org/10.1109/CVPR46437.2021.01409
23. Chen, R.J., Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Lipkova, J., et al.: Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell. 40, 865-878.e6 (2022).
24. Chen, R.J., Lu, M.Y., Weng, W.H., Chen, T.Y., Williamson, D.F.K., Manz, T., Shady, M., Mahmood, F.: Multimodal Co-Attention Transformer for Survival Prediction in Gigapixel Whole Slide Images. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3995–4005. IEEE, Montreal, QC, Canada (2021). https://doi.org/10.1109/ICCV48922.2021.00398
25. Goldman, M.J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., Banerjee, A., Luo, Y., Rogers, D., Brooks, A.N., Zhu, J., Haussler, D.: Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechnol. 38, 675–678 (2020).