applied sciences

MDPI

Article

Multiple Instance Classification for Gastric Cancer Pathological Images Based on Implicit Spatial Topological Structure Representation

Xu Xiang [1] and Xiaofeng Wu [2,*]

[1] School of Information Science and Technology, Fudan University, Shanghai 200433, China; 19210720042@fudan.edu.cn
[2] Research Center of Smart Networks and Systems, School of Information Science and Technology, Fudan University, Shanghai 200433, China
* Correspondence: xiaofengwu@fudan.edu.cn; Tel.: +86-021-6564-3757

Abstract: Gastric cancer is a malignant tumor with high incidence. Computer-aided screening systems for gastric cancer pathological images can contribute to reducing the workload of specialists and improve the efficiency of disease diagnosis. Due to the high resolution of images, it is common to divide the whole slide image (WSI) into a set of image patches with overlap before utilizing deep neural networks for further analysis. However, not all patches split from the same cancerous WSI contain information of cancerous issues. This restriction naturally satisfies the assumptions of multiple instance learning (MIL). Moreover, the spatial topological structure relationships between local areas in a WSI are destroyed in the process of patch partitioning. Most existing multiple instance classification (MIC) methods fail to take into account the topological relationships between instances. In this paper, we propose a novel multiple instance classification framework based on graph convolutional networks (GCNs) for gastric microscope image classification. Firstly, patch embeddings were generated by feature extraction. Then, a graph structure was introduced to model the spatial topological structure relationships between instances. Additionally, a graph classification model with hierarchical pooling was constructed to achieve this multiple instance classification task. To certify the effectiveness and generalization of our method, we conducted comparative experiments on two different modes of gastric cancer pathological image datasets. The proposed method achieved average fivefold cross-validation precisions of 91.16% and 98.26% for gastric cancer classification on the two datasets, respectively.

Keywords: gastric pathological images classification; graph convolutional networks; multiple instance learning

check for updates

Citation: Xiang, X.; Wu, X. Multiple Instance Classification for Gastric Cancer Pathological Images Based on Implicit Spatial Topological Structure Representation. *Appl. Sci.* **2021**, *11*, 10368. https://doi.org/10.3390/app112110368

Academic Editors: Cecilia Di Ruberto, Alessandro Stefano, Albert Comelli, Andrea Loddo and Lorenzo Putzu

Received: 9 October 2021
Accepted: 29 October 2021
Published: 4 November 2021

1. Introduction

Among all malignant tumors, the incidence and mortality of gastric cancer is in the forefront. Clinically, gastroscopy and biopsy for pathological diagnosis are still the gold standard methods for the examination of early gastric cancer. As the resources of digital gastric pathology images are being accumulated, many efforts have been made in designing computer-aided pathological image analysis systems to reduce the workload of pathologists and improve the efficiency of cancer diagnosis. Among all the pathological image analysis tasks, the most important is the classification task, which is the cornerstone and bottleneck of other in-depth studies, such as nuclei localization, gland segmentation, etc.

The early pathological image classification tasks are based on conventional machine learning algorithms. Common classifiers such as logistic regression (LR), support vector machine (SVM), K-nearest neighbor (KNN), etc. were used. Due to the high resolution of pathological images, some pre-processing should be performed prior to applying machine learning algorithms. The mainstream method is firstly to divide the whole pathological slide image into small local patches. Then, feature extraction and classification between cancer and non-cancer are performed in each local patch. Finally, the prediction results of these patches are integrated through certain methods, such as majority voting, to determine the classification results of image level or case level. At the feature extraction stage, meaningful features such as gray level co-occurrence matrix (GLCM), local binary pattern (LBP), histogram of oriented gradient (HOG), scale-invariant feature transform (SIFT), or certain well-designed, handcrafted features related to cell morphology and density can be utilized for subsequent pathology image analysis tasks.

Due to the high computing requirements in the process of handling high-resolution WSIs, extracting effective features to reduce data dimension is an important problem. However, handcrafted feature extraction relies heavily on professional domain knowledge and cannot guarantee self-discrimination and completeness, which restricts the application of conventional machine learning algorithms in pathological image analysis tasks. Recently, due to the remarkable performance of deep learning methods in the field of computer vision, most of the image recognition techniques have been replaced by deep learning. This is also true for pathological image analysis. When locations of abnormal or cancerous areas in WSIs are provided by pathologists, some state-of-the-art convolutional neural networks (CNNs) can be directly applied in the patch-level classification stage. Sharma et al. [1] proposed a well-designed CNN model for automatic classification of gastric carcinoma. Wang et al. [2] proposed a two-stage framework including a fully convolutional location network for discriminative instance selection and recalibrated multi-instance deep learning (RMDL) for gastric WSI classification. Additionally, pixel-wise annotations are necessary to train a representative location network. On the other hand, extracting informative features based on CNNs pre-trained on the ImageNet has become an effective method of data pre-processing. Some recent studies suggest ImageNet-trained CNNs are strongly biased towards recognizing texture information [3]. In a whole pathological slide, clustered cells with the same structure and function exhibit a similar textured appearance. Therefore, the pathological image is rather closer to a type of order-less texture-like image [4]. Multiple typical CNNs architectures have been taken into account to demonstrate that the internal layers of CNNs can act as feature extractors and generalize well to pathological images [5,6].

However, it is time-consuming and labor-intensive to ask experts to annotate a large number of high-resolution pathological images in detail. The scarcity of annotations has been a great challenge for the task of classifying pathological images. Most existing and available pathological image resources generally only have coarse-grained label information. For instance, only the image level label is given for a WSI with resolution of a million pixels. The benign and malignant tissues may both appear mixed in a complete pathological image labeled as cancerous. Thus, not all patches extracted from pathological images labeled as cancer contain class-specific information. When only a few regions provide the key information related to the cancerous class of the WSI and the remaining regions are more related to the benign class, standard supervised learning will fail due to the usage of a large number of mislabeled patches. Therefore, compared to the supervised method based on patch-wise classification, modeling the pathological image classification task as a weakly supervised problem with inexact supervision can be more appropriate [7]. In our study, the pathological image classification task was viewed as the multiple instance classification (MIC) problem.

The MIC methods can be categorized into three paradigms: Instance-Space, Bag-Space and Embedded-Space [8]. In the Embedded-Space paradigm, each bag is mapped to a single feature vector that represents the whole information about the bag. When the bag-level embeddings are obtained, the original MIC problem can be converted into a standard supervised classification problem. Standard classifiers such as SVM, KNN and neural networks can be applied, where each bag-level feature vector has its own label. To facilitate comparison, the fully connected layers were adopted as bag-level classifiers in our work. How to obtain bag-level embeddings that contain the global information is the key step in the Embedded-Space paradigm. Some studies have introduced multiple instance pooling layers to neural networks to solve the tasks related to pathological image analysis [9,10]. The MIL pooling layers integrate instance-level representations into bag-level embedding to serve the downstream task, such as histopathology image retrieval and classification. But the drawback of these fixed MIL pooling operators is clearly that they are pre-defined and unable to adapt to various specific tasks. To address this problem, an attention-based MIL pooling operator [11] is proposed, which can introduce trainable parameters to weight different instances and help obtain a more flexible and adaptive global representation. Yao et al. [12] combined MI-FCN and attention-based MIL pooling to perform the cancer survival prediction task. Lu et al. [13] selected representative regions according to their attention scores to train instance-level clustering layers to enhance the interpretability of the model. Inspired by attention-related works [11], the RMDL network proposed by Wang et al. [2] achieved image-level classification task by multi-scale feature fusion and recalibration of instance features according to the importance of each instance to the image label.

Nevertheless, most existing MIC algorithms including attention-related methods treat instances in each bag as independent and identically distributed without considering the spatial topological relationships of instances corresponding to the original bag [14]. The spatial topological relationships between local areas in a WSI were actually destroyed when the high resolution WSI was divided into patches. A graph is an important data structure suitable for processing unstructured data. By way of constructing a series of objects and their interrelations as nodes and edges, the graph structure has strong modeling representation for many real scenarios. In this paper, we designed an effective framework for high-resolution gastric pathological image classification tasks. We introduced graphs to model the spatial relationships between instances. With spatial topological structure information retained, the connections between nodes with similar features were constructed according to their distances in the feature space. We designed a graph convolutional network to solve this MIC problem. The node embeddings were fused hierarchically through the multi-layer graph convolution modules, and then a global representation containing spatial topological structure information and feature-space information between all instances was finally obtained for final prediction. The experimental results show that the proposed method achieved promising results on real gastric cancer pathological datasets compared with several existing MIC methods.

Overall, the main contributions of our work can be summarized as follows:

(1) We present an efficient framework composed of a feature extraction module based on ImageNet-trained CNNs and a multiple instance classification module based on GCNs for the classification task of gastric pathological WSIs;

(2) We construct the graph structure according to the similarity between the patch embeddings by implicitly fusing the information on their spatial topological structural relationships between instances. The proposed MIC module based on GCNs achieves information fusion in both physical space and feature space for all instances;

(3) We conduct experiments on two real high-resolution gastric pathological image datasets with different imaging mechanisms to prove the effectiveness and robustness of our proposed framework. To our knowledge, our work is the first to conduct experiments both on an H&E-stained pathological image dataset and a stimulated Raman scattering (SRS) microscope image dataset.

2. Background Knowledge

2.1. Multiple Instance Classification

Multiple instance learning (MIL) provides an elegant framework to address the weakly supervised learning problem with inexact supervision. Different from the general fully supervised problem that assumes each sample has a label, training samples $\mathrm{D} = \{(X_1, y_1), (X_2, y_2), \ldots, (X_m, y_m)\}$ in MIL are composed of bags, where each bag X_i bears a label y_i and consists of a set of instances $\{x_{i,1}, x_{i,2}, \ldots, x_{i,m_i}\}$. Bag X_i is assigned a positive label if there exists at least one positive instance, while it is assigned a negative label if it only contains negative instances. The goal of the classification task in MIL is to predict the labels for unseen bags, where bag labels are observed and instance labels are not observed in the training dataset.

$$y_i = \begin{cases} 1\ \exists x_{i,k}:\ x_{i,k} = 1 \\ 0\ \forall x_{i,k}:\ x_{i,k} = 0 \end{cases} \tag{1}$$

The MIC framework is naturally suitable for handling high-resolution pathological image classification tasks. For example, an alternative mapping is to regard the pathological images for malignant/benign patients as positive/negative bags and patches (or features of patches) extracted from the same image as instances in each bag. All the patches extracted from a pathological image with the negative label can be considered to contain benign tissue information. On the other hand, in a pathological image with the positive label, there must exist at least several patches containing malignant tissue information. The properties of pathological image classification are consistent with the assumption of the MIC framework.

2.2. Graph Convolutional Networks and Graph Classification

Inspired by the theories related to the graph convolution in graph signal processing, a set of graph neural networks (GNNs) based on graph convolution operations have been developed in recent years. Earlier methods based on spectral domain need to process the whole graph simultaneously and have high time complexity in matrix decomposition. Thus, recently, spatial graph convolution has been proposed and widely used. For the input/output graph signal matrix X/X', the graph convolution operation can be expressed as Equation (2), where θ are trainable filters, U is the orthogonal matrix composed of eigenvectors of the Laplacian matrix of the graph, and $\sigma(\cdot)$ represents nonlinear activation function.

$$X' = \sigma\Big(Udiag(\theta)U^T X\Big) = \sigma(\theta X) \tag{2}$$

Considering that the effective information in real graphs is usually contained in the low frequency band, a local filtering method was proposed [15]. It used the Chebyshev polynomial to approximate the graph filter coefficients and simplify the learnable parameters. The graph convolution layer is defined in Equation (3), where $\widetilde{L}_{sym}$ is symmetric normalized laplacian matrix. The calculation of $\widetilde{L}_{sym}X$ is equivalent to the first-order aggregation of the eigenvectors of neighboring nodes. Additionally, the parameterized weight matrix W is introduced to enhance the network fitting capability. Stacking multiple graph convolution layers can achieve the filtering ability of frequency response function in high order polynomial form.

$$X' = \sigma\Big(\widetilde{L}_{sym}XW\Big) \tag{3}$$

Specifically, tasks related to the graph data can be divided into node level, edge level and graph level. As an important graph level task, graph classification aims to learn a model from the graph to the corresponding label with consideration for the attributes of each node and the overall structure information of the graph. In graph classification tasks, the graph readout operation is usually adopted to generate the whole graph representation after the multi-layer message passing and state updating of each node. The standard approach based on the global pooling operator is to generate embedding for all nodes using a simple aggregation. However, common global pooling methods treat all the nodes equally, and thus are more suitable for handling small-scale graph data. Recently, many hierarchical pooling operators [16–18] have been developed that are suitable for more complex graph structures by compressing the information gradually to obtain the whole graph representation.

2.3. Differentiable Pooling

As one kind of hierarchical pooling method, differentiable pooling (DIFFPOOL) is a pooling mechanism based on graph collapse, which fuses the nodes in the same cluster to form the signal of a super-node in the next layer. Compared with global pooling methods, the DIFFPOOL method can capture the rich structure information better in various graph architectures. Real-world graphs tend to be more complex ones; hierarchical pooling methods can extract the complex hierarchical structure and generate a better modeling of the entire graph structure for downstream tasks.

An illustration of a DIFFPOOL module is shown in Figure 1. Nodes with the same color are divided into the same cluster. The graph collapses hierarchically into more and more sparse subgraphs, and eventually one super-node is formed that contains the whole graph information. Given the cluster assignment matrix $S^{(l)}$, adjacency matrix $A^{(l)}$ and node embedding matrix $Z^{(l)}$, a new coarsened adjacency matrix $A^{(l+1)}$ and a new matrix of embeddings $H^{(l+1)}$ for each of the nodes/clusters in the coarsened graph is generated as described in Equations (4) and (5), where n_{l+1} represents the number of nodes (or clusters) in the $l+1$ layer.

$$H^{(l+1)} = S^{(l)^T} Z^{(l)} \in R^{n_{l+1} \times d} \tag{4}$$

$$A^{(l+1)} = S^{(l)^T} A^{(l)} S^{(l)} \in R^{n_{l+1} \times n_{l+1}} \tag{5}$$

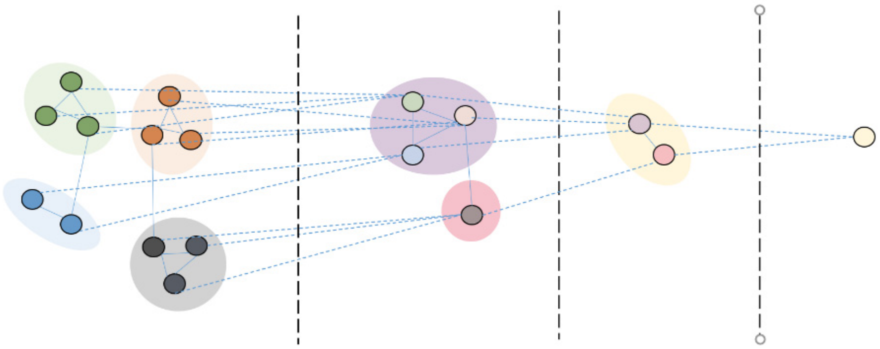

Figure 1. Differentiable pooling module.

Two independent GNN modules are used to generate $S^{(l)}$ and $Z^{(l)}$, as defined in Equations (6) and (7). Notice that two GNN modules input the same graph, but training parameters and execution functions are different. One GNN module is used to generate the node embedding, while the other GNN module is used to generate the probability that the input nodes assign to the n_{l+1} clusters. In the last DIFFPOOL module of the networks, the number of clusters is fixed as 1, and a final embedding is generated for downstream tasks.

$$Z^{(l)} = GNN_{l,embed}\left(A^{(l)}, H^{(l)}\right) \tag{6}$$

$$S^{(l)} = softmax\left(GNN_{l,pool}\left(A^{(l)}, H^{(l)}\right)\right) \tag{7}$$

3. Materials and Methods

3.1. Datasets

Our experiment was conducted on two different modes of gastric cancer pathological image datasets. The first dataset was sampled from the gastric cancer SRS image database, which was established in a joint project with Zhongshan Hospital and the Department of physics in Fudan University. As a novel label-free chemical imaging technique, many studies have proved the feasibility of the SRS microscope for imaging various pathological tissues. A total of 185 available gastric cancer SRS microscope images were obtained, including 90 benign tissue images and 95 malignant tissue images. Pixel resolution was 0.385 um/pixel. An example of gastric SRS microscope image is shown in Figure 2a.

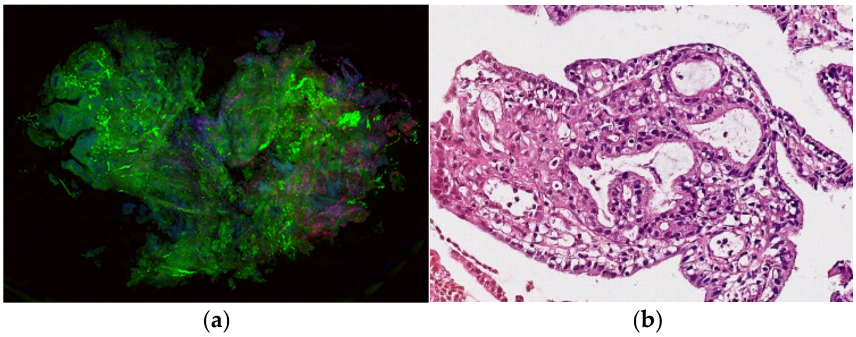

Figure 2. The example of a gastric cancer histopathology image. (**a**) SRS image. (**b**) H&E stained image.

The second dataset (Mars) was an H&E stained gastric cancer pathological image dataset, from the Jiangsu big data development and application competition on intelligent diagnosis of cancer risk. An example of H&E stained gastric pathological image in Mars dataset is shown in Figure 2b. There were 1032 positive samples and 1000 negative samples in the whole dataset. All images were obtained at 20× field of view. Positive samples were those in which cancer lesions existed in the image. Negative samples included normal gastric tissue images and gastritis tissue images. All data were desensitized in strict accordance with internationally accepted medical information desensitization standards to effectively guarantee data security and protect the privacy of patients.

3.2. Data Preprocessing

The datasets were split randomly into training, validation and test sets according to the proportion of 7:2:1. Firstly, we cropped the high-resolution pathological images into image patches of small size with 50% overlap. The whole structure of the tissue and the local characteristics of the nucleus should be taken into consideration when setting the size of image patches. For the SRS dataset, the original image was divided into patches with size of 450 × 450 pixels. For the Mars dataset, image patch size was set to 224 × 224 pixels.

Since there were background areas in the high-resolution pathological images that did not contain any tissue information, patches corresponding to those background areas were discarded. Specifically, original images were converted into binary images through the OSTU algorithm. The number of pixels containing tissues *num_tissue* was counted in the binary image patch of the same size and at the same position of the original image patch. The patches with the indicator *num_tissue* smaller than the threshold were filtered out. After the data preprocessing, the total numbers of available patches from the two datasets are shown in Table 1.

Table 1. Some indicators in data processing.

Dataset	WSI Number	Patch Resolution	Threshold	Patch Number
SRS	185	224 × 224 pixels	35,000 (17.3%)	68,387
Mars	2032	450 × 450 pixels	15,000 (29.9%)	396,539

3.3. Proposed Model of Multiple Instance Classification Based on GCNs

In the framework of multiple instance classification, the WSIs with the malignant/benign labels can be viewed as positive/negative bags. The patches containing malignant/benign tissue information can be regarded as positive/negative instances. In our experiment, all the image patches passed through CNNs for further feature extraction and dimension reduction. Thus, the instances in the MIC framework are replaced by patch features, which is illustrated in Figure 3.

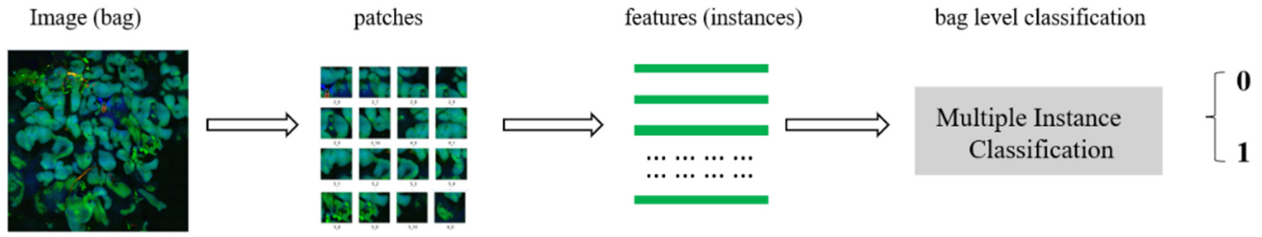

Figure 3. The flowchart of the multiple instance classification.

In order to solve this multiple instance classification problem from the perspective of graphs, we propose a novel algorithm based on GCNs for gastric cancer pathology image classification. The whole flowchart is illustrated in Figure 4. The first step is to map bag space into graph space, converting the multiple instance classification problem into a graph classification problem. An alternative heuristic strategy is: given a set of bags $[X_1, X_2, \ldots, X_N]$, each bag contains a different number of instances $\left[{x_1}^{(i)}, {x_2}^{(i)}, \ldots, {x_K}^{(i)}\right]$. We regard each bag in MIL as a graph, thus a bag of instances is converted into an undirected graph. The adjacency matrix can be derived with the following formula:

$$ {A_{mn}}^{(i)} = \begin{cases} 1\ if\ distance\left({x_m}^{(i)}, {x_n}^{(i)}\right) < \eta \\ 0\ otherwise \end{cases} \tag{8} $$

where distance $\left({x_m}^{(i)}, {x_n}^{(i)}\right)$ is the Euclidean distance between the m-th and n-th instance in i-th bag. The subscript represents the order of instances in the bag, and also implies that it corresponds to the spatial position relationships in the original image. η is the threshold to decide whether there is a connecting edge between two instances based on their distance. The task is converted to learn a graph classification model that maps from graph to the corresponding label with the node feature, edges and graph labels given.

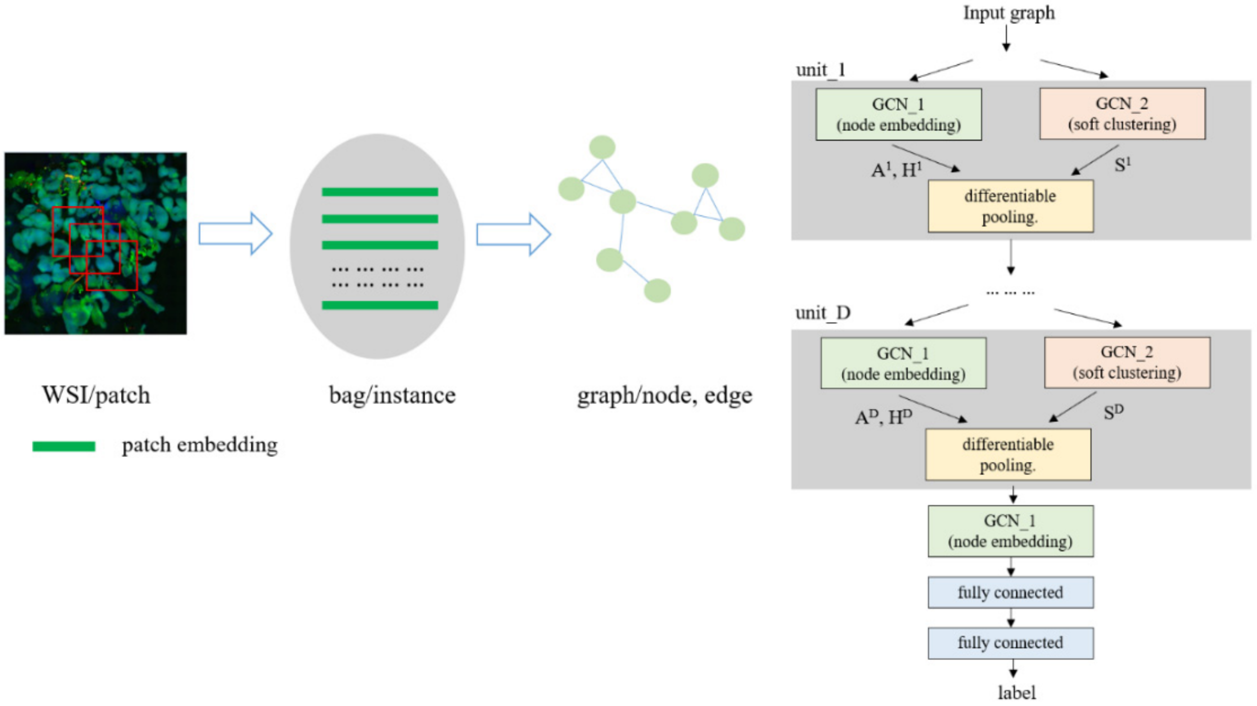

Figure 4. The flowchart of the proposed method.

After converting bags of instances into graphs, a classification model based on GCNs is constructed, as detailed in Figure 4. The network architecture is composed of stacks of units consisting of GNN modules and DIFFPOOL modules, one feature embedding GNN module and two fully connected layers. In the experiment, the GNN module was built on the GraphSAGE architecture [19]. Compared to a traditional GCN, GraphSAGE is a framework for inductive representation learning on large graphs. By sampling neighbor nodes randomly, it resolves the defect of poor flexibility and expansibility of traditional GCN and achieves the small batch distributed training of large-scale graph data. The node aggregation and embedding generation are illustrated in the following formula:

$$h_v^k = \sigma\left(W^k \cdot mean\left(h_u^{k-1},\ \forall u \,\epsilon\, \mathcal{N}(v) \cup \{v\}\right)\right) \tag{9}$$

where h_v^k is the representation of node v at the layer k and $\mathcal{N}(v)$ represents a set of neighbors of node v. $\sigma(\cdot)$ represents the nonlinear activation function; specifically, LeakyReLU [20] was adopted in the experiment. Additionally, the aggregation function adopts inductive aggregation.

The advantage of using a graph structure and graph model for the MIL task is the implicit expression of spatial topological relationships between instances in the process of constructing the adjacency matrix, which is transformed into the topological expression of graphs. The model based on GCNs can realize the feature fusion and transformation in the process of propagating node state, and simultaneously, hierarchically aggregates the structural information existing among nodes. Thus, a global representation containing feature information of all nodes and topology information of the whole graph can be generated for downstream tasks. The model adopts cross-entropy loss as loss function, as defined by Equation (10):

$$L(p(y_i|x), y_i) = -y_i \cdot log(p(y_i|x)) - (1 - y_i)log(1 - p(y_i|x)) \tag{10}$$

where y_i represents the true label of i-th graph, and $p(y_i|x)$ is its probability.

4. Results and Discussion

4.1. Experimental Environment and Setup

Our experimental platform consisted of a remote server with the Linux operating system, and the software and hardware environment included Python 3.6.9, CUDA 10.0, Pytorch-GPU 1.4.0, Pytorch-Geometric (PyG) 1.6.1 and a GeForce RTX 2080 Ti GPU. PyG is a geometric deep learning extension library based on Pytorch, which is suitable for handling graphs and other irregularly structured data. The construction and training of GNNs were implemented on PyG.

During the training process, the Adam optimization algorithm was used for parameter learning. The learning rate was initialized to 1×10^{-4}. The weight decay coefficient was set to 1×10^{-4}, and the batch size was set to 1. In order to avoid the loss curve oscillation during the later stage of training, a learning rate decay strategy was adopted with the loss on the training set as index. When this indicator did not decrease for two successive epochs, the learning rate decreased to half of the original. The maximum epoch number was set to 100.

All experiments adopted five folds and five runs of cross validation. The mean values of recall, precision and *F1_score* were utilized as evaluation indicators and their corresponding variance values were given. Recall rate reflects the probability that a positive sample is not missed. Precision rate reflects the probability that a positive sample is not wrongly classified. *F1_score* is the harmonic average of these as a comprehensive index. The detailed definition can be seen in the following formulas:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{12}$$

$$F1\_score = \frac{2}{\left(\frac{1}{\text{recall}}\right) + \left(\frac{1}{\text{precision}}\right)} \tag{13}$$

where *TP* represents the number of positive samples that were correctly classified. *FN* represents the number of negative samples that were wrongly classified, and *FP* represents the number of negative samples that were correctly classified.

4.2. Influence of Model Parameters

We investigated the performance under different settings of model parameters, including network depth D and the number of clusters C. A single GNN module contained a GraphSAGE layer, a batch normalization layer [21] and a LeakyReLU layer. One GNN module learned the representation of nodes in the graph; the other GNN module achieved the automatic cluster of nodes. A single layer of DIFFPOOL was added to integrate the nodes into the same cluster. Two GNN modules and a DIFFPOOL layer could be viewed as one unit as a whole. The network depth could be defined as the number of units stacked. Due to the low-pass filtering property of graph convolution, the stack of multiple graph convolution layers will lead to an over-smooth problem. D was chosen from {1, 2, 3, 4}. The clustering number of the next unit was selected as 25% of the previous unit. C was defined as the number of clusters in the last unit of the model. C was chosen from {2, 4, 6, 8, 10}.

Figure 5 shows that the classification performance tends to be saturated and slightly decreased with the increase in network depth D and number of clusters C. In the experiment, with the increase in D, the training time also presented an approximate linear growth.

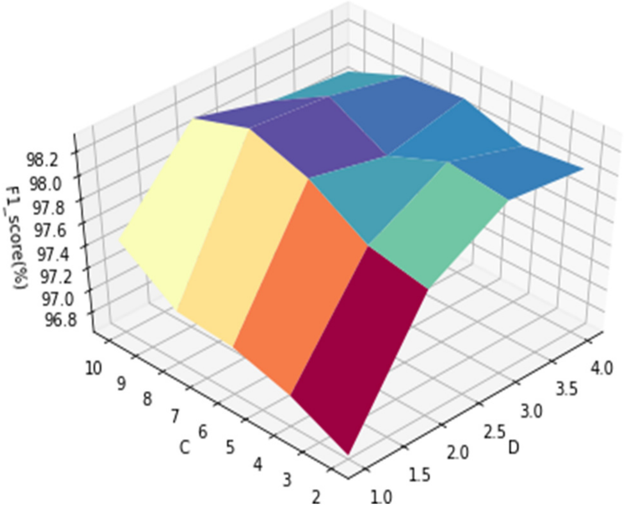

Figure 5. Performance with different model parameters.

4.3. Performance Comparison of Different Feature Extractors

The performance of six kinds of features was compared, and the experimental results are shown in Table 2. Model parameters were set as D = 2, C = 8. Experimental results were cross-validated by five folds and five runs. Average recall, precision, *F1_score* and their standard deviations were given as assessment indices. TX_fea was obtained by concatenating three typical traditional features: SIFT, LBP and the statistics of GLCM. VGG-16_fea [22], ResNet18_fea [23], DenseNet-121_fea [24] and EffientNet-B0_fea [25] were the activation vectors of four different ImageNet-trained CNNs. It can be observed that the features extracted by ImageNet-trained CNNs showed greater advantages than the traditional feature descriptors. This indicates that the features extracted by ImageNet-trained CNNs have better discrimination and integrity, and can be transferred to the pathological image analysis tasks.

Table 2. Performance of different features in Mars dataset.

Feature	Dimension	Recall (%)	Precision (%)	*F1_Score* (%)
TX_fea	408	94.78 ± 0.013	94.99 ± 0.013	94.59 ± 0.012
VGG-16_fea	4096	97.24 ± 0.012	97.24 ± 0.013	97.26 ± 0.012
ResNet-18_fea	512	98.14 ± 0.005	98.14 ± 0.005	98.16 ± 0.005
DenseNet-121_fea	1024	98.34 ± 0.009	98.34 ± 0.009	98.35 ± 0.008
EfficientNet-B0_fea	1280	97.69 ± 0.010	97.69 ± 0.010	97.70 ± 0.010

4.4. Comparisons with Other Multiple Instance Methods

To demonstrate the effectiveness of the proposed method, the algorithm was tested and compared with existing MIC methods on SRS and Mars datasets, respectively. The flowchart of different MIC methods is shown in Figure 6. To facilitate subsequent comparisons, all available patches were passed through the ImageNet-pretrained ResNet-18 model to obtain 512-d feature vectors. In the Embedded-Space paradigm MIC method, the crucial step was to fuse these instance level embeddings to obtain a bag level embedding that was suitable for applying to downstream tasks.

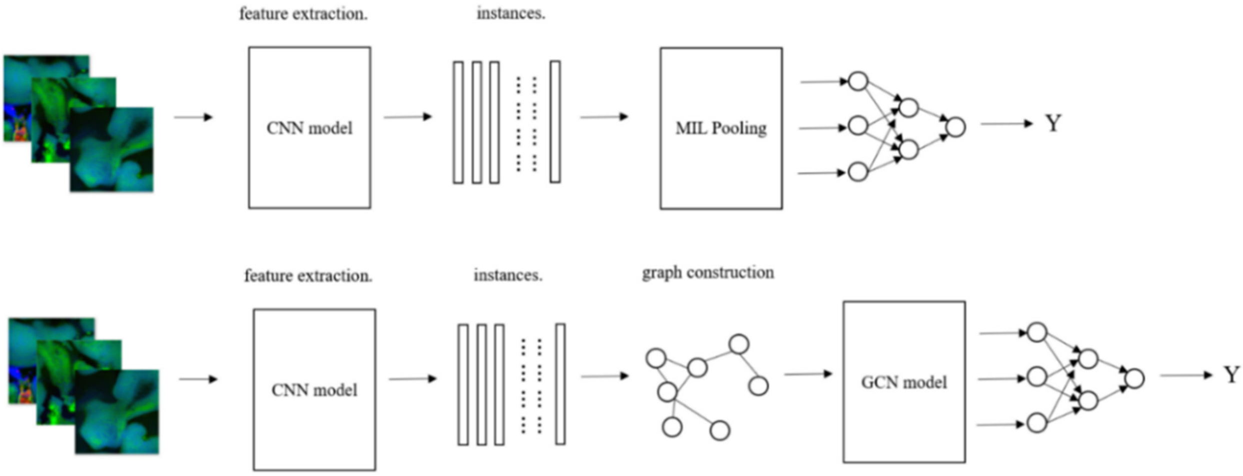

Figure 6. Flowchart of the different multiple instance classification methods.

In Tables 3 and 4, MIP (mean/max) represents utilizing common fixed MIL pooling operators. Bag level embedding was generated by operating mean or max pooling on each dimension for all instances in the same bag, as seen in Equations (14) and (15). Similarly, MIP (attention) represented utilizing an attention-based MIL pooling operator, which introduced trainable parameters **w** and **v** to allocate weights automatically for each instance by end-to-end training, as seen in Equation (16). Besides these classic MIP operators, we further compared with the RMDL network proposed by Wang et al. [2], which consists of local-global feature fusion, instance recalibration and multi-instance pooling modules. After the MIL pooling layer, the bag embedding was sent into fully connected layers for final classification. Compared with fixed ones, the attention-based MIL pooling method can be more flexible. It provides information about the contribution of each instance to the final prediction, which can help to find key instances. However, due to the scarcity of patch label information, extra errors may be introduced if the classifier cannot be trained sufficiently, which was confirmed in our experimental results. On the Mars dataset with more data, the performance of the attention-based MIL pooling method was 2% worse than that of the fixed max pooling method. This suggests that with the increase in the number of instances, the classifier becomes harder to train, resulting in the final performance degradation. Compared with the attention-based MIL pooling method, the RMDL method achieved almost the same performance on the SRS dataset and a slight improvement on the Mars dataset. Essentially, the RMDL method also adaptively assigned weights to instances by end-to-end training and fused instance features according to their scores. For these various existing methods, topological structural information is not considered to integrate into the final bag-level embedding in the process of instance feature fusion.

$$\mathrm{MIP(max)}: \ \mathbf{z} = \max\left(\mathbf{h_j}\right) \ j = 1, 2, \ldots, K \tag{14}$$

$$\mathrm{MIP(mean)}: \ \mathbf{z} = \frac{1}{K}\sum_{j=1}^{K} \mathbf{h_j} \ j = 1, 2, \ldots, K \tag{15}$$

$$\mathrm{MIP(attention)}: \ \mathbf{z} = \sum_{j=1}^{K} a_j \mathbf{h_j} \ a_j = \frac{\exp\left\{\mathbf{w}^{\mathrm{T}} \tan\mathrm{h}\left(\mathbf{V}\mathbf{h}_{\mathbf{j}}^{\mathrm{T}}\right)\right\}}{\sum_{i=1}^{K} \exp\left\{\mathbf{w}^{\mathrm{T}} \tan\mathrm{h}\left(\mathbf{V}\mathbf{h}_{\mathbf{i}}^{\mathrm{T}}\right)\right\}} \tag{16}$$

Table 3. Performance of different MIL methods on SRS dataset.

Method	Recall (%)	Precision (%)	*F1_Score* (%)
MIP (mean)	82.10 ± 0.051	82.70 ± 0.051	81.83 ± 0.051
MIP (max)	83.65 ± 0.044	84.82 ± 0.041	83.81 ± 0.046
MIP (attention)	84.37 ± 0.042	85.34 ± 0.038	84.47 ± 0.041
RMDL	84.47 ± 0.044	85.29 ± 0.042	84.23 ± 0.045
GCN (mean_pool)	89.08 ± 0.038	89.88 ± 0.033	89.67 ± 0.033
GCN (max_pool)	89.13 ± 0.046	90.20 ± 0.042	88.96 ± 0.049
GCN + DIFFPOOL	90.40 ± 0.032	91.16 ± 0.032	90.75 ± 0.034

Table 4. Performance of different MIL methods on Mars dataset.

Method	Recall (%)	Precision (%)	*F1_Score* (%)
MIP (mean)	92.40 ± 0.013	92.45 ± 0.013	92.40 ± 0.014
MIP (max)	95.48 ± 0.011	95.52 ± 0.011	95.45 ± 0.011
MIP (attention)	93.11 ± 0.009	93.15 ± 0.009	93.13 ± 0.009
RMDL	93.76 ± 0.008	93.81 ± 0.008	93.74 ± 0.009
GCN(mean_pool)	95.81 ± 0.008	95.83 ± 0.008	95.81 ± 0.008
GCN (max_pool)	97.70 ± 0.007	96.73 ± 0.007	97.70 ± 0.007
GCN + DIFFPOOL	98.24 ± 0.004	98.26 ± 0.004	98.24 ± 0.004

As mentioned in Section 2.2, pooling methods in the graph classification task include standard global pooling and hierarchical pooling methods. In order to compare with the GCN model with the standard global pooling method, extra experiments were conducted. As shown in Tables 3 and 4, GCN (mean_pool/max_pool) and GCN+DIFFPOOL methods refer to introducing a graph structure to solve the histopathological image classification problem. The former is based on a graph classification model with global pooling, which consists of three GCN layers and one global pooling layer, specifically. The latter adopts the network architecture proposed in the previous experiments, which is based on a graph classification model with DIFFPOOL module, one hierarchical pooling method. Similarly, fully connected layers were added for final bag level classification.

Experimental results show that classification performance of the proposed method has been evaluated. The proposed graph-based method showed superior performance compared with several MIL methods on two different gastric cancer pathological image datasets. Compared to the RMDL method, the graph-based method (GCN+DIFFPOOL) showed at least an increase of 4.5% in the final classification indicators. The proposed GCN+DIFFPOOL method achieved an average *F1_score* of 90.75% and 98.24% on the SRS and Mars datasets, respectively. When applying MIL to the pathological image analysis tasks, the overlapping patches were viewed as instances in bags. Therefore, there must exist underlying structural information between instances in the same bag. The structure of the graph can model the relationships between the instances, and the GNNs can capture this structure information better in a non-independent, identically distributed MIL problem. Compared with the general MIL method, MIL based on graphs showed better results on the different modes of gastric cancer pathological image datasets, which further demonstrates the feasibility and versatility of the proposed method.

To validate the introduction of spatial topological relationships between instances in the process of graph construction, extra experiments were conducted by randomly changing the corresponding relationships between the order of instances and the subscript of adjacency matrix $A_{N \times N}$. Experiment results are shown in Table 5. After randomly shuffling the order of instances in the process of graph construction, the *F1_score* indicator decreased by 2.4% and 1.7% on the SRS and Mars datasets, respectively. Since the order of instances reflects the spatial position relations of patches in the original image, the corresponding relationships between the order of instances and the subscript of adjacency matrix $A_{N \times N}$ were randomly changed, which was equivalent to destruction of the topological relationships between instances. The experiment reversibly verified that the proposed graph-based MIC model implicitly incorporates the spatial topological structure information between the instances when obtaining the global representation of the bag, thus leading to the improvement in overall classification performance.

Table 5. Validation experiments for spatial topological structure information.

Dataset	**Shuffle**	**Recall (%)**	**Precision (%)**	***F1_Score* (%)**
SRS	Before	90.40 ± 0.032	91.16 ± 0.032	90.75 ± 0.034
	After	88.86 ± 0.036	88.30 ± 0.032	88.33 ± 0.035
Mars	Before	98.24 ± 0.004	98.26 ± 0.004	98.24 ± 0.004
	After	96.54 ± 0.005	96.59 ± 0.005	96.51 ± 0.005

5. Conclusions

In this paper, we proposed a multiple instance classification framework for gastric cancer pathological image classification. The proposed framework was composed of ImageNet-pretrained CNNs and GCNs. The former was utilized to extract patch features in the light of its superior ability in tackling diverse datasets and obtaining effective global features. The latter achieved information fusion in both physical space and feature space for all instances in the same bag. The features extracted by ImageNet-pretrained CNNs had better representation and generalization ability compared with traditional operators. The GCNs with hierarchical pooling module can fuse node features hierarchically and implicitly combine the spatial topological information of instances into the global representation of the bag. The framework can effectively integrate the spatial topological relationships between patches, alleviating the problem that some existing multiple instance classification methods cannot take advantage of the structural information between instances. The experimental results tested on two real gastric cancer pathology image datasets showed that graph-based methods achieved superior performances compared with several currently available MIL approaches.

However, our method had limitations in providing instance-level predictions which is necessary for these classifiers to be translated into clinical practice. We envisage incorporating the instance-level prediction module into the whole framework while maintaining the superior classification performance of the proposed model in the future. Instance-level evaluation could further help specialists locate the areas crucial to final predictions, and enhance the interpretability of models based on deep learning algorithms.

Author Contributions: Conceptualization, X.X.; methodology, X.X.; software, X.X.; validation, X.X.; formal analysis, X.X.; data curation, X.X.; writing—original draft preparation, X.X.; writing—review and editing, X.X. and X.W.; supervision, X.W.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sharma, H.; Zerbe, N.; Klempert, I.; Hellwich, O.; Hufnagl, P. Deep Convolutional Neural Networks for Automatic Classification of Gastric Carcinoma Using Whole Slide Images in Digital Histopathology. *Comput. Med Imaging Graph.* **2017**, *61*, 2–13. [CrossRef] [PubMed]
2. Wang, S.; Zhu, Y.; Yu, L.; Chen, H.; Lin, H.; Wan, X.; Fan, X.; Heng, P.-A. RMDL: Recalibrated Multi-Instance Deep Learning for Whole Slide Gastric Image Classification. *Med. Image Anal.* **2019**, *58*, 101549. [CrossRef] [PubMed]
3. Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F.A.; Brendel, W. ImageNet-trained CNNs are Biased Towards Texture; Increasing Shape Bias Improves Accuracy and Robustness. *arXiv* **2018**, arXiv:1811.12231.
4. Komura, D.; Ishikawa, S. Machine Learning Approaches for Pathologic Diagnosis. *Virchows Arch.* **2019**, *475*, 131–138. [CrossRef] [PubMed]
5. Saxena, S.; Shukla, S.; Gyanchandani, M. Pre-Trained Convolutional Neural Networks as Feature Extractors for Diagnosis of Breast Cancer Using Histopathology. *Int. J. Imaging Syst. Technol.* **2020**, *30*, 577–591. [CrossRef]
6. Gupta, K.; Chawla, N. Analysis of Histopathological Images for Prediction of Breast Cancer Using Traditional Classifiers with Pre-Trained CNN. *Procedia Comput. Sci.* **2020**, *167*, 878–889. [CrossRef]
7. Zhou, Z.-H. A Brief Introduction to Weakly Supervised Learning. *Natl. Sci. Rev.* **2018**, *5*, 44–53. [CrossRef]
8. Amores, J. Multiple Instance Classification: Review, Taxonomy and Comparative Study. *Artif. Intell.* **2013**, *201*, 81–105. [CrossRef]
9. Conjeti, S.; Paschali, M.; Katouzian, A.; Navab, N. Deep Multiple Instance Hashing for Scalable Medical Image Retrieval. In Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI), Quebec City, QC, Canada, 10–14 September 2017; pp. 550–558.
10. Das, K.; Conjeti, S.; Roy, A.G.; Chatterjee, J.; Sheet, D. Multiple Instance Learning of Deep Convolutional Neural Networks for Breast Histopathology whole Slide Classification. In Proceedings of the IEEE 15th International Symposium on Biomedical Imaging (ISBI), Washington, DC, USA, 4–7 April 2018; pp. 578–581. [CrossRef]
11. Ilse, M.; Tomczak, J.M.; Welling, M. Attention-based Deep Multiple Instance Learning. *arXiv* **2018**, arXiv:1802.04712.
12. Yao, J.; Zhu, X.; Jonnagaddala, J.; Hawkins, N.; Huang, J. Whole Slide Images Based Cancer Survival Prediction Using Attention Guided Deep Multiple Instance Learning Networks. *Med. Image Anal.* **2020**, *65*, 101789. [CrossRef] [PubMed]
13. Lu, M.Y.; Williamson, D.F.K.; Chen, T.Y.; Chen, R.J.; Barbieri, M.; Mahmood, F. Data-Efficient and Weakly Supervised Computational Pathology on Whole-Slide Images. *Nat. Biomed. Eng.* **2021**, *5*, 555–570. [CrossRef] [PubMed]
14. Carbonneau, M.-A.; Cheplygina, V.; Granger, E.; Gagnon, G. Multiple Instance Learning: A Survey of Problem Characteristics and Applications. *Pattern Recognit.* **2018**, *77*, 329–353. [CrossRef]
15. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv* **2017**, arXiv:1609.02907.
16. Ying, R.; You, J.; Morris, C.; Ren, X.; Hamilton, W.L.; Leskovec, J. Hierarchical Graph Representation Learning with Differentiable Pooling. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Montreal, MT, Canada, 3–8 December 2018; pp. 4805–4815.
17. Cangea, C.; Velickovic, P.; Jovanovic, N.; Kipf, T.; Lio, P. Towards Sparse Hierarchical Graph Classifiers. *arXiv* **2018**, arXiv:1811.01287.
18. Diehl, F. Edge Contraction Pooling for Graph Neural Networks. *arXiv* **2019**, arXiv:1905.10990.
19. Hamilton, W.L.; Ying, R.; Leskovec, J. Inductive Representation Learning on Large Graphs. In Proceedings of the in Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 1025–1035.
20. Xu, B.; Wang, N.; Chen, T.; Li, M. Empirical Evaluation of Rectified Activations in Convolutional Network. *arXiv* **2015**, arXiv:1505.00853.
21. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the 32nd International Conference on International Conference on Machine Learning, Lille, France, 6 July 2015; Volume 35, pp. 448–456.
22. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
23. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
24. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
25. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the International Conference on Machine Learning (ICML), Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.