Multi-modal Knowledge Decomposition based Online Distillation for Biomarker Prediction in Breast Cancer Histopathology

Qibin Zhang[1], Xinyu Hao[1,2], Qiao Chen[2], Rui Xu[3], Fengyu Cong[1,2], Cheng Lu[4]✉, and Hongming Xu[1]✉

[1] School of Biomedical Engineering, Faulty of Medicine, Dalian University of Technology, Dalian, China
[2] Faculty of Information Technology, University of Jyvaskyla, Jyvaskyla, Finland
[3] School of Software Technology, Dalian University of Technology, Dalian, China
[4] Department of Radiology, Guangdong Provincial People's Hospital, Southern Medical University, Guangzhou, China
lucheng@gdph.org.cn; mxu@dlut.edu.cn

Abstract. Immunohistochemical (IHC) biomarker prediction benefits from multi-modal data fusion analysis. However, the simultaneous acquisition of multi-modal data, such as genomic and pathological information, is often challenging due to cost or technical limitations. To address this challenge, we propose an online distillation approach based on Multi-modal Knowledge Decomposition (MKD) to enhance IHC biomarker prediction in haematoxylin and eosin (H&E) stained histopathology images. This method leverages paired genomic-pathology data during training while enabling inference using either pathology slides alone or both modalities. Two teacher and one student models are developed to extract modality-specific and modality-general features by minimizing the MKD loss. To maintain the internal structural relationships between samples, Similarity-preserving Knowledge Distillation (SKD) is applied. Additionally, Collaborative Learning for Online Distillation (CLOD) facilitates mutual learning between teacher and student models, encouraging diverse and complementary learning dynamics. Experiments on the TCGA-BRCA and in-house QHSU datasets demonstrate that our approach achieves superior performance in IHC biomarker prediction using uni-modal data. Our code is available at https://github.com/qiyuanzz/MICCAI2025_MKD.

Keywords: Missing modality · Biomarker prediction · Histopathology.

1 Introduction

Tumor biomarkers, including estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2), play a critical role in breast cancer diagnosis, therapeutic decision-making, prognostic assessment, and disease monitoring [20]. However, determining biomarker status through IHC

staining is both time-consuming and costly [14]. With the increasing use of digital histopathology and rapid advancements in deep learning, predicting IHC biomarker status from H&E-stained whole slide images (WSIs) has emerged as a promising alternative [14,20]. This method not only offers a more efficient and cost-effective alternative to traditional IHC staining, but also holds significant potential for uncovering complex morphological features intricately linked to biomarker status [5].

In recent years, extensive research has focused on leveraging H&E-stained pathology images to predict tumor biomarker status. Early methods predominantly relied on fully supervised learning, which typically involved training patch-level classifiers based on annotations, with predictions aggregated across the entire WSI to infer patient-level biomarker status [9,14]. Since these methods required detailed pixel-level annotations by clinicians, they suffer from a time-intensive and laborious annotation process. The growing adoption of multiple instance learning (MIL) in computational pathology has driven a shift toward weakly supervised learning methods, offering a more efficient and scalable alternative. For instance, Lu et al. [13] proposed a novel method that utilizes graph convolutional neural networks to extract WSI-level representations for predicting HER2 status, enabling the model to capture the global biological geometric structure of entire slides. Wang et al. [20] employed CTransPath [21] as a feature extractor and designed a multi-label learning model capable of simultaneously predicting ER, PR, and HER2 biomarker statuses. This innovative approach highlights the growing potential of combining foundational models with MIL aggregation to enhance biomarker prediction performance.

Weakly supervised learning models demonstrate great potential in WSI classification, but their performance can often be improved by incorporating supplementary data. Recent studies have shown that joint training with genomic profiles and pathology slides can significantly enhance performance in tasks such as tumor subtyping, survival regression, and tumor grading [2,3,29,25,7]. For instance, Chen et al. [2] proposed a cross-modal attention mechanism to integrate pathology and genomic features, which enhances the model's ability to capture complex inter-modal relationships. Zhou et al. [29] developed two parallel encoder-decoder architectures to fuse intra-modal information and generate cross-modal representations, thus improving the model's representational capacity and predictive accuracy. Despite these advancements, clinical application of multi-modal learning is often hindered by the high costs and technical complexities associated with acquiring genomic and pathology data simultaneously.

In order to mitigate the need for expensive genomic data collection, we propose an approach that leverages multi-modal data during training while enabling inference using only pathology slides. To handle missing modalities during testing, we employ knowledge distillation (KD) [6], a widely-used technique for transferring knowledge from a multi-modal teacher to a uni-modal student [23]. However, several challenges remain when applying KD to our scenario: (1) Teacher models are often selected based on subjective experience rather than objective metrics, leading to suboptimal guidance and inconsistent distillation outcomes.

(2) According to the *Modality Focusing Hypothesis* (MFH) [26], the effectiveness of cross-modal KD depends on the retention of modality-general decisive features in the teacher model. A higher proportion of such features enhances student model performance, yet effective strategies to increase this proportion remain underexplored and challenging.

To address the aforementioned challenges, we propose an online KD approach based on multi-modal knowledge decomposition (MKD), designed to amplify the representation of modality-general decisive features while concurrently optimizing the efficacy of both student and teacher models. Our main contributions are: (1) We employ the MKD to enhance the transferability of modality-general decisive features from the teacher model. (2) We propose a robust and efficient online KD model that ensures the teacher and student models appropriately capture the relative relationships among samples, while fostering greater diversity and complementarity in dynamic learning. (3) Extensive experiments show that both the teacher and student models achieve state-of-the-art (SOTA) performance on both the public TCGA-BRCA and in-house datasets in biomarker prediction.

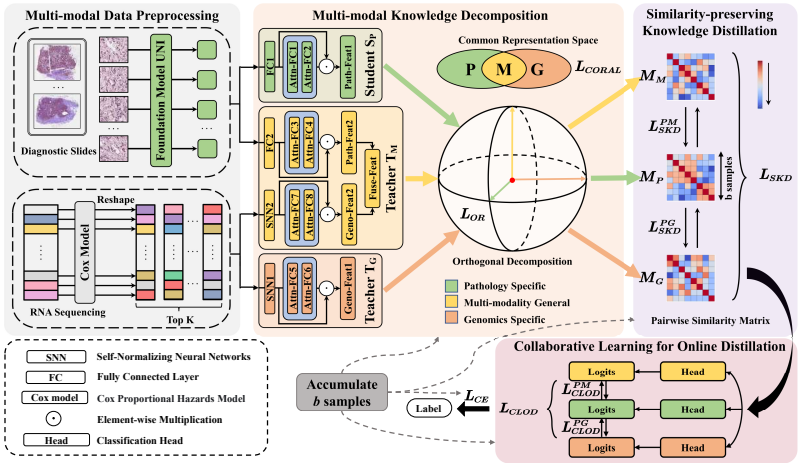

Fig. 1. Overview of our approach. Note that our approach includes four components: multi-modal data preprocessing (MDP), MKD, SKD, and CLOD.

2 Methods

Fig. 1 shows an overview of the proposed approach, including four components: MDP, MKD, SKD, and CLOD. The process starts by extracting genomic and pathomic features, then decomposing multi-modal knowledge into pathology-specific, modality-general, and genomics-specific features. SKD enables the pathology student model to learn sample relationships, while CLOD

fosters mutual learning between teacher and student models. During training, gradients accumulate over multiple samples before updating model parameters, allowing SKD to capture feature relationships. Details are provided below.

2.1 Multi-modal Data Preprocessing (MDP)

Given a WSI, it is divided into multiple tiles, with tissue tiles selected using the CLAM toolbox [12]. Feature embedding is then performed on tissue tiles using the UNI foundation model [1]. Let $P \in \mathbb{R}^{n_p \times d}$ denote the feature embedding for a WSI, where n_p is the number of tissue tiles, and d denotes the feature dimension. We hypothesize that genes associated with overall survival (OS) of patients are also associated with IHC biomarker status. Thus, the Cox proportional hazards model [4] is employed to identify the top K genes most relevant to patients' OS from genomic profiles, with the genomic features reshaped into a representation $G \in \mathbb{R}^{n_g \times d}$ based on their rankings, where $n_g = \lfloor K/d \rfloor$.

2.2 Multi-modal Knowledge Decomposition (MKD)

To thoroughly decompose and integrate knowledge, we develop two teacher and one student models, each tailored to capture distinct aspects of genomic profiles and pathology slides. These aggregators focus on collaboration, uniqueness, and general representation, facilitating a comprehensive understanding of multi-modal data. As observed in Fig. 1, pathology features P are first compressed using a fully connected layer, followed by further compression using the Attention-based MIL (ABMIL) [8] in the student model S_P, as expressed as:

$$z_p = \sum_{i}^{n_p} a_i P_i, \quad a_i = \frac{\exp\left\{W^T\left(\tanh\left(VP_i^T\right) \odot \text{sigmoid}\left(UP_i^T\right)\right)\right\}}{\sum_{j=1}^{n_p} \exp\left\{W^T\left(\tanh\left(VP_j^T\right) \odot \text{sigmoid}\left(UP_j^T\right)\right)\right\}}, \quad (1)$$

where $V, U \in \mathbb{R}^{n_p \times d}$ and $W \in \mathbb{R}^{d \times 1}$ are learnable linear projection matrices, and $\odot$ is an element-wise multiplication. Similarly, genomic features G undergo a two-step process in the teacher model T_G: they are first compressed via a Self-Normalizing Network (SNN) [10], followed by refinement with ABMIL. Meanwhile, we build a teacher model T_M, which fuse the global representations of two modalities learned by ABMIL using the Kronecker product [19]. By processing through the three distinct aggregators including pathology-specific, modality-general, and genomic-specific features, the multi-modal knowledge are systematically decomposed, enabling the extraction of an integrated and meaningful knowledge representation.

To advance knowledge distillation and enhance the model's generalization ability, we perform domain alignment on the decomposed knowledge. Specifically, we minimize the CORAL loss [16] between the decomposed knowledge, which is expressed as:

$$L_{CORAL} = \frac{1}{4d^2}\left(\|C_P^b - C_G^b\|_F^2 + \|C_P^b - C_M^b\|_F^2 + \|C_G^b - C_M^b\|_F^2\right), \quad (2)$$

where $\|\cdot\|_F$ denotes the Frobenius norm, and $C_j^b \in \mathbb{R}^{d\times d}$ is the covariance matrix for b cumulative samples, where $j \in \{P, G, M\}$. The L_{CORAL} loss reduces covariance differences across modalities, aligning features into a unified representation for compatible decomposed knowledge. To transfer decisive modality-general features to the student model S_P, we introduce a pairwise orthogonality constraint, promoting feature independence and ensuring each captures distinct, complementary information from different modalities. To enforce this, we introduce an orthogonal loss function as follows:

$$L_{OR} = |\langle z_p, z_g\rangle| + |\langle z_p, z_m\rangle| + |\langle z_g, z_m\rangle|\,, \tag{3}$$

where $\langle\cdot\rangle$ denote the dot product, and $|\cdot|$ is the absolute operation used to enforce pairwise orthogonality between inputs. $z_p, z_g, z_m \in \mathbb{R}^{1\times d}$ represent the features specific to pathology, genomics and multi-modal general features, respectively. This orthogonal loss reduces redundancy and promotes more robust feature representations across modalities. Consequently, our MKD loss is calculated as:

$$L_{MKD} = L_{CORAL} + \alpha L_{OR}, \tag{4}$$

where α is a hyperparameter that serves as a weighting parameter.

2.3 Similarity-preserving Knowledge Distillation (SKD)

To capture sample relationships, we introduce SKD [18] to guide the training of the student model S_P, ensuring input pairs with similar (or dissimilar) activations in the teacher network also produce similar (or dissimilar) activations in the student network. Our SKD loss is calculated as:

$$L_{SKD} = L_{SKD}^{PM} + L_{SKD}^{PG}, \tag{5}$$

where the L_{SKD}^{PM} loss for a single pair of student and teacher models is computed as:

$$L_{SKD}^{PM} = \frac{1}{b^2}\|\frac{Z_P^b \left(Z_P^b\right)^T}{\|Z_P^b \left(Z_P^b\right)^T\|_2} - \frac{Z_M^b \left(Z_M^b\right)^T}{\|Z_M^b \left(Z_M^b\right)^T\|_2}\|_F^2, \tag{6}$$

where $\|\cdot\|_2$ denotes row-wise L2 normalization, and Z_p^b, Z_M^b denote the pathological and multi-modal feature matrices for b concatenated samples. The loss L_{SKD}^{PG} is computed similarly as in Eq.(6). Notably, our comprehensive SKD loss effectively preserves the consistency of the activation similarity matrices M_M, M_P, M_G, enabling effective knowledge transfer within multi-modal feature space while maintaining their internal sample-specific structures.

2.4 Collaborative Learning for Online Distillation (CLOD)

To foster collaborative learning between teacher and student models, we adopt an online learning framework [28,30], which consolidates training into a single stage by treating all networks as peers. This framework facilitates symmetrical

knowledge sharing, allowing each network learn equally from others without being overly reliant on a predefined teacher model. Our CLOD loss is defined as:

$$L_{CLOD} = KL\left(p_P||p_M\right) + KL\left(p_M||p_P\right) + KL\left(p_P||p_G\right) + KL\left(p_G||p_P\right), \quad (7)$$

where p_P, p_M, p_G denote the probability distributions predicted by the pathology-specific, modality-general, and genomic-specific classification heads, respectively. The Kullback-Leibler (KL) divergence terms in Eq.(7) measure distribution discrepancies, promoting alignment and mutual knowledge sharing across networks. This collaborative setup fosters diverse learning dynamics and enables bidirectional knowledge exchange, improving overall performance. Therefore, the overall loss L of our online KD model is formulated as:

$$L = L_{CE} + L_{MKD} + L_{SKD} + L_{CLOD}, \quad (8)$$

where L_{CE} represents the overall cross-entropy loss calculated for two teacher models and one student model.

3 Experiments and Results

3.1 Datasets and Implementations

TCGA-BRCA: The TCGA-BRCA dataset provides a multi-omics resource, with cases having missing or low-quality genomic profiles or pathology slides excluded. For patients with multiple slides, one diagnostic slide was randomly selected. Genomic profiles were represented using log-transformed, Z-score normalized RNA-Seq expression values. In Fig. 2, the left three ring charts illustrate the distribution of patients across the ER, PR, and HER2 labels in this dataset, which are used for both internal training and testing cohorts.

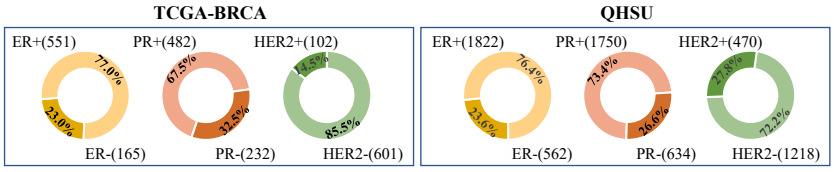

Fig. 2. Summary of public TCGA-BRCR and in-house QHSU cohorts.

QHSU: QHSU dataset includes 2384 H&E-stained WSIs, with each WSI corresponding to an individual breast cancer patient. IHC biomarker information for these patients was obtained from diagnostic records assessed by skilled pathologists. In Fig. 2, the right three ring charts illustrate the distribution of patients across the ER, PR, and HER2 labels. Since this in-house dataset contains only pathology slides, it is used as an external test set.

Evaluation & Implementation: We performed a 5-fold cross-validation on the TCGA-BRCA internal cohort and report the average test performance across all folds. The five trained models were then tested on the external QHSU cohort, with average results reported. Our model was implemented in Python using the PyTorch library, and trained on a workstation equipped with an NVIDIA GeForce RTX 4090 GPU. We used the AdamW optimizer with a learning rate of 2e-4, a weight decay of 1e-5, and a temperature of 4. The hyperparameter α in Eq.(4) was fine-tuned and set to 1/6. Model parameters were updated after accumulating gradients from 16 samples.

3.2 Experimental Results

Internal Comparison. The proposed model was compared against seven MIL methods [8,12,15,27,11,17,20], two KD approaches [22,24], and three multi-modal learning models [2,3,29] to evaluate its effectiveness and flexibility. As presented in Table 1, the internal comparisons on the TCGA-BRCA cohort reveal that our model, when using only pathology slides, achieves the best overall performance compared to SOTA methods, with a 2% improvement in AUC values and notable improvements in other metrics. This highlights the capability of the teacher model to effectively transfer critical and generalizable features to the student model, thereby enhancing its performance. In multi-modal testing, all models perform better than using pathology slides alone, particularly for HER2 prediction. Notably, our model consistently ranks first or second across various metrics, indicating that joint training of teacher and student models fosters mutual learning without compromising individual performance. In addition, our approach eliminates the need to explicitly validate the teacher's guidance, significantly reducing computational costs.

Table 1. Comparisons with SOTA methods on the TCGA-BRCA dataset. The bold and underlined fonts highlight the best and the second-best results, respectively.

Modality	Models	ER(%)			PR(%)			HER2(%)		
		AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
Patho.	ABMIL [8]	88.91	84.85	89.84	86.10	80.55	85.08	72.13	77.30	39.36
	CLAM [12]	89.20	85.69	90.68	85.33	<u>82.22</u>	<u>86.86</u>	70.51	80.56	37.13
	TransMIL [15]	87.77	<u>86.94</u>	<u>91.82</u>	81.37	77.50	83.04	66.23	75.60	33.45
	DTFD [27]	89.89	86.66	91.37	86.50	82.36	86.76	70.72	75.88	40.18
	WIKG [11]	88.05	84.44	90.10	85.54	78.88	70.37	68.50	74.89	38.30
	RRT [17]	89.35	83.33	88.28	<u>86.61</u>	80.55	85.31	68.08	77.44	28.13
	DAMLN [20]	89.58	85.97	91.39	86.38	81.80	86.54	69.92	**84.53**	22.37
	GEE [22]	<u>90.01</u>	84.30	89.37	86.24	81.52	85.81	70.06	74.60	<u>41.32</u>
	TDC [24]	89.35	83.33	88.28	84.75	73.05	77.01	<u>72.86</u>	77.87	**42.87**
	Ours	**93.31**	**88.47**	**92.58**	**88.65**	**83.75**	**88.14**	**74.56**	<u>81.56</u>	39.10
Multi.	MCAT [2]	<u>94.64</u>	**90.41**	**93.68**	90.37	<u>85.28</u>	88.75	<u>93.10</u>	84.96	65.06
	Porpoise [3]	92.64	89.86	93.28	**91.79**	**85.97**	**89.57**	92.80	<u>89.65</u>	<u>68.94</u>
	CMTA [29]	93.91	89.44	93.15	<u>91.51</u>	83.63	87.26	92.03	87.94	66.55
	Ours	**95.81**	<u>90.24</u>	<u>93.67</u>	91.39	<u>85.28</u>	<u>89.26</u>	**95.76**	**92.48**	**76.88**

Table 2. Comparisons with SOTA methods on the QHSU dataset.

Modality	Models	ER(%)			PR(%)			HER2(%)		
		AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
Patho.	ABMIL [8]	87.27	83.09	89.55	82.71	64.63	69.67	72.41	56.92	50.76
	CLAM [12]	86.83	82.71	89.16	80.93	64.01	68.73	73.67	60.91	49.43
	TransMIL [15]	84.74	81.24	88.21	78.33	71.42	78.48	68.95	58.83	42.49
	DTFD [27]	87.50	82.69	89.41	81.64	61.85	65.73	71.90	56.93	50.70
	WIKG [11]	82.79	84.44	90.10	79.04	58.48	60.23	70.38	52.50	49.26
	RRT [17]	85.31	81.44	88.65	79.09	67.13	72.98	71.81	57.17	49.78
	DAMLN [20]	86.02	82.45	88.58	81.18	70.34	76.57	73.40	**69.44**	42.71
	GEE [22]	88.31	83.30	87.95	82.94	70.40	76.17	71.09	49.53	48.28
	TDC [24]	87.77	78.48	84.31	83.40	59.19	62.72	68.88	59.38	51.29
	Ours	**89.00**	**84.97**	**90.45**	**84.36**	**74.01**	**79.95**	**74.12**	57.79	**52.74**

External Comparison. Table 2 shows the external comparison results on the QHSU dataset. The results reveal that models trained with knowledge distillation generally outperform MIL models trained solely on pathology slides in ER and PR predictions, indicating that knowledge distillation enables student models to learn more robust features. Notably, our model outperform all comparative methods on the external test set, demonstrating the effectiveness of our multi-modal knowledge decomposition in extracting general decisive features.

Table 3. Ablation study of MKD, SKD, CLOD modules on the TCGA-BRCA dataset.

Modality	Modules		ER(%)		PR(%)		HER2(%)	
	MKD	SKD+CLOD	AUC	ACC	AUC	ACC	AUC	ACC
Patho.			88.91	84.85	86.10	80.55	72.13	77.30
	✓		93.00	86.94	87.70	84.02	72.19	82.26
		✓	91.80	85.97	87.58	83.19	67.00	80.00
	✓	✓	**93.31**	**88.47**	**88.65**	**83.75**	**74.56**	**81.56**
Multi.			94.45	**90.97**	90.14	84.16	93.57	90.78
	✓		91.96	89.58	88.90	84.86	92.48	**94.18**
		✓	95.75	90.55	91.05	84.02	95.57	93.61
	✓	✓	**95.81**	90.24	**91.39**	**85.28**	**95.76**	92.48

Ablation Study. We conducted ablation experiments on the TCGA-BRCA dataset to assess the contributions of the MKD, SKD and CLOD modules to overall model performance. Table 3 presents evaluation results. It is observed that the MKD module greatly enhances the performance when utilizing pathology slides alone. In contrast, the SKD and CLOD modules, as knowledge distillation components, significantly improve the performance of multi-modal models. Overall, the inclusion of all three modules greatly facilitates mutual learning between the teacher and student models, leading to consistently enhanced performance across different evaluation metrics.

4 Conclusion

In this paper, we propose a multi-modal knowledge decomposition based online distillation method for predicting multiple IHC biomarkers from H&E-stained breast cancer WSIs. Our key innovation lies in the MKD module, which efficiently decomposes input features into pathology-specific, modality-general, and genomics-specific components, facilitating the transfer of both generalizable and decisive knowledge. Additionally, the SKD module enhances knowledge transfer across samples by preserving their internal structures, while the CLOD module fosters mutual learning and knowledge sharing between the teacher and student models. Experiments conducted on two datasets demonstrate that our method achieves superior performance in both uni-modal data testing. Our approach is highly flexible, as its inference supports pathology slides, genomics profiles, or both, depending on available modalities.

Acknowledgments. This work was supported in part by Liaoning Province Science and Technology Joint Program (2024-MSLH-065), and the Fundamental Research Funds for Central Universities (DUT25Z2514, DUT24YG201).

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

References

1. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (2024)
2. Chen, R.J., Lu, M.Y., Weng, W.H., Chen, T.Y., Williamson, D.F., Manz, T., Shady, M., Mahmood, F.: Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4025 (2021)
3. Chen, R.J., Lu, M.Y., Williamson, D.F., Chen, T.Y., Lipkova, J., Noor, Z., Shaban, M., Shady, M., Williams, M., Joo, B., et al.: Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell **40**(8), 865–878 (2022)
4. Cox, D.R.: Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) **34**(2), 187–202 (1972)
5. Gamble, P., Jaroensri, R., Wang, H., Tan, F., Moran, M., Brown, T., Flament-Auvigne, I., Rakha, E.A., Toss, M., Dabbs, D.J., et al.: Determining breast cancer biomarker status and associated morphological features using deep learning. Communications medicine **1**(1), 14 (2021)
6. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network (2015), https://arxiv.org/abs/1503.02531
7. Hou, W., Lin, C., Yu, L., Qin, J., Yu, R., Wang, L.: Hybrid graph convolutional network with online masked autoencoder for robust multimodal cancer survival prediction. IEEE Transactions on Medical Imaging **42**(8), 2462–2473 (2023)
8. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)

9. Kather, J.N., Heij, L.R., Grabsch, H.I., Loeffler, C., Echle, A., Muti, H.S., Krause, J., Niehues, J.M., Sommer, K.A., Bankhead, P., et al.: Pan-cancer image-based detection of clinically actionable genetic alterations. Nature cancer **1**(8), 789–799 (2020)
10. Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S.: Self-normalizing neural networks. Advances in neural information processing systems **30** (2017)
11. Li, J., Chen, Y., Chu, H., Sun, Q., Guan, T., Han, A., He, Y.: Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11323–11332 (2024)
12. 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)
13. Lu, W., Toss, M., Dawood, M., Rakha, E., Rajpoot, N., Minhas, F.: Slidegraph+: Whole slide image level graphs to predict her2 status in breast cancer. Medical Image Analysis **80**, 102486 (2022)
14. Naik, N., Madani, A., Esteva, A., Keskar, N.S., Press, M.F., Ruderman, D., Agus, D.B., Socher, R.: Deep learning-enabled breast cancer hormonal receptor status determination from base-level h&e stains. Nature communications **11**(1), 5727 (2020)
15. 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)
16. Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14. pp. 443–450. Springer (2016)
17. 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)
18. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1365–1374 (2019)
19. Van Loan, C.F.: The ubiquitous kronecker product. Journal of computational and applied mathematics **123**(1-2), 85–100 (2000)
20. Wang, M., Wang, T., Cong, F., Lu, C., Xu, H.: Double-tier attention based multi-label learning network for predicting biomarkers from whole slide images of breast cancer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 91–101. Springer (2024)
21. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification. Medical image analysis **81**, 102559 (2022)
22. Wu, K., Jiang, Z., Zhu, X., Shi, J., Zheng, Y.: Genomics-embedded histopathology whole slide image encoding for data-efficient survival prediction. In: Medical Imaging with Deep Learning (2024)
23. Xing, X., Chen, Z., Zhu, M., Hou, Y., Gao, Z., Yuan, Y.: Discrepancy and gradient-guided multi-modal knowledge distillation for pathological glioma grading. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 636–646. Springer (2022)

24. Xing, X., Zhu, M., Chen, Z., Yuan, Y.: Comprehensive learning and adaptive teaching: Distilling multi-modal knowledge for pathological glioma grading. Medical Image Analysis **91**, 102990 (2024)
25. 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)
26. Xue, Z., Gao, Z., Ren, S., Zhao, H.: The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation. arXiv preprint arXiv:2206.06487 (2022)
27. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y.: Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18802–18812 (2022)
28. Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4320–4328 (2018)
29. 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)
30. Zhu, X., Gong, S., et al.: Knowledge distillation by on-the-fly native ensemble. Advances in neural information processing systems **31** (2018)