ccRCC Metastasis Prediction via Exploring High-Order Correlations on Multiple WSIs

Huijian Zhou[1], Zhiqiang Tian[1], Xiangmin Han[2(✉)], Shaoyi Du[1], and Yue Gao[2]

[1] Xi'an Jiaotong University, 710049, China
zhj11055331@stu.xjtu.edu.cn, {zhiqiangtian,dushaoyi}@xjtu.edu.cn
[2] Tsinghua University, 100084, China
{hanxiangmin, gaoyue}@tsinghua.edu.cn

Abstract. Metastasis prediction based on gigapixel histopathology whole-slide images (WSIs) is crucial for early diagnosis and clinical decision-making of clear cell renal cell carcinoma (ccRCC). However, most existing methods focus on extracting task-related features from a single WSI, while ignoring the correlations among WSIs, which is important for metastasis prediction when a single patient has multiple pathological slides. In this case, we propose a multi-slice-based hypergraph computation (MSHGC) method for metastasis prediction, which considers the intra-correlations within a single WSI and cross-correlations among multiple WSIs of a single patient simultaneously. Specifically, intra-correlations are captured within both topology and semantic feature spaces, while cross-correlations are modeled between the patches from different WSIs. Finally, the attention mechanism is used to suppress the contribution of task-irrelevant patches and enhance the contribution of task-relevant patches. MSHGC achieves the C-index of 0.8441 and 0.8390 on two carcinoma datasets, outperforming state-of-the-art methods, which demonstrates the effectiveness of the proposed MSHGC.

Keywords: Metastasis prediction · Multiple WSIs · Cross-correlations · Hypergraph.

1 Introduction

Metastasis prediction [4, 16] based on histopathology whole slide images (WSIs) aims to predict the risk of cancer metastasis based on the patient's pathological slice. Most WSI-based methods only consider features from a single WSI, but a patient may have multiple WSIs, which can provide more information about the current condition of the cancer. As a result, we focus on metastasis prediction on multi-slice, whose key challenge is how to capture the task-relevant features within multiple gigapixel WSIs of each patient.

There has been much effort attempting to address this challenge. CLAM [13] applies multiple instance learning (MIL) and proposes an attention-based pooling function that can automatically identify subregions of high diagnostic value, which ignores the correlations between the patches. DeepGraphSurv [12] embeds

patches into semantic feature space and constructs a graph to model the relationships between patches according to semantic feature distance. Patch-GCN [1] treats the patches as a 2D point cloud and constructs the graph in the image space. However, both methods are developed based on GNNs, which can only capture pairwise interactions for representation learning. It is important to note that in the WSIs, interactions often occur among groups of three or more nodes and cannot be simply described by binary relationships.

Hypergraph neural networks [9, 19, 6] are the extension of GNNs to learn high-order correlations, in which hyperedges can connect more than two vertices. HGSurvNet [3] utilizes the hypergraph to model the complex correlations among the patches, which focuses on intra-correlations within a single WSI. However, due to the heterogeneity of tumors, different WSIs from the same patient often contain diverse information related to the disease, which is complementary. By amalgamating these complementary features, we can enrich the predictive signal, thereby refining the performance of survival prediction algorithms in clinical settings, while existing methods cannot model the high-order correlations among multiple WSIs, which limits the performance of the model.

In this paper, we propose a multi-slice-based hypergraph computation (MSHGC) framework for metastasis prediction based on multiple WSIs, in which the intra-correlations within a single WSI and the cross-correlations between the WSIs of each patient are both taken into consideration by utilizing multi-slice hypergraph convolution. For the intra-correlations, we only consider the patches from a single WSI, and both topological connection and semantic feature distances are utilized to construct an intra-hypergraph. For the cross-correlations, We consider the relationships between patches from different WSIs and construct the cross-hypergraph in semantic feature space. Finally, we utilize the self-attention mechanism [17] to assign weights to each patch for aggregating patch-level features into global features. The main contributions are summarized as follows:

- We propose a multi-slice-based hypergraph computation method for metastasis prediction on multiple WSIs. The proposed method effectively achieves high-order correlation modeling and learning based on multi-slice in a uniform framework.
- To obtain a more comprehensive task-specific global representation of a patient on multiple WSIs, we apply an attention mechanism to compute the self-attention score of each patch, which is guided by multi-slice high-order correlation, as its contribution to the metastasis risk prediction results.
- The proposed MSHGC is validated on two carcinoma datasets. The experimental results demonstrate that the proposed MSHGC outperforms the state-of-the-art methods, including graph-based, MIL-based, and hypergraph-based methods.

2 Method

The Fig.1 is the framework of MSHGC. Firstly, we construct two types of hypergraph, intra-hypergraph and cross-hypergraph, by multi-slice hypergraph

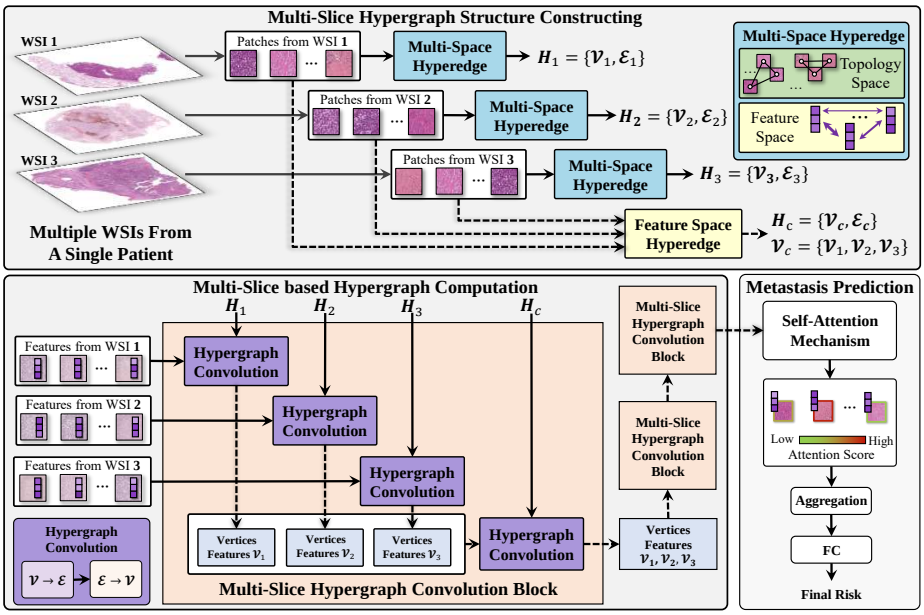

Fig. 1. The framework of MSHGC. 1) The multi-space hyperedge of the intra-hypergraph is constructed in topology space and feature space, while the cross-hypergraph is constructed in feature space. 2) Multi-slice hypergraph convolution block has four hypergraph convolution modules that contain two steps, message passing from $\mathcal{V}$ to $\mathcal{E}$ and $\mathcal{E}$ to $\mathcal{V}$.

structure constructing module. Intra-hypergraph is constructed in both topology space and feature space, while cross-hypergraph is only constructed in feature space. Secondly, the features of patches and hypergraphs are fed into a multi-slice hypergraph convolution block to capture the complex high-order correlations within and across WSIs. Finally, through the self-attention mechanism, we aggregate the weighted patch-level features to global representation, on which the metastasis risk is predicted by a fully connected layer.

2.1 Multi-Slice Hypergraph Structure Constructing

Given a sample patient P with M WSIs, a random patch sampling process is first implemented in each WSI to generate N patches, where N is a fixed sampled patch number. The patches are fed into a pre-trained ResNet model [7] to yield pathology semantic features. In that case, a patient can be represented by the features $\mathbf{X}_p = \{\mathbf{X}^1, ..., \mathbf{X}^M\}$, where $\mathbf{X}^i \in \mathbb{R}^{N \times C}$ denotes the features of i^{th} WSI and C denotes the feature dimension of a single patch.

To conduct the multi-slice hypergraph convolution block, two types of hypergraph are needed: intra-hypergraph and cross-hypergraph.

For intra-hypergraph, we only take patches sampled from the same WSI into consideration. To capture the high-order correlations between patches, we utilize both topological connections and semantic pathological features to generate multi-space hyperedge. To be specific, given a patch i as vertex v_i, the KNN method is used to find neighbor vertices for v_i in the topological space and semantic feature space to establish hyperedges respectively. The distance between vertices in the semantic feature space $d(v_i, v_j)$ and topological space $g(v_i, v_j)$ are defined by the following formulas:

$$d(v_i, v_j) = (\sum_{c=1}^{C} (\mathcal{F}_i[c] - \mathcal{F}_j[c])^2)^{\frac{1}{2}}, \tag{1}$$

$$g(v_i, v_j) = ((x_i - x_j)^2 + (y_i - y_j)^2)^{\frac{1}{2}}, \tag{2}$$

where v_i, $\mathcal{F}_i$ and (x_i, y_i) denote the i^{th} vertex, the features of i^{th} vertex and the center coordinates of the i^{th} vertex in WSI. For cross-hypergraph, all sampled patches are taken into consideration. However, as the topological distance between patches sampled from different WSIs is difficult to measure, the cross-hypergraph only constructs hyperedges in the semantic feature space by using the KNN method. It should be noted that in order to avoid interference among patches from the same WSI, we manually set the semantic feature distance among patches sampled from the same WSI to infinity.

A hypergraph can be represented by an incidence matrix $\mathbf{H} \in \mathbb{R}^{V \times E}$, where V and E denote the number of vertex and hyperedge respectively. The element in $\mathbf{H}$ is defined as

$$h(v, e) = \begin{cases} 1, v \in e \\ 0, v \notin e \end{cases}, \tag{3}$$

where v and e denote vertex and hyperedge, respectively. Finally, for patient P with M WSIs, we can get M intra-hypergraph $\mathbf{H}_{intra} \in \mathbb{R}^{N \times 2N}$ and a cross-hypergraph $\mathbf{H}_{cross} \in \mathbb{R}^{MN \times MN}$.

2.2 Multi-Slice Hypergraph Convolution Block

As shown in Fig.1, we design a multi-slice hypergraph convolution block to conduct the multi-slice-based hypergraph computation, which is based on the several hypergraph convolution operation[5]. The hypergraph convolution consists of three steps: 1) **Vertex Feature Reweighting.** The input vertex features, $\mathbf{X}_{in}$, are reweighted by multiplying with learnable parameters Θ. 2) **Message Passing from $\mathcal{V}$ to $\mathcal{E}$.** The reweighted vertex features F_v connected by hyperedges are integrated into hyperedge features $\mathrm{F}_e \in \mathbb{R}^{E \times C}$, which is achieved by multiplying $\mathbf{H}^\top \in \mathbb{R}^{E \times V}$ with $\mathrm{F}_v \in \mathbb{R}^{V \times C}$. 3) **Message Passing from $\mathcal{E}$ to $\mathcal{V}$.** Finally, the output vertex features $\mathbf{X}_{out} \in \mathbb{R}^{V \times C}$ are obtained by multiplying $\mathbf{H} \in \mathbb{R}^{V \times E}$ with $\mathrm{F}_e \in \mathbb{R}^{E \times C}$.

A multi-slice hypergraph convolution block consists of two stages: 1) intra-hypergraph computation and 2) cross-hypergraph computation. In particular, for a patient P with $\{\{\mathbf{X}^1, ..., \mathbf{X}^M\}, \{\mathbf{H}^1_{intra}, ..., \mathbf{H}^M_{intra}\}, \mathbf{H}_{cross}\}$, the intra-hypergraph

computation consists of M hypergraph convolution branches. The intra-hypergraph computation in the l^{th} layer can be formulated as:

$$\{\mathbf{X}^i_{intra}\}^M_{i=1} = \{(\mathbf{D}^i_v)^{-\frac{1}{2}}\mathbf{H}^i_{intra}(\mathbf{D}^i_e)^{-1}(\mathbf{H}^i_{intra})^\top(\mathbf{D}^i_v)^{-\frac{1}{2}}\mathbf{X}^i_{(l)}\Theta^i_{(l)}\}^M_{i=1}, \quad (4)$$

where $\mathbf{D}^i_e \in \mathbb{R}^{E\times E}$ and $\mathbf{D}^i_v \in \mathbb{R}^{V\times V}$ denote the diagonal degree matrix of hyperedges and vertices of $\mathbf{H}^i_{intra}$ respectively, i denotes the i^{th} branch. Then, the results of intra-hypergraph computation are concatenated and fed into a cross-hypergraph computation module that is similar to Eq.4 but replace $\mathbf{H}^i_{intra}$ with $\mathbf{H}_{cross}$. Finally, we add a residual block followed by a non-linear activation function $\delta(\cdot)$. A multi-slice hypergraph convolution block can be represented by:

$$\mathbf{X}_{(l+1)} = \delta(Hyconv(Concat(\{Hyconv(\mathbf{X}^i_{(l)}, \mathbf{H}^i_{intra})\}^M_{i=1}), \mathbf{H}_{cross}) + \mathbf{X}_{(l)}), \quad (5)$$

where $Hyconv(\cdot)$ denotes the hypergraph convolution operation, $Concat(\cdot)$ denotes concatenation operation. After several multi-slice hypergraph convolution blocks, the learned patient features $\mathbf{X}_{(n)} \in \mathbb{R}^{MN\times C}$ can be obtained.

2.3 Metastasis Prediction

The outputs of several multi-slice hypergraph convolution blocks are still patch-level features, which are not suitable for predicting the metastasis risk of the patient. The simplest way is to aggregate patch-level features into patient-level features by average pooling or max pooling operations, which is suboptimal in case each patch contains different information relevant to metastasis prediction. Therefore, we design an attention module to assign weights to each patch and aggregate patch-level features into patient-level features according to the weights. The process of generating attention can be defined as follows:

$$\mathbf{A} = L_1((\mathbf{X}_{(n)}\mathbf{W}_q)(\mathbf{X}_{(n)}\mathbf{W}_k)^\top\mathbf{W}_a), \quad (6)$$

where $\mathbf{W}_q \in \mathbb{R}^{C\times C/r}$, $\mathbf{W}_k \in \mathbb{R}^{C\times C/r}$ and $\mathbf{W}_a \in \mathbb{R}^{MN\times 1}$ are learnable matrices, r is the scale factor, $\mathbf{A} \in \mathbb{R}^{MN\times 1}$ is the attention matrix, $L_1(\cdot)$ denotes the L1 normalization. By multiplying $\mathbf{A}^\top \in \mathbb{R}^{1\times MN}$ with $\mathbf{X}_{(n)} \in \mathbb{R}^{MN\times C}$, we can obtain the patient-level features $\mathbf{X}_p \in \mathbb{R}^{1\times C}$ that be used to predict the metastasis risk by a fully connected layer.

To train the model, we use the negative Cox log partial likelihood loss function [20] as a supervision signal, formulated as follows:

$$\mathcal{L}_{NLL} = \sum_{i=1}^{P}\delta_i(-s^p_i + \log\sum_{j\in\{j:s^g_j\le s^g_i\}}\exp(s^p_j)), \quad (7)$$

where s^p_i and s^g_i denote the predicted result and ground truth, respectively. P is the number of comparable pairs. δ_i denotes whether the sample is censored.

Table 1. Datasets statistics. LMT and SMT denote the longest and shortest metastasis time of each dataset, respectively. C-rate denotes the proportion of censored samples.

Dataset	ccRCC_H1	ccRCC_H2
patients	111	92
WSIs	333	276
LMT	1593	1887
SMT	64	81
C-rate (%)	87.39	85.87

Table 2. Results for different methods on ccRCC_H1 and ccRCC_H2 measured by C-index.

Methods	ccRCC_H1	ccRCC_H2
MIL_Attention [8]	$0.5845_{\pm 0.262}$	$0.6808_{\pm 0.125}$
DTFD-MIL [21]	$0.6799_{\pm 0.156}$	$0.7704_{\pm 0.121}$
TransMIL [15]	$0.6331_{\pm 0.248}$	$0.8031_{\pm 0.135}$
CLAM [13]	$0.7419_{\pm 0.253}$	$0.7615_{\pm 0.175}$
DeepGraphSurv [12]	$0.6890_{\pm 0.193}$	$0.7747_{\pm 0.160}$
Patch-GCN [1]	$0.6963_{\pm 0.191}$	$0.7072_{\pm 0.204}$
TEA-Graph [11]	$0.6473_{\pm 0.250}$	$0.5804_{\pm 0.171}$
HGSurvNet [3]	$0.8072_{\pm 0.061}$	$0.8011_{\pm 0.135}$
MSHGC(Ours)	$\mathbf{0.8441}_{\pm 0.059}$	$\mathbf{0.8390}_{\pm 0.060}$

3 Experiments

3.1 Datasets

The proposed MSHGC was evaluated on two ccRCC datasets collected from the cooperative hospital. Each patient in the dataset has three WSIs. Table 1 summarizes the detailed statistical information for the datasets.

3.2 Methods for Comparison

The proposed methods are compared with eight typical state-of-the-art methods. The codes are reproduced based on the released codes or detailed introductions.

1) **MIL_Attention** [8] applies multiple instance learning (MIL) to train the model. Patches are regarded as instances, whose features are aggregated by attention mechanism to obtain the global features of the patient.
2) **TransMIL** [15] is a Transformer-based MIL method, which extracts task-relevant features more accurately by modeling correlations between instances.
3) **CLAM** [13] proposes an attention-based pooling function that can automatically identify subregions of high diagnostic value and aggregate patch-level features into slide-level representations.
4) **DTFD-MIL** [21] introduces the concept of pseudo-bags to virtually enlarge the number of bags, and proposes a double-tier MIL framework to extract the task-relevant features more effectively.
5) **DeepGraphSurv** [12] takes the global topological structures of WSIs into consideration and reduces the randomness of patch sampling by attention mechanism to increase model robustness.
6) **Patch-GCN** [1] treats the patches as a 2D point cloud and constructs graph structures in the image space. It hierarchically aggregates instance-level histology features by a context-aware graph convolutional network.
7) **TEA-Graph** [11] uses a graph attention network [18] (GAT) with positional embeddings to extract the context features around the patch by aggregating the neighborhoods of the patch with different attention scores.

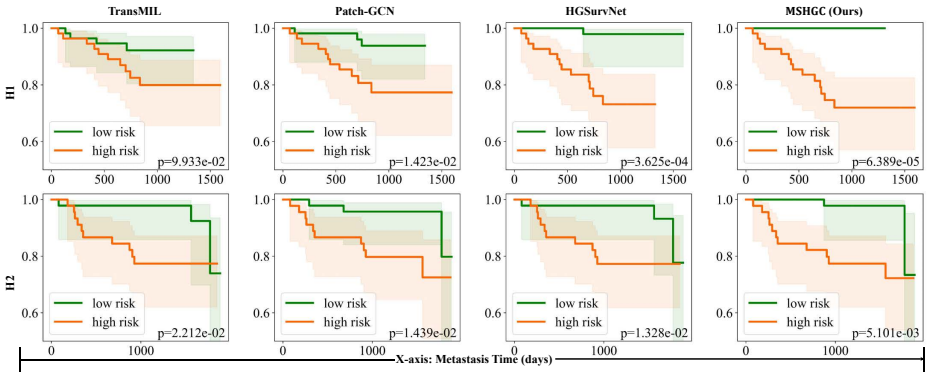

Fig. 2. The experimental results of four methods (TransMIL, Patch-GCN, HGSurvNet, and MSHGC) measured by KM-estimation curves of two datasets. More distinguish gaps between high- and low-risk curves correspond to better classification performance.

8) **HGSurvNet** [3] applies hypergraph to model the high-order correlations among the patches and combines the hyperedge features and node features to generate the global representation of WSIs.

3.3 Implementation

For each WSI, we first filter out the unnecessary white background by OTSU [14] algorithm. From each WSI, we randomly sample 2000 patches whose semantic features are extracted by ResNet-34 [7] model pre-trained on ImageNet [2], and the feature dimension is $x_i \in \mathbb{R}^{1\times 512}$. The number of multi-slice hypergraph convolution blocks is set to 3. For training, we use stochastic gradient descent with momentum 0.9 and weight decay 5×10^{-4} with a mini-batch size of 8. The training epoch is set to 100 for each fold. The initial learning rate is set to 0.01. The training process is implemented on an NVIDIA GeForce RTX 3090 GPU.

The five-fold cross-validation strategy is adopted for the proposed method and the compared methods. The results are presented as a composite of the mean values and standard errors obtained from the five-fold cross-validation process.

3.4 Results and Discussions

The experimental results of all the methods measured by C-index on the two datasets are presented in Table 2. MSHGC outperforms all compared methods, achieving the C-index of 0.8441 and 0.8390 on ccRCC_H1 and ccRCC_H2 respectively.

The proposed MSHGC outperformed MIL-based methods, as MIL_Attention, DTFD-MIL, and CLAM ignore the correlations between the patches, which is significantly important in metastasis prediction. While TransMIL takes the correlations between patches into consideration, the noise is also introduced due

Table 3. Ablation study: Comparison on different settings of MSHGC.

Aggregation			Intra-Cross HyConv		ccRCC_H1	ccRCC_H2
Mean	Max	Attention	Intra-HyConv	Cross-HyConv		
		✓		✓	$0.7005_{\pm 0.0221}$	$0.6563_{\pm 0.114}$
		✓	✓		$0.7914_{\pm 0.065}$	$0.7412_{\pm 0.066}$
✓				✓	$0.7364_{\pm 0.145}$	$0.6660_{\pm 0.192}$
	✓			✓	$0.7683_{\pm 0.180}$	$0.8242_{\pm 0.101}$
		✓		✓	$\mathbf{0.8441}_{\pm 0.059}$	$\mathbf{0.8390}_{\pm 0.060}$

to the full connection with patches that are irrelevant to the task. In contrast, MSHGC uses hypergraph convolution to perform feature fusion between patches with similar semantics, avoiding the introduction of too many task-irrelevant features.

Table 2 shows that MSHGC has superior performance than the state-of-the-art graph-based method [1, 11, 12], which demonstrates the limitations of pairwise correlations in modeling of high-order correlations between patches. HGSurvNet uses hypergraphs to model the correlations between patches and achieve better performance than MIL-based methods and graph-based methods, which indicates the superior performance of hypergraphs in modeling high-order correlations. However, HGSurvNet ignores the correlations among the WSIs of a single patient, which limits the performance of HGSurvNet. However, MSHGC not only models the high-order correlations within a single WSI but also captures the information among WSIs by multi-slice hypergraph convolution blocks.

Furthermore, The experimental results of four methods (TransMIL, Patch-GCN, HGSurvNet, and MSHGC) measured by KM-estimation curves [10] of two datasets are shown in Fig. 2. It can be seen that MSHGC has the most significant gap between the curves of the low- and high-risk groups, which denotes that MSHGC has better capability of binary risk classification. All the experimental results demonstrate that the proposed MSHGC method outperforms all the compared methods, including MIL-based methods, the graph-based method, and the latest hypergraph-based method.

3.5 Ablation Study

To evaluate the effectiveness of different components in MSHGC, We conducted ablation experiments on two datasets, whose results are shown in Table 3. Experimental results show that compared to using only intra-hypergraph convolution (Intra-HyConv) or cross-hypergraph convolution (Cross-HyConv), using both simultaneously achieves the best performance, which indicates the importance of intra- and cross-correlations. Furthermore, we apply average pooling, max pooling, and attention-based pooling as aggregation functions respectively. The attention-based method outperforms two other methods, which proves the effectiveness of the attention-based pooling function.

4 Conclusion

Despite WSI-based methods have made progress in recent years, few existing methods extract features based on multi-slice from a single patient. In this work, the proposed MSHGC captures the complex correlations within and across WSIs in a unified framework by innovative multi-slice hypergraph convolution block. To better extract intra-correlations, MSHGC considers both topological connections and semantic pathology features between patches to construct an intra-hypergraph. For constructing a cross-hypergraph, the relationship between patches from the same WSI is ignored to capture cross-correlations more accurately. Finally, MSHGC reduces the interference of task-irrelevant patches on metastasis prediction by the attention module to further improve model performance. Experimental results on two carcinoma datasets demonstrate that MSHGC outperforms the state-of-the-art methods for metastasis prediction.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant No. 62173269.

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., 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: MICCAI. pp. 339–349. Springer (2021)
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. IEEE (2009)
3. Di, D., Zou, C., Feng, Y., Zhou, H., Ji, R., Dai, Q., Gao, Y.: Generating hypergraph-based high-order representations of whole-slide histopathological images for survival prediction. IEEE Trans. Pattern Anal. Mach. Intell. **45**(5), 5800–5815 (2022)
4. Fan, K., Wen, S., Deng, Z.: Deep learning for detecting breast cancer metastases on wsi. In: Innovation in Medicine and Healthcare Systems, and Multimedia: Proceedings of KES-InMed-19 and KES-IIMSS-19 Conferences. pp. 137–145. Springer (2019)
5. Feng, Y., You, H., Zhang, Z., Ji, R., Gao, Y.: Hypergraph neural networks. In: AAAI. vol. 33, pp. 3558–3565 (2019)
6. Gao, Y., Ji, S., Han, X., Dai, Q.: Hypergraph computation. Engineering (2024)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
8. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: ICML. pp. 2127–2136. PMLR (2018)
9. Jiang, J., Wei, Y., Feng, Y., Cao, J., Gao, Y.: Dynamic hypergraph neural networks. In: IJCAI. pp. 2635–2641 (2019)
10. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. J. Am. Stat. Assoc **53**(282), 457–481 (1958)

11. Lee, Y., Park, J.H., Oh, S., Shin, K., Sun, J., Jung, M., Lee, C., Kim, H., Chung, J.H., Moon, K.C., et al.: Derivation of prognostic contextual histopathological features from whole-slide images of tumours via graph deep learning. Nat. Biomed. Eng. pp. 1–15 (2022)
12. Li, R., Yao, J., Zhu, X., Li, Y., Huang, J.: Graph CNN for survival analysis on whole slide pathological images. In: MICCAI. pp. 174–182. Springer (2018)
13. 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)
14. Otsu, N.: A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics **9**(1), 62–66 (1979)
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. Song, J.H., Hong, Y., Kim, E.R., Kim, S.H., Sohn, I.: Utility of artificial intelligence with deep learning of hematoxylin and eosin-stained whole slide images to predict lymph node metastasis in t1 colorectal cancer using endoscopically resected specimens; prediction of lymph node metastasis in t1 colorectal cancer. Journal of gastroenterology **57**(9), 654–666 (2022)
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
18. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. In: ICLR (2017)
19. Xia, L., Huang, C., Zhang, C.: Self-supervised hypergraph transformer for recommender systems. In: SIGKDD. pp. 2100–2109 (2022)
20. Zadeh, S.G., Schmid, M.: Bias in cross-entropy-based training of deep survival networks. IEEE Trans. Pattern Anal. Mach. Intell. **43**(9), 3126–3137 (2021)
21. 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: CVPR. pp. 18802–18812 (2022)