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

Yu Zhao\textsuperscript{1,2,*}, Fan Yang\textsuperscript{1,*}, Yuqi Fang\textsuperscript{3}, Hailing Liu\textsuperscript{1}, Niyun Zhou\textsuperscript{1}, Jun Zhang\textsuperscript{1}, Jiarui Sun\textsuperscript{1}, Sen Yang\textsuperscript{1}, Bjoern Menze\textsuperscript{2}, Xinjuan Fan\textsuperscript{4} and Jianhua Yao\textsuperscript{1}

\textsuperscript{1}Tencent AI Lab, \textsuperscript{2}Technical University of Munich, \textsuperscript{3}The Chinese University of Hong Kong, \textsuperscript{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
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 pooling [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 indiscriminate 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
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.

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
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:

$$L_{VAE} = L_{Pixel} + L_{KL} = -E_{q(h|x)}[\log(p(x|h))] + D_{KL}(q(h|x)||p(h)),$$

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:

$$L_{GAN} = \log(D(x)) + \log(1 - D(G(h)))$$

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 $L_{Pixel}$ 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:

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

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:

$$L_{Dis_l}^{Like} = -E_{q(h|x)}[\log(p(Dis_l(x)|h))]$$

Using $L_{Dis_l}^{Like}$ replacing $L_{Pixel}$, we can obtain the loss function of entire VAE-GAN [25]:

$$L = \lambda_{Dis} \cdot L_{Dis_l}^{Like} + \lambda_{KL} \cdot L_{KL} + \lambda_{GAN} \cdot L_{GAN}$$

where $\lambda_{Dis}$, $\lambda_{KL}$ and $\lambda_{GAN}$ are the hyperparameters 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...
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 $\{x^i_1, x^i_2, ..., x^i_{K_i}\}$. In our case, $x^i_j \in \mathbb{R}^D$ 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., $x^i_j$ 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 bags 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.

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

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

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 $H^i_k = (h^i_{o1}, ..., h^i_{oK})$, where $h^i_o$ indicates the percentage of instances in $X_i$ with feature $f_k$ located in the $o$th bin.

$$h^i_o = \frac{1}{K_i} \sum_{x^i_j[k]} f_o(x^i_j[k]), \quad (8)$$

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

### 3.3.2 Feature Evaluation

After obtaining the histograms of feature $f_k$ of all bags $\{H^i_1, ..., H^i_N\}$, we evaluate the importance of the feature by the MMD distance, which is defined as:

$$D(f_k) = \frac{1}{|G_P|} \sum_{X_i \in G_P} \phi(H^i_k) - \frac{1}{|G_N|} \sum_{X_i \in G_N} \phi(H^i_k), \quad (9)$$

where $G_P$ and $G_N$ are the groups of all positive bags and negative bags respectively and $\phi$ 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 $\{x^i_1, x^i_2, ..., x^i_{K_i}\}$ (K can variate for different bags.). First the adjacency
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:

$$A_{mn} = \begin{cases} 1 & \text{if } \text{dist}(x_i^p, x_i^q) < \gamma \\ 0 & \text{otherwise} \end{cases}$$

(10)

where \( \text{dist}(x_i^p, x_i^q) \) is the distance between \( p \)th and \( q \)th instance feature in bag \( i \). Here we use Euclidean distance to calculate \( \text{dist} \). \( \gamma \) determines whether there is an edge connecting two instances. \( \gamma = 0 \) represents there is no edge connecting \( x_i^p \) and \( x_i^q \) while \( \gamma = +\infty \) represents the input is a fully connected graph. At the same time, the instance features \( [x_i^1, x_i^2, ..., x_i^K] \) of bag \( i \) are considered to be the nodes of a graph. Then we obtain the graph of bag \( i \) as:

$$G_i = G(A_i, E_i)$$

(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 \( L = I - D^{-1/2}AD^{-1/2} \). \( D \) 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 \( \Lambda \), and \( \text{diag}(\Lambda) \) is the spectrum of graph Laplacian \( L \):

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

(12)

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

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

(13)

where \( \text{ReLU} \) 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 \( \text{ReLU} \) 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 \( \text{ReLU} \) 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:

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

(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:

\[ \text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \]  \hspace{1cm} (15)

\[ \text{Precision} = \frac{TP}{TP + FP} \]  \hspace{1cm} (16)

\[ \text{Recall} = \frac{TP}{TP + FN} \]  \hspace{1cm} (17)

\[ \text{F1-score} = \frac{2 \times (\text{Recall} \times \text{Precision})}{\text{Recall} + \text{Precision}} \]  \hspace{1cm} (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.

<table>
<thead>
<tr>
<th>Hyperparameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAE-GAN</td>
<td></td>
</tr>
<tr>
<td>Loss weight $\lambda_{\text{Dis}}$</td>
<td>1</td>
</tr>
<tr>
<td>Loss weight $\lambda_{\text{KL}}$</td>
<td>1</td>
</tr>
<tr>
<td>Loss weight $\lambda_{\text{GAN}}$</td>
<td>1</td>
</tr>
<tr>
<td>Feature Selection</td>
<td></td>
</tr>
<tr>
<td>Histogram bin number $N_b$</td>
<td>50</td>
</tr>
<tr>
<td>Feature selection rate</td>
<td>50%</td>
</tr>
<tr>
<td>GCN</td>
<td></td>
</tr>
<tr>
<td>Distance threshold $\gamma$</td>
<td>$0.5 \times \max_{x_p,x_q \in X} {\text{dist}(x_p,x_q)}$</td>
</tr>
</tbody>
</table>

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 $128 \times 128$ 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.

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

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

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

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


