Computer Methods and Programs in Biomedicine 229 (2023) 107268

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Biomedical signal and image processing methods

Interpretable classification of pathology whole-slide images using attention based context-aware graph convolutional neural network

Meiyan Liang[a,*], Qinghui Chen[a], Bo Li[b], Lin Wang[c,d], Ying Wang[a], Yu Zhang[a], Ru Wang[a], Xing Jiang[a], Cunlin Zhang[e]

[a] School of Physics and Electronic Engineering, Shanxi University, Taiyuan 030006, China
[b] Department of Rehabilitation Treatment, Shanxi Rongjun Hospital, Taiyuan 030000, China
[c] Shanxi Bethune Hospital, Shanxi Academy of Medical Sciences, Tongji Shanxi Hospital, Third Hospital of Shanxi Medical University, Taiyuan, 030032, China
[d] Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, 430030, China
[e] Beijing Key Laboratory for Terahertz Spectroscopy and Imaging, Key Laboratory of Terahertz, Optoelectronics, Ministry of Education, Capital Normal University, Beijing 100048, China

ARTICLE INFO

Article history:
Received 19 October 2022
Revised 23 November 2022
Accepted 23 November 2022

Keywords:
Computational pathology
Context-aware information
Graph convolution network
Interpretability
Whole slide image

ABSTRACT

Background and Objective: Whole slide image (WSI) classification and lesion localization within giga-pixel slide are challenging tasks in computational pathology that requires context-aware representations of histological features to adequately infer nidus. The existing weakly supervised learning methods mainly treat different locations in the slide as independent regions and cannot learn potential nonlinear interactions between instances based on i.i.d assumption, resulting in the model unable to effectively utilize context-ware information to predict the labels of WSIs and locate the region of interest (ROI).
Methods: Here, we propose an interpretable classification model named bidirectional Attention-based Multiple Instance Learning Graph Convolutional Network (ABMIL-GCN), which hierarchically aggregates context-aware features of instances into a global representation in a topology fashion to predict the slide labels and localize the region of lymph node metastasis in WSIs.
Results: We verified the superiority of this method on the Camelyon16 dataset, and the results show that the average predicted ACC and AUC of the proposed model after flooding optimization can reach 90.89% and 0.9149, respectively. The average accuracy and ACC score are improved by more than 7% and 4% compared with the existing state-of-the-art algorithms.
Conclusions: The results demonstrate that context-aware GCN outperforms existing weakly supervised learning methods by introducing spatial correlations between the neighbor image patches, which also addresses the 'accuracy-interpretability trade-off' problem. The framework provides a novel paradigm for the clinical application of computer-aided diagnosis and intelligent systems.

1. Introduction

Histopathological whole-slide image (WSI) classification and lesion localization within giga-pixel slide are challenging tasks in computational pathology that requires context-aware representations of histological features to adequately infer nidus. To date, manual inspection of histopathological slides remains the gold standard for severe disease diagnosis, especially for assessing malignancy progression, lesion staging, and further treatment of cancer-related regions [1]. However, it suffers from long inspection cycle and serious reliance on subjective interpretation, which cannot satisfy the demand of precision medicine. Therefore, it is an urgent need to develop artificial intelligent systems for automatic analysis of whole slide images (WSI).

In the field of computational pathology, deep learning has presented a great potential for real-time, objective, and reproducible clinical-grade predictions of giga-pixel WSIs [2–6]. However, deep learning based approaches require either manual annotation of giga-pixel WSIs using supervised leaning or only slide-level labels using weakly supervised learning. Supervised learning algorithms use ground truth correspondences to construct a deep model, which have achieved state-of-the-art (SOTA) performance across large number of computer vision tasks including image segmentation and lesion boundaries localization [7–10]. However, supervised learning algorithms should provide pixel-wise labeling of whole slide image. It is labor-intensive and time-consuming, which limits its application in realistic clinical scenarios.

* Corresponding author.
E-mail address: meiyanliang@sxu.edu.cn (M. Liang).

https://doi.org/10.1016/j.cmpb.2022.107268
0169-2607/

To address this issue, existing supervised learning methods mainly adopt multiple instance learning (MIL), which eliminate the laborious and time-consuming labeling process by assigning one label to the whole bag [11–14]. However, the MIL baseline model could only obtain the spatial positions of top-ranked instances [15–17], which is prone to miss most of the positive instances in the WSI. This is because these frameworks are designed under the instance independent and identical distribution (i.i.d.) hypothesis, which ignores the correlations between instances in a bag. To better utilize the context-aware information of the WSI, researchers consider introducing instance-level constraints [18–21], multi-scale features, feature representations and Transformer architectures into MIL-based models to capture local and global correlations between the instances of WSI, thereby improving the overall performance of the model.

Specifically, many researchers attempt to introduce instance-level constraints in MIL to improve the performance of the weakly supervised model. Nevertheless, in the iteration process, only top-k instances of WSIs are selected for feature aggregation before the output layer [22–24]. Therefore, it requires massive amounts of WSIs to build the model since only a fraction of the instances within each slide could actually participate in the training process. Moreover, multi-scale feature assisted MIL learning is proposed as an intuitive method to simulate the workflow of pathologists interpreting WSIs, which obtains context-aware features of the slide by extracting the patch features at multiple scales (magnifications) [25–27]. Likewise, multi-scale feature based methods also cause computational burden due to the large volume of data and are not always applicable in certain cases. Even some studies get conflicting conclusions when multi-scale features are introduced. For instance, Chen et al. [13] have demonstrated that multiple instance learning with multi-scale features does not always outperform single-scale approaches. The effective combination of multiple features is also a worth considering issue when multi-scale feature is applied in MIL learning [28–30]. Feature representation embedding is another perspective for WSI predictions, which is prone to maintain the original spatial locations of patches and avoid information redundancy [31–33]. In these studies, image patches of the WSI can be highly compressed and represented as feature vectors by contrastive learning or GAN-based methods. Afterwards, the feature vectors are spatially arranged to form a data cube according to the location of the source patches in the slide, which is then fed into the CNN-based network as a whole for interpretable classification using the attention mechanism. However, this approach cannot be utilized in realistic clinical scenarios because the large compression ratio or sparse feature representation of gigapixel WSI will lead to limited generalization ability for the model at test time or failure to provide useful slide-level interpretability.

Currently, the MIL-based approaches have achieved exciting improvements in WSI prediction and regions of interest (ROI) localization. However, almost all MIL-related methods are based on the i.i.d. assumption and are not fully applicable for the real WSI classification task. This is due to the fact that these methods cannot fully simulate the context-aware structure of image patches with specific spatial location in WSI by capturing the intrinsic relationship between different magnifications.

In the realistic scenarios, pathologists should consider both the feature representation of the region and the nonlinear interactions between the regions when making final diagnostic predictive determinations. Therefore, it would be much desirable to introduce the dependencies between the instances in the bag for multiple instance learning algorithm. Recently, some researchers attempt to use Transformer architecture to build the relationship between image patches to overcome the problem of irrelevance between instances caused by the i.i.d. assumption in MIL-based methods [34–36]. However, these methods are mainly based on the self-attention mechanism to establish the dependencies between image patches, which could cause a large burden for the computational device in WSI prediction [37–39]. To tackle this issue, a subset of image patches is sampled randomly from the WSI to form a sparse representation of the slide, and then fed into the Transformer architecture to simulate the image patches located in WSI. However, it will fall into the same dilemma as the instance-level constrained approach does [40].

Inspired by pathologists interpreting pathological slides using microscopes at different magnifications, we propose a context-aware graph convolution framework named bidirectional Attention-based Multiple Instance Learning Graph Convolutional Network (ABMIL-GCN), which not only considers the feature representation of each patch in the slide, but also provides potential non-linear interactions between the neighbor patches. This framework builds context-aware information of slides via patch feature embedding to form a WSI-graph representation for final diagnosis and prediction. It breaks the limitations of conventional i.i.d assumption about instances in a bag, which could greatly improve the performance of attention-based MIL pooling aggregation by bidirectional message passaging mechanism between the neighbor patches. By embedding context-aware information for each patch, it not only improves the overall prediction accuracy of WSI images, but also avoids information redundancy when applying multi-scale feature-based MIL learning. Furthermore, the non-linear interactions of the neighbor instances and spatial feature embedding constraints can effectively reduce false negative and false positive instances in slides, which provides a basis for locating ROIs with high accuracy.

The main scientific contributions are summarized as follows:

- Bidirectional ABMIL-GCN framework is proposed to simulate local- and global- topology structure of the pathology patches in whole slide images, which fully preserves the original spatial relations for the slides so that the model can capture structural correlation between the neighbor patches.
- Under the i.i.d. assumption, misclassification of individual patch may alter the prediction of slide labels, resulting in a large number of false negatives and false positives. Here, bidirectional ABMIL-GCN framework embeds the spatial location of neighboring instances for each patch, which can reduce the false negative and false positive rate of instances through inter-patch message passing mechanism.
- Flooding regularization is applied in bidirectional ABMIL-GCN framework to prevent further optimization of the training loss when it reaches a reasonably small value (called flood level). Therefore, the approach forces the training loss float around the flood level by setting a lower bound on it.
- The interpretability heat map of WSI can also be obtained by bidirectional ABMIL-GCN, which utilize the gradient of highly correlated node groups in the graph to obtain more accurate attention weights of the instances. Therefore, ROI in H&E-stained slides has good consistency with the pixel-wise annotated ground truth.

2. Methods

2.1. Slide-level graph construction

Given a training WSI dataset $D = \{X_i, Y_i\}_{i=1}^{N}$, where $X_i$and Y_i are the ith input WSI and its corresponding class label, $Y_i \in \{0, 1\}$ for binary classification, N is the number of WSI in the training process. To construct a graph G for each image in the training dataset,

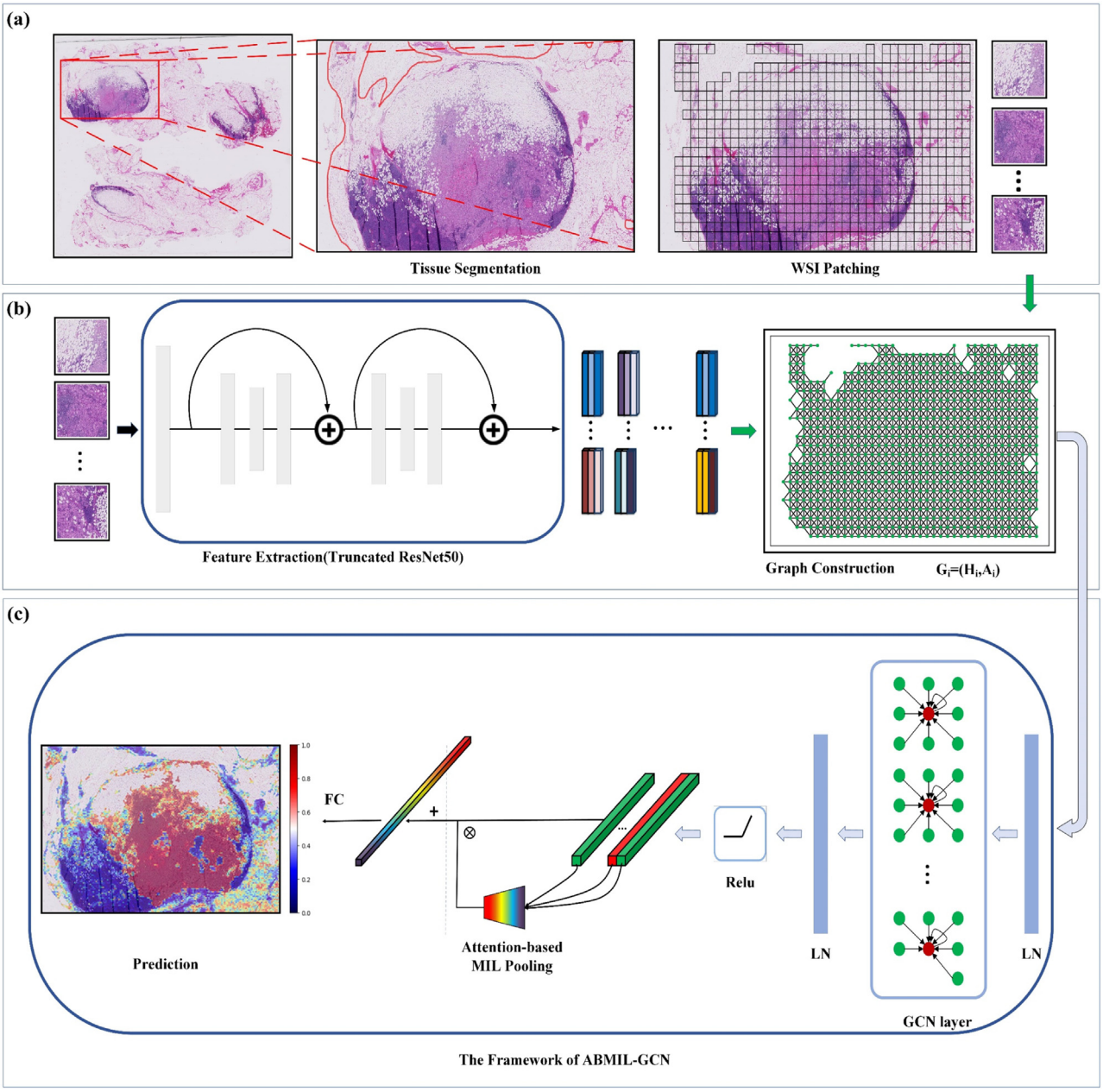

Fig. 1. The overview of the data reprocessing, ABMIL-GCN framework and interpretability. (a) The preprocessing of a whole slide image contains tissue foreground segmentation (left) and patching (right). (b) Image patches are encoded into a descriptive feature representations (left) by a truncated Resnet50, which can be viewed as nodes. These nodes are embedded in the graph to form a context-aware slide-level graph representation based on its spatial coordinates (right). Here, the slide-level graph representation can be expressed as $G_i = (H_i, A_i)$. (c) During the training and inference stages, the constructed graph passed through a graph convolutional layer to obtain the context-aware topology information of the WSI. An attention-based MIL pooling network is applied to aggregate patch-level information into slide-level representations, which are used for final diagnostic prediction. Specifically, the attention-based MIL pooling network ranks each region in the slide and assigns an attention score based on its relative importance to the slide-level label. Meanwhile, the attention scores can be visualized as a heat map to identify ROIs and interpret the morphology feature used for diagnosis.

the H&E-stained WSI is first down-sampled and converted into the HSV color space. Then, automatic foreground segmentation is performed for the WSI using Otsu's binarization on the saturation channel to separate H&E-stained tissue from the background. Thus, the foreground of the X_i can be obtained by magnification conversion and image registration. Afterwards, the foreground of WSI is segmented into non-overlapping image patches of size 256×256. Meanwhile the coordinates of the segmented patches are also preserved in this process. Fig. 1(a) shows the entire preprocessing process of the input WSI.

In weakly supervised learning, given a input whole slide image X_i and a corresponding label Y_i. Each slide X_i contains multiple instances, which can be represent as $X_i = \{x_{i,j}\}_{j=1}^{P}$. Where $x_{i,j}$ denote the jth patch (instance) of the slide X_i. Here, each instance $x_{i,j}$ of slide X_i implies a binary label $y_{i,j}$, which is not given exactly. P is the number of instances in the X_i, which varies widely for each WSI. According to multi-instance learning assumption, the relationship between instance label $y_{i,j}$ and slide label Y_i is as follows:

$$Y_i = \begin{cases} 0, if \sum_{j=1}^{P} y_{i,j} = 0, \\ 1, otherwise. \end{cases} \quad (1)$$

The instance $x_{i,j}$ can be feature extracted by a truncated ResNet50 and denoted as $h_{i,j} \in \mathbb{R}^{1\times1024}$ in the latent space. Therefore, the corresponding feature representation of the WSI can be expressed as $H_i = \{h_{i,j}\}_{j=1}^{P}$ mathematically. Here, $H_i \in \mathbb{R}^{P\times1024}$.

To utilize the contextual-aware information of the WSI in the embedding space, bidirectional ABMIL-GCN framework is proposed to improve the prediction performance of WSI. The idea of this method is to construct a graphical representation for a WSI. Follow this route, the contextual-aware information can be efficiently utilized to make predictions through the bidirectional information transfer mechanism between neighbor patches in the graph. Specifically, the foreground of the WSI is segmented into non-overlapping image patches called nodes, and $h_{i,j}$ is regarded as the feature representation for the node $x_{i,j}$. Each node is connected with the neighbor nodes called edges. The current image patch is associated with the surrounding image patches through the edges according to the preserved coordinates. Here, we models an 3 × 3 image receptive field of the WSI image by building an 8 nearest neighbor graph neural network for each patch. The slide-level graph representation can be expressed as $G_i = (H_i, A_i)$, which is shown in Fig. 1(b). where A_i is the adjacency matrix of graph representation $G_i$for the input slide X_i. where A_i is expressed as:

$$A_i^{j,j'} = \begin{cases} 1, if\ x_{i,j}\ is\ adjacent\ to\ x_{i,j'}, \\ 0, \quad otherwise. \end{cases} \tag{2}$$

where $j, j' \in \{1, 2, ..., P\}$.

2.2. The framework of ABMIL-GCN

As shown in Fig. 1(c), the ABMIL-GCN network mainly consists of a graph convolutional layer, layer normalization (*LN*) operation, a gated attention-based MIL pooling layer and a fully connected layer. Here, the graph convolutional layer [41] is the core of the framework because it treats different locations in the slide as context-aware regions and learns potential nonlinear interactions between instances. Graph convolutional layer encodes the graph structure to obtain the slide-level morphological feature representation of WSI. This means that we learn a GCN mapping function for a WSI-Graph G_i: $H_i \in \mathbb{R}^{P\times d_{in}} \rightarrow H_i^{out} \in \mathbb{R}^{P\times d_{out}}$, which iteratively aggregates node features in spatial neighborhoods and predicts the WSI through bidirectional information transfer between the nodes. In the framework, layer normalization [42] is also added at the output of a graph convolutional layer for each given WSI to accelerate model convergence, prevent gradient vanishing and improve generalization ability. In our case, the context-aware WSI feature representation H_i^{out} after GCN layer can be expressed as follows:

$$H_i^{out} = \left\{h_{i,j}^{out}\right\}_{j=1}^{P} = RELU\left(LN\left(\tilde{D}_i^{-1/2}\tilde{A}_i\tilde{D}_i^{1/2}LN(H_i)W_i\right)\right) \tag{3}$$

Where $H_i^{out} \in \mathbb{R}^{P\times512}$ represent the output graph representation of H_i, $h_{i,j}^{out} \in \mathbb{R}^{1\times512}$ is the jth element of H_i^{out}. Here, $\tilde{A}_i = A_i + I(\tilde{A}_i \in \mathbb{R}^{P\times P})$ denote the new adjacency matrix of the bidirectional GCN by adding an identity matrix I. $\tilde{D}_i^{j,j} = \sum_{j'} \tilde{A}_i^{j,j'}(\tilde{D}_i \in \mathbb{R}^{P\times P})$ is the degree matrix of the adjacency matrix $\tilde{A}_i$. $W_i \in \mathbb{R}^{1024\times512}$ denote the connection weights learned by the graph convolutional layer.

For whole slide image prediction and lesion localization without pixel-wise annotation, ABMIL-GCN uses a gated attention-based MIL pooling [43] to aggregate patch-level features into slide-level representations, as it provides the model with the flexibility of selectively aggregating information from multiple relevant node families to predict the slide-level labels.

Therefore, the slide-level feature representation for graph G_i can be given by:

$$z_i = \sum_{j=1}^{P} a_{i,j} h_{i,j}^{out} (z_i \in \mathbb{R}^{1\times512}) \tag{4}$$

where

$$a_{i,j} = \frac{exp\left\{w^T(tanh(Vh_{i,j}^{out\,T}) \odot sigm(Uh_{i,j}^{out\,T}))\right\}}{\sum_{j=1}^{P} exp\left\{w^T(tanh(Vh_{i,j}^{out\,T}) \odot sigm(Uh_{i,j}^{out\,T}))\right\}} \tag{5}$$

$a_{i,j} \in \mathbb{R}$ denotes the attention score of the node feature $h_{i,j}^{out}$ in WSI, and it can provide interpretable feature for WSI predictions via heat map. $w \in \mathbb{R}^{256\times1}$, $U \in \mathbb{R}^{256\times512}$ and $V \in \mathbb{R}^{256\times512}$ are learned parameter matrices. $\odot$ and $sigm(\cdot)$ represent element-wise multiplication operation and sigmoid function, respectively. The superscript T denotes the transpose operation. In the framework, gated attention-based MIL pooling introduces nonlinearities for the weakly supervised learning, which may potentially eliminate troublesome linearity generated by $tanh(\cdot)$.

After gated attention-based MIL pooling operation, the obtained slide-level feature representation z_i passes through a fully connected layer to make a prediction $\hat{Y}_i$ for the slideX_i. In the experiment, cross entropy loss is applied in implementation to minimize the K-L divergence between predicted label $\{\hat{Y}_i\}_{i=1}^{N}$and ground truth $\{Y_i\}_{i=1}^{N}$.

3. Experiments

3.1. Dataset and implement details

The experimental dataset is Camelyon16, which includes a total of 399 WSIs of sentinel lymph node from two independent data sets collected in Radboud University Medical Center (Nijmegen, The Netherlands), and the University Medical Center Utrecht (Utrecht, The Netherlands) . The dataset contains pixel-wise annotations for lymph node metastases in hematoxylin and eosin stained (H&E) whole slide images (WSIs), which is one of the largest annotated, public digital pathology datasets available. The total number of official training WSIs is 270, including 159 normal tissue slides and 111 lymph node metastases slides. The 270 slides can be split into training set and validation set according the ratio of 8.5:1.5. Afterwards, the performance of the model was tested using 129 slides, including 80 normal tissue slides and 49 lymph node metastasis slides. In the implementation, each WSI was cropped into approximately 44, 274 256 × 256 image patches at 40 × magnification, with some WSI having graph size as large as 142, 949 instances. Adam optimizer is applied in the implementation with the initialized learning rate of 1×10^{-4} and a weight decay of 1×10^{-6}. The cross-entropy loss is used in the iteration to minimize the divergence between the distribution of predicted class probability and the ground truth. Meanwhile, early stopping is also applied to avoid overfitting of the weakly supervised model. Our model was trained on a INVIDA RTX 3090 (24GB) for 100 epochs with a batch size of 1.

In the implementation, performance of early stopping highly depends on the iterative dynamics and is extremely sensitive to the randomness in the optimization process. This means that early stopping at the optimal epoch in a single training path does not necessarily perform well in another round of training. In [44], Kiryo et al. observed that overfitting can be occurred in weakly supervised learning when the empirical risk goes below zero. Therefore, a gradient ascent technique is proposed to maintain the empirical risk non-negative to prevent overfitting, which can be generalized and applied to weakly supervised settings. In our case,

flooding [45] is introduced to bidirectional ABMIL-GCN to stabilize performance of the model and further prevent overfitting. The idea of *flooding* is to add the *flood level* as a regularization term to the loss function, which can effectively prevent overfitting by setting a lower bound on the training loss and forcing the loss to maintain greater than or equal to *flood level* during training. This technique can also improve the accuracy of the test set and reduce the MSE of the classification risk in a certain condition. The effectiveness of the flooding is proved as follows:

For the any given input slide X and corresponding label Y, it is assumed that bidirectional ABMIL-GCN can fully simulate the mapping function between X and Y. Therefore, the classification risk can be defined as

$$R(g) := \mathrm{E}_{p(X,Y)}[l_{ce}(g(X), Y)] \tag{6}$$

Where $g(\cdot)$ is the score function, which transforms the input data X into predictions probability $\hat{Y}$, l_{ce} represents binary cross entropy loss, $p(X, Y)$ denotes an unknown joint probability distribution density function for each data point. $\mathrm{E}_{p(X,Y)}[\cdot]$ indicates the expectation of $(X, Y) \sim p(X, Y)$.

The goal of binary classification is to learn the function $g(\cdot)$ that minimizes the classification risk $R(g)$. However, it is difficult to evaluate $R(g)$ exactly as ground truth distribution of $p(X, Y)$ is unknown. Therefore, we minimize its empirical risk by calculating the average cross entropy loss of the training data instead:

$$\hat{R}(g) := \frac{1}{N}\sum_{i=1}^{N} l_{ce}(g(X_i), Y_i) \tag{7}$$

Here, N is the number of samples in the training set.

The empirical risk after using the *flooding* optimization is defined as:

$$\tilde{R}(g) = \left|\hat{R}(g) - b\right| + b \tag{8}$$

Where $\tilde{R}(g)$ denote the flooded empirical risk. $b$is so-called the *flood level*. When $\hat{R}(g) \geq b$, $\tilde{R}(g) = \hat{R}(g) \geq b$; Whereas $\hat{R}(g) < b$, $\tilde{R}(g) \geq b$; therefore, $\tilde{R}(g) \geq b$ is always satisfied in both cases. Here, b could also be regarded as the lower bound of the loss function when *flooding* regularization is applied.

Meanwhile, it can be inferred that the MSE of flooded empirical risk is not higher than that of the empirical risk without flooding.

$$MSE(\hat{R}(g) \geq MSE(\tilde{R}(g)) \tag{9}$$

Furthermore, if the flood level b further satisfying the condition that its value is between the original training loss and the test loss, the constraint on the MSE of the empirical risk will be more stringent [45]. Therefore, inequality (9) can be transformed into:

$$MSE(\hat{R}(g) > MSE(\tilde{R}(g)) \tag{10}$$

However, optimal flood level is unknown in advance. To determine the optimal value of b, an exhaustive hyper-parameter search was performed for the flood level with candidates selected from the interval of 0.00 to 0.14 with a fixed step size of 0.02.

Fig. 2 is the variation curve of training accuracy and AUC with the flood level based on the validation accuracy. The result in Fig. 2 shows that the training error of the models can be maintained at a relatively low value when different *flood level* is applied, indicating that *flooding* is applicable for GCN-based models. The marker placed on the flood level curve is the optimal value of b, which is selected as a regularization term in the implementation. In this case, b=0.10 is chosen by performing the exhaustive search in parallel models and between the original training loss and the test loss.

The relationship between the test loss and gradient amplitude of the training/test loss is visualized in Fig. 3. The markers '+' in Fig. 3 indicate the test loss of the proposed model without *flooding*, while the markers 'o' indicate the test loss of the model with *flooding* in the iteration. For each case, the color of plot becomes darker (yellow → green) as the training epoch proceeds. As shown in Fig. 3(a), the statistical gradient amplitude of the training loss for ABMIL-GCN model with flooding is significantly larger than that of the model without flooding, which indicates that flooding regularization can prevent the model from staying at local minima, thereby prone to achieve the optimal solution in the training process. Meanwhile, the test loss for ABMIL-GCN model with flooding can be preserved at a relatively low value compared to the model without flooding as the training proceeds, indicating that the flooding regularization can effectively avoid overfitting. Fig. 3(b) is the relationship between the test loss and gradient amplitude of the test loss in the case of with and without flooding regularization. The results showed that the gradient amplitude and the loss value of the ABMIL-GCN model with flooding are both smaller on the test set compared to the model without flooding. Meanwhile, the fluctuation of the model with flooding are relatively small than those of the model without flooding in most iterations for the test set. Moreover, the test loss of the model without flooding ascends rather than descents as the training proceeds. This further demonstrated that the model with flooding is more stable and generalizes better for external data.

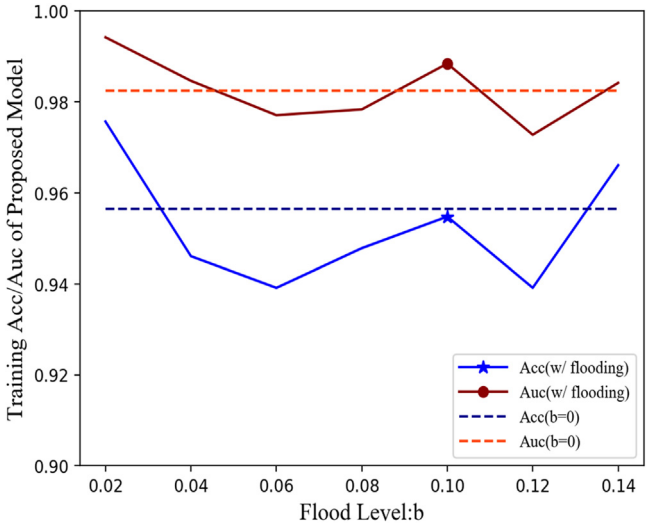

Fig. 2. The solid lines represent the variation curve of training accuracy (ACC) and AUC with the flood level b, respectively. The marker placed on the flood level curve is the optimal value of b. The horizontal dashed lines denote the ACC and AUC score without the flood level.

3.2. Results and discussion

Table 1 shows the results of bidirectional ABMIL-GCN model with and without flooding regularization on the Camelyon16 dataset. Here, we use the indicators such as accuracy (ACC) and area under curve (AUC) to evaluate the performance of the models. As can be seen in Table 1, the results showed that ABMIL-GCN model with flooding often improves test accuracy over the baseline without flooding. The average prediction ACC and AUC of the proposed model with flooding optimization can reach 90.89% and 0.9149, which improved by 2.67% and 2.18% comparing with the model without flooding regularization, respectively. Meanwhile, the standard deviations of ACC and AUC were both lower than 0.0033 and 0.0041 in 5 trials for bidirectional ABMIL-GCN with

Table 1
Comparison results of the baseline models on camelyon16 dataset.

Method	Accuracy	AUC	Precision	Sensitivity	Specificity	F-1 Score
Mean-pooling	0.5984±0.0312	0.5387±0.0032	0.4990±0.1102	0.2612±0.1113	0.8050±0.1169	0.3149±0.0846
Max-pooling	0.6419±0.0292	0.7011±0.0131	0.7642±0.1670	0.1973±0.1710	0.9375±0.0568	0.2640±0.1966
ABMIL[43]	0.8233±0.0114	0.8581±0.0134	0.9400±0.0119	0.5714±0.0342	**0.9775±0.0050**	0.7101±0.0255
CLAM[17]	0.8295±0.0085	0.8729±0.0124	0.9199±0.0540	0.6082±0.0271	0.9650±0.0278	0.7304±0.0076
ABMIL-GCN(w/o FL)	0.8822±0.0090	0.8931±0.0099	0.9050±0.0238	0.7714±0.0153	0.9500±0.0137	0.8326±0.0123
ABMIL-GCN(w/ FL)	**0.9089±0.0033**	**0.9149±0.0041**	**0.9463±0.0093**	**0.8061±0.0102**	0.9719±0.0054	**0.8705±0.0051**

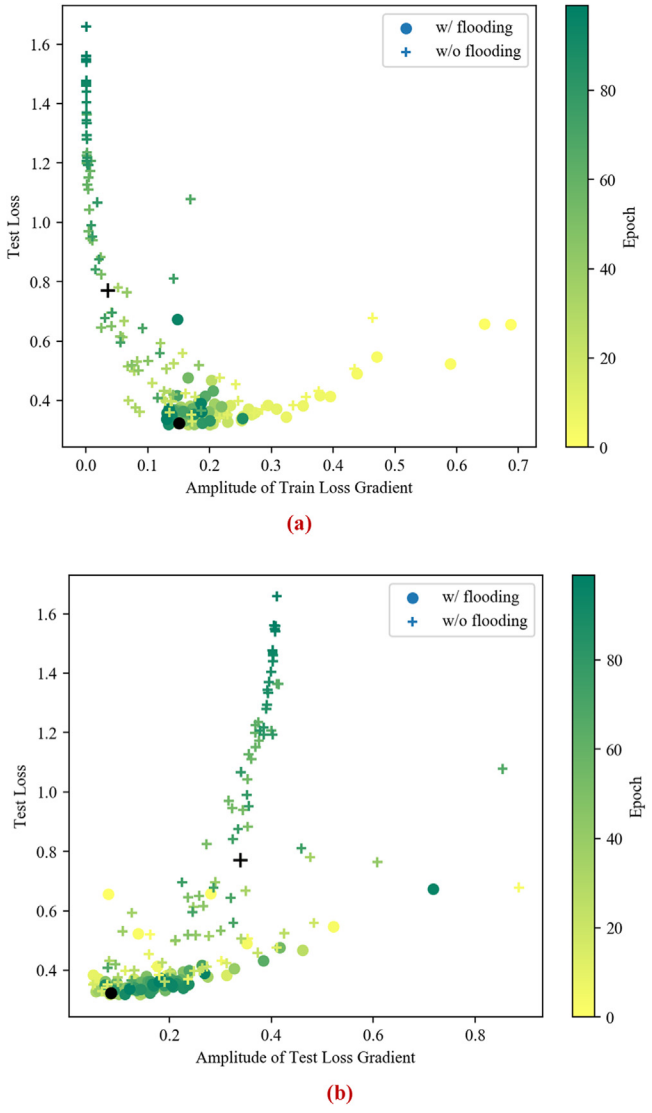

Fig. 3. (a) The relationship between the test loss and gradient amplitude of the training loss. (b) The relationship between the test loss and gradient amplitude of the test loss. The different shaped markers ('o' or '+') in the figures indicate the model with and without flooding regularization. The same shaped markers denote the training/testing epoch of a single weakly supervised model in the iteration. The color becomes darker (yellow → green) as the training iteration proceeds. The large black 'o' and '+' indicate the convergence point of the corresponding model.

flooding regularization. This means that the standard deviation of both ACC and AUC value for the ABMIL-GCN model with flooding is reduced to approximately 1/3 of the baseline, further demonstrating the stability of the model when flooding is applied.

Table 1 also compared the performance of ABMIL-GCN with other benchmark weakly supervised models, such as multiple instance learning (MIL) based Mean-pooling, Max-pooling, ABMIL [43] and CLAM [17] on the same data set. The results showed that the maximum accuracy of multiple instance learning with traditional pooling operators such as Mean-pooling and Max-pooling is not better than 65%. This is mainly because these methods do not consider the contribution of each image patch for the entire WSI when processing image patches simultaneously in a batch. Therefore, when attention-based MIL pooling (ABMIL) is applied, it can improve the performance of MIL model by weighted averaging of instances. While for CLAM, instance-level clustering constraints are combined with attention-based pooling operation to classify WSIs efficiently and accurately. Specifically, instance-level clustering constraints is applied to refine the feature space, and attention-based pooling is used to identify sub-regions of high diagnostic value based on the learned features of the WSI. Therefore, ABMIL and CLAM are more effective than traditional MIL based pooling operators for WSI prediction, such as Mean-pooling, Max-pooling. However, in real scenarios, image patches of each WSI have a fixed arrangement in space, which means that adjacent image patches tend to be spatially correlated. The frameworks based on the I.I.D assumption are not fully applicable for WSI prediction because they ignore the spatial information between image patches and treat each image patch as an individual. Therefore, the performance of ABMIL-GCN can be improved by establishing a bidirectional information transfer mechanism for the neighbor image patches in the WSI, which is remarkably better than that of the ABMIL and CLAM do.

To gain deep insight into the ABMIL-GCN model, the obtained latent feature representations of the slides in the test set are visualized by T-SNE (Fig. 4). In Fig. 4(a), the red dots and blue dots represent the slides with and without lymph node metastases, respectively. The result showed that each category is clustered together and the different categories have sharp boundaries in the latent space, which demonstrates the effectiveness of the model. Meanwhile, latent feature representations of these slides obtained by CLAM are visualized for comparison, as shown in Fig. 4(b). The result showed that the red dots and blue dots are clustered into three cluster centers in the latent space, and the two classes are mixed at the boundary. This is because CNN-based CLAM treats image patches at different positions in the slide as independent regions. Whereas the bidirectional ABMIL-GCN make predictions by capturing the contextual-aware dependencies of patches in each WSI. Furthermore, the false negative rate of the proposed model is 18.37%, which reduced by 20% compared with CLAM. While the false positive rate of the proposed model is 2.5%, which is on par with CLAM. This is due to the fact that each positive slide contains approximately less than 10% of the cancer area on average for Camelyon16 dataset. While only a portion of image patches are involved in training when CLAM is applied, resulting in the presence of a large number of negative areas affecting the prediction of positive slide. However, for bidirectional ABMIL-GCN, all the node in the constructed graph are participate the training process through the nonlinear interactions between them. Therefore, bidirectional ABMIL-GCN can greatly reduce the false negative rate and false positive rate, which are of great significance for clinical applications.

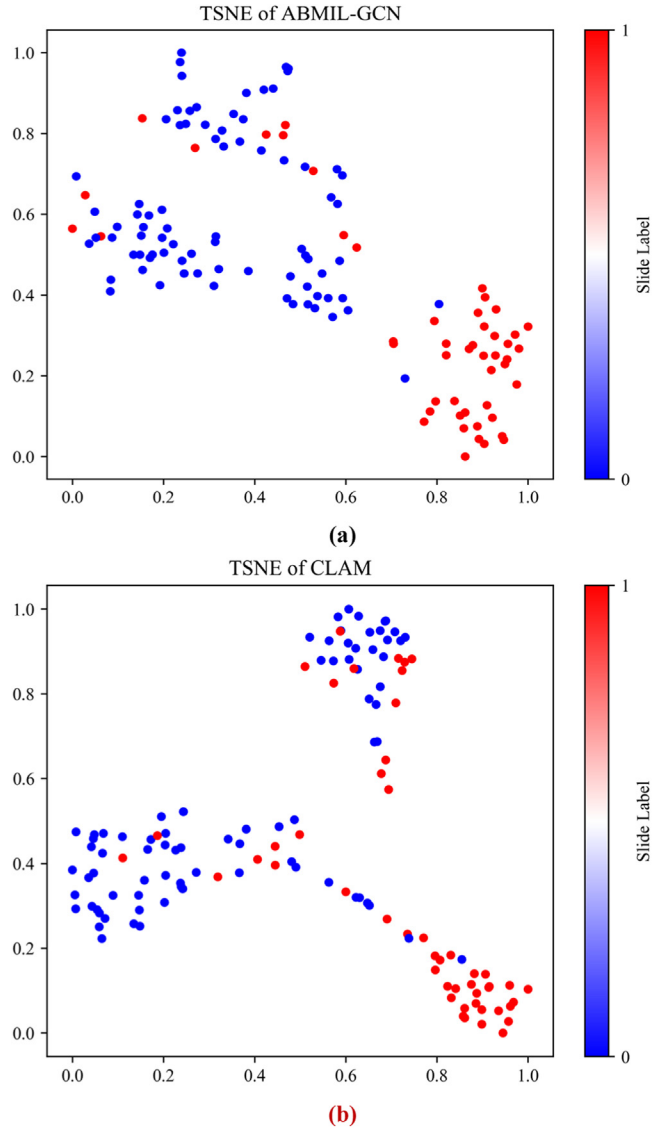

Fig. 4. (a) T-SNE feature embedding of WSI using ABMIL-GCN (with flooding).(b) T-SNE feature embedding of WSI using CLAM[17].

Fig. 5 shows the receiver operating characteristic curve (ROC) of the our bidirectional ABMIL-GCN (with flooding) model and other 5 baseline models including ABMIL-GCN (without flooding), Mean-pooling, Max-pooling, ABMIL and CLAM. In Fig. 5, we use the average performance of ROC curve over 5 trials as the final result to increase the confidence of each model. Here, all the models are under the same configuration environment. It can be observed that the area under the curve (AUC) for bidirectional ABMIL-GCN model with flooding can reach 0.92, which illustrates that the proposed model has outperformed other baseline models across almost all the possible classification thresholds, indicating the superior performance of the proposed model in whole slide image prediction.

Bidirectional ABMIL-GCN model makes prediction by considering potential nonlinear interactions between instances and aggregating the instances (nodes) features into slide-level based on the gated attention mechanism. Here, the input of the network is a graph representation embedded with all node features, which enables all instances in the WSI to participate in the training process. Therefore, pixel-wise attention heatmap can be visualized

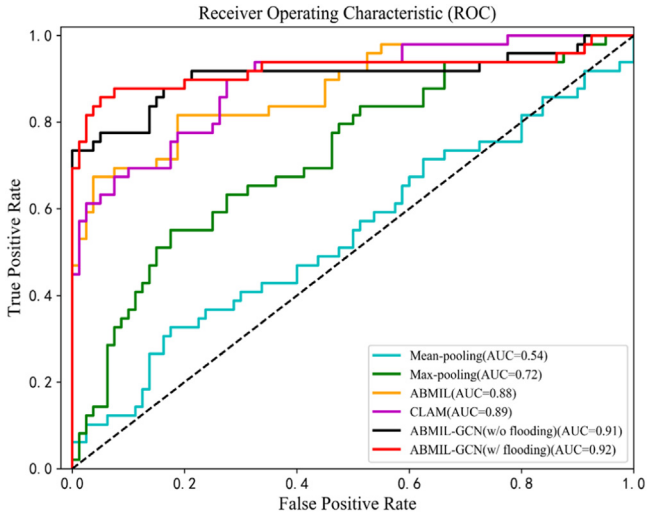

Fig. 5. Receiver Operating Characteristic (ROC).

and interpreted by converting the obtained attention scores of the instance into percentiles and mapping the normalized scores to their corresponding spatial location in the original slide. Fig. 6 showed the interpretability heatmaps generated by ABMIL-GCN model, state-of-the-art method and the pixel-wise annotations for the corresponding WSI. In Fig. 6, (a_1), (a_2), (b_1) and (b_2) are the interpretable heatmaps of WSI and zoomed in view of ROI, which are obtained by bidirectional ABMIL-GCN and CLAM, respectively. Fig. 6 (c_1) and (c_2) are the corresponding ground truth of the WSI image. In Fig. 6 (c_1) and (c_2), the area within the blue curve is the pixel-wise annotation of lymph node metastases, which is provided by computational pathology group in Medical Center of the Radboud University. As shown in Fig. 6 (a_1)-(c_2), it can be concluded that bidirectional ABMIL-GCN model is capable of providing with more true positive instances than CLAM. The heatmap obtained using bidirectional ABMIL-GCN could accurately delineate the boundary between tumor and normal, which have good consistency with human pathology expertise. This is mainly because graph networks could form effective context-aware relationships by establishing connections and interaction mechanisms between the current node and its neighbors. The architecture of graph feature representation embedding is able to maximize the optimization performance of gated attention-based MIL pooling through a bidirectional information transfer mechanism between adjacent instances, thereby avoiding the appearance of false negative and false positive instances. Fig. 6 (d_1)-(f_2) is a positive slide which contains only a fraction of positive instances. It is predicted as positive and the ROI can be localized by heatmap in Fig. 6 (d_1) and (d_1). It can be seen that the obtained ROI has good consistency with the ground truth. However, it is predicted to be negative when CLAM is applied. Therefore, bidirectional ABMIL-GCN can effectively reduce the false positive rate of WSI prediction by introducing context-ware information of the slide, which is an essential indicator for application in realistic scenarios.

However, the proposed ABMIL-GCN also have limitations in representing contextual information of whole slides. For instance, the degree of correlation between a node and surrounding nodes in the graph is not always the same in practice. While the ABMIL-GCN model regards the surrounding patches of each patch as equal contributions, which affects the performance of the model. Therefore, the next step is to extend our work to add weights of surrounding nodes for each node in the graph to further improve the performance of the model.

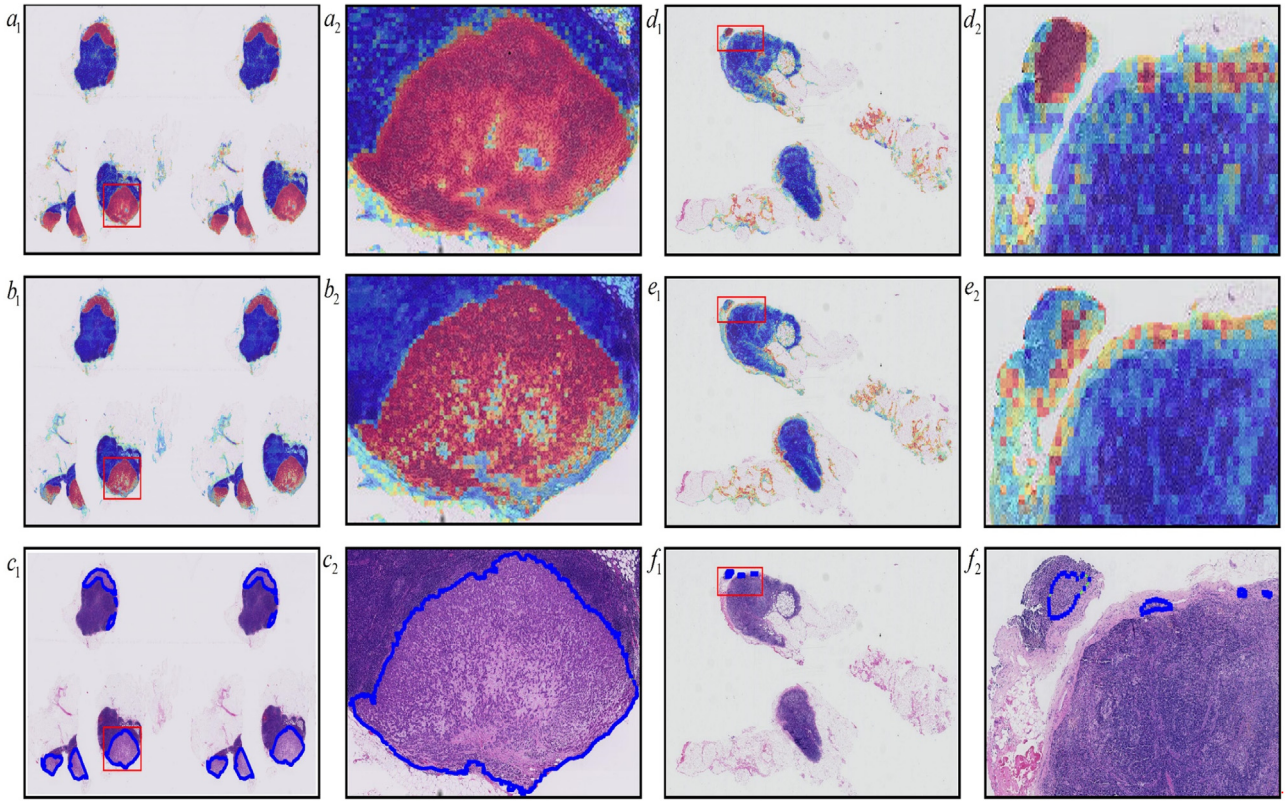

Fig. 6. Interpretable heatmaps of the WSIs. (a_1), (a_2), (d_1) and (d_2) are the obtained interpretable heatmaps of WSI and corresponding zoom-in view of them using bidirectional ABMIL-GCN (with flooding regularization). (b_1), (b_2), (e_1) and (e_2) are the obtained interpretable heatmaps of WSI and corresponding zoom-in view of them using CLAM. (c_1) and (f_1) are pixel-wise annotation of lymph node metastases in hematoxylin and eosin stained (H&E) whole slide images. (c_2) and (f_2) are zoomed in view of (c_1) and (f_1), respectively.

4. Conclusions

Real-time, objective, and accurate WSI prediction and ROI localization are of great significance for the diagnosis and treatment of critical illness. The existing weakly supervised learning methods are mainly based on the assumption of i.i.d, which regard the patches at different positions in the WSI as independent regions, resulting in the model unable to effectively utilize context-aware information to predict WSI tags and locate ROI. Therefore, bidirectional ABMIL-GCN framework is proposed to simulate context-aware topology structure of the pathological slide through the combination of graphical feature representation embedding and gated attention-based pooling. The results indicate that bidirectional ABMIL-GCN could not only achieve higher prediction accuracy with flooding regularization, but also provide human-interpretable features with localization heatmap. The average prediction ACC and AUC of the proposed model after flooding optimization can reach 90.89% and 0.9149 on Camelyon16 dataset, respectively. The corresponding standard deviation of ABMIL-GCN model is lower than 0.0033 and 0.0041, which outperform the state-of-the-art algorithms. Particularly, bidirectional ABMIL-GCN can greatly reduce the false negative rate of WSI prediction, which is of great significance for clinical diagnosis. The superior performance of this framework provides a new paradigm for high-precision prediction and interpretable ROI localization of whole slide images in computational pathology.

Declaration of Competing Interest

The authors declared that they have no conflicts of interest in this work.

Acknowledgments

This work is supported in part by National Natural Science Foundation of China under Grant 11804209, Natural Science Foundation of Shanxi Province under Grant 201901D211173, Scientific and Technological Innovation Programs of Higher Education Institutions in Shanxi under Grant 2019 L0064, and Natural Science Foundation of Shanxi Province under Grant 202103021223411.

References

[1] N. Marini, S. Marchesin, S. Otálora, et al., Unleashing the potential of digital pathology data by training computer-aided diagnosis models without human annotations, NPJ Digit. Med. 5 (1) (2022) 1–18.
[2] C. Sun, B. Li, G. Wei, et al., Deep learning with whole slide images can improve the prognostic risk stratification with stage III colorectal cancer, Comput. Methods Progr. Biomed. (2022) 106914.
[3] J. Lou, J. Xu, Y. Zhang, et al., PPsNet: an improved deep learning model for microsatellite instability high prediction in colorectal cancer from whole slide images, Comput. Methods Progr. Biomed. 225 (2022) 107095.
[4] C. Zhou, Y. Jin, Y. Chen, et al., Histopathology classification and localization of colorectal cancer using global labels by weakly supervised deep learning, Comput. Med. Imaging Graph. 88 (2021) 101861.
[5] H. Qu, M. Zhou, Z. Yan, et al., Genetic mutation and biological pathway prediction based on whole slide images in breast carcinoma using deep learning, NPJ Precis. Oncol. 5 (1) (2021) 1–11.
[6] G. Bueno, M.M. Fernandez-Carrobles, L. Gonzalez-Lopez, et al., Glomerulosclerosis identification in whole slide images using semantic segmentation, Comput. Methods Progr. Biomed. 184 (2020) 105273.
[7] M. Van Rijthoven, M. Balkenhol, K. Siliņa, et al., HookNet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images, Med. Image Anal. 68 (2021) 101890.
[8] Pedersen A., Smistad E., Rise T.V., et al. Hybrid guiding: a multi-resolution refinement approach for semantic segmentation of gigapixel histopathological images. arXiv:2112.03455 [eess.IV] (2021).
[9] Feng Y., Hafiane A., Laurent H. A deep learning based multiscale approach to segment cancer area in liver whole slide image. arXiv:2007.12935 [eess.IV] (2020).

[10] R. Feng, X. Liu, J. Chen, et al., A deep learning approach for colonoscopy pathology WSI analysis: accurate segmentation and classification, IEEE J. Biomed. Health Inform. 25 (10) (2020) 3700–3708.
[11] M.Y. Lu, T.Y. Chen, D.F.K. Williamson, et al., AI-based pathology predicts origins for cancers of unknown primary, Nature 594 (7861) (2021) 106–110.
[12] J. Yao, X. Zhu, J. Jonnagaddala, et al., Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks, Med. Image Anal. 65 (2020) 101789.
[13] H. Chen, X. Han, X. Fan, et al., Rectified cross-entropy and upper transition loss for weakly supervised whole slide image classifier, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, 2019, pp. 351–359.
[14] Tu M., Huang J., He X., et al. Multiple instance learning with graph neural networks. arXiv:1906.04881 [cs.LG] (2019).
[15] G. Campanella, M.G. Hanna, L. Geneslaw, et al., Clinical-grade computational pathology using weakly supervised deep learning on whole slide images, Nat. Med. 25 (8) (2019) 1301–1309.
[16] Y. Zhao, F. Yang, Y. Fang, et al., Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4837–4846.
[17] M.Y. Lu, D.F.K. Williamson, T.Y. Chen, et al., Data-efficient and weakly supervised computational pathology on whole-slide images, Nat. Biomed. Eng. 5 (6) (2021) 555–570.
[18] S. Wang, Y. Zhu, L. Yu, et al., RMDL: recalibrated multi-instance deep learning for whole slide gastric image classification, Med. Image Anal. 58 (2019) 101549.
[19] G. Xu, Z. Song, Z. Sun, et al., Camel: a weakly supervised learning framework for histopathology image segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 10682–10691.
[20] M. Lerousseau, M. Vakalopoulou, M. Classe, et al., Weakly supervised multiple instance learning histopathological tumor segmentation, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, 2020, pp. 470–479.
[21] P. Chikontwe, M. Kim, S.J. Nam, et al., Multiple instance learning with center embeddings for histopathology classification, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, 2020, pp. 519–528.
[22] S. Cheng, S. Liu, J. Yu, et al., Robust whole slide image analysis for cervical cancer screening using deep learning, Nat. Commun. 12 (1) (2021) 1–10.
[23] F. Kanavati, G. Toyokawa, S. Momosaki, et al., Weakly-supervised learning for lung carcinoma classification using deep learning, Sci. Rep. 10 (1) (2020) 1–11.
[24] J.D. Ianni, R.E. Soans, S. Sankarapandian, et al., Tailored for real-world: a whole slide image classification system validated on uncurated multi-site data emulating the prospective pathology workload, Sci. Rep. 10 (1) (2020) 1–12.
[25] N. Hashimoto, D. Fukushima, R. Koga, et al., Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 3852–3861.
[26] B. Li, Y. Li, K.W. Eliceiri, Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 14318–14328.
[27] T. Xiang, Y. Song, C. Zhang, et al., DSNet: a dual-stream framework for weakly-supervised gigapixel pathology image analysis, IEEE Trans. Med. Imaging (2022).
[28] N. Marini, S. Otálora, F. Ciompi, et al., Multi-scale task multiple instance learning for the classification of digital pathology images with global annotations, in: Proceedings of the MICCAI Workshop on Computational Pathology. PMLR, 2021, pp. 170–181.
[29] Thandiackal K., Chen B., Pati P., et al. Differentiable zooming for multiple instance learning on whole-slide images. arXiv:2204.12454 [cs.CV] (2022).
[30] Hou W., Yu L., Lin C., et al. H2-MIL: exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2022.
[31] D. Tellez, G. Litjens, J. van der Laak, et al., Neural image compression for gigapixel histopathology image analysis, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2) (2019) 567–578.
[32] M. Shaban, R. Awan, M.M. Fraz, et al., Context-aware convolutional neural network for grading of colorectal cancer histology images, IEEE Trans. Med. Imaging 39 (7) (2020) 2395–2405.
[33] M. Lerousseau, M. Vakalopoulou, E. Deutsch, et al., SparseConvMIL: sparse convolutional context-aware multiple instance learning for whole slide image classification, in: Proceedings of the MICCAI Workshop on Computational Pathology. PMLR, 2021, pp. 129–139.
[34] Z. Shao, H. Bian, Y. Chen, et al., Transmil: transformer based correlated multiple instance learning for whole slide image classification, Proceedings of Advances in Neural Information Processing Systems 34 (2021) 2136–2147.
[35] A. Myronenko, Z. Xu, D. Yang, et al., Accounting for dependencies in deep learning based multiple instance learning for whole slide imaging, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, 2021, pp. 329–338.
[36] Zheng Y., Gindra R.H., Green E.J., et al. A graph-transformer for whole slide image classification. arXiv:2205.09671 [cs.CV] (2022).
[37] R.J. Chen, M.Y. Lu, W.H. Weng, et al., Multimodal co-attention transformer for survival prediction in gigapixel whole slide images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 4015–4025.
[38] H. Li, F. Yang, Y. Zhao, et al., DT-MIL: deformable transformer for multi-instance learning on histopathological image, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, 2021, pp. 206–216.
[39] Y. Zhao, Z. Lin, K. Sun, et al., SETMIL: spatial encoding transformer-based multiple instance learning for pathological image analysis, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, 2022, pp. 66–76.
[40] W. Li, V.D. Nguyen, H. Liao, et al., Patch transformer for multi-tagging whole slide histopathology images, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, 2019, pp. 532–540.
[41] Kipf T.N., Welling M. Semi-supervised classification with graph convolutional networks. arXiv:1609.02907 [cs.LG] (2016).
[42] Ba J.L., Kiros J.R., Hinton G.E. Layer normalization. arXiv:1607.06450 [stat.ML] (2016).
[43] M. Ilse, J. Tomczak, M. Welling, Attention-based deep multiple instance learning, in: Proceedings of the International conference on machine learning. PMLR, 2018, pp. 2127–2136.
[44] R. Kiryo, G. Niu, M.C. Du Plessis, et al., Positive-unlabeled learning with non-negative risk estimator, Proceedings of Advances in neural information processing systems 30 (2017).
[45] Ishida T., Yamane I., Sakai T., et al. Do we need zero training loss after achieving zero training error?. arXiv:2002.08709 [cs.LG] (2020).