Predicting Lymph Node Metastasis Using Histopathological Images Based on Multiple Instance Learning with Deep Graph Convolution

Yu Zhao[1,2*], Fan Yang[1*], Yuqi Fang[3], Hailing Liu[4], Niyun Zhou[1], Jun Zhang[1], Jiarui Sun[1], Sen Yang[1], Bjoern Menze[2], Xinjuan Fan[4] and Jianhua Yao[1]
[1]Tencent AI Lab, [2]Technical University of Munich, [3]The Chinese University of Hong Kong
[4]Sixth Affiliated Hospital of Sun Yat-sen University

Abstract

Multiple instance learning (MIL) is a typical weakly-supervised learning method where the label is associated with a bag of instances instead of a single instance. Despite extensive research over past years, effectively deploying MIL remains an open and challenging problem, especially when the commonly assumed standard multiple instance (SMI) assumption is not satisfied. In this paper, we propose a multiple instance learning method based on deep graph convolutional network and feature selection (FS-GCN-MIL) for histopathological image classification. The proposed method consists of three components, including instance-level feature extraction, instance-level feature selection, and bag-level classification. We develop a self-supervised learning mechanism to train the feature extractor based on a combination model of variational autoencoder and generative adversarial network (VAE-GAN). Additionally, we propose a novel instance-level feature selection method to select the discriminative instance features. Furthermore, we employ a graph convolutional network (GCN) for learning the bag-level representation and then performing the classification. We apply the proposed method in the prediction of lymph node metastasis using histopathological images of colorectal cancer. Experimental results demonstrate that the proposed method achieves superior performance compared to the state-of-the-art methods.

1. Introduction

Recently, weakly-supervised learning (WSL) has gained greater attention in the machine learning field since it significantly reduces the workload of human annotation. Multi-instance learning (MIL) is a typical weakly-supervised learning [48], which has been widely employed in different tasks, including object detection [37, 38, 18], semantic segmentation [43, 33], scene classification [42, 19], medical diagnosis[5, 31], etc. In the MIL task, the training dataset is composed of bags, where each bag contains a set of instances. The goal of MIL is to learn a model for predicting the bag label. Different from conventional fully-supervised machine learning problems, where each instance has a confident label, only the bag-level label is available in MIL. Furthermore, instances in a bag are not necessarily relevant, sometimes even providing confusing information. For example, some instances do not contain discriminative information related to the bag class, or they are more related to other classes of bags.[2]

Based on which level the discriminative information is at (instance-level or bag-level) and how the relevant information is extracted (implicitly or explicitly), MIL algorithms can be categorized into three groups, i.e., instance-space paradigm, bag-space paradigm, and embedded-space paradigm [41, 2]. The instance-space paradigm tends to focus on local information, which learns instance classifier at the first stage and then achieves the bag-level classifier by simply aggregating instance-level results. These instance-space methods are mostly based on the standard multiple instance (SMI) assumption [27], i.e., a bag is positive only if it contains at least one positive instance and otherwise is negative [46, 3, 30]. However, this key-instance based SMI assumption is inappropriate in applications where the classification is based on the global bag information instead of an individual instance. The bag-space paradigm and embedded-space paradigm, on the other hand, extract discriminative information from the whole bag. The difference between these two paradigms lies in the way to exploit the bag-level information. The bag-space paradigm implicitly utilizes bag-to-bag distance/similarity, while the embedded-space paradigm explicitly embeds the information of a bag into a feature space [41].

In this paper, we propose a novel embedded-space multiple instance learning method with feature selection and graph convolutional network for image classification. The method has three major components: instance-level feature

*Yu Zhao and Fan Yang contributed equally and Jianhua Yao is the corresponding author (jianhuayao@tencent.com).

extraction, instance-level feature selection, and bag-level classification. Our method is developed for the prediction of lymph node metastasis using histopathological images of colorectal cancer. Lymph node metastasis (LNM) from colorectal cancer is a major factor in patient management and prognosis [13, 9, 39]. Patients diagnosed with LNM should undergo lymph node dissection surrounding the colon region [7]. This research has great clinical value since LNM pre-surgical detection indicates the necessity of lymph node dissection to prevent further spreading. This is a challenging task and we tackle it in the following aspects. (1) The size of a whole slide image (WSI) is usually very large (around 100000×50000 pixels in our case). Given the current computational resource, it is infeasible to load the WSI into the deep neural networks. Therefore we divide the WSI into a set of image patches (512×512 pixels) and treat the WSI as a bag of patches. In this way, the prediction problem is formalized as an embedded-space multiple instance learning task. (2) To the best of our knowledge, there is no prior work or knowledge that indicates useful features for metastasis prediction. Similar as other image classification works, we employ deep neural networks to automatically extract latent features from the image [26]. Moreover, in our case where the instance labels are not available, conventional methods usually utilize a pre-trained model (e.g., trained on ImageNet) as the feature extractor. However, the domain gap between natural scene images and histopathological images may compromise the performance of the pre-trained model on histopathological images [34]. To solve this problem, we develop a combination model of variational autoencoder and generative adversarial network (VAE-GAN) to train the feature extractor in a self-supervised way. (3) Generating effective representation for all instances in a bag is not a trivial problem. Various methods such as max pooling, average pooling, log-sum-exp pooling [41], dynamic pooling [44], and adaptive pooing [29] have been proposed. However, these operators are either non-trainable or too simple. In this paper, we apply the graph convolutional network (GCN) for generating the bag representation, which is fully trainable. (4) Features extracted from the instances are redundant. We propose a novel feature selection approach to remove the indiscriminative instance-level features.

To summarize, the contributions of this paper include:

1) We propose an embedded-space deep multiple instance learning method with GCN for the prediction of lymph node metastasis in colorectal cancer on histopathological images. To the best of our knowledge, this is the first method tackling this challenging clinical problem.

2) We design a VAE-GAN model to generate instance and use the trained encoder component of the VAE-GAN as the feature extractor. With this setting, we can train the feature extractor in a self-supervised way, without knowing the instance label.

3) We develop a novel feature selection approach working on instances to select discriminative features for final bag representation. The proposed method utilizes a histogram to build a bag-level representation of this feature and then use the maximum mean discrepancy (MMD) [14, 40] of the obtained bag-level representation between positive and negative bags to evaluate the feature significance.

4) We apply a GCN for generating the bag representation and bag-level classification, which is fully trainable.

2. Related Work

2.1. Deep Multiple Instance Learning

Combined with deep features, multiple instance learning has shown great representation power in recent studies [42, 36, 41, 6, 16, 17]. Wu *et al*. [42] utilized a MIL neural network to simultaneously learn the object proposals and text annotation. Sun el al. [36] proposed a weakly supervised MIL network for object recognition on natural images, which solved the inaccurate instance label problem in data augmentation. Hou *et al*. [17] showed that a decision classifier based on MIL can boost the performance in classifying glioma and non-small-cell lung carcinoma by aggregating instance-level predictions. In this paper, we follow this line of research and employ MIL to solve a challenging clinical problem of predicting lymph node metastasis from histopathological images of the primary tumor region.

2.2. Bag Representation

In a MIL task, generating bag representation is a crucial step. Pooling methods such as max pooling, average pooling, and log-sum-exp pooling [41] are typically adopted in this step. However, these pooling methods are not trainable which may limit their applicability. A novel dynamic pooling method iteratively renewing bag information from instance was proposed in [44]. Kraus *et al*. [24] proposed a noisy-and pooling layer against outliers, which demonstrated promising results in microscopy images. Zhou *et al*. [47] proposed an adaptive pooling method that can be dynamically adjusted to various classes in video caption tagging. These methods are partly trainable with restricted flexibility. Ilse *et al*. [19] proposed a fully trainable method utilizing the attention mechanism to allocate weights to instances. However, this method considers the bag representation as a weighted sum of instance features, which is just a linear combination. Different from the approaches mentioned above, our work builds the bag representation with

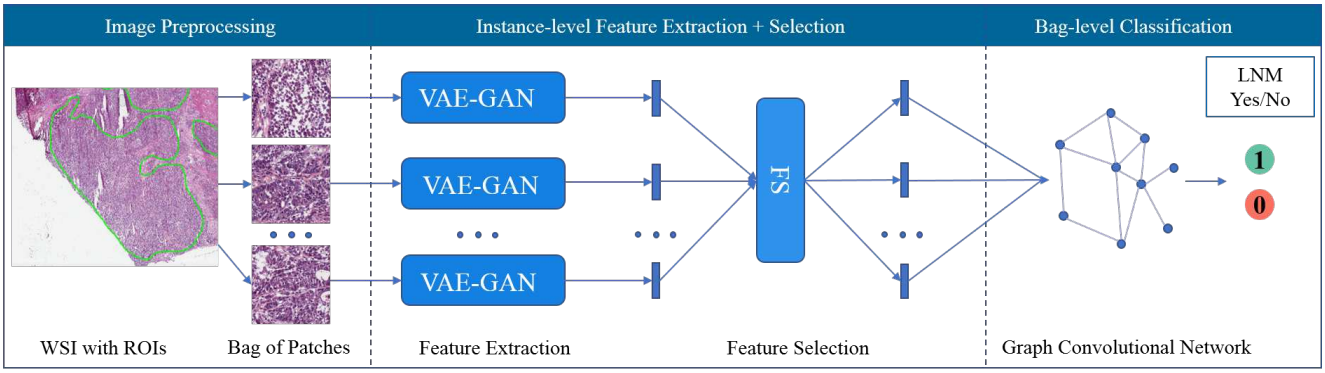


Figure 1. The overall framework of the proposed approach. It consists of image preprocessing, instance-level feature extraction, instance-level feature selection, and bag-level classification. VAE-GAN works as an instance-level feature extractor. The feature selection procedure selects discriminative instance-level features. The GCN is used to synthesize selected instance-level features, generate bag representation and perform the final classification.

a GCN, which is fully trainable and integrates the patch information into a complex and high-level representation.

2.3. Pathological Image Analysis

In clinical practice, pathology image analysis is the gold standard for cancer diagnosis. Nowadays, the development of deep neural networks has made many breakthroughs in automatic pathology image analysis and assisted diagnosis [50, 5, 21]. As mentioned above, MIL is naturally suitable for pathological image analysis due to the vast image. Authors in [21] recently reported an instance-space MIL method applied in the prediction of microsatellite instability (MSI). In the first step of this method, they assigned the bag label to its patches and then trained a ResNet with the patch-label pairs. In the second stage, the trained ResNet was used to generate the patch-level prediction and then all these patch-level prediction results are aggregated with the majority voting strategy. Another instance-space approach is the Whole Slide Histopathological Images Survival Analysis framework (WSISA) [50] for survival prediction. This method first unsupervisely clustered patches into different clusters and then selected useful clusters by evaluating patch-level classification performance. After that, it aggregated patch-level features from selected clusters to make the patient-level prediction. The success of the above two methods implies that these tasks meet the SMI assumption. However, these instance-space MIL methods have their limitations. They are suitable for the tasks where discriminative information is considered to lie at the instance level and there exist key instances whose labels are strongly related to bag-level labels. The above conditions do not hold in our problem.

Recently, Campanella *et al*. [5] proposed an embedded-space MIL method with applying a recurrent neural network (RNN) to integrate patch information extracted from the WSI. They treated each WSI as a bag and considered all the patches of the WSI as sequential inputs to the RNN. This model was trained on three huge datasets and successfully classified the sub-types of three cancers. Meanwhile, attention-based deep multiple instance learning is another currently proposed embedded-space MIL method [19]. It achieved the state-of-the-art performance on classifying epithelial cells in colon cancer by training on large-scale data. These two methods utilize different approaches to integrate instance information for bag representation, which are both end-to-end trainable. However, the end-to-end method requires the network to extract the instance features and generate bag representation simultaneously with only bag-level classification error as supervision, which makes the network hard to train, especially when lacking sufficient training data. Therefore, we propose a feature selection component in our method to remove the redundant and unhelpful features to alleviate the workload of the network for generating the bag representation and performing the bag-level classification. Considering the specialty of our problem, we also equip our MIL method with GCN to take advantage of the structure information among instances.

3. Methods

3.1. Overview

The overall framework of the proposed method is illustrated in Fig. 1. The whole pipeline consists of four steps: image preprocessing, instance-level feature extraction with VAE-GAN (3.2), instance-level feature selection (3.3), and bag-level classification with graph convolutional network (3.4). In the image preprocessing step, the tumor areas manually annotated by pathologists are selected as the regions of interest (ROIs) referring to the manually annotated labels obtained from clinical experts in our team. Then, the ROIs are divided into non-overlapping patches of size 512×512. The details of the other three components are illustrated in

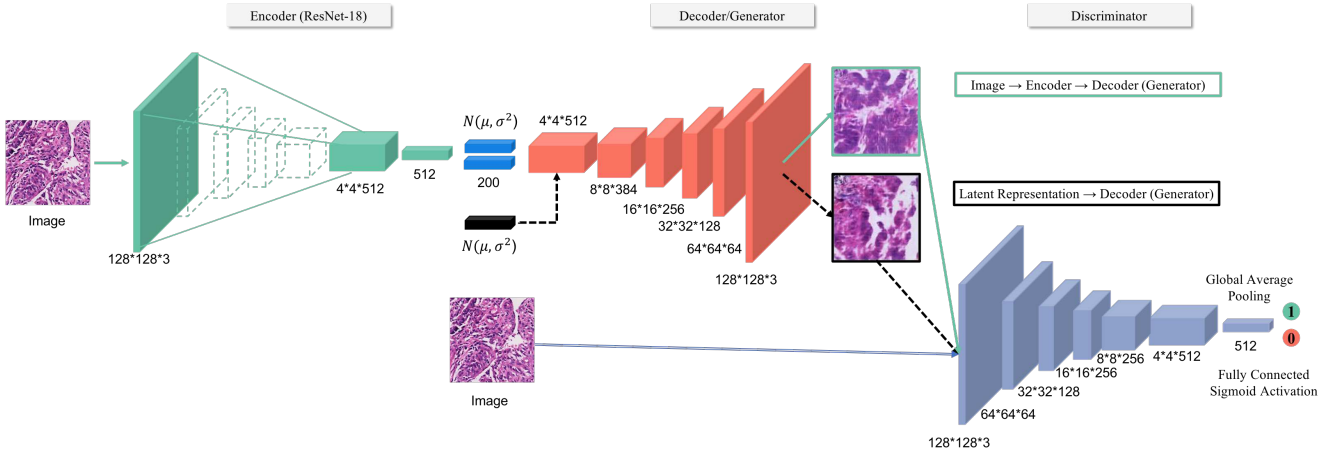


Figure 2. The architecture of the VAE-GAN. The ResNet-18 is used as the encoder of the VAE. The decoder of VAE and generator of GAN share the same component in VAE-GAN.

the remainder of this section.

3.2. VAE-GAN

A variational autoencoder (VAE) [23] is comprised of an encoder which encodes input data x to a latent representation h, and a decoder which decodes the latent representation h back to the original data-space. In order to regularize the encoder of the VAE, a prior over the latent distribution $p(h)$ is usually imposed. In this work, we use the Gaussian distribution, i.e. $N \sim (0, I)$ to regularize the encoder. The VAE loss [23, 25] is formulated as:

$$\begin{aligned}\mathscr{L}_{VAE} &= \mathscr{L}^{pixel}_{LLike} + \mathscr{L}_{KL} \\ &= -E_{q(h|x)}[log(p(x|h))] + D_{KL}(q(h|x)||p(h)),\end{aligned} \tag{1}$$

where D_{KL} denotes the Kullback-Leibler divergence.

A generative adversarial network (GAN) consists of a generator network G which aims to map the latent representation h to data space, and a discriminator network D which aims to distinguish the generated fake data from the real data. The loss of GAN is defined as:

$$\mathscr{L}_{GAN} = log(D(x)) + log(1 - D(G(h))) \tag{2}$$

A VAE-GAN is a combination of a VAE and a GAN, where the decoder of VAE and generator of GAN share the same component [25]. In a VAE-GAN architecture, the original VAE reconstruction error $\mathscr{L}^{pixel}_{LLike}$ is replaced by the reconstruction error expressed in the GAN discriminator. To be specific, let $Dis_l(x)$ denote the hidden representation of the l^{th} layer of the discriminator. A Gaussian observation model for $Dis_l(x)$ with mean $Dis_l(\tilde{x})$ and identity covariance is introduced:

$$p(Dis_l(x)||h) = N(Dis_l(x)|Dis_l(\tilde{x}), I) \tag{3}$$

where $\tilde{x} \sim Decoder(h)$ is the sample from the decoder of x, then the reconstruction error in the GAN discriminator can be denoted as follows:

$$\mathscr{L}^{Dis_l}_{LLike} = -E_{q(h|x)}[logp(Dis_l(x)|h)] \tag{4}$$

Using $\mathscr{L}^{Dis_l}_{LLike}$ replacing $\mathscr{L}^{pixel}_{LLike}$, we can obtain the loss function of entire VAE-GAN [25]:

$$\mathscr{L} = \lambda_{Dis} * \mathscr{L}^{Dis_l}_{LLike} + \lambda_{KL} * \mathscr{L}_{KL} + \lambda_{GAN} * \mathscr{L}_{GAN} \tag{5}$$

where λ_{Dis}, λ_{KL} and λ_{GAN} are the hyperparamters of the VAE-GAN loss.

Different from the conventional goal of GANs, the main function of the VAE-GAN in our work is for training or fine-tuning the encoder component which will be used as an instance-level feature extractor. The detailed architecture of our VAE-GAN can be found in Fig. 2. The widely utilized ResNet-18 [15] acts as the encoder. The decoder of VAE and generator of GAN share the same component, which incorporates five up-sampling stacks. Each up-sampling stack contains a transposed convolution followed by batch normalization and the rectified linear unit (ReLU) as the activation function. The discriminator is made up of five down-sampling stacks. In each encoder stack, there is one convolution layer followed by the batch normalization and LeakyReLU activation function.

3.3. Feature Selection

The feature selection procedure chooses the most discriminative instance-level features for generating the bag representation. This step is especially important in medical image analysis tasks due to lack of training data. Removing redundant or irrelevant features can also alleviate the workload and simplify the following learning task. Unlike most feature selection problems where there are feature-label pairs, in our task, the instance-level feature has no

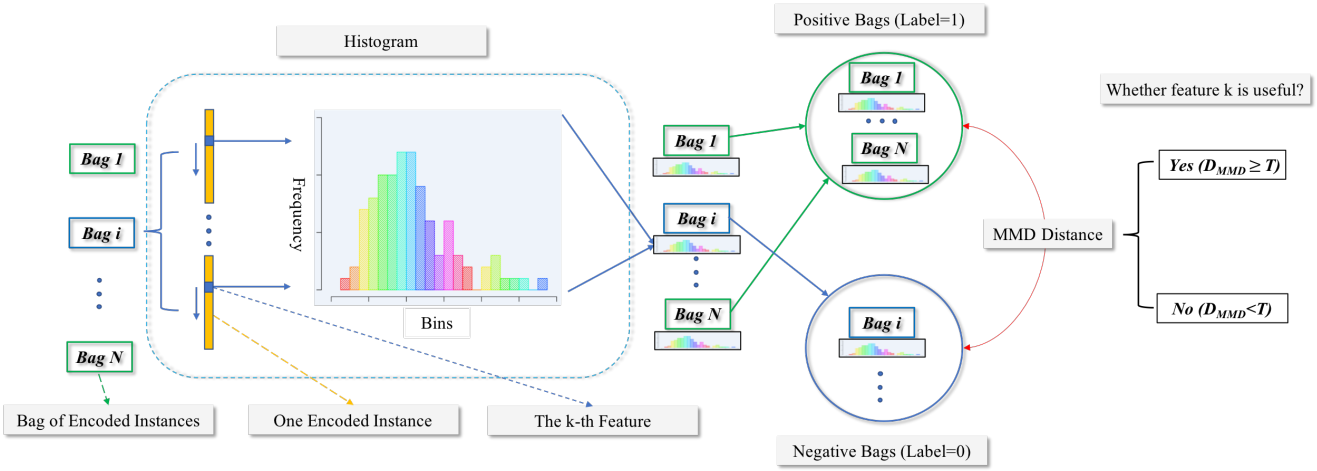


Figure 3. The pipeline of the feature selection component. Each bag has a various number of instances (features vectors). When evaluating the discriminating value of a feature for bag-level classification, the histogram acts as a bridge connecting the instance-level feature and the bag-level label. For instance, when evaluating the k-th feature, feature k is chosen as the representation of the instance. A histogram of this feature is calculated in each bag and then these histograms are utilized as the representations of a bag. After that, The MMD distance are calculated between positive bags and negative bags using these bag representations. The feature is regarded as discriminative if the MMD distance is large.

associated label and is just assigned a bag-level label. We need to build a bridge between the extracted instance feature and the bag label. As demonstrated in Fig. 3, we utilize the histogram [2] as the bridge and the maximum mean discrepancy [4] as the criterion to evaluate the feature importance.

Assume we have N bag-label pairs denoted as $\{X_1, X_2, ..., X_N\}$ and $\{Y_1, Y_2, ..., Y_N\}$, where the i^{th} bag contains K_i instances represented as $\{x_1^i, x_2^i, ..., x_{K_i}^i\}$. In our case, $x_j^i \in \mathbb{R}^D$ is the j^{th} extracted instance features of i^{th} bag and $Y_i \in \{0, 1\}$ denotes whether there exists LNM in the bag. To make it easier to express, we use $F = [f_1, f_2, ..., f_D]$ to represent the extracted instance features, i.e, x_j^i is a sampler of F. Our goal is to evaluate the importance of each extracted instance feature $f_k = F[k]$, which is achieved through the following two steps: (1) Generate a histogram of every feature in each bag with N_b bins of equal widths (which reflects the distribution of the feature in a bag). (2) Use the histogram as the bag representation and calculate the difference of obtained histograms between positive and negative labels to assess the discriminating value of a feature for classification.

3.3.1 Histogram Generation

For feature f_k, we calculate the maximum value and minimum value of this feature among all instances in all bags.

$$f_k^{max} = max\{x_j^i[k]\}, (i = 1, ..., N, j = 1, ..., K_i) \quad (6)$$

$$f_k^{min} = min\{x_j^i[k]\}, (i = 1, ..., N, j = 1, ..., K_i) \quad (7)$$

Then, we divide the range $[f_k^{min}, f_k^{max}]$ into N_b bins of equal widths and map each bag X_i into a histogram $H_k^i = (h_1^{i,k}, ..., h_{N_b}^{i,k})$, where h_o^i indicates the percentage of instances in X_i with feature f_k located in the o^{th} bin.

$$h_o^{i,k} = \frac{1}{K_i} \sum_{x_j^i \in \mathbf{X_i}} f_o(x_j^i[k]), \quad (8)$$

where $o = 1, ..., N_b$ and $j = 1, ..., K_i$. $f_o(x_j^i[k]) = 1$ if $x_j^i[k]$ is located in the o^{th} bin, otherwise $f_o(x_j^i[k]) = 0$.

3.3.2 Feature Evaluation

After obtaining the histograms of feature f_k of all bags $\{H_k^1, ..., H_k^N\}$, we evaluate the importance of the feature by the MMD distance, which is defined as:

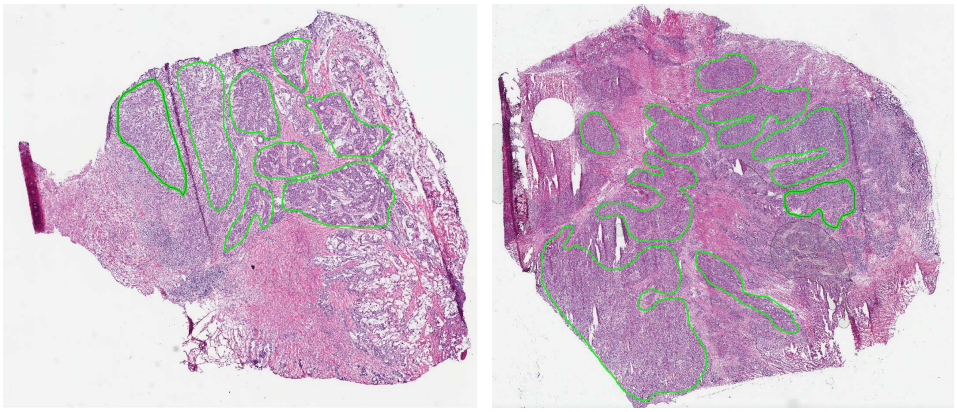
$$D(f_k) = \| \frac{1}{|G_P|} \sum_{X_i \in G_P} \phi(H_k^i) - \frac{1}{|G_N|} \sum_{X_j \in G_N} \phi(H_k^j) \|, \quad (9)$$

where G_P and G_N are the groups of all positive bags and negative bags respectively and ϕ is a mapping function. Bigger MMD distance means it is easier to discriminate the positive group from the negative group.

3.4. GCN-based Multiple Instance Learning

3.4.1 Graph Construction

We formulate our proposed network, i.e. GCN-based MIL as follows. Similar to [49], we utilize a heuristic approach to construct a graph from a bag of instance features $[x_1^i, x_2^i, ..., x_K^i]$ (K can variate for different bags.). First the adjacency



(a) Positive sample (b) Negative sample

Figure 4. Examples illustrating the whole slide image (WSI) from the cancer genome atlas (TCGA) dataset. The regions within the green contours are the colorectal cancer regions which are annotated by clinical experts. Sub-figure (a) shows a positive sample, i.e. WSI from patient with Lymph node metastasis (LNM). Sub-figure (b) demonstrates a negative sample.

matrix A can be obtained:

$$A_{mn} = \begin{cases} 1 & if \quad dist(x_p^i, x_q^i) < \gamma \\ 0 & otherwise \end{cases} \tag{10}$$

where $dist(x_p^i, x_q^i)$ is the distance between p^{th} and q^{th} instance feature in bag i. Here we use Euclidean distance to calculate $dist$. γ determines whether there is an edge connecting two instances. γ = 0 represents there is no edge connecting x_p^i and x_q^i while γ = $+\infty$ represents the input is a fully connected graph. At the same time, the instance features [x_1^i, x_2^i, ..., x_K^i] of bag i are considered to be the nodes of a graph. Then we obtain the graph of bag i as:

$$G_i = G(A_i, E_i) \tag{11}$$

where $A_i \in \{0,1\}^{K \times K}$ represents the adjacency matrix, $E_i \in \mathbb{R}^{K \times D}$ means node feature matrix constructed from a bag of X_i (D is the feature dimension).

3.4.2 Spectral Graph Convolution

Given a graph $G = (V, E)$, its normalized graph Laplacian $L = I - D^{-1/2} A D^{-/2}$. D is the degree matrix of G and A is the adjacency matrix mentioned above. Following the work in [11], we formulated kernel as a M^{th} order polynomial of diagonal Λ, and $diag(\Lambda)$ is the spectrum of graph Laplacian L:

$$g\theta(\Lambda^M) = \sum_{m=0}^{M-1} \theta_m \Lambda^m \tag{12}$$

Spectral convolution on graph G with vertex features $X \in \mathbb{R}^{N \times F}$ as layer input can be further obtained [10]:

$$Y = ReLU(g_\theta(L^M)X) \tag{13}$$

where ReLU is the commonly used activation function, $Y \in \mathbb{R}^{N \times F}$ is a graph of the identical number of vertices with convolved features. Chebyshev expansion is used to approximate $g_\theta(L)$ in order to accelerate filtering [11].

3.4.3 Network Architecture

Our GCN-based MIL network employs three stacked graph convolution layers (3.4.2), each followed by a ReLU activation and a self-attention graph pooling layer [28], to generate the bag representation (node embedding of the graph). After that, two fully connected layers with ReLU and a sigmoid activation function are utilized to achieve the bag-level classification. The categorical cross-entropy loss is used to optimize the network, which is defined as:

$$L = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} \delta(y_i = c) \log(P(y_i = c)) \tag{14}$$

where N denotes the data number and C represents the categories number. $\delta(y_i = c)$ is the indicator function and $P(y_i = c)$ is the predicted probability by the model.

4. Results and Discussion

4.1. Dataset

In this study, the Colon Adenocarcinoma (COAD) cohort of the Cancer Genome Atlas (TCGA) dataset [20] is used to evaluate our proposed method. This publicly released dataset contains 425 patients with colorectal cancer. For each patient, a H&E-stained histology WSI is acquired from the tumor region. Based on the clinical tumor node metastasis (TNM) staging information, these patients can be categorized into two groups: one without lymph node metastasis (patients in N_0 stage) and one with lymph node

metastasis (patients in stage from N_1 to N_4). There are 174 positive samples (patients with LNM) and 251 negative samples (patients without LNM) in the dataset. Fig. 4 shows one positive sample and one negative sample. Even experienced doctors and pathologists in our team cannot distinguish the patient with or without LNM if only relying on the histopathological image.

4.2. Evaluation

The area under the receiver operating characteristic curve (ROCAUC) together with the accuracy, precision, recall and F1-score are used to evaluate the performance of our proposed method and the state-of-the-art approaches. Specifically, these metrics are defined as:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{15}$$

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

$$\text{F1-score} = \frac{2 * (Recall * Precision)}{Recall + Precision} \tag{18}$$

Where TP, FP, TN, and FN represent the True Positive, False Positive, True Negative and False Negative respectively. Among these, ROCAUC is more comprehensive when comparing the performance of different methods.

Table 1. Hyperparameters of the proposed methods.

Hyperparameters	Value
VAE-GAN	
Loss weight λ_{Dis}	1
Loss weight λ_{KL}	1
Loss weight λ_{GAN}	1
Feature Selection	
Histogram bin number N_b	50
Feature selection rate	50%
GCN	
Distance threshold γ	$0.5 * \max_{x_p, x_q \in X} \{dist(x_p, x_q)\}$

4.3. Experimental Setup

In our experiment, the entire dataset (425 samples) is randomly divided into a training set (354 samples) and a test set (71 samples) in a ratio of 5:1. We perform five-fold cross-validation on the training dataset for parameter tuning purposes. We implement the VAE-GAN and GCN with PyTorch [32] and PyTorch Geometric library [12]. We randomly initialize the VAE-GAN under the default setting of PyTorch and resize the input images to 128×128 pixels. The Adam optimizer [22] is used to train both the VAE-GAN and GCN. To tackle the class imbalance problem during the bag-level classification stage, we employ the "WeightedRandomSampler" strategy [32] to prepare each training batch. Other related hyperparameters of each stage of the proposed method are given in Table 1.

4.4. Ablation Study

To evaluate the effectiveness of different components in the proposed method, we conduct ablation studies. We experiment on the following configurations: (A) Our proposed method: VAE-GAN + FS + GCN. (B) VAE-GAN + GCN: our framework without feature selection. (C) Pre-trained ResNet + FS + GCN: our framework but using pre-trained ResNet-18 (on ImageNet) as the instance-level feature extractor. (D) Pre-trained ResNet + GCN: using the pre-trained ResNet-18 (on ImageNet) as instance-level feature extractor and then using the GCN for bag-level classification directly without feature selection. (E) Pre-trained ResNet + GCN (end-to-end): This configuration is similar to (D), while the difference lies in that (E) is an end-to-end network, i.e, the back-propagated loss from the GCN can guide the training of the instance-level feature extraction network (ResNet-18).

The results of all these configurations are illustrated in Table 2. Comparing (A) to (B) and (C) to (D), it is noted that the use of the proposed feature selection procedure improves ROCAUC by 3.3% and 3.0% respectively. Similarly, when comparing (A) to (C) and (B) to (D), the VAE-GAN results in 7.3% and 7.1% performance gain. The comparison between (A) and (E) shows the two-stage method performs better than the end-to-end approach. Although the end-to-end configuration (E) can extract instance-level features and generate bag representation together, tackling both tasks simultaneously imposes heavy workload to optimize the whole network with only bag-level classification supervision, especially when lacking sufficient training data. Therefore, the two-stage method which separately handles instance-level feature extraction and bag-level representation is more appropriate for our task.

4.5. Comparison with State-of-the-Art Methods

Table 3 demonstrates the comparison between our proposed method and other state-of-the-art methods including (1) T-stage + LR, (2) Histomics + Histogram [8], (3) ResNet + voting [21], (4) WSISA [45], (5) ResNet + RNN [5], (5) Attention based MIL [19]. We reimplement all the previous methods based on the literatures and open source codes. From the table, we can observe that our approach outperforms these methods.

T-stage + LR and Histomics + Histogram [8] are two machine learning algorithms based on hand crafted features. T-

Table 2. The results of the ablation study.

Method \ Metric	Accuracy	Precision	Recall	F1-score	ROCAUC
(A) VAE-GAN + FS + GCN (OUR)	**0.6761**	**0.575**	0.7931	**0.6667**	**0.7102**
(B) VAE-GAN + GCN	0.5775	0.4902	**0.8621**	0.6250	0.6773
(C) Pre-trained ResNet + FS + GCN	0.4225	0.4032	**0.8621**	0.5495	0.6371
(D) Pre-trained ResNet + GCN	0.5634	0.4792	0.7931	0.5974	0.6067
(E) Pre-trained ResNet + GCN (End-to-End)	0.4648	0.4182	0.7931	0.5476	0.6010

Table 3. Comparisons between our proposed method and the state-of-the-art approaches.

Method \ Evaluation	Accuracy	Precision	Recall	F1-score	ROCAUC
Our	**0.6761**	0.575	0.7931	**0.6667**	**0.7102**
T-stage + LR	0.6357	**0.7143**	0.1887	0.2985	0.6471
Histomics + Histogram [8]	0.6124	0.5484	0.3208	0.4048	0.6157
ResNet + Voting [21]	0.5891	0.5	0.3208	0.3908	0.5824
WSISA [45]	0.5969	0.5152	0.3208	0.3953	0.5792
ResNet + RNN [5]	0.4109	0.4109	**1**	0.5824	0.5
Attention based MIL [19]	0.5891	0.5	0.3208	0.3908	0.5457

stage is a factor describing the invasion depth of the tumor into the intestinal wall [1]. Recent researches report that the depth of tumor invasion is related to the lymph node metastasis [35]. The T-stage + LR method utilizes the T-stage information as the feature and adopts the logistic regression to predict lymph node metastasis of patients in colorectal cancer. Histomics + Histogram method [8] extracts cell morphologic features including nucleus shape, intensity, texture, and the spatial relationship between nuclei as features. It utilizes a histogram to analyze the cell distribution in the WSI and uses the lasso regression [8] to predict the prognosis of patients.

The T-stage + LR method is only based on the T-stage feature, which lacks the information of local cancer cell texture and global histology of the tumor. The Histomics + Histogram method utilizes specifically-designed features, which limits its extension ability. For instance, the cell morphologic features maybe not suitable for the LNM prediction on colorectal cancer because cancer cells are similar and do not change during the propagation.

As mentioned in section 2.3, the ResNet + voting method [21] and WSISA are typical instance-space MIL methods. These methods work well if the discriminative information is considered to lie at the instance level and there exist key instances which are strongly related to the bag-level labels. However, the conditions do not hold in our task and therefore these two methods perform poorly.

The performances of the ResNet + RNN and Attention-based MIL methods are inferior in the LNM prediction task compared to our proposed method. This might be due to three reasons: (1) Extracting instance-level features and generating bag representation together impose heavy workload on the end-to-end network. (2) The network has learnt many unhelpful features which should be removed before generating the bag representation. (3) The RNN and attention mechanism are not as good as the GCN in the bag-level classification stage.

5. Conclusion

In this paper, we investigate a challenging clinical task of automatic prediction of lymph node metastasis using histopathological images of colorectal cancer. To achieve this, we develop a deep GCN-based MIL method combined with a feature selection strategy. Experimental results demonstrate that our method benefits from our proposed components, including (1) VAE-GAN for instance-level feature extraction, (2) instance-level feature selection and (3) GCN-based MIL for bag representation and bag-level classification. Our approach shows superior performance compared to the state-of-the-art methods. In the future, it would be meaningful to develop a unified GCN model for performing a joint instance selection and instance-level feature selection with the weak label in a bag level.

References

[1] Mahul B Amin, Frederick L Greene, Stephen B Edge, Carolyn C Compton, Jeffrey E Gershenwald, Robert K Brookland, Laura Meyer, Donna M Gress, David R Byrd, and

David P Winchester. The eighth edition ajcc cancer staging manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA: a cancer journal for clinicians*, 67(2):93–99, 2017. 8

[2] Jaume Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial intelligence*, 201:81–105, 2013. 1, 5

[3] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In *Advances in neural information processing systems*, pages 577–584, 2003. 1

[4] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006. 5

[5] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019. 1, 3, 7, 8

[6] Marc-André Carbonneau, Eric Granger, and Ghyslain Gagnon. Bag-level aggregation for multiple-instance active learning in instance classification problems. *IEEE transactions on neural networks and learning systems*, 30(5):1441–1451, 2018. 2

[7] George J Chang, Miguel A Rodriguez-Bigas, John M Skibber, and Virginia A Moyer. Lymph node evaluation and survival after curative resection of colon cancer: systematic review. *Journal of the National Cancer Institute*, 99(6):433–441, 2007. 2

[8] Jun Cheng, Jie Zhang, Yatong Han, Xusheng Wang, Xiufen Ye, Yuebo Meng, Anil Parwani, Zhi Han, Qianjin Feng, and Kun Huang. Integrative analysis of histopathological images and genomic data predicts clear cell renal cell carcinoma prognosis. *Cancer research*, 77(21):e91–e100, 2017. 7, 8

[9] Kenneth SH Chok and Wai Lun Law. Prognostic factors affecting survival and recurrence of patients with pt1 and pt2 colorectal cancer. *World journal of surgery*, 31(7):1485–1490, 2007. 2

[10] Fan RK Chung and Fan Chung Graham. *Spectral graph theory*. Number 92. American Mathematical Soc., 1997. 6

[11] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*, pages 3844–3852, 2016. 6

[12] Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019. 7

[13] Leonard L Gunderson, Daniel J Sargent, Joel E Tepper, Norman Wolmark, Michael J O'Connell, Mirsada Begovic, Cristine Allmer, Linda Colangelo, Steven R Smalley, Daniel G Haller, et al. Impact of t and n stage and treatment on survival and relapse in adjuvant rectal cancer: a pooled analysis. *Journal of Clinical Oncology*, 22(10):1785–1796, 2004. 2

[14] Philip Haeusser, Thomas Frerix, Alexander Mordvintsev, and Daniel Cremers. Associative domain adaptation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[16] Le Hou, Dimitris Samaras, Tahsin M Kurc, Yi Gao, James E Davis, and Joel H Saltz. Efficient multiple instance convolutional neural networks for gigapixel resolution image classification. *arXiv preprint arXiv:1504.07947*, page 7, 2015. 2

[17] Le Hou, Dimitris Samaras, Tahsin M Kurc, Yi Gao, James E Davis, and Joel H Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2424–2433, 2016. 2

[18] Fang Huang, Jinqing Qi, Huchuan Lu, Lihe Zhang, and Xiang Ruan. Salient object detection via multiple instance learning. *IEEE Transactions on Image Processing*, 26(4):1911–1922, 2017. 1

[19] Maximilian Ilse, Jakub M Tomczak, and Max Welling. Attention-based deep multiple instance learning. *arXiv preprint arXiv:1802.04712*, 2018. 1, 2, 3, 7, 8

[20] Cyriac Kandoth, Michael D McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie, Qunyuan Zhang, Joshua F McMichael, Matthew A Wyczalkowski, et al. Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471):333, 2013. 6

[21] Jakob Nikolas Kather, Alexander T Pearson, Niels Halama, Dirk Jäger, Jeremias Krause, Sven H Loosen, Alexander Marx, Peter Boor, Frank Tacke, Ulf Peter Neumann, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature medicine*, page 1, 2019. 3, 7, 8

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[24] Oren Z Kraus, Jimmy Lei Ba, and Brendan J Frey. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics*, 32(12):i52–i59, 2016. 2

[25] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015. 4

[26] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 2

[27] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y Ng. Efficient sparse coding algorithms. In *Advances in neural information processing systems*, pages 801–808, 2007. 1

[28] Junhyun Lee, Inyeop Lee, and Jaewoo Kang. Self-attention graph pooling. In *International Conference on Machine Learning*, pages 3734–3743, 2019. 6

[29] D. Liu, Y. Zhou, X. Sun, Z. Zha, and W. Zeng. Adaptive pooling in multi-instance learning for web video annotation. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 318–327, Oct 2017. 2

[30] Guoqing Liu, Jianxin Wu, and Zhi-Hua Zhou. Key instance detection in multi-instance learning. In *Asian Conference on Machine Learning*, pages 253–268, 2012. 1

[31] Siyamalan Manivannan, Caroline Cobb, Stephen Burgess, and Emanuele Trucco. Subcategory classifiers for multiple-instance learning and its application to retinal nerve fiber layer visibility classification. *IEEE transactions on medical imaging*, 36(5):1140–1150, 2017. 1

[32] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 7

[33] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1796–1804, 2015. 1

[34] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning with applications to medical imaging. *Advances in neural information processing systems*, 2019. 2

[35] TJ Saclarides, Achyut K Bhattacharyya, C Britton-Kuzel, D Szeluga, and SG Economou. Predicting lymph node metastases in rectal cancer. *Diseases of the colon & rectum*, 37(1):52–57, 1994. 8

[36] Miao Sun, Tony X Han, Ming-Chang Liu, and Ahmad Khodayari-Rostamabad. Multiple instance learning convolutional neural networks for object recognition. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 3270–3275. IEEE, 2016. 2

[37] Peng Tang, Xinggang Wang, Angtian Wang, Yongluan Yan, Wenyu Liu, Junzhou Huang, and Alan Yuille. Weakly supervised region proposal network and object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 352–368, 2018. 1

[38] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. C-mil: Continuation multiple instance learning for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2199–2208, 2019. 1

[39] Hao Wang, Xian-Zhao Wei, Chuan-Gang Fu, Rong-Hua Zhao, and Fu-Ao Cao. Patterns of lymph node metastasis are different in colon and rectal carcinomas. *World Journal of Gastroenterology: WJG*, 16(42):5375, 2010. 2

[40] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018. 2

[41] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *Pattern Recognition*, 74:15–24, 2018. 1, 2

[42] Jiajun Wu, Yinan Yu, Chang Huang, and Kai Yu. Deep multiple instance learning for image classification and auto-annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3460–3469, 2015. 1, 2

[43] Gang Xu, Zhigang Song, Zhuo Sun, Calvin Ku, Zhe Yang, Cancheng Liu, Shuhao Wang, Jianpeng Ma, and Wei Xu. Camel: A weakly supervised learning framework for histopathology image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10682–10691, 2019. 1

[44] Yongluan Yan, Xinggang Wang, Xiaojie Guo, Jiemin Fang, Wenyu Liu, and Junzhou Huang. Deep multi-instance learning with dynamic pooling. In *Asian Conference on Machine Learning*, pages 662–677, 2018. 2

[45] Jiawen Yao, Xinliang Zhu, and Junzhou Huang. Deep multi-instance learning for survival prediction from whole slide images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 496–504. Springer, 2019. 7, 8

[46] Qi Zhang and Sally A Goldman. Em-dd: An improved multiple-instance learning technique. In *Advances in neural information processing systems*, pages 1073–1080, 2002. 1

[47] Yizhou Zhou, Xiaoyan Sun, Dong Liu, Zhengjun Zha, and Wenjun Zeng. Adaptive pooling in multi-instance learning for web video annotation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 318–327, 2017. 2

[48] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2017. 1

[49] Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li. Multi-instance learning by treating instances as non-iid samples. In *Proceedings of the 26th annual international conference on machine learning*, pages 1249–1256. ACM, 2009. 5

[50] Xinliang Zhu, Jiawen Yao, Feiyun Zhu, and Junzhou Huang. Wsisa: Making survival prediction from whole slide histopathological images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7234–7242, 2017. 3