Complex & Intelligent Systems (2025) 11:218
https://doi.org/10.1007/s40747-025-01779-y

ORIGINAL ARTICLE

Transformer-based multiple instance learning network with 2D positional encoding for histopathology image classification

Bin Yang[1,2] · Lei Ding[2] · Jianqiang Li[2] · Yong Li[2] · Guangzhi Qu[3] · Jingyi Wang[2] · Qiang Wang[2] · Bo Liu[4]

Received: 31 October 2023 / Accepted: 5 January 2025

Abstract

Digital medical imaging, particularly pathology images, is essential for cancer diagnosis but faces challenges in direct model training due to its super-resolution nature. Although weakly supervised learning has reduced the need for manual annotations, many multiple instance learning (MIL) methods struggle to effectively capture crucial spatial relationships in histopathological images. Existing methods incorporating positional information often overlook nuanced spatial correlations or use positional encoding strategies that do not fully capture the unique spatial dynamics of pathology images. To address this issue, we propose a new framework named TMIL (Transformer-based Multiple Instance Learning Network with 2D positional encoding), which leverages multiple instance learning for weakly supervised classification of histopathological images. TMIL incorporates a 2D positional encoding module, based on the Transformer, to model positional information and explore correlations between instances. Furthermore, TMIL divides histopathological images into pseudo-bags and trains patch-level feature vectors with deep metric learning to enhance classification performance. Finally, the proposed approach is evaluated on a public colorectal adenoma dataset. The experimental results show that TMIL outperforms existing MIL methods, achieving an AUC of 97.28% and an ACC of 95.19%. These findings suggest that TMIL's integration of deep metric learning and positional encoding offers a promising approach for improving the efficiency and accuracy of pathology image analysis in cancer diagnosis.

Keywords Weakly supervised training · Image classification · Multiple instance learning

Introduction

Whole Slide Image (WSI) is a digital image obtained by scanning the tissue on a slide, which has been successfully used in medical diagnosis, education, and research [1–3]. With the advancements in WSI technology and deep learning, medical image analysis techniques have achieved further success [4–7]. Due to gigapixel resolution of WSI, traditional deep learning models cannot be trained directly, so the common approach is to divide the whole WSI into numerous patches and train patch-wise classification with patch-level labels. In clinical data, a WSI with gigapixel resolution may be divided into thousands of patches, and the labelling of such a large number of patches would consume a lot of time for pathologists. To overcome this difficulty, weakly supervised WSI classification techniques [5, 8–12] use only WSI-level labels

✉ Bo Liu
b.liu@massey.ac.nz

Bin Yang
yangbin@emails.bjut.edu.cn

Lei Ding
dinglei@emails.bjut.edu.cn

Jianqiang Li
lijianqiang@bjut.edu.cn

Yong Li
li.yong@bjut.edu.cn

Guangzhi Qu
gqu@oakland.edu

Jingyi Wang
wangjingyi@emails.bjut.edu.cn

Qiang Wang
wangqiang1997@emails.bjut.edu.cn

[1] Center for Strategic Assessment and Consulting, Academy of Military Science, Beijing, China

[2] Faculty of Information Technology, Beijing University of Technology, Beijing, China

[3] Computer Science and Engineering Department, Oakland University, Rochester, USA

[4] School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand

Published online: 21 March 2025

to train patch-level classifiers, which are mostly based on the Multiple Instance Learning (MIL) framework. Specifically, the model treats each image as a bag and splits the bag into numerous patches, each of which is considered an instance. If one instance in the bag is positive (e.g., lesion), the bag is marked as positive, otherwise it is negative.

In the field of computer vision, multiple instance learning has made great progress. However, the weakly supervised WSI classification based on multi-instance learning still suffers from numerous challenges. On one hand, the amount of WSI data as supervisory information is too small, which would lead to severe overfitting and fall into local optimal solutions during the training process. On the other hand, the relative positions of entities within an image provide essential prior information that is often underutilized. Traditional MIL models predominantly concentrate on the relationships between instances via attention mechanisms but tend to overlook or inadequately leverage the spatial relationships critical to interpreting histopathological images. Although some studies have attempted to integrate relative positional information through either fixed or learnable positional embeddings, these approaches frequently fail to capture the intricate spatial correlations among instances. Moreover, they may employ positional encoding strategies that inadequately represent the unique spatial dynamics characteristic of pathology images.

In this paper, a Transformer-based Multiple Instance Learning Network with 2D positional encoding framework (TMIL) is proposed to address these challenges. TMIL randomly splits the WSI (bag) into numerous pseudo-bags to build more training samples, and then introduces deep metric learning to provide richer supervised information. Specifically, TMIL extracts the instance with the highest probability value from the pseudo-bag. A new dataset is created to train both an instance-level classification model and a deep metric learning model, using pseudo-bag labels. The concatenated feature vector from both models forms the final instance representation. In addition, TMIL has designed a 2D positional encoding module (PEM) based on the Transformer to model the positional information between instances. The PEM module takes the row and column vectors of each instance as input and outputs a positional encoding vector, which is finally added to the representation that has passed through the self-attention module.

In summary, the main contributions of this paper are:

- This paper proposes a model named TMIL for the task of classification of pathological images. TMIL uses deep metric learning to provide richer supervised information for the training of the model to mitigate the overfitting phenomenon.
- After getting the feature representation of each patch, TMIL designs a 2D positional encoding module based on Transformer to model the location information between the patches. TMIL replaces the Transformer's one-dimensional location data with two-dimensional patch information using row and column vectors. These vectors are then modeled with a self-attention mechanism, enabling the network to focus on positional correlations between patches.
- The TMIL model was evaluated on a colorectal adenoma dataset. The results show that the TMIL model outperforms other MIL models on the classification task with an AUC of 97.28% and an ACC of 95.19%. Ablation experiments show that deep metric learning and two-dimensional positional encoding structures can significantly improve results.

Related work

Weakly-supervised WSI classification

With the development of deep learning, many histopathological image recognition tasks [13, 14] have achieved remarkable success. Rakhlin [13] used multiple deep neural networks and gradient boosting trees for pathological diagnosis of breast images. Chen [14] segmented the glands of colon images in a multi-task learning framework. However, these approaches process only the regions of interest (ROI) in histopathological images, which are selected by pathologists. These ROIs are smaller than entire pathology images and are scaled simply for model input. There are only a small number of ROI selected by pathologists, so it is convenient to implement patch-level or even pixel-level labelling work. Therefore, most of the approaches can be categorized into fully supervised learning methods.

In recent years, there has been a dramatic increase in interest in WSI-level pathology analysis, which is more relevant than ROI-level analysis. Due to gigapixel resolution of WSI, traditional deep learning models cannot be trained directly. A simple compression of WSI will lose a lot of potential information and details. Therefore, the ROI level pathology analysis is not applicable to WSI samples. In fact, how to feed the gigapixel resolution WSI into the model is a challenging topic. The current approaches of WSI-level pathology analysis [15] require pathologists to label the patch-level images, which leads to expensive costs. In the absence of patch-level labels, a weakly supervised model [11, 16, 17] can be designed to classify WSI using image-level labels. Xiang [11] integrated information at both high magnification (local) and low magnification (regional) levels to conduct WSI analysis, and then designed a Dual-Stream Network to predict WSI labels. Wang [16] used patch-based full convolutional networks to generate feature representations and explored

different context-aware block selection and feature aggregation strategies. Tellez et al. [17] used a neural network trained in an unsupervised fashion to compress gigapixel images and trained a convolutional neural network to predict image-level labels.

Multiple instance learning

Multiple instance learning has been widely used in weakly supervised pathology image diagnosis [18] and other fields [19, 20]. Multiple instance learning formulation point that each WSI is considered as a bag that contains many patches considered as instances. A bag is labelled as positive if any of its instances is positive.

In general, multiple instance learning can be divided into two categories.

The first category is the instance-based MIL model [21, 22], which assigns a pseudo-label to each instance according to certain rules during the training phase, and the instance-level data are constructed to train the instance-level encoder; while in the inference phase, the model outputs the top k instances with the highest scores, which are used to aggregate the final results. FS-GCN-MIL [21] consists of three parts, including a feature extractor based on a self-supervised learning mechanism for extracting feature representations of instances, a new instance-level feature selection method, and a bag-level classifier based on graph convolutional networks.

The second category is embedding-based MIL models [23–26], where the core point is to aggregate the instance representation into bag representation by an aggregator (e.g., max-pooling), which eventually is used for bag level classification. DeepAttnMISL [23] was designed to learn instance features from WSI and aggregate WSI-level information by introducing attention-based MIL pooling, experiments show that attention-based aggregation is more flexible and adaptive than other aggregation techniques. Most approaches focus on instance selection and aggregation, thus ignoring instance relations. A new framework [24] is proposed to jointly learn instance-level and bag-level embeddings, and use central loss to reduce intra-class variation. A multi-instance learning model [26] is proposed that can be easily plugged into a Vision transformer and effectively improves the model performance for downstream image classification tasks. In addition, some existing [27–29] works have explored incorporating relative position information through fixed or learnable positional embeddings. Bontempo et al. [27] introduce a Graph Neural Network that precedes the MIL framework, designed to enhance the representation of WSI structure through capturing the mutual spatial correlations of instances across multiple scales, both within and between scales. Zhao et al. [28] employ spatial-encoding-transformer layers within their aggregation module, which enhances instance representation by simultaneously incorporating information from both neighboring and globally correlated instances, with a novel joint absolute-relative position encoding scheme to augment context information encoding capabilities.

Attention and self-attention in deep learning

The attention mechanism is used to enhance useful signals, which originated from machine translation tasks in natural language processing and has now been widely used in various tasks in computer vision, including target detection, image classification, and segmentation [30]. In recent years, attention-based mechanisms for medical image analysis have also received attention [31–36].

The combination of attentional mechanisms and multi-instance learning has made an important contribution to the diagnosis of super resolution pathological images. ABMIL [31] introduces an attention mechanism into a multi-instance model while explaining the extent to which patch contributes to the classification results of pathological images. DSMIL [32] first trains a feature extractor for pathology images using self-supervised contrast learning, and then generates feature representations of instance by a multi-scale feature fusion mechanism that deploys a standard maximum pooling layer to identify the highest scoring instance (called key instance), and measures the distance between other instance and key instance to calculate the attention score. TransMIL [33] uses the Transformer to mine contextual information around individual patches and correlation information between different patches.

Methods

Overview of the TMIL model

Figure 1 illustrates the overall structure of the Transformer-based Multiple Instance Learning Network with 2D positional encoding framework (TMIL) designed in this paper. TMIL processes super-resolution pathology images using a two-stage training approach.

In the first stage, TMIL first segments the super resolution pathology images into numerous patches, and according to the concept of multiple instance learning, the super resolution pathology images are termed bag, and each patch under the bag is termed instance. The instances extracted from the bag are used to construct a new data-set, using the labels of the bag as pseudo-labels for the instances. Two ResNet models are trained on this dataset for classification tasks and metric learning tasks, respectively, each producing a 256-dimensional feature vector. These models are modified versions of the standard ResNet-50 architecture. As for

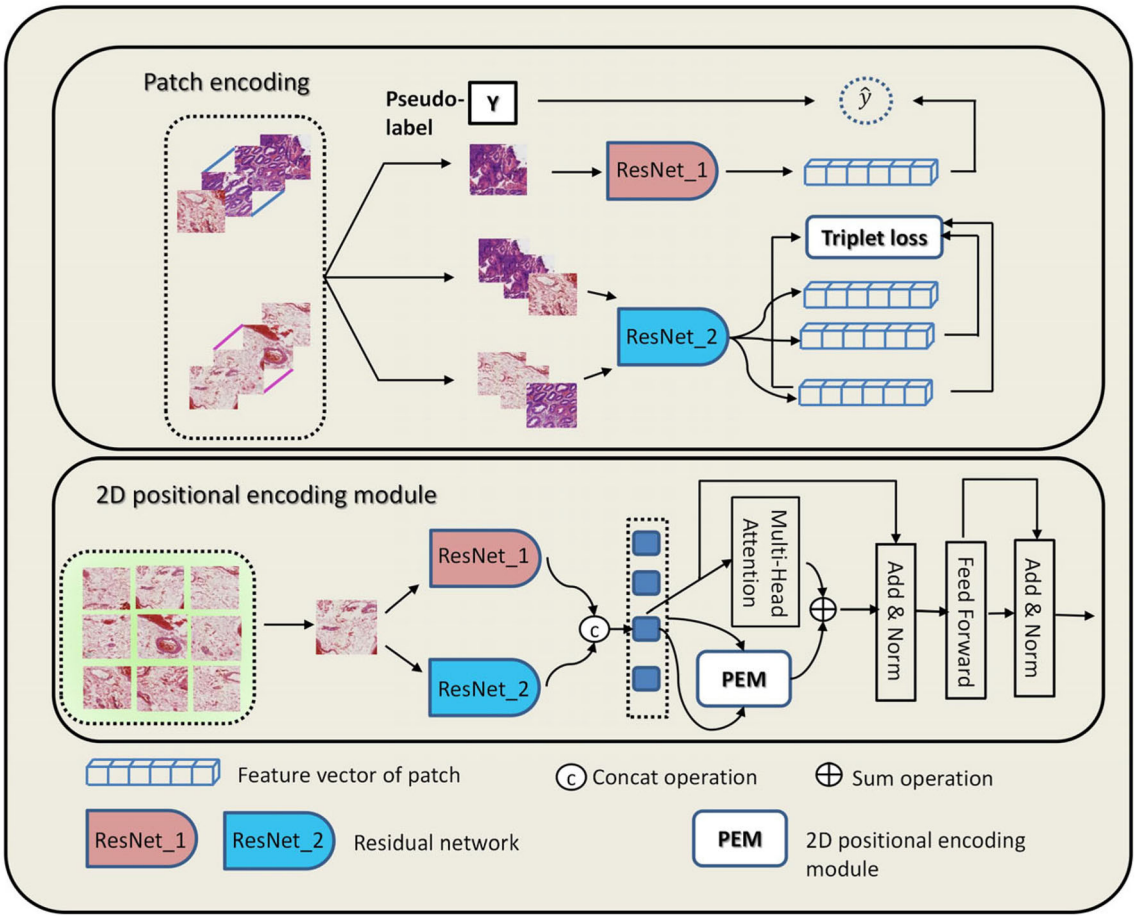

Fig. 1 Overview of TMIL model

ResNet_2 shown in Fig. 1, after the GAP layer of ResNet-50, we add a fully connected layer with 256 outputs that projects the representations into a space where metric learning loss is computed. The output of this layer will serve as the desired 256-dimensional feature vector. For ResNet_1, we add another fully connected layer on top of the output from ResNet_2 with a signal output that projects the representations into a space where binary cross-entropy loss is computed.

In the second stage, TMIL will use the patch feature representation obtained in the first stage as the input to the Transformer, and a two-dimensional positional encoding module (PEM) is designed to model the position between instances. After the feature representation passes through the multi-head attention module, it is summed with the position vector passed through the PEM to obtain the feature vector with position information. It's worth noting that our dataset includes slides with a maximum of 644 tiles each, allowing us to process every tile via the Transformer module with full self-attention, unlike approaches like TransMIL that use approximations due to memory limits.

In the inference phase, the process is similar to the second phase of training. First, the feature vectors of all patches are generated using the two ResNet trained in the patch encoding phase, and then the feature vectors are fed into the Transformer with two-dimensional position encoding module, finally the bag-level classification results are output.

Patch encoding

The traditional patch encoding mechanism for multiple instance learning has two approaches: one utilizes only the instance with the highest probability of cancer in each bag to train the classification task and uses the embedding vector at the end of the classification network as the feature vector of the instance. The other approach involves training through self-supervised contrast learning. The first approach uses the bag labels as pseudo-labels for the extracted instances. However, the limited number of bags results in a restricted number of extracted instances, which can increase the risk of overfitting during the training of the classification model. The second approach, while capable of using all instances

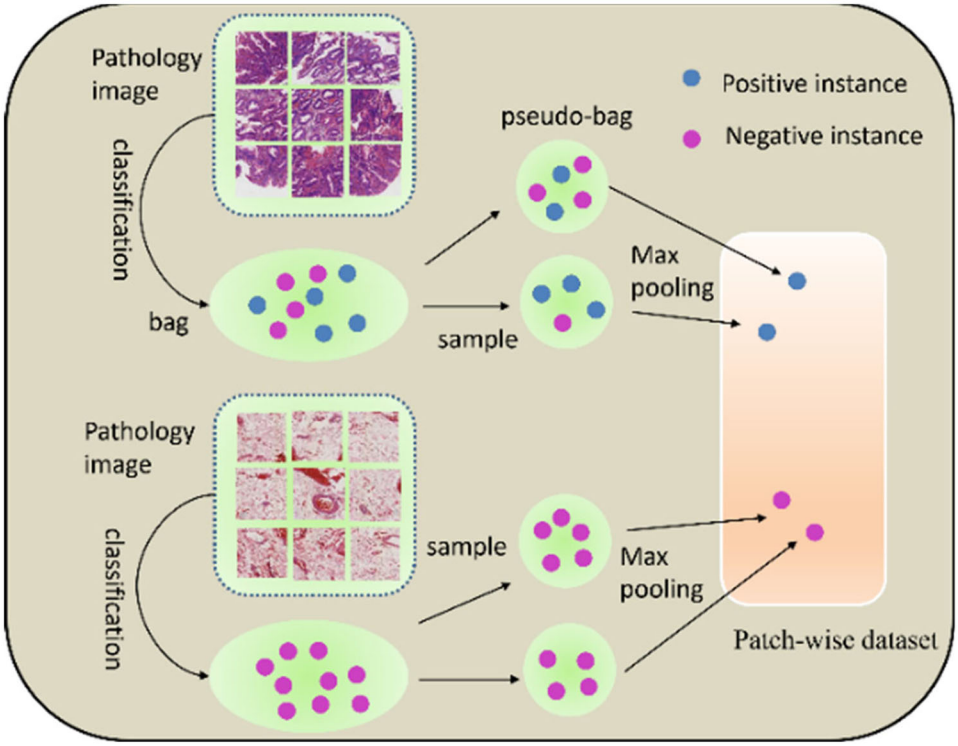

Fig. 2 Schematic diagram of the pseudo-bag mechanism

to train the encoder without requiring pseudo-labels, comes with stringent hardware and training period requirements.

In contrast, this paper combines pseudo-bag mechanism and metric learning to expand the number of instances, and train triplet loss function to make similar samples as close as possible in feature space and dissimilar samples as far as possible in feature space.

Constructing datasets using pseudo-bag. Figure 2 shows how to construct a patch dataset by the pseudo-bag mechanism. The original pathology images will be pre-processed into a large number of patches. TMIL constructs a new pseudo-bag by randomly selecting M instances without putting them back, using the labels of the original pathology images as the pseudo-labels of the pseudo-bags. When the original pathology image is negative, all patches of that pathology image are negative, and therefore the pseudo-bags are also negative. However, when the pathology image is positive, then some patches of that pathology image are necessarily positive, but some patches may be negative. There are pseudo-bags with the wrong label.

The size of the parameter M controls the degree of impact of this problem. When M is larger, it means that the more instances in the pseudo-bag, the lower the probability of labelling errors, but the number of pseudo-bags will decrease; similarly, the smaller M is, the fewer instances, the larger the number of pseudo-bags, but the higher the probability of labelling errors.

After constructing the pseudo-bag, the instance with the highest probability of cancer can be extracted from the pseudo-bag to form a new instance-level dataset. For the pseudo-bag with positive label, TMIL extracts the instance with the highest cancer probability, assigning it a positive pseudo-label. However, for pseudo-bag with a negative label, TMIL uses all instances of the pseudo-bag to balance the number of positive and negative samples in the final dataset. This strategy avoids the impact of the category imbalance problem on the encoder. In addition, this strategy allows us to concentrate the training process on the most informative instances, potentially enhancing the model's ability to accurately classify pathology images. Given a slide that contains N patches, the total number of instance-level dataset that can be generated from this slide is quantified by the floor division of N by M, denoted as $\lfloor \frac{N}{M} \rfloor$, where M represents the number of instances per pseudo-bag.

Feature encoding based on *depth metric learning*. In this paper, we introduce a metric learning model based on the traditional multiple instance learning.

The idea of metric learning is similar to self-supervised contrast learning in that similar samples are as close as possible in the feature space, while different samples are as far away as possible. The difference is that the training samples for self-supervised contrast learning do not have labels. Thus, in self-supervised contrast learning, feature augmentation is performed on each sample. Here, samples are considered similar only to their own augmented versions, while all other

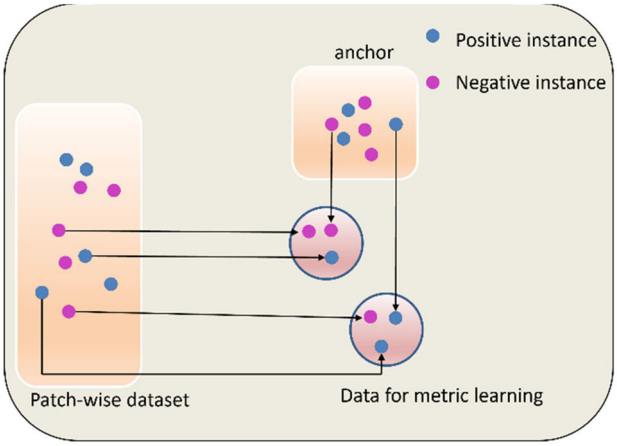

Fig. 3 The process of constructing metric learning data

samples are treated as negatives. Metric learning is a type of supervised learning where each sample has a label. This label information aids in identifying data similar to the sample. Furthermore, having label information as a priori knowledge greatly enhances model training.

Figure 3 illustrates the process of constructing a dataset for training metric learning. In this experiment, each instance is called an anchor, and an instance with the same pseudo-label is taken as a positive sample of the anchor, and an instance with a different pseudo-label is taken as a negative sample of the anchor.

The dataset is processed through a residual network to obtain three embedding vectors for each instance. On one hand, the embedding vector of the anchor goes through a fully connected layer for classification, and the anchor's predicted value is pseudo-labeled to compute the cross-entropy loss for training the classification task. On the other hand, the embedding vectors of the three instances are used to calculate the triplet loss to train the metric learning task. These tasks produce two residual networks, each outputting a 256-dimensional feature vector, and the concatenation of these two vectors forms the final feature vector for the instance. Each anchor instance is encoded through the residual network, which is trained to recognize and respond to the nuances of pathological data. The classification task focuses on differentiating between the classes represented in the dataset. The cross-entropy loss evaluates how well the predicted pseudo-labels align with the anchor's assigned pseudo-label, providing feedback to the model to fine-tune its parameters for better accuracy.

In parallel, the metric learning can compute the distances between the anchor's embedding vector and those of both positive and negative instances. The triplet loss function encourages the model to minimize the distance between the anchor and positive instances while maximizing the distance from the negative instances. This results in a feature space where instances of the same class are clustered together, and those of different classes are separated, facilitating the task of classification.

The dual tasks of classification and metric learning work concurrently, each contributing to a comprehensive representation of instances in the feature space. By utilizing both a cross-entropy loss for classification and a triplet loss for metric learning, the TMIL approach harnesses the benefits of supervised learning while also embedding a deeper understanding of instance similarities and differences within its architecture. This dual strategy is designed to be robust against overfitting, as it provides multiple avenues for the model to learn from the data, thus enabling it to develop a more generalizable and robust understanding of the features relevant for diagnostic tasks.

The specific equation is described as follows:

X is defined as the instance dataset:

$$X = \left\{\{x_1, y_1\}, \ldots, \{x_i, y_i\}, \ldots, \left\{x_k, y_k\right\}\right\}. \tag{1}$$

x_i is the i-th instance, y_i is the pseudo-label corresponding to the instance, and k denotes the total number of instances.

On the one hand, the dataset is used for the binary classification task, here using the cross-entropy loss function:

$$Lce = \sum_{i=1}^{k} y_i * \log\widetilde{y_i} + (1 - y_i) * \log\left(1 - \log\widetilde{y_i}\right) \tag{2}$$

$$\widetilde{y_i} = \sigma(W * F_1(x_i) + b) \tag{3}$$

σ is sigmoid function, W is a fully connected layer parameter, b is a fully connected layer bias. $F_1(*)$ is the classification encoder, which output 1×256 dimensional vector.

On the other hand, a deep metric learning model can be trained so that the model can learn to vector spaces of different dimensions, here we use the triplet loss function Lp:

$$Lp = \sum_{i=1}^{k}\left[\left|\left|F_1(x_i^a) - F_2\left(x_i^p\right)\right|\right|_2^2 - \left|\left|F_1(x_i^a) - F_2(x_i^n)\right|\right|_2^2 + \text{margin}\right] \tag{4}$$

x_i denotes the i-th instance, which model consider as an anchor and write as x_i^a; a instance from the dataset with the same label of x_i, which is written as x_i^p; and an instance from the dataset with different labels of x_i, which is written as x_i^n; $F_1(*)$ is the metric learning encoder. Sample pairs with matching labels are termed positive pairs, while those with different labels are termed negative pairs. Finally, margin is an adjustable parameter. This loss function allows similar examples to be closer in vector space and different examples to be further in vector space.

Each instance is fed into two encoders to obtain a 1×256-dimensional feature vector, and the feature vector is stitched together to obtain the final feature vector of the instance.

Bag classification

The multi-instance learning based on the embedding paradigm assumes that the instances in the same bag conform to independent homogeneous distribution, which violates the basic understanding of images. The self-attention mechanism in Transformer ensures that each instance can focus on the global instance information, thus circumventing the aforementioned assumption. TMIL employs the self-attention mechanism to capture correlations between instances. It also introduces a two-dimensional positional encoding module to address the Transformer's limitation in modeling two-dimensional positional relationships.

Instance relevance information. Figure 4 shows the process of modeling the instance relevance information by the self-attention mechanism. The model first encodes each patch using two encoders, then combines the outputs into a 512-dimensional feature vector, and finally inputs the feature vector into a multi-headed attention model, ensuring global attention to all instance information.

The pathology image is defined as a bag $B = \{x_1, \ldots, x_n\}$, where n is the number of instance and x_i is the i*th* instance of the bag. The 1×512 dimensional feature vector e_i of each instance is first obtained using the residual network, and then the vector e_i is passed through the fully connected layer to obtain q_i, k_i and v_i, which represent the query vector, the key vector, and the value vector respectively. The query vector is dotted with the key vectors of all other instances, the dotted product result is used as the attention score, and finally the value vectors are summed using the attention score.

$$q_i = W_1 * e_i + b_1,\ Q = \text{concat}\left[q_i,\ \ldots,\ q_n\right] \tag{5}$$

$$k_i = W_2 * e_i + b_2,\ K = \text{concat}[k_i,\ \ldots,\ k_n] \tag{6}$$

$$v_i = W_3 * e_i + b_3,\ V = \text{concat}[v_i,\ \ldots,\ v_n] \tag{7}$$

where concat[*] is the vector concatenation operation, which preserve and integrate different aspects of the relationships between instance vector in the input sequence and thereby preserve the expressive power of the model; d is the dimension of q_i; W_1, W_2 and W_3 are the parameters of the three fully connected layer; b_1, b_2 and b_3 are the biases of the three fully connected layers, respectively.

The above is the mathematical formula of the self-attention mechanism, while Transformer usually uses the multi-headed self-attention mechanism, which divides three vectors into multiple sub-vectors, and then enters the self-attentive module separately, and finally concatenation the output of the module together.

PEM. Figure 5 illustrates the two-dimensional position encoding module designed in TMIL. This module inputs the position indexes (row and column numbers) of all patches from a digital pathology image into the Transformer. It then derives the row and column number vectors from the patch position index and combines them to produce a new 512-dimensional position vector. The row number vector and column number vector are randomly initialized before the model training, and their values are updated by the gradient. The 2D position encoding module eventually feeds this position vector into another multi-headed attention mechanism to complete the modeling of position correlations between instances.

We define the maximum number of rows of the patches of the pathology image as R and the maximum number of columns as C. The matrix rows $\in R^{R*512}$ represents the position features of rows and the matrix cols $\in R^{C*512}$ represents the position features of columns. For the i-th instance, whose row number in the 2D image is r and column number is c, the position information of the example p_i:

$$p_i = rows[r] + cols[c] \tag{8}$$

where the matrices rows and cols are the parameters needed to be learned, which are initialized randomly in the first stage and continuously updated with gradients later as the model is trained. The p_i is input to the multi-headed self-attention module to obtain the encoding position information P.

The vector Z with location information is obtained by adding the feature vector and the location vector.

$$Z = E + P \tag{9}$$

Here E, P and Z are matrices, and each row corresponds to a feature vector of an instance. E is the relevance information, and P is the location information.

Finally, the feature vector Z with location information is sequentially passed through the residual structure and feed-forward networks, and finally a 1×512 dimensional feature vector can be obtained, which can be used as the feature representation of pathological images (bag).

In natural language processing, the Transformer typically consists of 12 sub-modules, including multi-head attention mechanisms which is essential for considering information from different positions of the input sequence and modeling complex dependencies. However, due to the limited number of medical images available, training a model with such a large number of parameters could lead to overfitting. Therefore, to mitigate this risk, this paper employs only one sub-module of the Transformer.

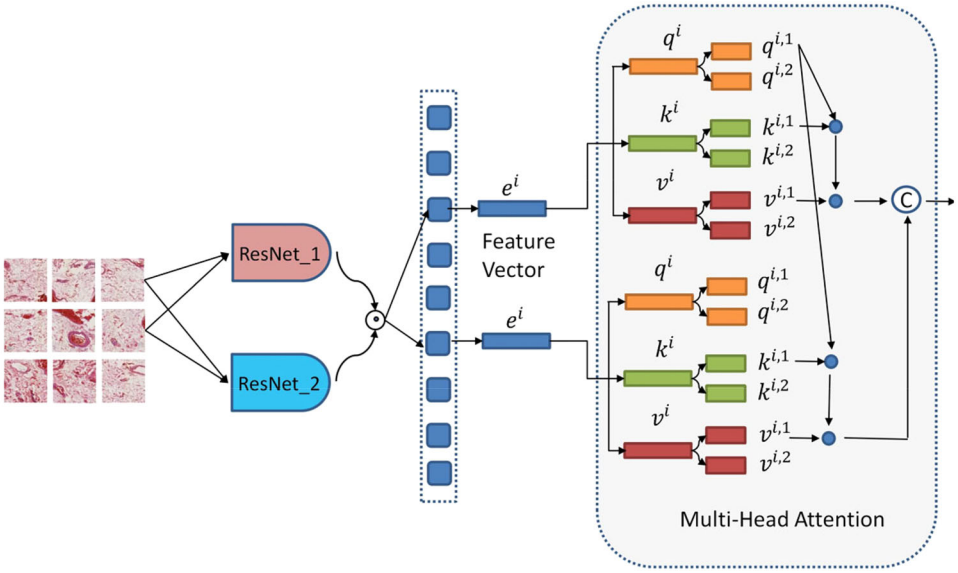

Fig. 4 Schematic diagram of instance correlations

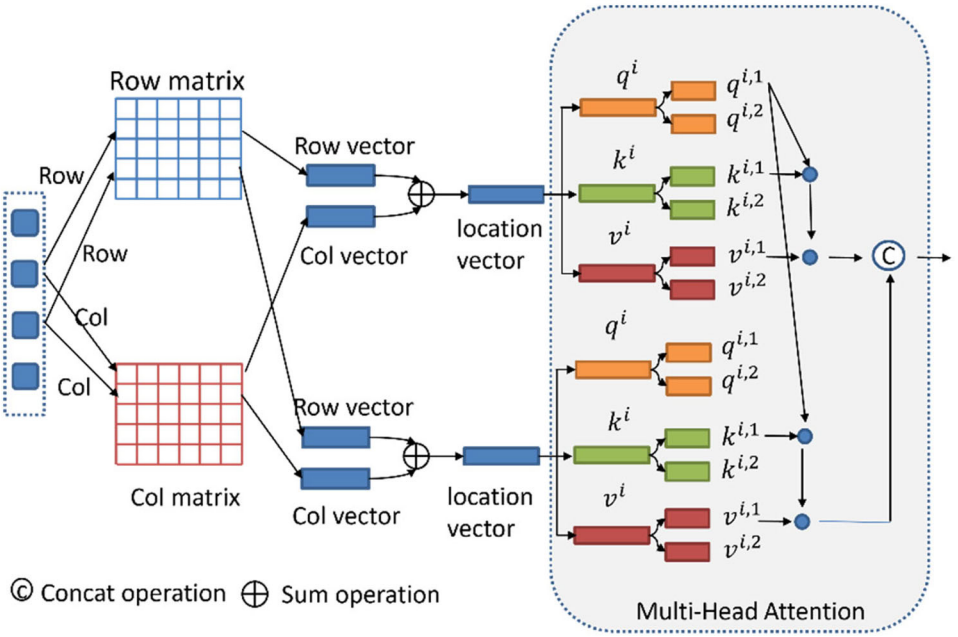

Fig. 5 Schematic diagram of the 2D position encoding module

Define LN as layer regularization and MLP as fully connected network:

$$Z_1 = LN(Z + E) \tag{10}$$

$$Z_2 = LN(\mathrm{MLP}(Z_1) + Z_1) \tag{11}$$

Transformer uses a mean pooling layer to aggregate the feature vectors of the instance into a feature vector of the pathology image (bag), and finally a fully connected layer to obtain the disease probability value y_p for this pathology image:

$$y_p = \sigma(W^* \ \mathrm{meanPooling}(Z_2) + b) \tag{12}$$

Based on y_p and the threshold value, we can determine whether the pathological image is diseased or not.

In the model training phase, the cross-entropy loss function is calculated using the predicted probability value y_p and the true label:

$$Loss = \sum_i y_i \ ^* \log y_{p,i} + (1 - y_i) * \log\left(1 - y_{p,i}\right) \tag{13}$$

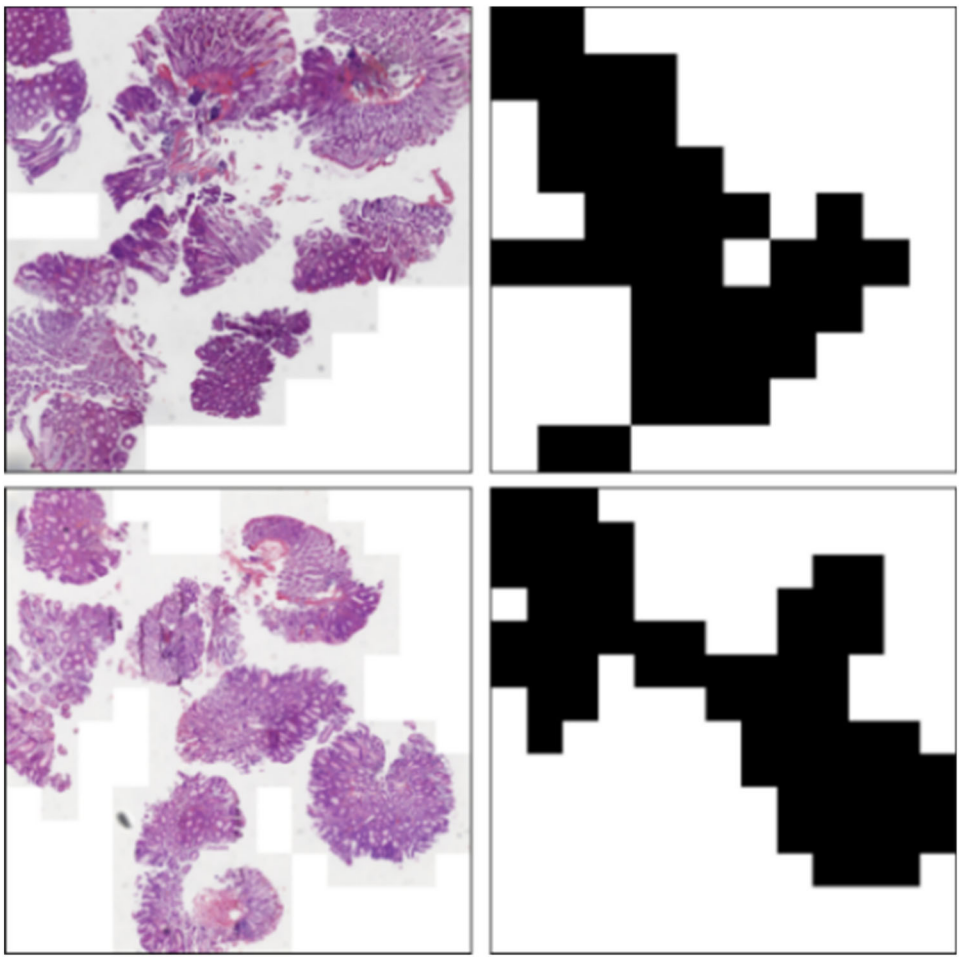

Fig. 6 Colorectal adenoma dataset. The left column shows the original histopathological slides from the colorectal adenoma dataset. The right column contains the ground truth masks, where black regions indicate cancerous areas

Experiment

Experimental setup

In this paper, a weakly supervised classification model is evaluated using a colorectal adenoma dataset [37], which is collected and labeled by pathologists at the People's Liberation Army General Hospital. In the dataset, 1 denotes the pathology image containing adenoma, while 0 indicates its absence. The colorectal adenoma dataset consists of multiple patches with the size of 1280×1280. After filtering each pathology image, the number of remaining patches ranges between 11 and 644. For the purpose of model training and evaluation, the dataset is divided into training and testing sets with a ratio of 6:4. Specifically, the training set consists of 155 patches, including 98 labeled as 1 and 57 labeled as 0. The testing set includes 104 patches, with 66 labeled as 1 and 38 labeled as 0.

In addition, the proposed method is implemented using the PyTorch framework. All experiments are conducted on a workstation equipped with a single NVIDIA RTX 5000 GPU with 16 GB of memory. The programming language used is Python 3.8. The system is running on Ubuntu 18.04 with CUDA version 10.2 and cuDNN 7.6.5.

Figure 6 shows the specific form of the colorectal adenoma dataset, with the left column of the figure showing the result of stitching the remaining patches after filtering one pathology image, and the right column of the figure showing the result of stitching the remaining patches labels. It should be noted that this paper investigates how to perform weakly supervised classification of large size pathology images when only image-level labels are available. Thus, only the image-level labels from this dataset are used for training, ignoring the patch-level labels.

The paper uses ACC (Accuracy) and AUC (Area Under Curve) scores to evaluate the classification performance in colorectal adenoma dataset.

Define TP to denote the number of samples with actual positive cases and predicted positive cases; TN to denote the number of samples with actual negative cases and predicted negative cases; FN to denote the number of samples with actual positive cases and predicted negative cases; and FP to denote the number of samples with actual negative cases and predicted positive cases.

Table 1 Comparison results of metric learning models

Methods	ACC	AUC
ABMIL	0.8654	0.9023
ABMIL_WC	0.9038	0.9274
TransMIL	0.9135	0.9542
TransMIL_WC	0.9135	0.9685
TMIL_NC	0.9327	0.9653
TMIL	**0.9519**	**0.9728**

The best result is highlighted in bold

Accuracy means the ratio of the number of correctly classified samples to the total number of samples.

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FN}+\text{FP}} \tag{14}$$

AUC is expressed as the area under the ROC (receiver operating characteristic curve). AUC is not affected by unbalanced data sets, nor by the distribution of test samples.

To compare with the most recent similar work [38], we conducted a detailed comparative analysis in the BreakHis [39] dataset. The BreakHis dataset is a collection of breast cancer histopathological images, which are microscopic images of breast tumor tissue. The dataset contains 7,909 images of 82 patients, with different magnifying factors (40X, 100X, 200X, and 400X). The images are divided into two main groups: benign tumors and malignant tumors. The dataset also contains four subtypes of benign tumors: adenosis (A), fibroadenoma (F), phyllodes tumor (PT), and tubular adenoma (TA); and four subtypes of malignant tumors: ductal carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC) and papillary carcinoma (PC). Each image filename stores information about the image itself, such as the method of biopsy, the tumor class, the tumor type, the patient identification, and the magnification factor. Additionally, following the work [37], we also used precision, recall and F1-score as evaluation metrics.

Ablation experiments

Classical multi-instance learning outputs a feature representation of each patch by a classifier. In this paper, we add a depth metric learning module to optimize the feature space of the instances. To verify the effectiveness of this module, the impact of the depth metric learning module on the model is discussed in this paper in three networks, ABMIL, TransMIL, and TML, respectively.

As shown in Table 1, the models ABMIL_WC and TransMIL_WC denote ABMIL and TransMIL with deep metric learning, respectively, and TMIL_NC denotes TMIL without deep metric learning. For ABMIL, the ACC metric improves from 0.8654 to 0.9038 and the AUC metric improves from 0.9023 to 0.9274 after adding the feature vector of metric learning; for TransMIL, the accuracy metric remains unchanged and the AUC metric improves from 0.9542 to 0.9685; for TMIL, the ACC metric improves from 0.9327 to 0.9519, and the AUC metric improved from 0.9653 to 0.9728. The above data verify that when the feature vectors trained by deep metric learning are stitched with the feature vectors trained by the multi-instance model based on the classification task, the new feature vectors obtained have a significant improvement on the model effect.

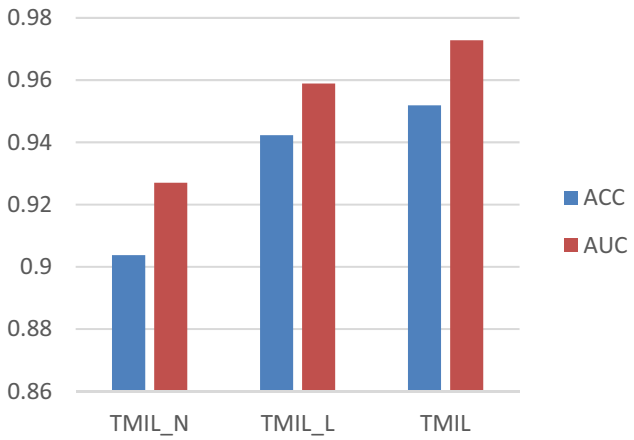

Fig. 7 Comparison results of different location encoding module

For the first time, TransMIL uses a self-attention mechanism to explore correlations between instances and also introduces a PPEG module to account for location information. Unlike TransMIL's two layers, the TMIL proposed in this paper uses only one layer of self-attention mechanism. Meanwhile, the PPEG module of TransMIL converts one-dimensional examples into two dimensions and then uses convolutional neural network to explore location information. While in this paper, TMIL directly models the location vector using the self-attention module, influenced jointly by the instance's row and column number vectors.

Figure 7 shows the effect of the 2D position encoding module proposed in this paper on the effect of the model. TMIL_N indicates the result of completely removing the position information, and TMIL_L indicates the result of using linear position encoding. It can be seen that the instance position information has a significant improvement on the effect, and also the PEM module proposed in this paper is better than the traditional Transformer's linear position encoding method.

To explore how the choice of M affects our algorithm's performance, we expand our ablation study by varying M in further experiments. Figure 8 provides a visual depiction of the model's performance as influenced by varying the number of pseudo-bags, M. Notably, the model reaches

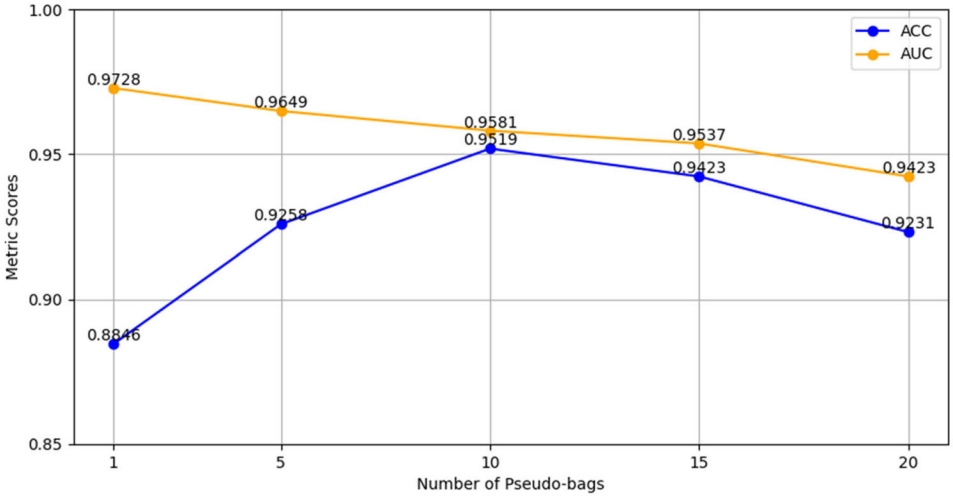

Fig. 8 Comparison results of different numbers of M in the pseudo bag

its highest ACC of 0.9519 and AUC of 0.9649 when M is set to 5. This suggests a sweet spot where each pseudo-bag contains enough instances to represent the slide's pathological features effectively without introducing too much noise or redundancy. As M increases to 10, both ACC and AUC experience a slight drop, with ACC decreasing to 0.9423 and AUC to 0.9581, indicating that while the model still performs well, there may be a diminishing return on adding more instances per pseudo-bag. Further increasing M to 15 leads to a more noticeable decline in ACC to 0.9423 and AUC to 0.9537, reinforcing the idea that excessively large pseudo-bags may impair the model's discriminative capability. This quantitative analysis highlights that M significantly affects the model's accuracy and its generalization ability, illustrating a clear trend that an optimally sized M is crucial for model performance.

Experimental result

Compare with *other* WSI *methods*. The specific comparative model chosen for this paper is as follows:

- ABMIL adds the attention mechanism to the multi-instance model for the first time, and the computed attention scores are used in the aggregation layer.
- DSMIL juxtaposes a multi-instance framework based on the example paradigm and a multi-instance framework based on the embedding paradigm in a single model. The encoder is trained by contrast learning in the first stage, and the instance classifier and bag classifier are concatenated in a dual-stream architecture in the later stage to focus more on the correlation between the highest scoring instance and the remaining instances.

Table 2 Results of weakly supervised classification

Methods	ACC	AUC
ABMIL	0.8654	0.9023
DSMIL	0.9231	0.9605
TransMIL	0.9135	0.9542
ViT	0.9423	0.9589
IBMIL	0.9327	0.9612
TMIL	**0.9519**	**0.9728**

The best result is highlighted in bold

- TransMIL introduces the Transformer to the multi-instance model for the first time. The model considers correlation between different instances and does not conform to the assumption of independent identical distribution.
- ViT introduces Transformer to image classification tasks for the first time. The original ViT is not directly applicable to the classification of super resolution pathology images. To make ViT applicable to medical scenarios, the input of ViT is modified to the embedding vector of patches in this paper. The purpose of comparing this model is to verify the superiority of Transformer in instance correlation modeling.
- IBMIL [40] is a new method that utilizes backdoor adjustment for interventional training to mitigate biases from bag-level contextual priors, offering a unique approach orthogonal to traditional bag MIL methods.
- TMIL is a Transformer multi-instance learning network based on two-dimensional location information proposed in this paper.

Table 2 shows the comparison results of TMIL and several existing methods. From this table, it can be seen that the Transformer network based on two-dimensional location

Table 3 Results of the p-value on test dataset

Methods	P-value	
	ACC	AUC
ABMIL	7.185e-13	4.124e-11
DSMIL	1.415e-6	3.149e-5
TransMIL	4.773e-10	4.034e-7
ViT	1.358e-4	7.891e-8
IBMIL	4.590e-7	9.514e-7

information proposed in this paper outperforms the traditional models based on attention mechanism in both ACC metrics and AUC metrics. ABMIL shows the lowest performance among the methods listed, with an ACC of 0.8654 and an AUC of 0.9023. The IBMIL method, despite having a lower Accuracy (ACC) at 0.9327 compared to ViT's 0.9423, achieves a notably higher Area Under the Curve (AUC) score of 0.9612, surpassing ViT's 0.9589 by approximately 0.0023 points. This suggests that IBMIL, with its novel approach utilizing backdoor adjustment for interventional training to mitigate biases from bag-level contextual priors, excels in distinguishing between positive and negative samples more effectively than ViT. The ACC value of TMIL is 95.19% and the AUC is 97.28%, which is an improvement of 0.96% in ACC and 1.39% in AUC compared with ViT. This experimental data demonstrates that TMIL is still able to perform pathology diagnosis on very large resolution pathology images and outperforms existing weakly supervised classification models with only image-level labels.

In this study, we employ the p-value [41] to assess the differences in performance between various methods and TMIL across ten iterations, using the standard that a p-value below 0.05 indicates a statistically significant difference. To select the appropriate statistical test, we first applied the Shapiro–Wilk test to assess the normality of the performance metrics (ACC and AUC). As the results showed non-normal distributions, we used the Wilcoxon signed-rank test, a non-parametric method, to compute the p-values reported in Table 3. The results in Table 3 indicate that all methods have p-values significantly lower than the 0.05 threshold for both ACC and AUC, demonstrating that TMIL significantly outperforms other approaches.

After completing the weakly supervised classification of super resolution pathology image, TMIL can also obtain the label of each patch in that pathology image. In the patch encoding phase, this paper provides pseudo-labels for the instance according to the pseudo-bag mechanism, thus directly training an instance-level classifier. Therefore, after completing the classification task of the pathological images, this paper can also directly use the instance-level classifier to predict each patch in the pathological images, and the output of the classifier will be used for the visualization of the lesion regions. The specific visualization effect is shown in Fig. 9.

Comparison with Similar Works. TMIL is a two-stage method which is more suitable for large-scale WSI images classification. In order to adapt to the multi-class BreakHis dataset, we replace the first stage with a pre-trained Swin Transformer as the backbone to extract feature representations of instances, and then modify the final classification fully connected layer of the bag-level classifier to obtain eight classification results. In each training epoch, we use down sampling to ensure consistent sampling numbers and reduce overfitting from imbalanced data.

In our study, we compare the classification results using Swin Transform alone with those obtained when it is employed as the feature extractor in the first stage. Table 4 shows the multi-class results of the two methods without considering the magnification factor. As shown in the last row of Table 4, our approach yields superior outcomes, with an improvement of 2.6% (from 95.25 to 97.82), 1.9% (from 95.25 to 97.82), and 2.2% (from 95.25 to 97.82) in precision, recall, and F1-score, respectively. However, comparing the results of each category, only the PT and MC classes show slightly better metrics than our method; for all other categories, our approach was the most effective. The LC subcategory sees a significant improvement in Precision (from 82.53 to 94.81) when using our method compared to the Swin method. The DC subcategory, which has the highest sample number (692), also outperform Swin in all three metrics. In general, our model achieves the highest results for magnification-independent classification. This is significant because multiclass classification can be more challenging than binary classification due to the increased number of classes. This showcases the robustness and versatility of your model.

We further evaluated the performance of the two methods using ROC curves. As depicted in Fig. 10, most categories demonstrated favorable ROC performance, with TA, PT, DC and PC improving by 1%, 1%, 1% and 1%, respectively. Notably, the AUC value for LC and MC increased by 2% (from 0.95 to 0.97) and 3% (from 0.94 to 0.97). This underscores the exceptional performance of our method for these categories.

Table 5 shows the multi-class results of the recent work ViT-DeiT and the method proposed in this paper with magnification-dependent classification. Overall, the performance figures obtained by our method at different magnifications are quite consistent, remaining around 98%, which indicates that our model is well suited for classification with this dataset. At 40X, Oure method shows a precision of 98.48%, recall of 98.57%, and an F1-score of 98.65%. While these numbers are slightly lower than the ViT-DeiT method, they are still remarkably high and indicate that the TMIL model is very adept at this magnification. TMIL starts to

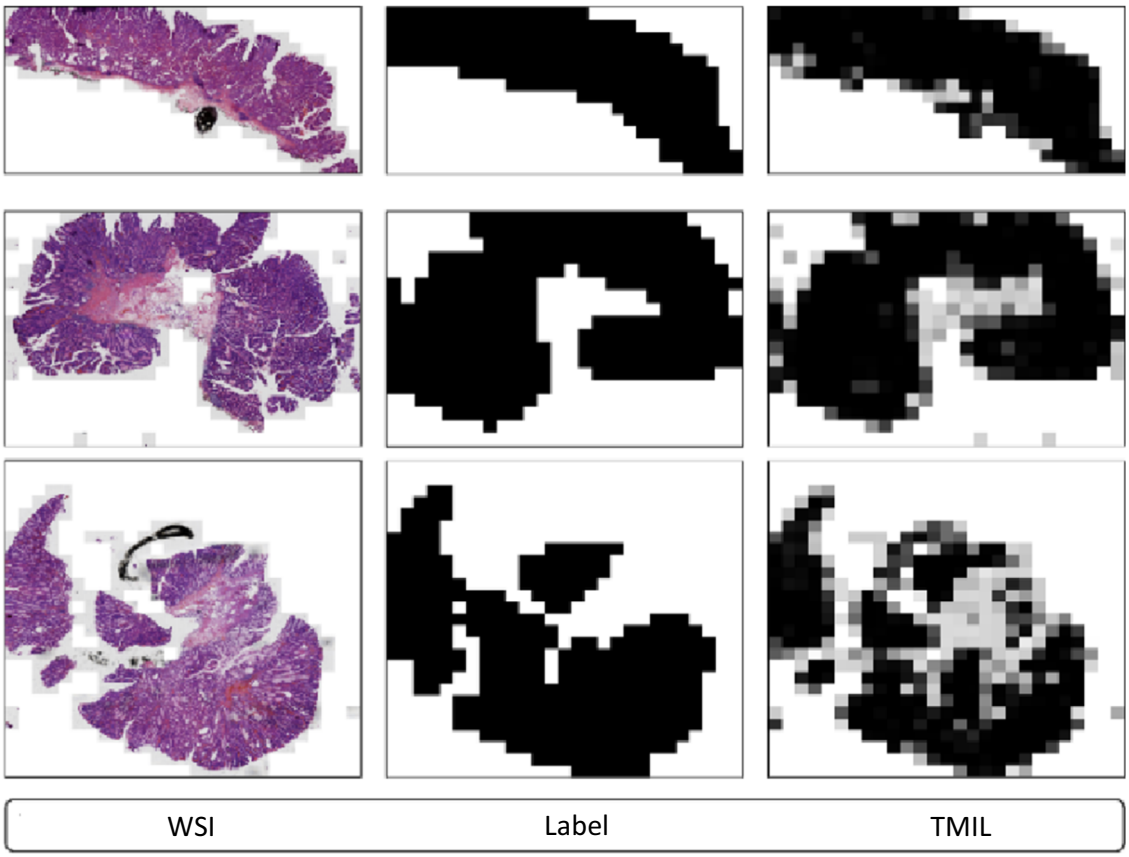

Fig. 9 A visualization of pathology images**.** The first column (WSI) contains the original whole slide images. The second column (Label) shows expert-annotated cancerous regions. The third column (TMIL) presents the predicted cancerous regions by the TMIL model

Table 4 Comparison of multiple classification results in each class

Sub category	Precision (%)		Recall (%)		F1-score (%)		Sample number
	Swin	Ours	Swin	Ours	Swin	Ours	
A	97.90	97.93	98.90	99.47	98.90	98.70	91
F	95.60	98.25	96.07	96.98	95.84	97.61	204
TA	95.69	1.00	96.73	98.33	96.21	99.15	92
PT	98.10	95.98	97.36	99.17	98.23	97.55	114
DC	95.15	97.24	96.53	99.89	95.83	98.55	692
LC	82.53	94.81	81.88	93.09	82.21	93.94	127
MC	99.32	99.01	9312	91.26	96.12	94.98	160
PC	96.49	9933	97.34	98.68	96.91	99.00	113
macro avg	95.25	**97.82**	95.23	**97.11**	95.22	**97.43**	

The best result is highlighted in bold

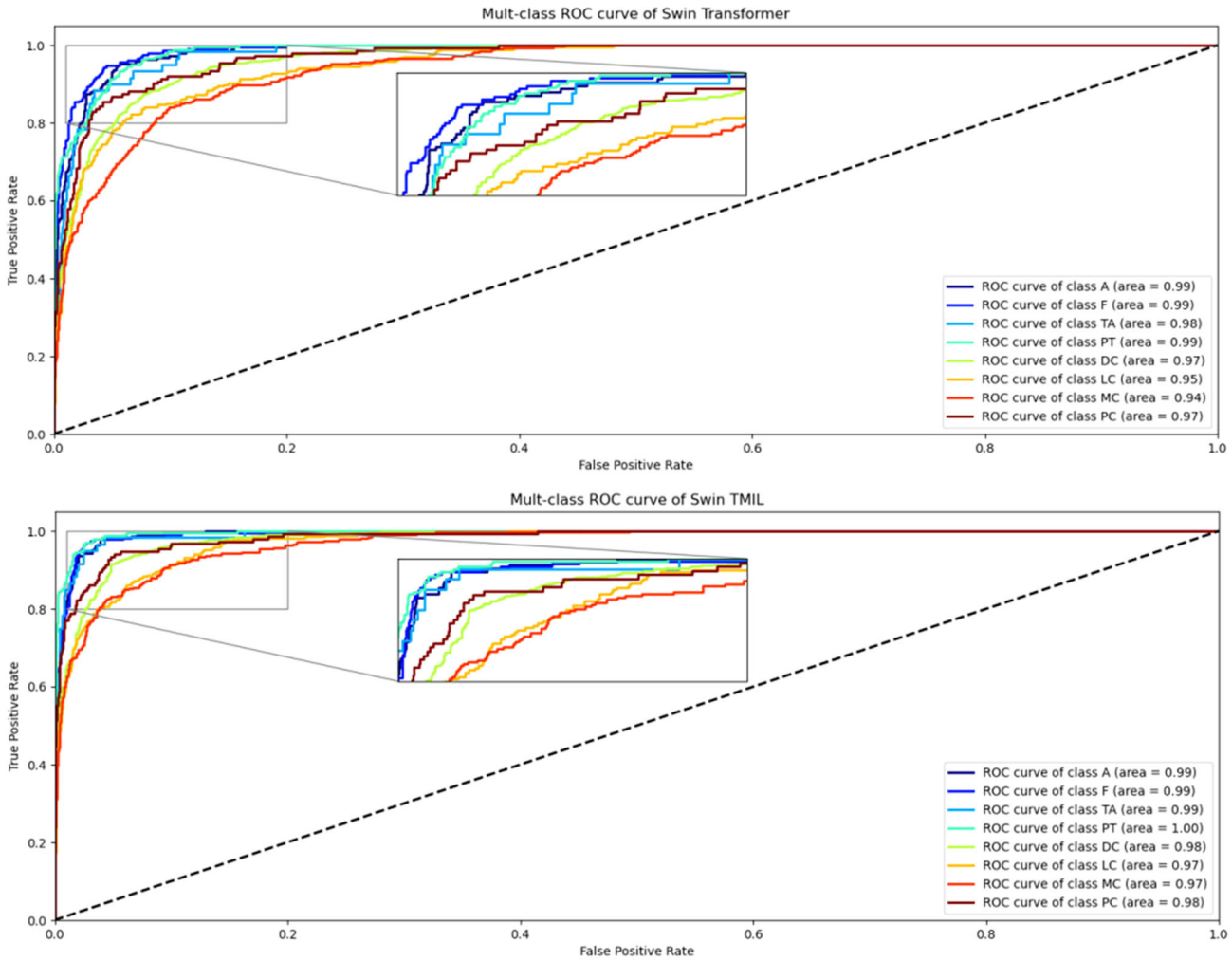

Fig. 10 Class-wise ROC curves for Swin transform (up) and Swin structure based TMIL (down)

Table 5 Comparison of multi-class classification performance with magnification-dependent classification

Methods	Magnification	Precision (%)	Recall (%)	F1-score (%)
ViT-DeiT	40X	99.38	99.46	99.40
	100X	98.31	98.51	98.35
	200X	98.31	98.27	98.23
	400X	98.57	98.78	98.65
TMIL(Ours)	40X	98.65	98.41	98.39
	100X	**98.69**	98.27	**98.36**
	200X	**98.48**	**98.57**	**98.65**
	400X	98.23	97.40	98.42

The best result is highlighted in bold

catch up and even surpasses ViT-DeiT in some metrics as magnification increases. This suggests TMIL might be more robust or adaptable to higher magnifications. TMIL overtakes ViT-DeiT slightly in precision, recall, F1-score at 200X by 0.17%, 0.30%, 0.42%. It is notable that while TMIL's Recall dips at 400X, its F1-score remains competitive, suggesting a balanced trade-off between Precision and Recall.

Strengths and *limitations*. The Transformer-based Multiple Instance Learning Network with 2D positional encoding (TMIL) introduces several innovative features that set it apart. Among its most notable strengths are its advanced 2D positional encoding module, which offers a more accurate modeling of spatial relationships compared to traditional methods. This, combined with the self-attention mechanism, ensures that each instance can focus on global instance information, capturing intricate relationships between instances. Furthermore, TMIL's two-stage training approach is adept at handling the complexities of super-resolution pathology images. Moreover, the metric learning and pseudo-bags mechanism expands the number of instances and reduces the risk of overfitting, a common challenge in deep learning.

However, while TMIL presents several benefits, it also comes with certain limitations. Firstly, the performance of MIL largely depends on the representation and quality of the data. If the instances within a bag are not very representative or are noisy, the model's performance might be affected. Secondly, while the introduction of the 2D positional encoding module in TMIL offers advantages, it also introduces biases, especially if the positional information is not always relevant or accurate. Its ability to generalize across diverse and larger datasets remains to be tested. Thirdly, compared to traditional supervised learning, MIL has a higher label ambiguity. Since only the bag's label is known and not the exact label for each instance within the bag, this can lead to instability in training the model, although our pseudo-bag mechanism can mitigate this effect. Lastly, the use of a single layer of self-attention, compared to some other methods with layered approaches, might have its own set of challenges in capturing deeper relationships.

Conclusion

The complexity and diversity of medical images make it exceptionally difficult to analyze them, and precise methods are required to extract useful information and features from them. The small amount of data and the difficulty of annotation all pose great challenges for medical image analysis. Weakly supervised learning can use less annotation information to train models, thus effectively reducing the annotation workload. In this paper, we propose a weakly supervised classification algorithm (TMIL) based on multiple instance learning to analyze medical images. TMIL enables the classification of pathological images without patch-level annotations, addressing the challenge of directly training models on super-resolution pathological images. The metric learning model was proposed by TMIL to enhance the feature representation of image patches, and the improved 2D position encoding module can fully exploit the 2D position information of images while ensuring the excellent instance relevance modeling capability of Transformer.

In this paper, TMIL is validated on a colorectal adenoma dataset, which achieves an ACC metric of 95.19% and an AUC metric of 97.28%. The experimental results show that the metric learning model and the improved two-dimensional positional encoding module effectively improve the model performance. Meanwhile, the results on BreakHis dataset show that TMIL is a robust and versatile method, showing strong performance across different magnifications. Its ability to maintain performance, especially in magnification-independent evaluations, suggests a broader applicability in real-world scenarios. Given its strengths, TMIL presents a promising approach for image classification tasks, especially in the domain of weakly supervised classification of medical image.

Funding National Natural Science Foundation of China, 62076015, Bo Liu.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Alotaibi A, Alafif T, Alkhilaiwi F, et al (2023) ViT-DeiT: an ensemble model for breast cancer histopathological images classification. In: 1st International Conference in Advanced Innovation on Smart City, ICAISC 2023 Proceedings.
2. Bi Q, Qin K, Li Z et al (2020) A multiple-instance densely-connected ConvNet for aerial scene classification. IEEE Trans Image Process. https://doi.org/10.1109/TIP.2020.2975718

3. Campanella G, Hanna MG, Geneslaw L et al (2019) Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med. https://doi.org/10.1038/s41591-019-0508-1
4. Chen H, Han X, Fan X, et al (2019) Rectified cross-entropy and upper transition loss for weakly supervised whole slide image classifier. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).
5. Chen H, Qi X, Yu L, Heng PA (2016) DCAN: deep contour-aware networks for accurate gland segmentation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.
6. Chen PHC, Gadepalli K, MacDonald R et al (2019) An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. Nat Med. https://doi.org/10.1038/s41591-019-0539-7
7. Chikontwe P, Kim M, Nam SJ, et al (2020) Multiple Instance Learning with Center Embeddings for Histopathology Classification. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).
8. Cornish TC, Swapp RE, Kaplan KJ (2012) Whole-slide imaging: routine pathologic diagnosis. Adv. Anat. Pathol. 19(3):152–9
9. Ghaffari Laleh N, Muti HS, Loeffler CML et al (2022) Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. Med Image Anal. https://doi.org/10.1016/j.media.2022.102474
10. Hanna MG, Parwani A, Sirintrapun SJ (2020) Whole slide imaging: technology and applications. Adv. Anat. Pathol. 27(4):251–9
11. Hashimoto N, Fukushima D, Koga R, et al (2020) Multi-scale domain-adversarial multiple-instance CNN for Cancer Subtype Classification with Unannotated Histopathological Images. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.
12. Hashimoto N, Takagi Y, Masuda H et al (2023) Case-based similar image retrieval for weakly annotated large histopathological images of malignant lymphoma using deep metric learning. Med Image Anal. https://doi.org/10.1016/j.media.2023.102752
13. Hou L, Samaras D, Kurc TM, et al (2016) Patch-based convolutional neural network for whole slide tissue image classification. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.
14. Ilse M, Tomczak J, Welling M (2018) Attention-based deep multiple instance learning. In: International conference on machine learning. PMLR. pp 2127–2136
15. Kanavati F, Toyokawa G, Momosaki S et al (2020) Weakly-supervised learning for lung carcinoma classification using deep learning. Sci Rep. https://doi.org/10.1038/s41598-020-66333-x
16. Kang H, Yang M, Zhang F et al (2023) Identification lymph node metastasis in esophageal squamous cell carcinoma using whole slide images and a hybrid network of multiple instance and transfer learning. Biomed Signal Process Control. https://doi.org/10.1016/j.bspc.2023.104577
17. Law MT, Yu Y, Urtasun R, et al (2017) Efficient multiple instance metric learning using weakly supervised data. In: Proceedings 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017.
18. Li B, Li Y, Eliceiri KW (2021) Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.
19. Lin H, Chen H, Graham S et al (2019) Fast ScanNet: fast and dense analysis of multi-Gigapixel whole-slide images for cancer metastasis detection. IEEE Trans Med Imaging. https://doi.org/10.1109/TMI.2019.2891305
20. Pantanowitz L, Valenstein PN, Evans AJ et al (2011) Review of the current state of whole slide imaging in pathology. J Pathol Inform. https://doi.org/10.4103/2153-3539.83746
21. Pinckaers H, Van Ginneken B, Litjens G (2022) Streaming convolutional neural networks for end-to-end learning with multi-megapixel images. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2020.3019563
22. Rakhlin A, Shvets A, Iglovikov V, Kalinin AA (2018) Deep convolutional neural networks for breast cancer histology image analysis. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).
23. Shao Z, Bian H, Chen Y, et al (2021) TransMIL: transformer based correlated multiple instance learning for whole slide image classification. In: Advances in Neural Information Processing Systems.
24. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations, ICLR 2015 Conference Track Proceedings.
25. Spanhol FA, Oliveira LS, Petitjean C, Heutte L (2016) A dataset for breast cancer histopathological image classification. IEEE Trans Biomed Eng. https://doi.org/10.1109/TBME.2015.2496264
26. Tellez D, Litjens G, Van Der Laak J, Ciompi F (2021) Neural image compression for gigapixel histopathology image analysis. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2019.2936841
27. Bontempo G, Bolelli F, Porrello A et al (2024) A graph-based multi-scale approach with knowledge distillation for WSI classification. IEEE Trans Med Imaging. 43(4):1412–1421
28. Zhao Y, Lin Z, Sun K, et al (2022) Setmil: spatial encoding transformer-based multiple instance learning for pathological image analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention.
29. Pisula J I, Bozek K (2022) Language models are good pathologists: using attention based sequence reduction and text pretrained transformers for efficient wsi classification. arXiv preprint arXiv:2211.07384
30. Wang C, Wu Y, Wang C et al (2022) Attention-based multiple-instance learning for Pediatric bone age assessment with efficient and interpretable. Biomed Signal Process Control. https://doi.org/10.1016/j.bspc.2022.104028
31. Wang X, Chen H, Gan C et al (2020) Weakly supervised deep learning for whole slide lung cancer image analysis. IEEE Trans Cybern. https://doi.org/10.1109/TCYB.2019.2935141
32. Wang Y, Zhang J, Kan M, et al (2020) Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.
33. Xiang T, Song Y, Zhang C et al (2022) DSNet: a dual-stream framework for weakly-supervised gigapixel pathology image analysis. IEEE Trans Med Imaging. https://doi.org/10.1109/TMI.2022.3157983
34. Xu G, Song Z, Sun Z, et al (2019) CAMEL: A weakly supervised learning framework for histopathology image segmentation. In: Proceedings of the IEEE International Conference on Computer Vision.
35. Yao J, Zhu X, Jonnagaddala J et al (2020) Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. Med Image Anal. https://doi.org/10.1016/j.media.2020.101789
36. Yu S, Ma K, Bi Q et al (2021) MIL-VT: multiple instance learning enhanced vision transformer for fundus image classification. In: Lecture Notes in Computer Science (including subseries Lecture

Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, New York
37. Zhang H, Meng Y, Qian X, et al (2021) A regularization term for slide correlation reduction in whole slide image analysis with deep learning. In: Proceedings of Machine Learning Research.
38. Zhang H, Meng Y, Zhao Y, et al (2022) DTFD-MIL: double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.
39. Zhao Y, Yang F, Fang Y, et al (2020) Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.
40. Lin T, Yu Z, Hu H, et al (2023) Interventional bag multi-instance learning on whole-slide pathological images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
41. Hollander M, Sethuraman J (2001) Nonparametric statistics: rank-based methods. Elsevier, New York