# Using multi-scale convolutional neural network based on multi-instance learning to predict the efficacy of neoadjuvant chemoradiotherapy for rectal cancer

Dehai Zhang, Yongchun Duan, Jing Guo*, Yaowei Wang*, Yun Yang*, Zhenhui Li*, Kelong Wang, Lin Wu, and Minghao Yu

**Structured Abstract—**Background: At present, radical total mesorectal excision after neoadjuvant chemoradiotherapy is crucial for locally advanced rectal cancer. Therefore, the use of histopathological images analysis technology to predict the efficacy of neoadjuvant chemoradiotherapy for rectal cancer is of great significance for the subsequent treatment of patients. Methods: In this study, we propose a new pathological images analysis method based on multi-instance learning to predict the efficacy of neoadjuvant chemoradiotherapy for rectal cancer. Specifically, we proposed a gated attention normalization mechanism based on the multilayer perceptron, which accelerates the convergence of stochastic gradient descent optimization and can speed up the training process. We also proposed a bilinear attention multi-scale feature fusion mechanism, which organically fuses the global features of the larger receptive fields and the detailed features of the smaller receptive fields and alleviates the problem of pathological images context information loss caused by block sampling. At the same time, we also designed a weighted loss function to alleviate the problem of imbalance between cancerous instances and normal instances. Results: We evaluated our method on a locally advanced rectal cancer dataset containing 150 whole slide images. In addition, to verify our method's generalization performance, we also tested on two publicly available datasets, Camelyon16 and MSKCC. The results show that the AUC values of our method on the Camelyon16 and MSKCC datasets reach 0.9337 and 0.9091, respectively. Conclusion: Our method has outstanding performance and advantages in predicting the efficacy of neoadjuvant chemoradiotherapy for rectal cancer.

***Index Terms*—**Pervasive computing, neoadjuvant chemoradiotherapy, internet of things, pathological images, rectal cancer.

***Clinical and Translational Impact Statement*—**This study aims to predict the efficacy of neoadjuvant chemoradiotherapy for rectal cancer to assist clinicians quickly diagnose and formulate personalized treatment plans for patients.

## I. Introduction

COLORECTAL cancer is the third leading cause of cancer globally, and rectal cancer accounts for about 30%-35% of colorectal cancer cases [1], [2]. More than 100,000 people worldwide are diagnosed with rectal cancer each year, 70% of which are locally advanced rectal cancer (LARC). Neoadjuvant chemoradiation (NCRT) and radical surgery are the best treatments recommended for locally advanced rectal cancer [3] because preoperative radiotherapy and chemotherapy can reduce recurrence and improve survival [4]. Many studies have shown that the response to neoadjuvant therapy affects the prognosis, especially the pathological complete response has a good effect on the prognosis [5]–[7]. Approximately 15% to 27% of patients will show a pathological complete response (PCR) to neoadjuvant chemoradiotherapy [8], [9]. However, a considerable number of patients will not respond to neoadjuvant therapy. Based on the resection specimens, these non-responders can be defined as patients with no changes in tumor regression after neoadjuvant therapy. These non-responders may benefit little from neoadjuvant therapy but still have related toxicity. More importantly, tumor progression may occur in some patients during

The data part of this work is supported by Yunnan Cancer Hospital.

Corresponding author: Jing Guo, Yaowei Wang, Yun Yang and Zhenhui Li. Jing Guo is with the School of Information Science and Engineering, Yunnan University, Kunming, China (e-mail:guojing@mail.ynu.edu.cn). Yaowei Wang is with the School of Software, Yunnan University, Kunming, China (e-mail:12019202407@mail.ynu.edu.cn). Yun Yang is with the Key Laboratory in Software Engineering of Yunnan Province, Kunming, China (e-mail:yangyun@ynu.edu.cn). Zhenhui Li is with the Third Affiliated Hospital of Kunming Medical University, Yunnan Cancer Hospital, Yunnan Cancer Center, Kunming, China (e-mail:lizhenhui621@qq.com).

Dehai Zhang and Yongchun Duan are co-first authors of the article.

Dehai Zhang, Yongchun Duan, Kelong Wang, and Minghao Yu are with the School of Software, Yunnan University, Kunming, China.

Lin Wu is with the Department of Radiology, Yunnan Cancer Hospital, Kunming, China.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JTEHM.2022.3156851, IEEE Journal of Translational Engineering in Health and Medicine

treatment. Therefore, it is very important to accurately predict non-response before implementing neoadjuvant therapy to develop a personalized treatment plan, including avoiding overtreatment and timely selection of alternative treatments. In addition, it can also help patients avoid the risks and uncertain consequences of surgery. However, in addition to the pathological evaluation after neoadjuvant therapy, there is currently no reliable method to accurately divide patients into non-good response (non-GR) groups and good response (GR) groups. Due to the heterogeneity of tumors [10], accurately predicting a non-response to neoadjuvant therapy remains challenging. Therefore, it is of great significance to predict the efficacy of neoadjuvant chemoradiotherapy for rectal cancer using histopathological images. In this study, we established a deep learning model based on the pathological images of patients with locally advanced rectal cancer to accurately predict the patient's response to neoadjuvant chemoradiotherapy so as to assist clinicians in formulating personalized treatment plans for patients.

Histopathological images analysis is an important step in cancer or disease diagnosis. In recent years, the diagnosis of cancer has been improved by using a deep learning based histopathological images analysis framework [11]. These advances have promoted the progress of diagnostic and computational methods for gastrointestinal diseases [12], [13]. However, many issues remain to be addressed in the field, including the variability of histopathological features across diseases, limited data, and high resolution of whole slide images (WSIs), making it difficult for the model to use WSIs for training directly. To match traditional input sizes for traditional feed-forward CNN models, typical WSIs would need to be down-sampled by a factor of ~40×, resulting in the loss of cellular and structural details, which are critical for prediction. To overcome this bottleneck, state-of-the-art methods for WSIs analysis adopt a two-stage approach [14]–[21]. First, patches are extracted by using block sampling on high-resolution WSIs, and the patches are used to train the CNN model, and the patches are encoded as prediction scores or low-dimensional feature vectors. Second, learn an aggregation model to integrate the slice-level information obtained for the entire slide prediction. Although the two-stage training method has achieved good results in the field of pathological images analysis, it still has defects; that is, the parameters in the feature extraction stage cannot be updated, resulting in the inability to obtain the specific features of pathological tissues. Therefore, recently, some researchers proposed an end-to-end framework to solve this problem [22], [23], but the end-to-end training method has the problem of very slow training.

In addition to the above problems, in the study of predicting the efficacy of neoadjuvant chemoradiotherapy for rectal cancer, there are the following limitations and challenges: First of all, since pathological images are all gigapixel-level data, the data training process is slow. Secondly, in pathological images analysis, the pathological images are usually first divided into slices of the same size, and then each slice is sent to the model for separate processing, and finally, the predictions of all patches are aggregated to obtain patient-level predictions, but this method often leads to the loss of global information of pathological images. Additionally, in the pathological image of the tumor, the normal tissue is much larger than the cancer tissue. After the slice sampling, the number of normal tissue slices is much larger than the number of cancer tissue slices, which will cause the problem of imbalance between positive and negative instances. Finally, the traditional pathological image analysis methods often require pathologists to manually divide the decision boundary between tumor and normal tissue, which is time-consuming and laborious, and greatly increases the burden on oncologists. Therefore, how to use pathological images without pixel-level annotation to predict the efficacy of neoadjuvant chemoradiotherapy for rectal cancer is also the limitation and challenges faced by this study.

To address these challenges, we propose a new based on weakly-supervised learning classification framework to classify rectal cancer histological images inspired by the previous work [24], [25], which uses multi-scale features. The experiments revealed that the proposed method could be directly applied to WSIs classification without delineating the decision boundary of cancerous tissues, greatly reducing the labeling burden of pathologists and outperforming the traditional multi-instance learning (MIL) methods. The main contributions of this paper are as below:

1) We designed a gated attention weight normalization mechanism based on the multilayer perceptron by reparameterizing the weight vectors in a gated attention network that decouples the length of those weight vectors from their direction, thereby speeding up the training process and the convergence speed of stochastic gradient descent.
2) We propose a bilinear attention multi-scale feature fusion mechanism to alleviate the problem of global information loss. This mechanism makes full use of the given vision-language information by learning the bilinear attention distributions of pathological images and can better integrate the global features provided by the larger receptive field patches and the detailed features provided by the smaller receptive field patches.
3) We designed a weighted loss function to alleviate the problem of instance imbalance by optimizing the instance-level and bag-level loss functions simultaneously.
4) Our method directly uses pathological images without pixel-level annotations for training and has excellent performance on both the private dataset and the publicly available Camelyon16 and MSKCC datasets.

We applied the proposed method to predict the efficacy of neoadjuvant chemoradiotherapy for rectal cancer and proved the outstanding performance of this method. We also verified the proposed algorithm on the Camelyon16 and MSKCC public datasets. The experimental results show that

the proposed method performs better than the conventional method on the weakly labeled dataset without any processing. The rest of this paper is organized as follows. Section II briefly reviews related work of the early diagnosis of rectal cancer and histopathology images analysis. In Section III, we describe our proposed method in detail. The experiments and comparison results are given in Section IV, followed by discussions and conclusions in Section V.

## II. Related Works

Our research is mainly to directly use the pathological tissue images that do not delineate the decision boundary of cancerous tissue to predict the pathological response of neoadjuvant chemoradiotherapy for rectal cancer. In this section, we briefly reviewed the latest advances in the prediction of pathological response of neoadjuvant chemoradiotherapy for rectal cancer and the related work of deep learning models for pathological images analysis.

**Prediction of pathological response to neoadjuvant chemoradiotherapy in rectal cancer.** Currently, the research on the pathological response of neoadjuvant chemoradiotherapy of rectal cancer mainly uses the radiomic characteristics of rectal cancer and the histological characteristics of WSIs to predict. Radiomics is a combination of quantitative images analysis and machine learning methods, and it is considered a form of AI. Histological features are quantitative images features that can provide tumor intensity, shape, size, volume, and texture features. Different imaging modes (such as MRI, CT, PET, etc.) can be used as the basis for feature extraction in imaging omics. All the features extracted from the images are "radiomics", and those feature sets with the predictive value selected after feature selection are usually called "radiomic signature". At present, the basic function of radiomics is to quantitatively analyze tumor regions of interest through a large number of radiomics characteristics, which can provide valuable diagnostic, prognostic or predictive information. Its purpose is to explore and use these information resources to develop diagnostic, predictive, or prognostic imaging omics models to support personalized clinical decision-making and improve individualized treatment options [26]–[28]. Histopathology is the gold standard of clinical tumor diagnosis, directly related to the development of treatment and the evaluation of prognosis. Recently, many studies have shown that it is feasible and effective to use the features of digital pathological images to predict the treatment response of neoadjuvant chemoradiotherapy [29], [30]. Therefore, this study is of great research significance.

**Deep learning models for WSIs analysis.** At present, deep learning methods have been widely used to predict the prognosis of various cancers. For example, some researchers use deep convolutional neural networks to assess the human tumor microenvironment and directly predict the prognosis from histopathological images [31]. At the same time, there is also evidence that deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer[32]. In addition, some scholars use deep learning based tissue analysis to predict the prognosis of colorectal cancer [33]. In recent years, pathological image analysis methods based on deep learning mainly rely on multi-instance learning and weakly supervised learning. Since only coarse-grained labels are available in pathological images analysis problems, after down-sampling WSIs into patches, it is not easy to model the patches with real labels. We can characterize this problem as an inexact supervision problem, so many studies use multi-instance learning to solve this problem. In the two-class multi-instance learning, the bag is marked as a positive bag only if it contains at least one positive instance; otherwise, it is a negative bag. In pathological images analysis research, the researcher regards the down-sampled patches as instances and WSIs as bags. MIL has been successfully applied to the classification of histopathological images [16], [25], [34]. Recently, authors in [14] proposed a MIL-RNN aggregation operator comprising patch-level training, top-k instance selection, and RNN-based aggregation for patient-level prediction. In addition, some researchers use multi-scale convolutional layers on pre-trained CNNs to capture scale-invariant patterns and use top-k pools to aggregate feature maps for patient-level prediction. Also, some researchers [24] proposed a weight normalization method by reparameterizing the weight vectors in a neural network that decouples the length of those weight vectors from their direction, thereby accelerating the convergence speed of stochastic gradient descent. Authors in [25] proposed a clustering-constrained attention multiple instance learning method (CLAM) that uses attention-based learning to automatically identify sub-regions of high diagnostic value to classify the whole slide images accurately. Furthermore, authors in [35] proposed bilinear attention networks (BAN) that find bilinear attention distributions to utilize given vision-language information seamlessly. At the same time, some researchers have also explored the application of ensemble learning and transfer learning in medical research and have achieved outstanding results[36]–[39]. At present, although these algorithms have achieved good results in different tasks, they rely too much on the way pathologists manually divide the decision-making boundary of pathological tissues [40], [41]. The motivation of our proposed method is to directly learn pathological images without any pre-processing by introducing attention modules for use in complex WSIs analysis tasks.

## III. Methods

### A. Attention-based Weight Normalization MIL

Compared with inputting a series of individually labeled instances, under the MIL framework, the input is a series of labeled "bags", and each "bag" includes many instances, and these instances have permutation invariance. In multiple instance learning, the images are described as a "bag": $X = \{x1, \ldots, xN\}$, each $x$i is a feature vector extracted from the corresponding i-th region in the images (instance), and N is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JTEHM.2022.3156851, IEEE Journal of Translational Engineering in Health and Medicine

the number of regions (instance) where the images are segmented. Take two-classification as an example. A bag contains multiple instances. If all instances are marked negative, the bag is negative; otherwise, the bag is positive. The current state-of-the-art attention-based MIL method [42] uses a permutation invariant aggregation operator called "Attention-based MIL pooling". Encode each instance *x*i as a low-dimensional embedding through CNN, then use the attention mechanism to assign attention scores to each low-dimensional embedding (instance), and use their weighted average to generate bag-level prediction z. Let $H = \{h1, \ldots, hK\}$ be a bag containing K low-dimensional embeddings, and then the bag-level prediction is defined as follows:

$$z = \sum_{k=1}^{K} a_k h_k \tag{1}$$

Where $a_k$ is defined as follows:

$$a_k = \frac{\exp\left\{w^{\mathrm{T}}\left(\tanh\left(Vh_k^{\mathrm{T}}\right)\odot \mathrm{sigm}\left(Uh_k^{\mathrm{T}}\right)\right)\right\}}{\sum_{j=1}^{K}\exp\left\{w^{\mathrm{T}}\left(\tanh\left(Vh_j^{\mathrm{T}}\right)\odot \mathrm{sigm}\left(Uh_j^{\mathrm{T}}\right)\right)\right\}} \tag{2}$$

Where $w\in \mathrm{R}^{\mathrm{L}\times 1}$, $V\in \mathrm{R}^{\mathrm{L}\times \mathrm{M}}$ and $U\in \mathrm{R}^{\mathrm{L}\times \mathrm{M}}$ are trainable parameters. Attention-based MIL pooling is trainable and allows the network to identify discriminative instances. Since our problem only contains slide-level labels, using the attention mechanism in the MIL pooling will help achieve better results. Although the mil pool based on attention can help us find examples closely related to diagnosis, this method has a clear disadvantage. That is, it does not consider the optimization of attention weight.

Gated attention weight normalization mechanism. We propose to normalize the attention weight by weight normalization[24] and use the multi-layer perceptron to calculate the attention score that accelerates the convergence of the stochastic gradient descent optimization and alleviates the problem of "vanishing gradient". Specifically, we decompose $U\in \mathrm{R}^{\mathrm{L}\times \mathrm{M}}$ and $V\in \mathrm{R}^{\mathrm{L}\times \mathrm{M}}$ into a parameter vector *m* and a parameter scalar *g*:

$$U = \frac{g_u}{\| m_u \|} m_u, V = \frac{g_v}{\| m_v \|} m_v \tag{3}$$

In summary, we can get:

$$a_k = \frac{\exp\left\{w^{\mathrm{T}}\left(M_k \odot N_k\right)\right\}}{\sum_{j=1}^{K}\exp\left\{w^{\mathrm{T}}\left(M_j \odot N_j\right)\right\}} \tag{4}$$

Among these:

$$M_i = \tanh\left(\frac{g_v}{\| m_v \|} m_v \mathrm{h}_i^{\mathrm{T}}\right), \quad N_i = \mathrm{sigm}\left(\frac{g_u}{\| m_u \|} m_u \mathrm{h}_i^{\mathrm{T}}\right) \tag{5}$$

During training, *m* and *g* were updated, respectively. The experiments indicated that the proposed Gated attention weight normalization mechanism is superior to the traditional attention-based MIL method. The normalization mechanism of gated attention weight is shown in Fig. 1.

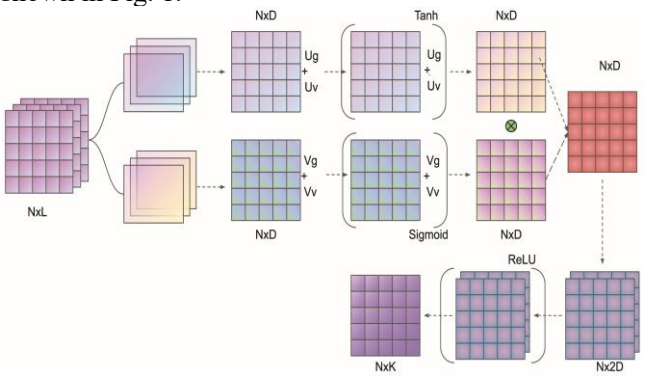


Fig. 1. Overview of MLP-based attention gating weight normalization. Among them, Ug, Uv, and Vg, Vv represent network parameters. N, L, D, and K respectively represent the number of patches included in each slide, the input feature dimension, the hidden layer dimension, and the number of classes.

### *B. Bilinear Attention Multi-Scale Feature Fusion*

To alleviate the problem of contextual information loss caused by single-scale sampling, we propose a bilinear attention multi-scale feature fusion mechanism. Firstly, we extract 20× and 5× magnification patches from pathological images by sliding window strategy. Then, the attention matrix of WSIs under the different field of view sampling is obtained through feature extraction and attention mechanism. Finally, we get the final WSIs attention matrix by bilinear transformation. The bilinear transformation follows the following formula:

$$M = M_{small} * A * M_{big} + b \tag{6}$$

Among these, $M_{small}$ is the attention matrix corresponding to the small field of view patches, $M_{big}$ is the attention matrix corresponding to the big field of view patches, $A$ is the weight, $b$ is the bias. Attention-based bilinear multi-scale feature fusion mechanism can effectively improve the classification performance of pathological images.

### *C. Network Framework*

We follow the feature extraction and classification process shown in Fig. 2. In this study, we use the Otsu binarization algorithm to separate tissue regions, and use the residual neural network (ResNet18)[43] as feature extractors, then use the method proposed in this study to aggregate the patch-level predictions into slide-level predictions. Since the tumor area is much smaller than the normal tissue area, the number of positive and negative instances is seriously unbalanced, which leads to the wrong division of instances in the model. Therefore, in the instance classification stage, we introduce dual focal loss[44] to alleviate this problem. The loss function of the classification network is defined as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JTEHM.2022.3156851, IEEE Journal of Translational Engineering in Health and Medicine



follows:

$$L_{slide} = w_1 L_{bag} + w_2 L_{1_{instance}} + w_3 L_{2_{instance}} \quad (7)$$

Where $L_{bag}$ means cross entropy loss, $L_{1\ instance}$ means cross entropy loss and $L_{2\ instance}$ means dual focal loss. We use the weighted loss function as the objective function of network optimization to obtain better performance.

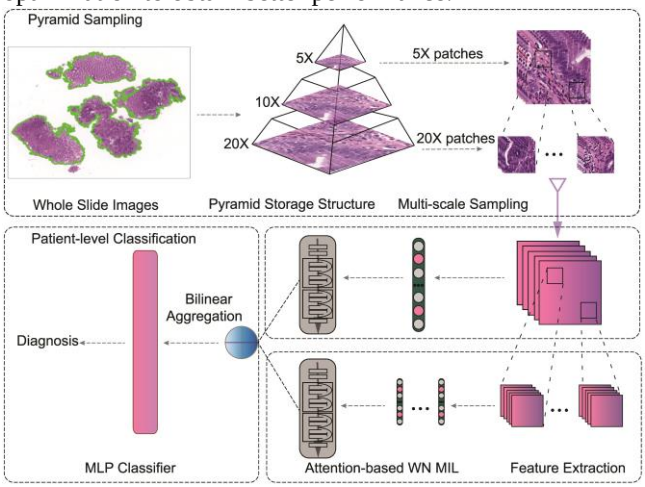


Fig. 2. Overview of our model. Patches extracted from each magnification of the WSIs are used for feature extractor separately. The trained feature extractors are used to compute embeddings of patches. The bilinear attention multi-scale feature fusion mechanism is used to connect WSIs embeddings of different scales to obtain the input of the final classifier.

## IV. Experiment and Results

### A. Data Description

We demonstrated our method and compared it with the standard method several times. Our locally advanced rectal cancer (LARC) dataset consists of 150 high-resolution WSIs from Yunnan cancer hospital. At present, the most commonly used clinical pathology evaluation criteria include TNM staging, histological subtype classification, tumor regression grade, pathological morphology, etc. In this study, the tumor regression grade was used as the qualitative standard for the treatment response of rectal cancer after neoadjuvant chemoradiotherapy. The data were first stained by H&E and then labeled by pathologists according to the regression grading system (TRG) of the American Joint Commission on Cancer (AJCC). There are four TRG groups: TRG 0, no residual tumor cells; TRG 1, single or small tumor remnant; TRG 2, part of the cancer tissue remains; TRG 3, a large number of cancer cells remain. Then, according to the AJCC TRG system, the treatment response is divided into two groups: good response (GR, TRG 0-1) and non-good response (non-GR, TRG 2-3). That is, in the model training stage, we use the pathological image of rectal cancer as the original data, extract the histopathological features such as the intensity, shape, and texture features of the tumor, and use the treatment response after neoadjuvant chemoradiotherapy as a label to establish a deep learning model. Then, the model can be used to predict the patient's response to neoadjuvant chemoradiotherapy accurately. We divided the training set, validation dataset, and test dataset with the ratio of 60%-20%-20%. The specific data are shown in TABLE I.

We also reported the performance of our model on the publicly available Camelyon16[11] and MSKCC[14] breast cancer metastasis detection datasets. The Camelyon16 dataset contains 399 WSIs, including 270 for training and 129 for testing. The MSKCC dataset contains 130 WSIs, including 100 for training and 30 for testing. We extracted two groups of pathological patches with an area of more than 10% from each WSIs according to 20× and 5× magnification for standby. In the Camelyon16 and MSKCC datasets, although the tumor area has complete pixel-level annotations on each slide, to verify the effectiveness of our weakly supervised model, we ignore pixel-level annotations during training and only consider slide-level labels. The specific data are shown in TABLE II and TABLE III.

TABLE I
THE LARC DATASET DISTRIBUTION

| | Total | GR | non-GR |
|---|---|---|---|
| Training | 90 | 44 | 46 |
| Validation | 30 | 15 | 15 |
| Testing | 30 | 15 | 15 |

TABLE II
THE CAMELYON16 DATASET DISTRIBUTION

| | Total | Normal | Tumor |
|---|---|---|---|
| Training | 270 | 159 | 111 |
| Testing | 129 | 80 | 49 |

TABLE III
THE MSKCC DATASET DISTRIBUTION

| | Total | Normal | Tumor |
|---|---|---|---|
| Training | 100 | 76 | 24 |
| Testing | 30 | 18 | 12 |

### B. Results

In this section, we briefly introduce the evaluation results of our proposed algorithm. For a comprehensive evaluation, we employed two standard metrics for evaluating classification quality: area under the receiver operating characteristic curve (AUC) and accuracy. We carefully conducted ablation experiments and observed positive results. The ablation experiment aims to explore the role of the gated attention weight normalization mechanism based on the multilayer perceptron, the multi-scale feature fusion method based on bilinear attention, and the weighted loss function in improving the performance of pathological images classification. We first conducted experiments on the prediction dataset of neoadjuvant chemoradiotherapy for rectal cancer collected from Yunnan Cancer Hospital. Beyond that, to verify the generalization ability of the model, we compare the proposed algorithm with CLAM[25], MIL-CE[22], MIL-RNN[14], DSMIL[45] four state-of-the-art algorithms are evaluated on Camelyon16 and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JTEHM.2022.3156851, IEEE Journal of Translational Engineering in Health and Medicine

MSKCC datasets. We reimplemented all the previous methods based on the literature and open source code. The backbone of the CNN used in our proposed model is ResNet18[43].

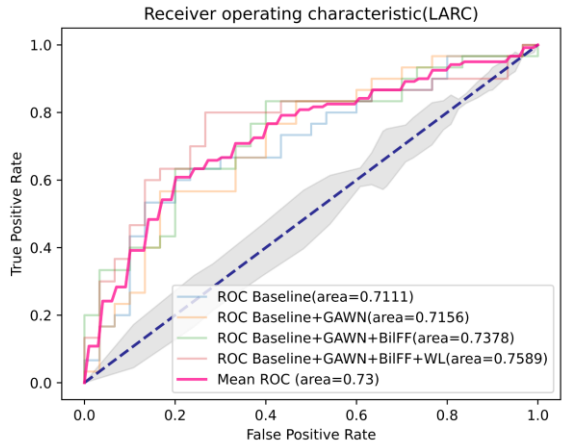


Fig. 3. Ablation experiment on LARC dataset, among them, GAWN represents the gated attention weight normalization mechanism, BiFF represents the bilinear attention multi-scale feature fusion mechanism, and WL represents the weighted loss function.

On the LARC dataset, we only use 90 pathological images as training data to obtain a 0.7589 AUC score on the test data set, achieving performance comparable to that of pathologists. Fig. 3 and TABLE IV shows the performance of our method on the LARC dataset. It can be seen from Fig. 4 that our model is easier to learn the best parameters than the traditional model, and the number of iterations is significantly reduced, which proves the effectiveness of our gated attention weight normalization mechanism.

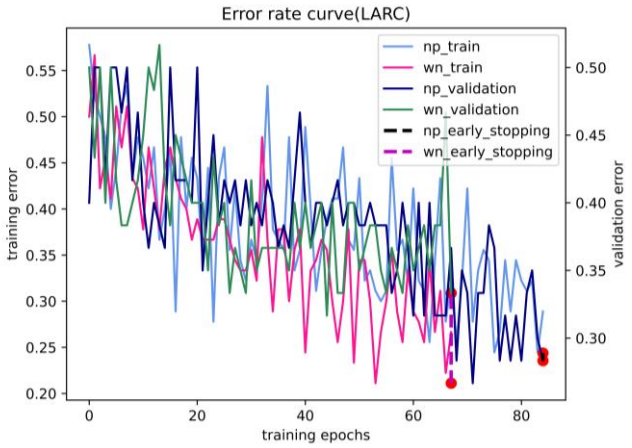


Fig. 4. The training and validation error for LARC using different network parameterizations. Where np and wn represent normal parameters and weight normalization parameters, respectively. The red dot is the early stop.

On the Camelyon16 and MSKCC datasets, by training only the slide-level labels, the AUC scores of our method on the test dataset reached 0.9337 and 0.9091, respectively. Among the other three models, DSMIL and MIL-RNN have the best performance, with AUC scores of 0.9225 and 0.8846, respectively.

TABLE IV
RESULTS ON LARC DATASET

| | | LARC | |
|---|---|---|---|
| | Scale | Accuracy | AUC |
| MIL-RNN[14] | Single | 0.6833 | 0.7044 |
| MIL-CE[22] | Single | **0.7000** | **0.7300** |
| DSMIL[45] | Single | 0.6982 | 0.7142 |
| CLAM[25] | Single | 0.6833 | 0.7111 |
| Ours | Single | 0.6667 | 0.7156 |
| DSMIL-LC[45] | Multiple | 0.7136 | 0.7358 |
| Ours | Multiple | **0.7167** | **0.7589** |

TABLE V
RESULTS ON CAMELYON16 DATASET

| | | Camelyon16 | |
|---|---|---|---|
| | Scale | Accuracy | AUC |
| MIL-RNN[14] | Single | 0. 8062 | 0. 8064 |
| MIL-CE[22] | Single | 0. 8000 | 0. 8400 |
| DSMIL[45] | Single | 0. 8682 | 0. 8944 |
| CLAM[25] | Single | 0. 8682 | 0. 8895 |
| Ours | Single | **0.8837** | **0.9163** |
| DSMIL-LC[45] | Multiple | 0. 8992 | 0. 9165 |
| Ours | Multiple | **0.9225** | **0.9337** |

TABLE VI
RESULTS ON MSKCC DATASET

| | | MSKCC | |
|---|---|---|---|
| | Scale | Accuracy | AUC |
| MIL-RNN[14] | Single | **0.8697** | **0.8944** |
| MIL-CE[22] | Single | 0.7250 | 0.7500 |
| DSMIL[45] | Single | 0.8532 | 0.8703 |
| CLAM[25] | Single | 0.8461 | 0.8636 |
| Ours | Single | 0.8646 | 0.8863 |
| DSMIL-LC[45] | Multiple | 0.8782 | 0.8956 |
| Ours | Multiple | **0.8846** | **0.9091** |

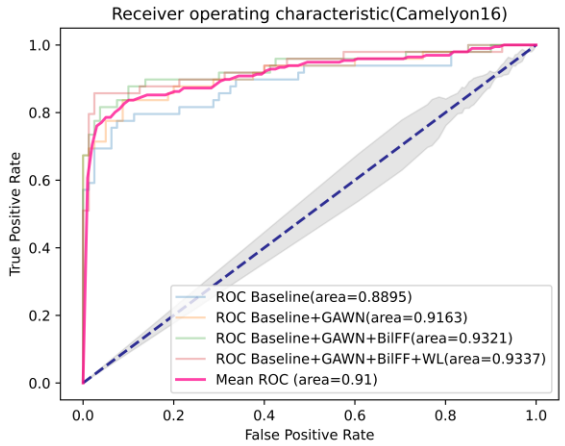


Fig. 5. Ablation experiment on Camelyon16 dataset, among them, GAWN represents the gated attention weight normalization mechanism, BiFF represents the bilinear attention multi-scale feature fusion mechanism, and WL represents the weighted loss function.

In contrast, our performance is quite competitive. Fig. 5 and Fig. 6 show the results of the ablation study of our algorithm on the two datasets of Camelyon16 and MSKCC.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JTEHM.2022.3156851, IEEE Journal of Translational Engineering in Health and Medicine



We observe that our method can accurately identify patches with tumor areas and give them higher attention weights (see TABLE V and TABLE VI). Compared with traditional attention-based multi-instance learning methods, our method accelerates the convergence of stochastic gradient descent optimization (see Fig. 7 and Fig. 8). Experiments show that our method is more advanced than other algorithms and has excellent performance on both private and publicly available datasets.

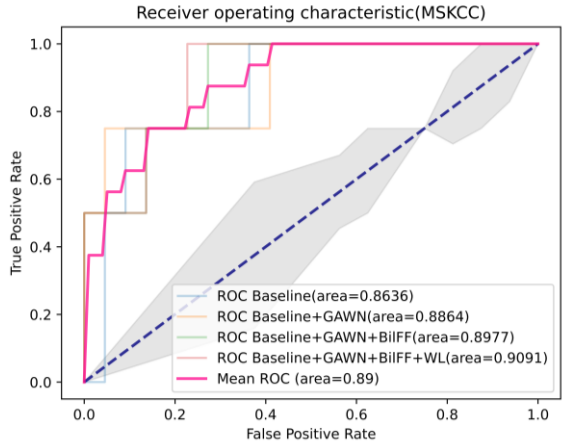


Fig. 6. Ablation experiment on MSKCC dataset, among them, GAWN represents the gated attention weight normalization mechanism, BiFF represents the bilinear attention multi-scale feature fusion mechanism, and WL represents the weighted loss function.

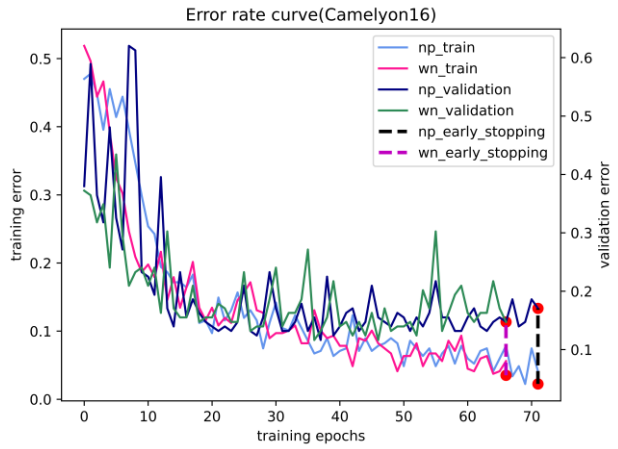


Fig. 7. The training and validation error for Camelyon16 using different network parameterizations. Where np and wn represent normal parameters and weight normalization parameters, respectively. The red dot is the early stop.

## V. Discussions and Conclusion

In this study, to simulate the actual diagnosis process of pathologists and provide personalized treatment plans for patients with rectal cancer, we propose a new classification framework for weakly-supervised pathological images. The framework only uses patient-level labels to predict the response to neoadjuvant chemoradiotherapy for rectal cancer, without the need for pathologists to manually outline the decision boundary between cancerous tissues and normal tissues, which greatly reduces the burden of labeling for pathologists.

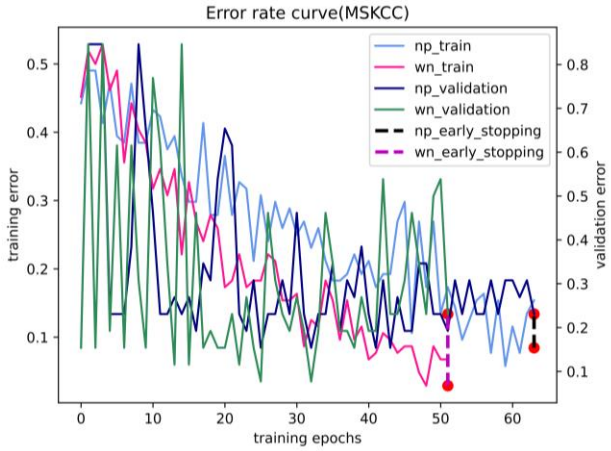


Fig. 8. The training and validation error for MSKCC using different network parameterizations. Where np and wn represent normal parameters and weight normalization parameters, respectively. The red dot is the early stop.

In addition, it can be seen from the result figures of the ablation experiment that each method we propose can effectively improve the performance of the model. Among them, the gated attention weight normalization mechanism speeds up the training process and effectively alleviates the problem of gradient diffusion. The bilinear attention multi-scale feature fusion mechanism can better integrate pathological image features, which can effectively alleviate the problem of global information loss caused by block sampling of WSIs. Finally, because the cancerous area in the pathological images is much larger than the normal area, it is easy to cause the imbalance of positive and negative instances after block sampling, and the weighted loss function just alleviates this problem. Experimental results show that our algorithm is superior to other weakly supervised learning algorithms on the private dataset and publicly available Camelyon16 and MSKCC datasets. Therefore, the method we propose is an effective method to predict the efficacy of neoadjuvant chemoradiotherapy for rectal cancer using histopathological images, and also has an excellent performance in breast cancer metastasis detection. In the future, we will further explore the feasibility of our method in other cancer prognosis studies and apply this technology to the clinic to assist clinicians in rapid diagnosis.

## Acknowledgment

We would like to acknowledge the financial support provided by the Chinese Natural Science Foundation, under Grant: 61876166 and 61663046; Yunnan Provincial Major Science and Technology Special Plan Project, under Grant: 202002AD080001; Yunnan Basic Research Program for Distinguished Young Youths Project, under Grant: 202101AV070003; Yunnan provincial major science and technology special plan projects: digitization research and application demonstration of Yunnan characteristic industry, under Grant: 202002AD080001; Yunnan Cancer Hospital and also by the Open Foundation of Key Laboratory in Software Engineering of Yunnan Province under Grant No. 2020SE304.

## REFERENCES


[1] Siegel, Rebecca L., et al. "Colorectal cancer statistics, 2020." CA: a cancer journal for clinicians 70.3 (2020): 145-164.

[2] Glynne-Jones, R., et al. "Rectal cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment, and follow-up." Annals of Oncology 28 (2017): iv22-iv40.

[3] Scott, N. A., et al. "Preoperative neoadjuvant therapy for curable rectal cancer – Reaching a consensus 2008." Colorectal Disease 11.3 (2009): 245-248.

[4] Sauer, Rolf, et al. "Adjuvant versus Neoadjuvant Radiochemotherapy for Locally Advanced Rectal Cancer A Progress Report of a Phase-III Randomized Trial (Protocol CAO/ARO/AIO-94)." Strahlentherapie und Onkologie 177.4 (2001): 173-181.

[5] Park, In Ja, et al. "Neoadjuvant treatment response as an early response indicator for patients with rectal cancer." Journal of clinical oncology 30.15 (2012): 1770.

[6] Patel, Uday B., et al. "Magnetic resonance imaging – detected tumor response for locally advanced rectal cancer predicts survival outcomes: MERCURY experience." Journal of Clinical Oncology 29.28 (2011): 3753-3760.

[7] Fokas, Emmanouil, et al. "Tumor regression grading after preoperative chemoradiotherapy as a prognostic factor and individual-level surrogate for disease-free survival in rectal cancer." JNCI: Journal of the National Cancer Institute 109.12 (2017): djx095.

[8] Sanghera, P., et al. "Chemoradiotherapy for rectal cancer: an updated analysis of factors affecting pathological response." Clinical Oncology 20.2 (2008): 176-183.

[9] Maas, Monique, et al. "Long-term outcome in patients with a pathological complete response after chemoradiation for rectal cancer: a pooled analysis of individual patient data." The lancet oncology 11.9 (2010): 835-844.

[10] Bedard, Philippe L., et al. "Tumour heterogeneity in the clinic." Nature 501.7467 (2013): 355-364.

[11] Bejnordi, Babak Ehteshami, et al. "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer." Jama 318.22 (2017): 2199-2210.

[12] van der Sommen, Fons, et al. "Machine learning in GI endoscopy: practical guidance in how to interpret a novel field." Gut 69.11 (2020): 2035-2045.

[13] Syed, Sana, and Ryan W. Stidham. "Potential for standardization and automation for pathology and endoscopy in inflammatory bowel disease." Inflammatory Bowel Diseases 26.10 (2020): 1490-1497.

[14] Campanella, Gabriele, et al. "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images." Nature medicine 25.8 (2019): 1301-1309.

[15] Wang, Xi, et al. "Weakly supervised deep learning for whole slide lung cancer image analysis." IEEE transactions on cybernetics 50.9 (2019): 3950-3962.

[16] Hou, Le, et al. "Patch-based convolutional neural network for whole slide tissue image classification." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[17] Li, Shaohua, et al. "Multi-instance multi-scale CNN for medical image classification." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2019.

[18] Yao, Jiawen, Xinliang Zhu, and Junzhou Huang. "Deep multi-instance learning for survival prediction from whole slide images." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2019.

[19] Chen, Hanbo, et al. "Rectified cross-entropy and upper transition loss for weakly supervised whole slide image classifier." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2019.

[20] Zhu, Xinliang, et al. "Wsisa: Making survival prediction from whole slide histopathological images." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[21] Xu, Yan, et al. "Parallel multiple instance learning for extremely large histopathology image analysis." BMC Bioinformatics 18.1 (2017): 1-15.

[22] Chikontwe, Philip, et al. "Multiple instance learning with center embeddings for histopathology classification." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2020.

[23] Xie, Chen, et al. "Beyond classification: Whole slide tissue histopathology analysis by end-to-end part learning." Medical Imaging with Deep Learning. PMLR, 2020.

[24] Salimans, Tim, and Durk P. Kingma. "Weight normalization: A simple reparameterization to accelerate training of deep neural networks." Advances in neural information processing systems 29 (2016): 901-909.

[25] Lu, Ming Y., et al. "Data-efficient and weakly supervised computational pathology on whole-slide images." Nature Biomedical Engineering 5.6 (2021): 555-570.

[26] Zhou, Xuezhi, et al. "Radiomics-based pre-therapeutic prediction of non-response to neoadjuvant therapy in locally advanced rectal cancer." Annals of surgical oncology 26.6 (2019): 1676-1684.

[27] Liu, Zhenyu, et al. "Radiomics analysis for evaluation of pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer." Clinical Cancer Research 23.23 (2017): 7253-7262.

[28] Shao, Lizhi, et al. "Multiparametric MRI and whole slide image-based pretreatment prediction of pathological response to neoadjuvant chemoradiotherapy in rectal cancer: a multicenter radiopathomic study." Annals of surgical oncology 27.11 (2020): 4296-4306.

[29] Skrede, Ole-Johan, et al. "Deep learning for prediction of colorectal cancer outcome: a discovery and validation study." The Lancet 395.10221 (2020): 350-360.

[30] Zhang, Fang, et al. "Predicting treatment response to neoadjuvant chemoradiotherapy in local advanced rectal cancer by biopsy digital pathology image features." Clinical and translational medicine 10.2 (2020): e110.

[31] Kather, Jakob Nikolas, et al. "Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study." PLoS medicine 16.1 (2019): e1002730.

[32] Kather, Jakob Nikolas, et al. "Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer." Nature medicine 25.7 (2019): 1054-1056.

[33] Bychkov, Dmitrii, et al. "Deep learning based tissue analysis predicts outcome in colorectal cancer." Scientific Reports 8.1 (2018): 1-11.

[34] Sudharshan, P. J., et al. "Multiple instance learning for histopathological breast cancer image classification." Expert Systems with Applications 117 (2019): 103-111.

[35] Kim, Jin-Hwa, Jaehyun Jun, and Byoung-Tak Zhang. "Bilinear attention networks." arXiv preprint arXiv:1805.07932 (2018).

[36] Yang, Yun, et al. "Two-Stage Selective Ensemble of CNN via Deep Tree Training for Medical Image Classification." IEEE Transactions on Cybernetics (2021).

[37] Yang, Yun, and Jianmin Jiang. "Adaptive bi-weighting toward automatic initialization and model selection for HMM-based hybrid meta-clustering ensembles." IEEE transactions on cybernetics 49.5 (2018): 1657-1668.

[38] Yang, Yun, and Jianmin Jiang. "Bi-weighted ensemble via HMM-based approaches for temporal data clustering." Pattern Recognition 76 (2018): 391-403.

[39] Yang, Yun, et al. "Multi-source transfer learning via ensemble approach for initial diagnosis of alzheimer's disease." IEEE Journal of Translational Engineering in Health and Medicine 8 (2020): 1-10.

[40] Zhao, Yu, et al. "Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[41] Wang, Shujun, et al. "RMDL: Recalibrated multi-instance deep learning for whole slide gastric image classification." Medical image analysis 58 (2019): 101549.

[42] Mormont, Romain, Pierre Geurts, and Raphaël Marée. "Comparison of deep transfer learning strategies for digital pathology." Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2018.

[43] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[44] Sazzad Hossain, Md, Andrew P. Paplinski, and John M. Betts. "Adaptive Class Weight based Dual Focal Loss for Improved Semantic Segmentation." arXiv e-prints (2019): arXiv-1909.

[45] Li, Bin, Yin Li, and Kevin W. Eliceiri. "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.