dMIL-Transformer: Multiple Instance Learning Via Integrating Morphological and Spatial Information for Lymph Node Metastasis Classification

Yang Chen, Zhuchen Shao, Hao Bian, Zijie Fang, Yifeng Wang, Yuanhao Cai, Haoqian Wang, Guojun Liu, Xi Li, and Yongbing Zhang

***Abstract*—Automated classification of lymph node metastasis (LNM) plays an important role in the diagnosis and prognosis. However, it is very challenging to achieve satisfactory performance in LNM classification, because both the morphology and spatial distribution of tumor regions should be taken into account. To address this problem, this article proposes a two-stage dMIL-Transformer framework, which integrates both the morphological and spatial information of the tumor regions based on the theory of multiple instance learning (MIL). In the first stage, a double Max-Min MIL (dMIL) strategy is devised to select the suspected top-K positive instances from each input histopathology image, which contains tens of thousands of patches (primarily negative). The dMIL strategy enables a better decision boundary for selecting the critical instances compared with other methods. In the second stage, a Transformer-based MIL aggregator is designed to integrate all the morphological and spatial information of the selected instances from the first stage. The self-attention mechanism is further employed to characterize the correlation between different instances and learn the bag-level representation for predicting the LNM category. The proposed dMIL-Transformer can effectively deal with the thorny classification in LNM with great visualization and interpretability. We conduct various experiments over three LNM datasets, and achieve 1.79%-7.50% performance improvement compared with other state-of-the-art methods.**

***Index Terms*—Histopathological image analysis, multiple instance learning, whole slide image, lymph node metastasis, Transformer.**

Manuscript received 14 September 2022; revised 7 May 2023; accepted 2 June 2023. Date of publication 13 June 2023; date of current version 6 September 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62031023, in part by the Shenzhen Science and Technology Project under Grants JCYJ20200109142808034 and GXWD20220818170353009, and in part by Guangdong Special Support under Grant 2019TX05X187. *(Yang Chen, Zhuchen Shao, and Hao Bian, contributed equally to this work.) (Corresponding author: Yongbing Zhang.)*

Yang Chen, Zhuchen Shao, Hao Bian, Zijie Fang, Yuanhao Cai, and Haoqian Wang are with the Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518000, China (e-mail: cy21@mails.tsinghua.edu.cn; szc21@mails.tsinghua.edu.cn; bianh21@mails.tsinghua.edu.cn; fzj22@mails.tsinghua.edu.cn; cyh20@mails.tsinghua.edu.cn; wanghaoqian@tsinghua.edu.cn).

Yifeng Wang and Yongbing Zhang are with the Harbin Institute of Technology, Shenzhen 518000, China (e-mail: wangyifeng@stu.hit.edu.cn; ybzhang08@hit.edu.cn).

Guojun Liu is with the Computer Science and Technology, Harbin Institute of Technology, Harbin 150000, China (e-mail: hitliu@hit.edu.cn).

Xi Li is with the Gastroenterology, Peking University Shenzhen Hospital, Shenzhen 518036, China (e-mail: lixi122188@sina.com).

Digital Object Identifier 10.1109/JBHI.2023.3285275

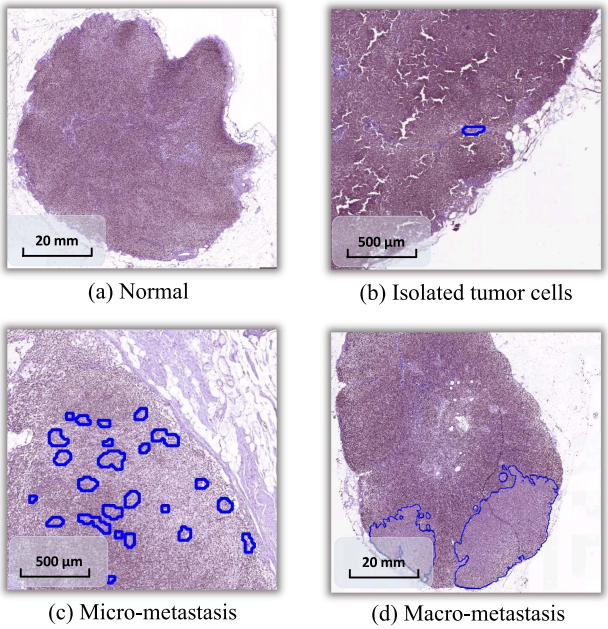
I. Introduction

THE study of histopathological slide is the gold standard for cancer diagnosis and prognosis. Since the whole slide image (WSI) scanner was approved for clinical use by the FDA in 2017 [1], deep learning has presented an opportunity in digital pathology image analysis [2], [3], [4]. According to the tumor, node, metastasis (TNM) staging system in the 8th edition of AJCC [5], the pN-staging of breast cancer needs to detect lymph nodes metastasis (LNM). As shown in Fig. 1, LNM categories include Normal (0 mm), Isolated tumor cells (ITC, metastasis $<$ 0.2 mm), Micro-metastasis (Micro, 0.2 mm $<$ metastasis $<$ 2.0 mm) and Macro-metastasis (Macro, metastasis $>$ 2.0 mm), while a WSI may exceed 25 mm in size and lacks detailed pixel-level annotation. Therefore, the detection of LNM poses a great challenge.

In clinical, WSI typically has gigapixels (e.g., 100000 $\times$ 50000), and it is a common practice to remove the background of the image by the OTSU threshold [6] and split the image into tens of thousands of patches in the data preprocessing stage. Previous approaches [7] require a pathologist to manually annotate the slide at the pixel-level for supervised learning, which is tedious, time-consuming, and expensive. Therefore, most WSIs only have slide-level labels, and the primary approach takes the diagnostic analysis of WSI as a weakly-supervised learning problem [8].

Current weakly supervised WSI classification methods usually follow the MIL framework [9], where the WSI is taken as a bag, and each bag contains tens of thousands of instances (patches). However, there are two problems in the application of classical MIL methods to LNM classification:

- WSIs in the ITC and Micro contain only a few positive regions, which causes the embedded-level MIL methods to get a bag-level representation with excessive noise information. Meanwhile, current instance-level MIL methods



(a) Normal (b) Isolated tumor cells

(c) Micro-metastasis (d) Macro-metastasis

Fig. 1. Representative samples of different LNM categories, where (a)(b) are considered as negative LNM and (c)(d) are considered as positive LNM.

can not accurately select the critical cancerous instances from WSI, resulting in poor classification performance.

- WSIs in both Micro and Marco categories contain positive regions [10], whose main difference lies in the spatial distribution of tumor cells. Therefore, the classical MIL methods can not learn the spatial correlation information between instances because they are based on the independent and identically distributed (I.I.D.) assumption of instances, which only utilize morphological characteristics of the instances.
- To design more precise therapies, pathologists need to subdivide the kind of metastasis into ITC, Micro, or Micro during actual clinical diagnosis. However, in previous LNM classification tasks, WSI was merely categorized as metastasis or normal.

Therefore, it would be much desirable to introduce spatial information in multi-category LNM classification. Recently, numerous studies [11], [12] have shown that Transformer can effectively exploit the spatial information between instances. Since Dosovitskiy et al. [13] proposed the Vision Transformer, it has received a great deal of attention in the field of computer vision [14]. Adding position encoding offers the Transformer the ability to correlate spatial information. For 2D images, the input embedding of the Transformer usually needs to be spatially continuous. However, since the critical instances of WSI are spatially discrete, the traditional block-based spatial position encoding approach cannot effectively encode the discrete instances.

To overcome the above challenges, in this article, we propose a two-stage dMIL-Transformer framework. In the first stage, a double Max-Min MIL (dMIL) strategy is proposed to select the critical instances in WSIs, which takes into account the true-positive and true-negative instances in the positive bag and the true-negative and false-positive instances in the negative bag. This strategy trains the instance selector by critical instances with pseudo labels inherited from the bags. In the second stage, we propose a Transformer-based MIL aggregator for bag-level LNM classification, Transformers can leverage spatial and rank information efficiently to model the relationship between instances and better distinguish hard-to-predict categories. The main contributions of this article are summarized as follows:

- We propose a MIL framework for the multi-category LNM classification for the first time. In contrast to prior research, this study not only distinguishes between metastasis and Normal but also subdivides the type of metastasis.
- We devise a dMIL strategy based on the instance-level MIL approach to select the critical instances in the first stage of the proposed framework, which is more comprehensive than the traditional MIL strategy.
- We firstly integrate the morphological, spatial, and malignancy rank information of the critical instances to subdivide LNM subtypes. Then, we utilize the Transformer to aggregate the above three embeddings based on the embedding-level MIL approach.
- Our extensive experiments over three LNM datasets demonstrate the generality, effectiveness, and efficiency of the proposed dMIL-Transformer, which outperforms the state-of-the-art MIL methods in general. Especially, the traditional MIL algorithm tends to misclassify the ITC and Micro, while the dMIL-Transformer can effectively distinguish these hard-to-predict categories.

II. Related Work

In this section, we introduce the fundamental paradigms of MIL, MIL-based deep learning algorithms, and the applications of MIL to histopathological images. Then, we introduce the self-attention and position encoding in computer vision.

A. MIL for WSI Classification

MIL was proposed by [15] for solving weakly-supervised learning problems. Since WSI contains gigapixels and usually has only slide-level annotations [16], it is an ideal solution for applying MIL to WSI classification [17], [18]. Amores [19] divided MIL into three paradigms: bag-level, instance-level and embedding-level approach. The current applications of these three paradigms in pathological images include critical instance selection and bag classification.

For critical instance selection, only the instance-level MIL methods can directly classify instance [20]. Chikontwe et al. [21] suggested that selecting critical instances can improve the performance of MIL, and Campanella et al. [22] proposed a Max-Max strategy to select critical instances. Xu el al. [23] proposed a cMIL strategy to pay extra attention to the false-positive instances in the negative bags. Lerousseau et al. [24], [25] proposed a $\alpha\beta$MIL strategy to focus on the negative instances in the positive bag additionally. However, only a tiny cancerous region in LNM means severe class imbalance, resulting in poor performance for these methods.

For bag classification, embedding-level MIL methods usually perform better than instance-level MIL methods [20].

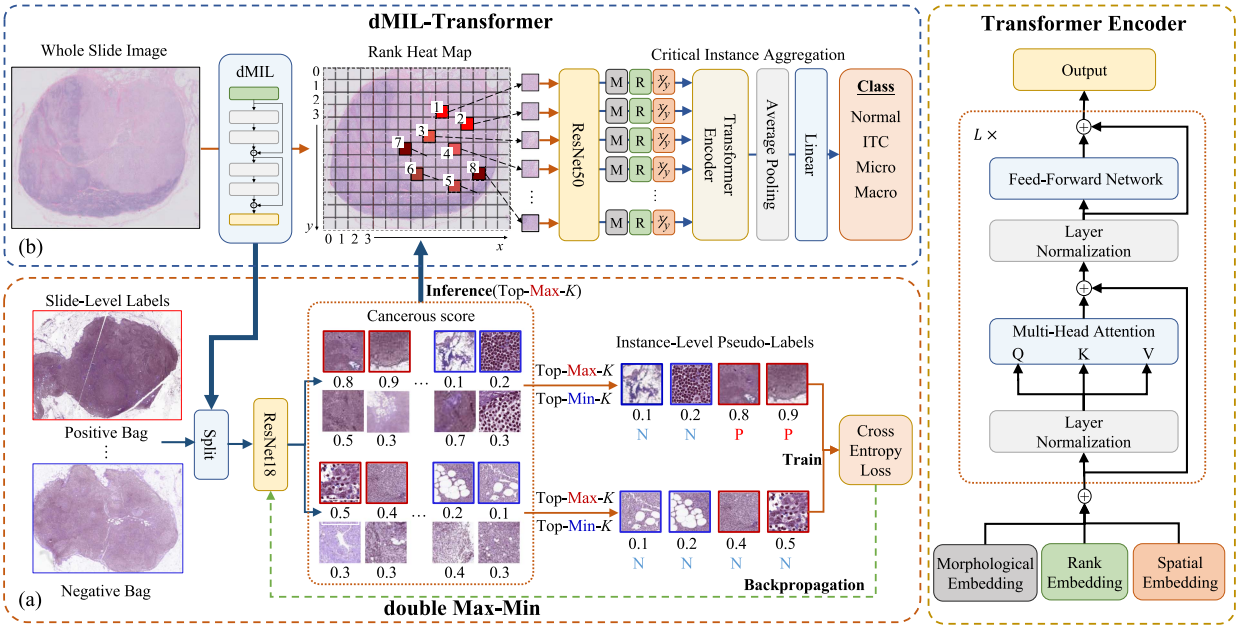


Fig. 2. Proposed dMIL-Transformer framework. It includes two stages: (a) Using double Max-Min strategy to train the ResNet18, which is then utilized to select top Max K cancerous-scoring instances (N means negative, P means positive); (b) Using ResNet50 to extract morphological embedding of instances, and then aggregating the instance embedding through a Transformer-based MIL aggregator. The input of the Transformer includes the morphological, spatial, and rank embeddings of the instance.

Pooling-based methods, including Max-pooling [26] and Mean-pooling [27] are utilized traditionally. Ilse et al. [28] proposed attention-based MIL to effectively balance the interpretability of instance and the performance of the model. Shi et al. [29] introduced an attention mechanism in bag loss for MIL to further improve the performance. Li et al. [30] proposed a patch Transformer to predict multiple tags of WSI simultaneously. However, the above methods ignore the correlation between instances, making these methods difficult to model the spatial structure information between instances [31]. To overcome this problem, Li et al. [32] proposed DSMIL to consider the correlation between the highest-scoring positive instance and all the remaining instances. Campanella et al. [22] proposed MIL-RNN, which firstly extracted the critical instances by an instance-level MIL and then aggregated the selected instances by a recurrent neural network (RNN). However, RNN has the problem of gradient disappearance, making it difficult to process long sequences, when WSIs contain amounts of instances. Shao et al. [33] leveraged Transformer to implicitly model the context information of an instance. Zhang et al. [34] applied Grad-CAM in ABMIL to select the key instances and aggregated them by a double-tier MIL. However, these methods cannot directly utilize spatial information between instances.

B. Position Encoding and Self-Attention

Since Transformer, which employs self-attention and position encoding, was proposed by Vaswani et al. [35], it has achieved great success in natural language processing. Recently, Dosovitskiy et al. [13] proposed vision Transformer to model the global information, which firstly splits images into patches and then employs a Transformer encoder. Chu et al. [11] suggested that one of the critical problems of Transformer is how to use the position information efficiently. Recent research shows that Transformer has promising applications for medical images. Yun et al. [36] proposed to embed Transformer into U-Net for lesion region segmentation in medical images. Li et al. [30] proposed a multi-head attention network for tumor subtype segmentation, and Gao et al. [37] attempted to aggregate cell nuclei features based on Transformer for cancer subtype classification.

However, due to the large number of patches in WSI, it is challenging to feed all patches into the Transformer encoder simultaneously. It is better to select critical instances to represent the whole bag information. Nevertheless, this will lack critical instances' continuous spatial information, which is vital for learning the correlation between instances. Therefore, it is a great challenge to encode the discrete spatial information to apply Transformer in WSI.

III. Method

A. Overview of the dMIL-Transformer

The framework of our proposed dMIL-Transformer is shown in Fig. 2, which consists of two ResNets [38] and a Transformer encoder. The whole processing pipeline consists of two stages. In the first stage, an instance-level binary classifier is trained using a designed dMIL strategy to select the K instances with the highest cancerous score. In the second stage, features of the critical instances are extracted using ResNet50, and the spatial and rank information of instances are encoded by sinusoidal position embedding, respectively. The three embeddings are further fused, and the Transformer establishes the correlation

between the instance embeddings to obtain the bag embedding, which is then used to predict the bag-level labels.

B. Problem Formulation

In MIL, the set of training samples are considered as a bag $B = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, with n being the number of instances in the bag. x_i denotes the ith instance in the bag, y_i denotes the ith instance label in the bag, and Y is defined as the bag label.

In the first stage, the selection of K instances is an instance-level binary MIL classification problem, since the Normal and ITC are considered as negative while Micro and Macro are considered as positive. For binary MIL classification, Y is defined as:

$$Y = \begin{cases} 0, \text{iff } \Sigma y_i = 0, y_i \in \{0, 1\}, \\ 1, \text{ otherwise.} \end{cases} \tag{1}$$

In the second stage, the multi-classification of LNM is defined as:

$$Y = \begin{cases} 0, \text{iff } \Sigma y_i = 0, y_i \in \{0, 1\}, \\ c_i, \text{ otherwise,} \end{cases} \tag{2}$$

where $c_i \in N^+$ represents the categories of ITC, Micro and Macro based on the diameter of metastasis.

Here a scoring function S is further defined for predicting the bag label Y:

$$Y = S(\mathcal{X}) = g(\underset{x \in \sigma(\mathcal{X})}{\Omega} f(x)), \tag{3}$$

where Ω represents the feature aggregation operator, σ represents the instance selection operator, $\mathcal{X}$ indicates the set of instances in the bag, f is used to encode instance embedding, and g is used to predict bag labels. Here, f is defined as:

$$f(x) = w_1 E_m(x) + w_2 E_s(x) + w_3 E_r(x), \tag{4}$$

where E_m denotes the morphological embedding, E_s denotes the spatial embedding, E_r denotes the rank embedding, and w_i denotes the weight factor, with $w_1 = 1$, $w_2 = 0.5$, and $w_3 = 1$. Obviously, $f(x)$ can be further considered to integrate morphological, spatial and rank embeddings of instances.

C. dMIL Strategy for Critical Instance Selection

Since WSI contains tens of thousands of patches, it will consume enormous computing resources by directly aggregating all the instances. So it is a common practice to select the critical instances as a subset of the bag to aggregate bag-level features [22]. However, the previous strategy cannot train a good decision boundary to select instances, especially the false positive instances will be wrongly selected.

To address this problem, we train a binary instance-level MIL classifier to obtain the instance selection operator σ_{dMIL}. For a given scoring function (ResNet18) and a bag, we select K instances with the highest scores and the K instances with the lowest scores. Then we provided the pseudo labels for all the selected instances as:

$$\sigma_{dMIL}(\mathcal{X}) = \begin{cases} p_{x_i^{(+)}} = 1, & x_i^{(+)} \in \sigma_{Max}\left(\mathcal{X}^{(+)}, K\right), \\ p_{x_i^{(+)}} = 0, & x_i^{(+)} \in \sigma_{Min}\left(\mathcal{X}^{(+)}, K\right), \\ p_{x_i^{(-)}} = 0, & x_i^{(-)} \in \sigma_{Max}\left(\mathcal{X}^{(-)}, K\right), \\ p_{x_i^{(-)}} = 0, & x_i^{(-)} \in \sigma_{Min}\left(\mathcal{X}^{(-)}, K\right), \end{cases} \tag{5}$$

where p_{x_i} denotes the pseudo label of the ith instance, $+/-$ denotes the bag label as positive/negative, $\sigma_{Max}(\cdot, K)$ denotes the K instances with the highest score in each bag, and $\sigma_{Min}(\cdot, K)$ denotes the K instances with the lowest score in each bag.

Denote $\hat{y}_i$ as the probability of a positive prediction for the ith instance using ResNet18, then $Loss_{patch}$ is defined as:

$$\hat{y}_i = \text{ResNet18}(x_i), \qquad x_i \in \sigma_{dMIL}(\mathcal{X}), \tag{6}$$

$$Loss_{patch} = -\sum_{\sigma_{dMIL}} (p_i \log \hat{y}_i + (1 - p_i) \log(1 - \hat{y}_i)). \tag{7}$$

The detailed critical instance selector is described in Algorithm 1, where the weights of ResNet18 and the pseudo labels of the selected instances in each bag are updated in each epoch. Fig. 3 shows different instance selection strategies in MIL. The traditional MIL method uses the Max-Max strategy, which only considers the positive instances with the highest scores in the positive bag and the negative instances with the highest scores in the negative bag. However, even a positive bag still contains many negative instances, and traditional MIL methods cannot achieve satisfactory results. Conversely, the dMIL strategy proposes a comprehensive approach by considering both positive and negative instances in the positive bags as well as negative and false positive instances in the negative bags.

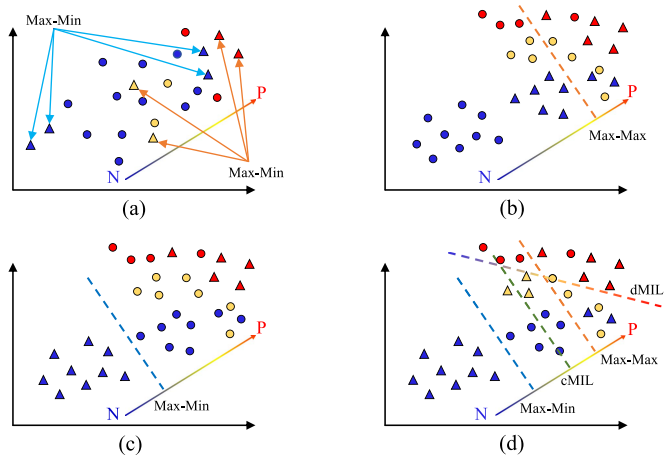


Fig. 3. dMIL strategy diagram. Blue indicates instance in negative bag, yellow/red indicates negative/positive instance in positive bag. We use triangles to represent selected instances, and circles to represent unselected instances. Each dotted line represents the decision boundary of the classifier trained by the selected instance. (a) The dMIL strategy takes into account the most positive and negative instances from both positive and negative bags to determine the decision boundary. (b) Max-Max strategy is the decision boundary of baseline MIL [22], (c) cMIL strategy is the decision boundary based on the combination of Max-Max and Max-Min proposed by [23], (d) and our classifier trained based on dMIL strategy has better decision boundary.

Algorithm 1: Critical instance selector using dMIL strategy.

Input: The bag $\mathcal{X}$ with instances $x_1, \ldots, x_n$
Output: Critical instances $\hat{x}_1, \ldots, \hat{x}_K$
for *epoch* $\in [0:1:nums\_epoch]$ **do**
  %1. Inference step: predicting instance score.
  for $i \in [1, n]$ **do** $\hat{y}_i \leftarrow \text{ResNet18}(x_i)$;
  $\hat{x}_{max,1}, \ldots, \hat{x}_{max,K} \leftarrow Max(\hat{y}_1, \ldots, \hat{y}_n)$ ◁ get the K critical instances with the highest score
  $\hat{x}_{min,1}, \ldots, \hat{x}_{min,K} \leftarrow Min(\hat{y}_1, \ldots, \hat{y}_n)$ ◁ get the K critical instances with the lowest score
  if $\mathcal{X}$ *is positive* **then**
    $p_{x_{max,i}} \leftarrow 1$ ◁ pseudo label is set to 1
    $p_{x_{min,i}} \leftarrow 0$ ◁ pseudo label is set to 0
  else if $\mathcal{X}$ *is negative* **then**
    $p_{x_{max,i}} \leftarrow 0$ ◁ pseudo label is set to 0
    $p_{x_{min,i}} \leftarrow 0$ ◁ pseudo label is set to 0
  end
  %2. Training step: updating network parameters.
  $Loss_{patch} = -\sum_j \left(p_j \log \hat{y}_j + (1 - p_j) \log (1 - \hat{y}_j)\right)$
end
return $\hat{x}_1, \ldots, \hat{x}_K \leftarrow \hat{x}_{max,1}, \ldots, \hat{x}_{max,K}$

D. Transformer-Based MIL Aggregator

Applying the trained ResNet18 by dMIL strategy, we select the top-K instances with the highest scores, i.e. $\{x_1, x_2, \ldots, x_K\} \in \sigma_{Max}(\mathcal{X})$ for each bag. And then $\{x_1, x_2, \ldots, x_K\}$ is fed into a Transformer encoder, which functions as the aggregator operator Ω in (3) to obtain bag-level embedding. The Transformer implementation includes multi-head self-attention (MSA) and feed-forward network (FFN), which are defined as:

$$\mathbf{X}^0 = [f(x_1); f(x_2); \ldots; f(x_K)], \tag{8}$$

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}_V, \tag{9}$$

$$head = \text{SA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D_k}}\right)\mathbf{V}, \tag{10}$$

$$\text{MSA}(\mathbf{X}) = \text{Concat}(head_1, \ldots, head_h)\,\mathbf{W_O}, \tag{11}$$

$$\mathbf{X}_*^{\ell-1} = \text{MSA}(\text{LN}(\mathbf{X}^{\ell-1})) + \mathbf{X}^{\ell-1}, \ell = 1 \ldots L, \tag{12}$$

$$\mathbf{X}^{\ell} = \text{FFN}(\text{LN}(\mathbf{X}_*^{\ell-1})) + \mathbf{X}_*^{\ell-1}, \ell = 1 \ldots L, \tag{13}$$

where $\mathbf{X}^0 \in \mathbb{R}^{K \times D}$, $\mathbf{W_Q} \in \mathbb{R}^{D \times D_q}$, $\mathbf{W_K} \in \mathbb{R}^{D \times D_k}$, $\mathbf{W_V} \in \mathbb{R}^{D \times D_v}$, $\mathbf{W_O} \in \mathbb{R}^{hD_v \times D}$, $head_i \in \mathbb{R}^{K \times D_v}$, D is the dimension of instance embedding, $\frac{1}{\sqrt{D_k}}$ is the scale factor, SA is self-attention layer, L is the number of MSA blocks, h is the number of heads in each MSA block, and Layer Normalization (LN) is applied before every FFN and MSA block. FFN is typically processed by a two-layer linear transformation with an activation GELU [39] in between. A residual function is added for each of the sub-layers.

Next, we use a linear layer as g indicated in (3) and average pooling layer to predict the bag label $\hat{Y}$:

$$\hat{Y} = \text{Linear}\left(AvgPool\left(\mathbf{X}^L\right)\right). \tag{14}$$

Finally, we calculate the cross-entropy $Loss_{bag}$ between $\hat{Y}$ and Y as indicated in (2) to update the network by:

$$Loss_{bag} = -\sum (Y \log \hat{Y} + (1 - Y) \log(1 - \hat{Y})). \tag{15}$$

1) Morphological Embedding: The morphological embedding E_m of each patch is extracted from the ImageNet pre-trained ResNet50 model:

$$E_m = \text{ResNet50}(x), x \in \sigma_{Max}(\mathcal{X}), \tag{16}$$

where $E_m \in \mathbb{R}^D$ is the feature after the global average pooling layer, with D being the dimension of instance embedding.

2) Spatial Embedding: We introduce spatial embedding to distinguish the tumor spatial difference between Micro-metastasis and Macro-metastasis categories. Since the K critical instances selected in WSI are spatially discrete, the learnable position embedding in discrete space cannot cover the whole space. Consequently, the sinusoidal position embedding is employed as:

$$(p_x, p_y) = \text{coord}(x), x \in \sigma_{Max}(\mathcal{X}), \tag{17}$$

$$E_{s,i} = \begin{cases} \sin\left(\frac{p_x}{10000^{i/d}}\right), & i = 0, 2, \ldots, d-2, \\ \cos\left(\frac{p_x}{10000^{(i-1)/d}}\right), & i = 1, 3, \ldots, d-1, \\ \sin\left(\frac{p_y}{10000^{(i-d)/d}}\right), & i = d, d+2, \ldots, 2d-2, \\ \cos\left(\frac{p_y}{10000^{(i-1-d)/d}}\right), & i = d+1, d+3 \cdots 2d-1, \end{cases} \tag{18}$$

where the function coord extracts the corresponding geometric coordinates of an instance, (p_x, p_y) denotes the absolute position of the pixel center of the patch. To deal with the problem of tremendous pixel numbers in each WSI, (p_x, p_y) will be scaled by $\frac{1}{4 \times patchsize}$ to compute the spatial embedding. To encode (p_x, p_y) in E_s simultaneously, we set $d = D/2$, and use the first d dimensions to encode p_x and the last d dimensions to encode p_y.

3) Rank Embedding: To better focus on the most suspected positive instance when classifying Normal and ITC categories, we introduce a rank embedding E_r. The ranking information of the top-K instances is encoded through an 1-D sinusoidal position embedding, which can be described as:

$$r = \text{rank}(x), x \in \sigma_{Max}(\mathcal{X}), \tag{19}$$

$$E_{r,i} = \begin{cases} \sin\left(\frac{r}{10000^{i/D}}\right), & i = 0, 2, \ldots, D-2, \\ \cos\left(\frac{r}{10000^{(i-1)/D}}\right), & i = 1, 3, \ldots, D-1, \end{cases} \tag{20}$$

where rank denotes the rank of each instance cancerous score within each bag.

IV. Experiment

To validate the superiority of our dMIL-Transformer, we carried out various experiments on LNM classification over a variety of public cohorts. Comparison results, ablation study, and interpretability analysis are provided in this section.

A. Dataset

To test the effectiveness of our dMIL-Transformer in LNM classification problems, we evaluate it over all public breast

TABLE I
DATASET DESCRIPTION

Dataset \ Category	Negative		Positive		Sum
	Normal	ITC	Micro	Macro	
CAMELYON17	318	36	59	87	500
Train Dataset	191	27	43	39	300
Valid Dataset	63	3	7	27	100
Test Dataset	64	6	9	21	100
CAMELYON16	80		27	22	129
SLN-Breast	94		36		130

cancer LNM datasets as we know. The exact number of datasets is described in Table I.

We use CAMELYON17 [40] as the development cohort, because it has the most subdivided LNM subtype for WSIs. For CAMELYON17, since its test dataset is not publicly available, we only use the training dataset (500 WSIs). For the CAMELYON16 [40], we use the test dataset to evaluate the three classification performance, including: Normal, Micro-metastasis, and Macro-metastasis. We also use the SLN-Breast released by The Cancer Imaging Archive [41] for the binary classification.

B. Experiment Setup and Evaluation Metrics

For different datasets, it should be noted that CAMELYON17 contains four categories, namely Normal, ITC, Micro, and Macro. The samples are randomly divided into training, validation, and testing datasets in ratio of 6:2:2. However, CAMELYON16 test dataset takes ITC and Normal as negative, consequently CAMELYON16 only has 3 categories. Furthermore, SLN-Breast only has binary labels indicating whether a WSI has metastasis. For fair comparison, we train and test our dMIL-Transformer as a four classification model over CAMELYON17. In addition, we combine the Normal and ITC as negative category over CAMELYON17 and train our dMIL-Transformer as a three classification model, which is then tested over CAMELYON16 test dataset. For SLN-Breast, the Normal and ITC are treated as negative with Micro and Macro combined as positive, making the dMIL-Transformer be a binary classification model. All the competing methods adopt the same settings.

For objective comparison, we evaluate the performance of LNM classification primarily via macro-average area under the curve (AUC). Besides, the bag-level Recall, F1-Score, Precision, and Accuracy are also considered, where threshold is 0.5 for binary classification. The LNM classification is rather unbalanced and AUC is always applied to evaluate unbalanced samples, therefore, AUC is the most important. Recall is also significant in the task as undetected diagnosis is more harmful than a misdiagnosis in LNM classification. However, as F1-Score, Precision, and Accuracy are easily affected by data imbalance, they are less significant in LNM classification.

C. Hardware and Software Implementation

All experiments are conducted on a computing server equipped with three NVIDIA 3090 24 G Graphics Processing Units (GPUs), two Intel Xeon(R) Silver 4210R Central Processing Units (CPUs), 512 GB of memory, and 200 TB of storage space. In the first stage, each WSI contains an average of 8,000 patches, with the processing time for each WSI being approximately 20 s. In the second stage, it is only necessary to predict the candidate critical instances within each WSI, taking about 1 s for each WSI.

The model is implemented with the open source libraries PyTorch 1.8 [42] and PyTorch-lightning 1.5.0, and the model evaluation metrics are implemented with TorchMetrics 0.6.0. Horizontal flipping, vertical flipping, and 90° random rotation are utilized for data augmentation. In the first stage, the mini-batch size is 512, and learning rate is 0.0002. In the second stage, the mini-batch size is 10, and learning rate is 0.0002. The Ranger optimizer [43] is used, and the maximum number of training epochs in both of the two stages is 30, using early stop strategy. The numbers of selected instance K are 100 in four classification, 50 in binary classification, and 200 in three classification, respectively. The number of MSA blocks L is 3. Based on previous studies [44], [45], ResNet [38] has good performance and reasonable computational cost for image classification task. For a fair comparison, all the methods use ImageNet pre-trained ResNet50 as feature extractor and use the same training hyper-parameters.

D. WSIs Preprocessing

WSI is a high-resolution digital slide created by panoramic scanning technology, capturing detailed tissue samples for use in pathological diagnosis and research. Under 20X magnification (0.5μm/pixel), WSI usually contains gigapixels (e.g., 100000×50000), so it cannot be directly used as the input of the neural network. The usual practice is to read the WSI image through the OpenSlide software and divide it into tens of thousands patches for subsequent processing. Due to the different parameters of various scanners, not all WSIs have 20X magnification. For those WSIs at 40X magnification, it is necessary to downsample to 20X magnification to ensure that the extracted patches have the same size of receptive field. The resolution of each patch is set to 224×224 based on experience, and then patches with less than 20% of the tissue area were removed by the OTSU method.

E. Baseline

The baseline methods include one-stage and two-stage methods, and our method belongs to two-stage method. One-stage methods means that directly aggregating all instances in WSI to represent the bag, two-stage methods means that selecting critical instances firstly, and then aggregating the selected instances to obtain a bag-level representation. The results of all baseline methods are performed under the same experimental settings.

The one-stage baseline methods we select are as follows:

- **ABMIL** [28] uses an attention-based network to give each instance an importance score and then aggregates all instances to obtain a bag-level score. **Gated-ABMIL** [28] is a better method based on ABMIL using gating mechanism.
- **CLAM** [44] is an improved method, on the basis of ABMIL, by introducing clustering-constrained loss, where **CLAM-SB** is based on single-branch-attention, and **CLAM-MB** is based on multi-branch-attention.

TABLE II
COMPARISON RESULTS OVER CAMELYON17 DATASET

Method	AUC	Recall	F1-Score	Precision	Accuracy
ABMIL [28]	0.8394	0.5595	0.5682	0.6196	0.8600
Gated-ABMIL [28]	0.8442	0.5595	0.5761	**0.6792**	0.8600
CLAM-SB [44]	0.8493	0.5317	0.5311	0.5894	0.8500
CLAM-MB [44]	0.7954	0.6151	**0.6303**	0.6620	**0.8800**
DSMIL [32]	0.7748	0.4684	0.4392	0.4146	0.8100
Loss-Attention [29]	0.8013	0.5000	0.4759	0.4561	0.8700
TransMIL [33]	0.8867	0.6112	0.6104	0.6125	0.8700
MIL-Center [21]	0.7507	0.5441	0.5602	0.6176	0.8000
MIL-RNN [22]	0.8041	0.4616	0.4472	0.4382	0.6500
DTFD-MIL [34]	0.8418	0.6151	**0.6303**	0.6620	0.8700
MIL [22]+TRF	0.8585	0.5519	0.5448	0.5396	0.8200
cMIL [23]+TRF	0.8607	0.5530	0.4879	0.4979	0.6600
$\alpha\beta$MIL [24]+TRF	0.8047	0.5762	0.5330	0.5164	0.7600
dMIL-Transformer	**0.9046**	**0.6552**	0.6170	0.5942	0.8700

The bold values represent the best metrics and underline the second best.

TABLE III
COMPARISON RESULTS OVER CAMELYON16 DATASET

Method	AUC	Recall	F1-Score	Precision	Accuracy
ABMIL [28]	0.7847	0.6061	0.5807	0.5758	0.7596
Gated-ABMIL [28]	0.7706	0.6033	0.5950	0.6891	0.7596
CLAM-SB [44]	0.7646	0.5909	0.5713	0.5758	0.7519
CLAM-MB [44]	0.7882	0.5825	0.6016	0.6960	0.7519
DSMIL [32]	0.7424	0.5303	0.5198	0.5632	0.7209
Loss-Attention [29]	0.7360	0.5758	0.5570	0.5693	0.7442
TransMIL [33]	0.8417	0.6305	0.6646	0.7865	0.7596
MIL-Center [21]	0.7586	0.5798	0.5724	0.6234	0.7364
MIL-RNN [22]	0.6774	0.5212	0.5571	0.6316	0.6279
DTFD-MIL [34]	0.7987	0.6279	0.6322	0.7354	0.7751
MIL [22]+TRF	0.7843	0.5495	0.5523	0.6766	0.7209
cMIL [23]+TRF	0.8047	0.5879	0.6218	0.7646	0.7519
$\alpha\beta$MIL [24]+TRF	0.7749	0.6569	0.6625	0.6845	0.6821
dMIL-Transformer	**0.9167**	**0.6952**	**0.7122**	**0.8638**	**0.8140**

The bold values represent the best metrics and underline the second best.

TABLE IV
COMPARISON RESULTS OVER SLN-BREAST DATASET

Method	AUC	Recall	F1-Score	Precision	Accuracy
ABMIL [28]	0.8596	**0.8505**	**0.8756**	0.9153	**0.9076**
Gated-ABMIL [28]	0.8484	0.8452	0.8667	0.8988	0.9000
CLAM-SB [44]	0.8582	0.8121	0.8342	0.8697	0.8769
CLAM-MB [44]	0.8537	0.8452	0.8667	0.8988	0.9000
DSMIL [32]	0.8452	0.8259	0.8462	0.8766	0.8846
Loss-Attention [29]	0.8124	0.7813	0.7694	0.7609	0.8076
TransMIL [33]	0.8384	**0.8505**	**0.8756**	0.9153	**0.9076**
MIL-Center [21]	0.7432	0.4946	0.4170	0.3605	0.7153
MIL-RNN [22]	0.8168	0.5978	0.4241	0.6393	0.4307
DTFD-MIL [34]	0.8602	**0.8505**	**0.8756**	0.9153	**0.9076**
MIL [22]+TRF	0.8425	0.7538	0.7070	0.7064	0.7307
cMIL [23]+TRF	0.8421	0.8226	0.8517	0.9038	0.8923
$\alpha\beta$MIL [24]+TRF	0.8699	0.7571	0.7027	0.7071	0.7231
dMIL-Transformer	**0.8936**	0.8141	0.8483	**0.9178**	0.8923

The bold values represent the best metrics and underline the second best.

- **DSMIL** [32] is a method using dual stream networks, and it focuses on the relationship between the most suspected positive instance and other instances.
- **Loss-Attention** [29] introduces the attention mechanism to the loss function by an auxiliary fully connected layer.
- **TransMIL** [33] introduces a correlative-MIL paradigm and utilizes the Transformer network to aggregate instances embedding.

The two-stage baseline methods we select are as follows:

- **MIL-RNN** [22] firstly selects instances by Max-Max strategy, and then aggregates the selected instances by RNN.
- **MIL-Center** [21] is an improved method that introduces center loss to the instance selector on the basis of MIL-RNN, utilizing pyramid-CNN to aggregate the features of the selected instances.
- **DTFD-MIL** [34] introduces the concept of pseudo bags to alleviate the limited number of WSIs and formulate a double-tier MIL framework by the instance probability derivation.

MIL/cMIL/$\alpha\beta$MIL/dMIL+Transformer (TRF) mean using different instance selection strategies and the same Transformer-based aggregator, and the instance selection strategies are considered as follows:

- **MIL** [22] selects the instances with highest score in the positive and negative bag by Max-Max strategy.
- **cMIL** [23] and **$\alpha\beta$MIL** [24], based on **MIL**, additionally consider the instances with the lowest scores in the positive and negative bag, respectively.
- **dMIL**(ours) simultaneously selects the instance with the highest and lowest scores in both positive and negative bags based on double Max-Min strategy.

F. Comparison Result

As shown in Table II, Table III and Table IV, we **bold** the best metrics and underline the second best. Our proposed dMIL-Transformer achieves 1.79% improvement over CAEMLYON17 in AUC. Further, the performance of dMIL-Transformer is tested on the external cohorts CAEMLYON16 and SLN-Breast, and our method has good generalization performance, which achieves 7.50%/2.61% improvement in AUC, respectively. For other metrics, dMIL-Transformer also showed considerable competitiveness compared with other methods. Particularly for Recall, high Recall rate is crucial in clinical since false negatives may result in severe repercussions.

For one-stage methods, the results of attention-based methods are not satisfactory. This is because all the above attention-based methods, except DSMIL and TransMIL, ignore the correlation between instances, which misclassifies the ITC as Normal and Micro as Macro. Although DSMIL pays attention to the morphological correlation between the highest-scoring instance and the remaining instances, it ignores the correlation of spatial information between different instances. For TransMIL, it aggregates all instances in the bag simultaneously, but ignores the importance between instances. In contrast, our proposed dMIL-Transformer can achieve 1.79%/7.50%/3.40% in AUC higher than the one-stage methods over CAMELYON17, CAMELYON16, and SLN-Breast datasets, respectively.

In terms of the two-stage methods, firstly for different instance selectors, the MIL strategy ignores the lowest scoring instances in the WSI, making it difficult to identify false positive regions in tumor tissue. The cMIL and $\alpha\beta$MIL strategy ignore the negative regions in the positive and negative bag, respectively. In contrast, the dMIL strategy utilizes more comprehensive

TABLE V
COMPARISON OF NUMBER OF TOP-K INSTANCES ON CAMELYON17

Instance Number	AUC	Recall	F1-Score	Precision	Accuracy
$K = 25$	0.8584	0.4960	0.4533	0.4177	0.8400
$K = 50$	0.8981	0.6311	0.6073	0.5868	0.8600
$K = 100$ ✓	**0.9046**	0.6552	**0.6170**	**0.5942**	**0.8700**
$K = 200$	0.8675	**0.6591**	0.6113	0.5788	0.8500

The bold values represent the best metrics and underline the second best.

TABLE VI
COMPARISON OF TRANSFORMER LAYER NUMBERS ON CAMELYON17

Layer Number	AUC	Recall	F1-Score	Precision	Accuracy
1 Layers	0.8875	0.6274	0.5895	0.5644	0.8300
2 Layers	0.8886	0.6473	0.5951	0.5712	0.8400
3 Layers ✓	**0.9046**	**0.6552**	**0.6170**	0.5942	**0.8700**
6 Layers	0.8847	0.5397	0.5323	0.5854	0.8500
12 Layers	0.8630	0.5561	0.5720	**0.6322**	0.8500

The bold values represent the best metrics and underline the second best.

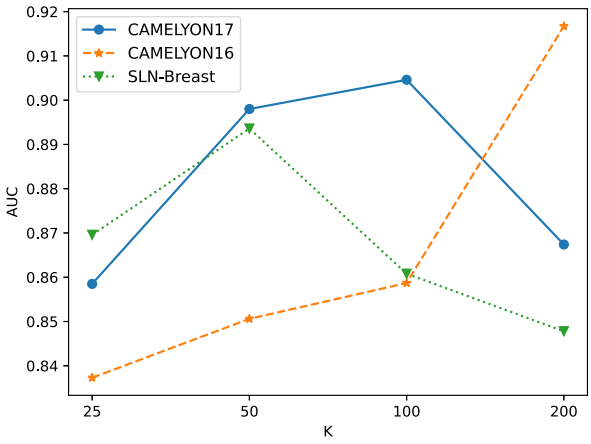


Fig. 4. AUC scores of the different K over three LNM datasets.

information from the WSI, and more details are shown in Fig. 6. In the second step, we employ the Transformer-based aggregator to better utilize the ranking and spatial information of the instances compared to the MIL-Center, MIL-RNN, and DTFD. Therefore, our method achieves 4.39%/11.20%/2.37% improvement in AUC over CAMELYON17, CAMELYON16, and SLN-Breast datasets, respectively.

G. Ablation Study

To further illustrate the effects of the number of selected critical instances, Transformer layers, and the contribution of spatial and rank embeddings in the dMIL-Transformer, a series of ablation studies are conducted. All ablation experiments are conducted over CAMELYON17.

1) Effects of Critical Instance Number: The number of selected critical instances, K, is an essential hyperparameter in dMIL-Transformer. Here we set $K = 25$, 50, 100, 200 while fixing the other hyperparameters. The ablation results are shown in Table V. As depicted, dMIL-Transformer with $K = 100$ significantly outperforms other settings of K, reaching an AUC of 0.9046. There are mainly two reasons. Firstly, for a smaller K (e.g., $K = 25, 50$), the instance selector is unable to choose sufficient positive instances to reflect the ratio difference in both Micro and Macro, considering the ratio of positive to negative instances among the selected critical instances is an important factor. Secondly, for a larger K (e.g., $K = 200$), too many false-positive pseudo labels may be generated by mistake in the first step, resulting in the degradation of the performance, since the number of positive instances is usually dozens in ITC and Micro.

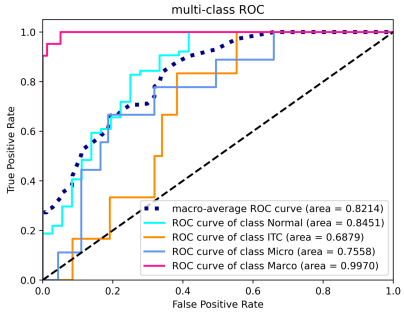
Therefore, for various LNM classification tasks, the selection of K is largely dependent on its characteristics. As illustrated in Fig. 4, when simply considering the binary LNM classification task, spatial information is not necessary to distinguish whether the metastasis is Micro or Macro. Therefore, selecting a smaller K is preferable to prevent the introduction of an excessive number of false positive instances. However, it is worth noting that when K is too small, the quantity of information carried by the selected instances is insufficient, leading to a decrease in performance. To perform better on the LNM three-classification task, the K should be larger to distinguish Micro vs. Marco, because larger K is conducive to identify the different spatial distribution of patches. When executing the most refined LNM four-classification task, K should be moderate to trade-off the performance of ITC vs. Normal and Micro vs. Marco.

2) Effects of Transformer Layer Number: Different numbers of Transformer layers will lead to various performance of the instance aggregator. As show in Table VI, a too deep Transformer not only leads to huge computation resource consumption, but also causes overfitting. While the feature aggregation ability will be limited if the Transformer has too few layers. It can be seen that selecting a three-layer Transformer shows the best comprehensive performance over CAMELYON17. Therefore, we keep this parameter setting in the external validation and in-house cohorts.

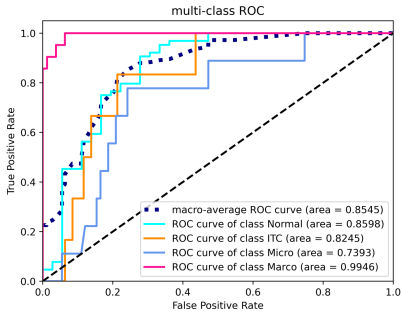
3) Effects of Spatial and Rank Embeddings: As shown in Fig. 5, by comparing 'wo/E_r', 'wo/E_s' and 'wo/E_s&E_r', it can be seen that both rank embedding and spatial embedding improve the performance. Moreover, the best performance can be achieved when fusing spatial embedding and rank embedding. Among them, rank embedding is more important than spatial embedding in LNM classification. Because the ITC category only contains very few positive instances, which is similar to the Normal category and particularly difficult to classify. Rank embedding can encode the suspected positive degree of the selected instance, which can help the Transformer focus on the ITC category more efficiently. The role of spatial embedding is to further facilitate the distinction between Micro- and Macro-Metastasis, since the spatial distribution is more dispersed for Macro-Metastasis while more centralized for Micro-Metastasis.

H. Interpretability Analysis

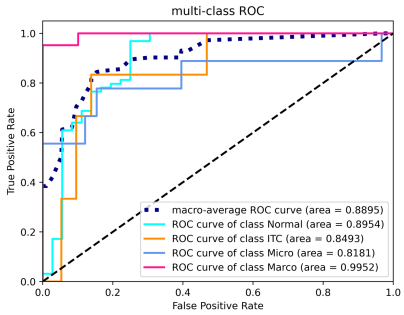
Fig. 6 depicts a Macro-metastasis WSI. The red box shows that our dMIL strategy selects the true-positive instances more accurately than other instance selectors. The blue box shows that other strategies incorrectly select many false-positive instances



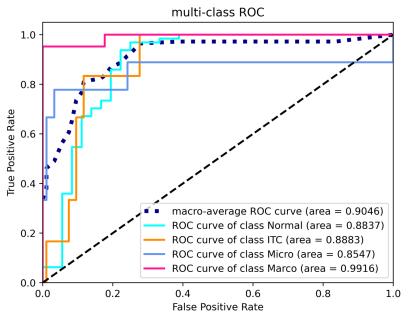
(a) w/o Es&Er



(b) w/o Er



(c) w/o Es



(d) dMIL-Transformer

Fig. 5. Comparison of spatial embedding and rank embedding on CAMELYON17.

as critical instances, while our dMIL strategy seldom selects the false-positive instances. Obviously, Fig. 6 shows that our dMIL selection strategy is more beneficial in selecting the critical instance and has good interpretability. In addition, our dMIL-Transformer can help the pathologists figure out the region of tumor during operation and facilitate the clinical determination of performing a lymph dissection, thereby reducing the risk of a secondary surgery.

V. Discussion

WSI is a type of high-resolution images digitized from tissue slides. By utilizing computer vision and deep learning algorithms for WSI analysis in LNM classification tasks, more accurate cancer diagnoses and prognoses can be achieved. In clinical practice, accurate staging of lymph node metastasis helps physicians better understand the patient's prognosis and formulate more appropriate treatment plans, ultimately improving patient survival rates and quality of life.

In our study, we mainly found that for WSI analysis, it is essential to not only focus on the histomorphological features of the images but also to consider the spatial distribution of tumors within the images. Consequently, we encoded spatial coordinates of critical instances as additional inputs to the Transformer model, enabling the network to learn spatial features of specific categories, thereby enhancing classification capabilities. Moreover, in LNM classification tasks, ITC and Micro-metastasis categories only contain a small number of tumor regions. If all tissue regions are considered simultaneously, the model may receive a large amount of redundant information during the input stage, which could adversely affect the classification performance of the model. Therefore, we found that selecting critical instances can effectively alleviate this problem. Our proposed dMIL strategy selects top-K critical instances to retain crucial information for WSI classification.

However, our study still has limitations. The number of critical instances is determined experimentally, and different numbers are required for LNM classification problems with varying category numbers. As shown in Fig. 7, we present a case of unsuccessful classification, where the dMIL-Transformer incorrectly classifies an ITC as a Normal category, which may impact subsequent treatment plans. Although our dMIL strategy can identify tumor regions, it also selects numerous non-tumor areas. We speculate the cause of failure may be due to the pre-determined number of candidate critical instances based on ablation studies, without the ability to dynamically adjust the number of critical instances for each WSI. Consequently, when analyzing WSIs with very small tumor areas, it may still erroneously select an excess of negative regions. We suggest that future work could employ specific algorithms to automatically determine the required number of critical instances for each WSI.

VI. Conclusion

This article proposes a subtle two-stage dMIL-Transformer framework for weakly-supervised classification of lymph node metastasis (LNM). In the first stage, to further take advantage of the true-negative instances in positive bags and the false-positive instances in negative bags, a double Max-Min MIL (dMIL) strategy is proposed. Interpretability analyses prove that the proposed dMIL strategy can obtain a better decision boundary and reduce misclassifications compared with traditional MIL methods. In the second stage, a Transformer-based MIL aggregator is designed to further improve the performance of bag classification by integrating the morphological, spatial, and malignancy rank information among different instances. In the second stage,

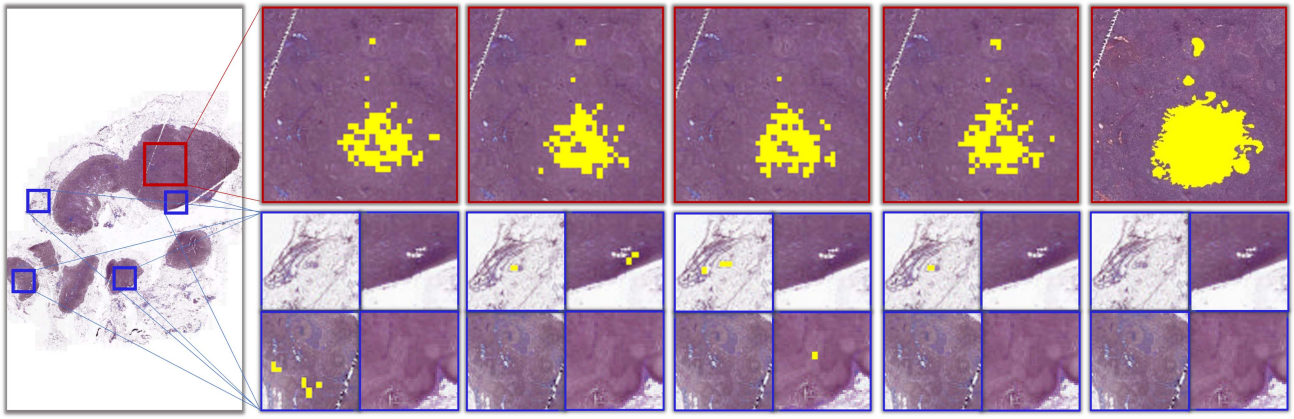


Fig. 6. Interpretability and visualization of critical instance selection. Red box indicates the region containing positive instances, the blue box indicates the negative region, and the yellow region indicates ground truth and the positive instances predicted by the network. Compared with the previous instance selection strategies, our proposed dMIL strategy can select the critical instances more accurately.

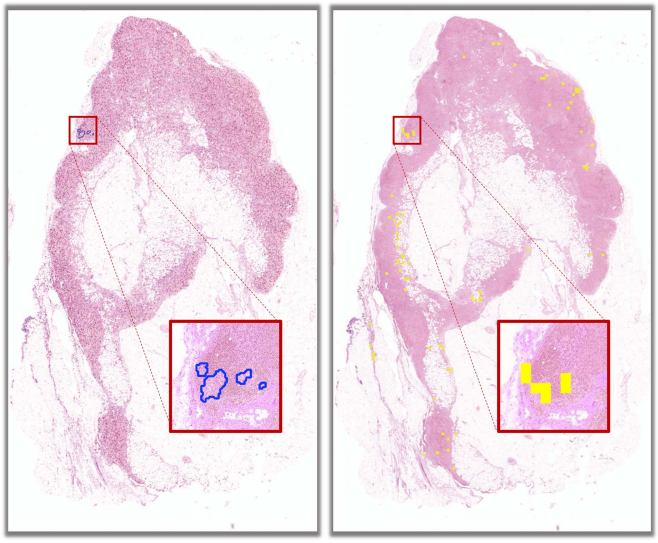


Fig. 7. Fail cases of dMIL-Transformer on LNM classification. (a) represents a slide of the ITC category, with the red box highlighting a magnified image of the tumor region. (b) Yellow block represents the spatial distribution of critical instances selected by our dMIL strategy on the WSI.

the Transformer-based MIL aggregator is proposed to further improve the performance of bag classification by integrating the morphological, spatial, and rank information among different instances. Ablation studies demonstrate the effectiveness of the methods proposed in the two stages, and their combination can obtain the best performance. Besides, we conduct extensive experiments over three public LNM classification datasets in breast cancer, and our method achieves the state-of-the-art performance in general. Experimental results validate that traditional MIL algorithms often misclassify ITC and Micro categories as Negative and Macro, while dMIL-Transformer with spatial and rank information can effectively distinguish these hard-to-predict categories.

Although K is tailored according to the characteristic of the LNM classification task in dMIL-Transformer and achieves a satisfactory performance, the selection of K can be automated in the future to increase the generability of the framework. Besides, all experiments are conducted with $20\times$ magnification, which makes the training time of the dMIL strategy relatively longer. In the future, we hope to directly extract the morphological features of the image by self-supervised methods, which may improve training efficiency.

REFERENCES

[1] A. J. Evans et al., "US food and drug administration approval of whole slide imaging for primary diagnosis: A key milestone is reached and new questions are raised," *Arch. Pathol. Lab. Med.*, vol. 142, pp. 1383–1387, 2018.
[2] G. Litjens et al., "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.
[3] M. Veta et al., "Predicting breast tumor proliferation from whole-slide images: The TUPAC16 challenge," *Med. Image Anal.*, vol. 54, pp. 111–121, 2019.
[4] C.L. Srinidhi, O. Ciga, and A. L. Martel, "Deep neural network models for computational histopathology: A survey," *Med. Image Anal.*, vol. 67, 2021, Art. no. 101813.
[5] M. B. Amin et al., "The eighth edition AJCC cancer staging manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging: The Eighth edition AJCC cancer staging manual," *CA: A Cancer J. Clinicians*, vol. 67, pp. 93–99, 2017.
[6] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.
[7] B. Lee and K. Paeng, "A robust and effective approach towards accurate metastasis detection and pn-stage classification in breast cancer," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2018, pp. 841–850.
[8] Y. Xu, J.-Y. Zhu, E. I.-C. Chang, M. Lai, and Z. Tu, "Weakly supervised histopathology cancer image segmentation and classification," *Med. Image Anal.*, vol. 18, pp. 591–604, 2014.
[9] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *Nat. Sci. Rev.*, vol. 5, pp. 44–53, 2018.
[10] X. Wang et al., "Predicting gastric cancer outcome from resected lymph node histopathology images using deep learning," *Nat. Commun.*, vol. 12, 2021, Art. no. 1637.
[11] X. Chu, B. Zhang, Z. Tian, X. Wei, and H. Xia, "Do we really need explicit position encodings for vision transformers?," 2021, *arXiv:2102.10882*.

[12] B. Heo, S. Yun, D. Han, S. Chun, J. Choe, and S. J. Oh, "Rethinking spatial dimensions of vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vi.*, 2021, pp. 11936–11945.
[13] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," presented at the International Conference on Learning Representations, Oct. 2020.
[14] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–41, 2022.
[15] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 1998, pp. 570–576.
[16] K. Das, S. Conjeti, J. Chatterjee, and D. Sheet, "Detection of breast cancer from whole slide histopathological images using deep multiple instance CNN," *IEEE Access*, vol. 8, pp. 213502–213511, 2020.
[17] F. Aeffner et al., "Introduction to digital image analysis in whole-slide imaging: A white paper from the digital pathology association," *J. Pathol. Inform.*, vol. 10, no. 1, 2019, Art. no. 9.
[18] Y. Zhao et al., "Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4836–4845.
[19] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artif. Intell.*, vol. 201, pp. 81–105, 2013.
[20] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognit.*, vol. 77, pp. 329–353, 2018.
[21] P. Chikontwe, M. Kim, S. J. Nam, H. Go, and S. H. Park, "Multiple instance learning with center embeddings for histopathology classification," in *Medical Image Computing and Computer Assisted Intervention MICCAI 2020*, vol. 12265, A. L. Martel, P. D. Abolmaesumi, D. Stoyanov Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, eds. Cham, Switzerland: Springer International Publishing, 2020, pp. 519–528.
[22] G. Campanella et al., "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Med.*, vol. 25, pp. 1301–1309, 2019.
[23] G. Xu et al., "CAMEL: A weakly supervised learning framework for histopathology image segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 10682–10691.
[24] M. Lerousseau et al., "Weakly supervised multiple instance learning histopathological tumor segmentation," in *Medical Image Computing and Computer Assisted Intervention MICCAI 2020*, (Lecture Notes in Computer Science), A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, eds. Cham, Switzerland: Springer International Publishing, 2020, pp. 470–479.
[25] M. Lerousseau et al., "Weakly supervised pan-cancer segmentation tool," in *Medical Image Computing and Computer Assisted Intervention MICCAI 2021*, M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, eds. Cham, Switzerland: Springer International Publishing, 2021, pp. 248–256.
[26] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie, "Deep multi-instance networks with sparse label assignment for whole mammogram classification," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2017, pp. 603–611.
[27] J. Feng and Z.-H. Zhou, "Deep MIML network," in *Proc. 31th AAAI Conf. Artif. Intell.*, San Francisco, California, USA: AAAI Press, Feb. 2017, pp. 1884–1890.
[28] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2127–2136.
[29] X. Shi, F. Xing, Y. Xie, Z. Zhang, L. Cui, and L. Yang, "Loss-based attention for deep multiple instance learning," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, pp. 5742–5749.
[30] W. Li, V.-D. Nguyen, H. M. Liao Wilder, K. Cheng, and J. Luo, "Patch transformer for multi-tagging whole slide histopathology images," in *Medical Image Computing and Computer Assisted Intervention*, vol. 11764, D. T. Shen Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, eds. Cham, Switzerland: Springer International Publishing, 2019, pp. 532–540.
[31] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li, "Multi-instance learning by treating instances as non-i.i.d. samples," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 1–8.
[32] B. Li, Y. Li, and K. W. Eliceiri, "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14318–14328.
[33] Z. Shao et al., "TransMIL: Transformer based correlated multiple instance learning for whole slide image classification," in *Proc. 35th Conf. Neural Inf. Process. Syst.*, 2021, pp. 2136–2147.
[34] H. Zhang et al., "DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18802–18812.
[35] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[36] B. Yun, Y. Wang, J. Chen, H. Wang, W. Shen, and Q. Li, "SpecTr: Spectral transformer for hyperspectral pathology image segmentation," 2021, *arXiv:2103.03604*.
[37] Z. Gao et al., "Instance-based vision transformer for subtyping of papillary renal cell carcinoma in histopathological image," in *Proc. 24th Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2021, pp. 299–308.
[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[39] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)" 2020, *arXiv:1606.08415*.
[40] G. Litjens et al., "1399 h&e-stained sentinel lymph node sections of breast cancer patients: The CAMELYON dataset," *GigaScience*, vol. 7, Jun. 2018, Art. no. giy065.
[41] K. Clark et al., "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," *J. Digit. Imag.*, vol. 26, no. 6, pp. 1045–1057, 2013.
[42] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process.g Syst.*, 2019, vol. 32.
[43] L. Wright, "Ranger - a synergistic optimizer," 2019. [Online]. Available: https://github.com/lessw2020/Ranger-Deep-Learning-Optimizer
[44] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomed. Eng.*, vol. 5, no. 6, pp. 555–570, Mar. 2021.
[45] X. Liu et al., "Development of prognostic biomarkers by TMB-Guided WSI analysis: A two-step approach," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 4, pp. 1780–1789, Apr. 2023.