THERAPEUTIC ADVANCES in
Gastroenterology

Original Research

Development and validation of a deep learning-based approach to predict the Mayo endoscopic score of ulcerative colitis

Ther Adv Gastroenterol

2023, Vol. 16: 1–13

DOI: 10.1177/17562848231170945

Jing Qi*, Guangcong Ruan*, Yi Ping, Zhifeng Xiao, Kaijun Liu, Yi Cheng, Rongbei Liu, Bingqiang Zhang, Min Zhi, Junrong Chen, Fang Xiao, Tingting Zhao, Jiaxing Li, Zhou Zhang, Yuxin Zou, Qian Cao, Yongjian Nian and Yanling Wei

Abstract

Background: The ulcerative colitis (UC) Mayo endoscopy score is a useful tool for evaluating the severity of UC in patients in clinical practice.

Objectives: We aimed to develop and validate a deep learning-based approach to automatically predict the Mayo endoscopic score using UC endoscopic images.

Design: A multicenter, diagnostic retrospective study.

Methods: We collected 15120 colonoscopy images of 768 UC patients from two hospitals in China and developed a deep model based on a vision transformer named the UC-former. The performance of the UC-former was compared with that of six endoscopists on the internal test set. Furthermore, multicenter validation from three hospitals was also carried out to evaluate UC-former's generalization performance.

Results: On the internal test set, the areas under the curve of Mayo 0, Mayo 1, Mayo 2, and Mayo 3 achieved by the UC-former were 0.998, 0.984, 0.973, and 0.990, respectively. The accuracy (ACC) achieved by the UC-former was 90.8%, which is higher than that achieved by the best senior endoscopist. For three multicenter external validations, the ACC was 82.4%, 85.0%, and 83.6%, respectively.

Conclusions: The developed UC-former could achieve high ACC, fidelity, and stability to evaluate the severity of UC, which may provide potential application in clinical practice.

Registration: This clinical trial was registered at the ClinicalTrials.gov (trial registration number: NCT05336773)

Plain language summary

Why was this study done?

The development of an auxiliary diagnostic tool can reduce the workload of endoscopists and achieve rapid assessment of ulcerative colitis (UC) severity.

What did the researchers do?

We developed and validated a deep learning-based approach to automatically predict the Mayo endoscopic score using UC endoscopic images.

What did the researchers find?

The model that was developed in this study achieved high accuracy, fidelity, and stability, and demonstrated potential application in clinical practice.

What do the findings mean?

Deep learning could effectively assist endoscopists in evaluating the severity of UC in patients using endoscopic images.

Keywords: ulcerative colitis, Mayo endoscopy score, deep learning, vision transformer

Received: 16 November 2022; revised manuscript accepted: 4 April 2023.

Correspondence to:
Yanling Wei
Department of Gastroenterology, Chongqing Key Laboratory of Digestive Malignancies, Daping Hospital, Army Medical University (Third Military Medical University), 10 Changjiang Branch Road, Chongqing, 400042, China
lingzi016@126.com

Yongjian Nian
Department of Digital Medicine, School of Biomedical Engineering and Imaging Medicine, Army Medical University (Third Military Medical University), Chongqing, 400038, China
yjnian@tmmu.edu.cn

Qian Cao
Department of Gastroenterology, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, 310016, China
caoq@zju.edu.cn

Jing Qi
Yuxin Zou
Department of Digital Medicine, School of Biomedical Engineering and Imaging Medicine, Army Medical University, Chongqing, China

Guangcong Ruan
Yi Ping
Zhifeng Xiao
Kaijun Liu
Yi Cheng
Department of Gastroenterology, Chongqing Key Laboratory of Digestive Malignancies, Daping Hospital, Army Medical University (Third Military Medical University), Chongqing, China

Rongbei Liu
Zhou Zhang
Department of Gastroenterology, Sir Run Run Shaw Hospital,

Zhejiang University School of Medicine, Hangzhou, China

Bingqiang Zhang
Department of Gastroenterology, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China

Min Zhi
Junrong Chen
Department of Gastroenterology, Guangdong Provincial Key Laboratory of Colorectal and Pelvic Floor Diseases, The Sixth Affiliated Hospital of Sun Yat-Sen University, Guangzhou, China

Fang Xiao
Department of Gastroenterology, Tongji Hospital of Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

Tingting Zhao
Jiaxing Li
School of Basic Medicine, Army Medical University (Third Military Medical University), Chongqing, China

*The authors contributed equally.

Introduction

Ulcerative colitis (UC) is a type of inflammatory bowel disease (IBD) characterized by chronic inflammation and ulcers in the colon and rectum on endoscopy. Common symptoms include diarrhea, abdominal pain, and cramps.[1] The worldwide incidence of UC is increasing, especially in newly industrialized countries. The annual incidence of UC ranges from 8.8 to 23.1 per 100,000 person-years in North America, 1.7 to 57.9 per 100,000 person-years in Northern Europe, and 7.3 to 17.4 in Oceania.[2,3] Endoscopy is the core basis of the current management of UC.[4,5] Endoscopic remission is both a long-term treatment target in daily clinical practice and a key component in clinical trials for the regulatory approval of novel therapeutic agents.[6,7] Accurate evaluation of the severity of UC in patients is helpful for guiding clinical decision-making.[8] Note that the Mayo endoscopic score is the most prevalent, with details of image features collected during endoscopy and clinical symptoms reported by patients, of which the UC endoscopic score dominates.[9] Currently, the Mayo endoscopic score is widely employed as the standard of severity of UC under endoscopy; however, large errors are possible due to the subjective consciousness and lack of experience attributed to clinical evaluations of physicians, which may delay disease diagnosis and even miss the period of changing to the best treatment.[10] Moreover, endoscopic evaluation requires training, and the evaluation results often differ between two endoscopists. Thus, validated endoscopic indices enabling standardized, reproducible, and uniform reporting are essential for clinical practice and comprise an integral part of the clinical trial landscape.

Recent studies have suggested roles for artificial intelligence and deep learning in various fields including endoscopic field. For example, the development of deep learning-based auxiliary diagnostic tools to identify lesions in endoscopic images or videos has become a major research focus.[11,12] Presently, deep learning technology has been gradually applied to the diagnosis of digestive system diseases due to its advantages of high efficiency and accuracy (ACC), such as the classification of gastric cancer invasion degree[13] and the identification of intestinal diseases,[14] showing good diagnostic performance and significantly reducing the work intensity of endoscopists. Ozawa[15] was the first to evaluate the performance of a convolutional neural network (CNN) in UC image recognition, the developed model based on GoogLeNet achieved an area under the curve (AUC) of 0.98 in distinguishing Mayo 0 or Mayo 1 disease from Mayo 2 or Mayo 3 disease; however, the performance consistency between the deep model and endoscopists was not assessed. Subsequently, Stidham et al.[16] demonstrated that their model performed similarly to experienced human reviewers. Considering that UC severity ratings have not been analyzed at the individual level, Bhambhvani and Zamora[17] conducted a three-level classification study (Mayo 1, Mayo 2, and Mayo 3) to explore the utility of CNNs in grading UC, but their sample size was too small with only 777 endoscopic images being included. Becker et al.[18] trained ResNet50 to perform multiple binary tasks; however, this study lacked four-level classification tasks. Note that these existing studies are limited to existing CNNs and generally lack validation of model generalization. Furthermore, the finite acceptance domain of the convolution operator makes it difficult to model remote dependence, and its static weight cannot flexibly adapt to the input content.[19] A vision transformer (ViT),[20] composed of attention modules, compensates for the above deficiency of CNNs and performs as well as or even better than CNNs in many computers' vision tasks, such as classification, segmentation and object detection. Since Mayo endoscopic scoring of UC images is a more fine-grained classification task, and lesions with different classes are usually similar, it can easily cause confusion. Fortunately, ViT is able to capture more discriminative feature information from UC endoscopic images by modeling remote dependencies, which is beneficial for improving the classification performance. Note that Qi proposed a pyramid hybrid feature fusion framework to predict Mayo endoscopic score,[21] which had a dual-branch hybrid architecture with ResNet50 and a pyramid ViT; its disadvantage was the lack of comprehensive performance evaluation for the proposed framework. In this study, to assist in evaluating the severity of UC in patients, the UC-former was developed based on ViT with a constructed loss function to predict the Mayo endoscopic score using UC endoscopic images, and both internal validation and multicenter validation were introduced to evaluate the performance of the UC-former; moreover, the performance of the UC-former was also compared with that of endoscopists. Experimental results demonstrated the efficiency of the UC-formers in assessing UC severity.

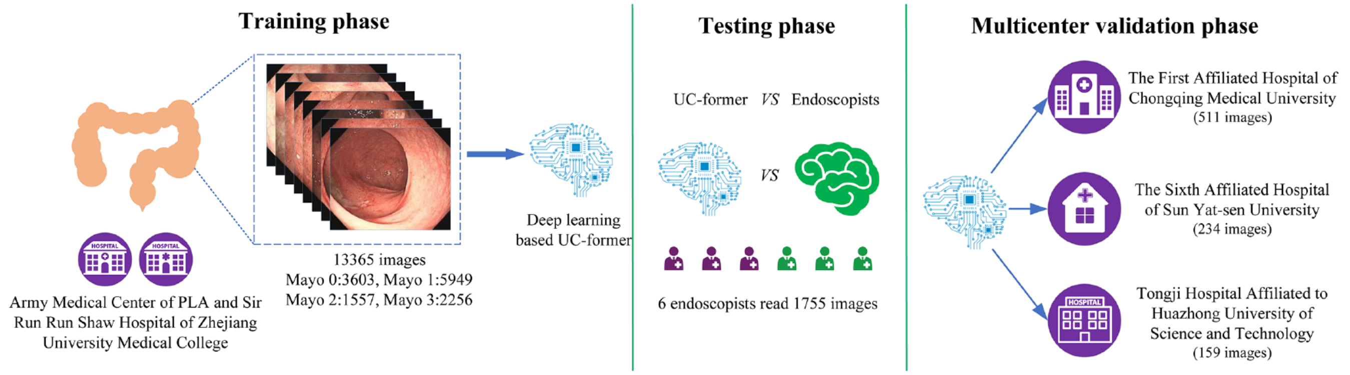
Training phase
Army Medical Center of PLA and Sir Run Run Shaw Hospital of Zhejiang University Medical College
13365 images
Mayo 0:3603, Mayo 1:5949
Mayo 2:1557, Mayo 3:2256
Deep learning based UC-former
Testing phase
UC-former VS Endoscopists
VS
6 endoscopists read 1755 images
Multicenter validation phase
The First Affiliated Hospital of Chongqing Medical University (511 images)
The Sixth Affiliated Hospital of Sun Yat-sen University (234 images)
Tongji Hospital Affiliated to Huazhong University of Science and Technology (159 images)

Figure 1. Graphic abstract of the study.

Methods

Study design

This multicenter, diagnostic retrospective study was carried out in five hospitals in China. Patients with UC who underwent colonoscopy between 1 January 2018 and 31 December 2021 were identified from the Daping Hospital of Army Medical University (Army Medical Center of PLA), Sir Run Run Shaw Hospital of Zhejiang University, The First Affiliated Hospital of Chongqing Medical University, The Sixth Affiliated Hospital of Sun Yat-Sen University, and Tongji Hospital of Huazhong University of Science and Technology (Figure 1).

This study was approved by the Ethics Committee of Army Medical Center of PLA and was performed according to the Declaration of Helsinki. For patients whose colonoscopy images were stored in retrospective databases at each participating hospital, informed consent was exempted by the institutional review boards of the participating hospitals. The study protocol was approved by the clinical trial (ClinicalTrials.gov, ID: NCT05336773). The reporting of this study conforms to the STROBE statement.[22] In order to protect the patient's privacy and rights, we have ensured that the patient's personal information is deleted so that the identity of the patient may not be ascertained in any way.

Patients

Subjects were patients aged 18–72 years who had UC, and UC disease activity was assessed using the Mayo endoscopic score. The clinical manifestations of the enrolled patients with UC showed typical lesions. Patients with IBD unclassified were excluded. There were no exclusion criteria at the level of the input datasets.

Image quality control and dataset

All colonoscopy examinations were performed, usually with the patient under sedation, by well-trained endoscopists from the gastroenterology department using high-definition colonoscopes (CV290SL, Olympus Medical Systems, Tokyo, Japan). Colonoscopy records included a written description and a scheme representing the colon where the different lesions (frank erythema, aphtha, superficial and deep ulcerations, pseudopolyp, and stenosis) were displayed for each colonic segment (rectum and sigmoid, descending, transverse, and ascending colon).

We collected 15120 images of 768 patients from Army Medical Center of PLA (Chongqing, China) and Sir Run Run Shaw Hospital of Zhejiang University (Zhejiang, China) from January 2018 to December 2021. The dataset was randomly divided, including 13365 images (671 cases) in the training set and 1755 images (97 cases) in the internal test set (Supplementary Table 1).

In addition, 511 images (42 cases) from The First Affiliated Hospital of Chongqing Medical University, 234 images (45 cases) from The Sixth Affiliated Hospital of Sun Yat-Sen University, and 159 images (11 cases) from Tongji Hospital affiliated to Huazhong University of Science and Technology were selected as the external test sets.

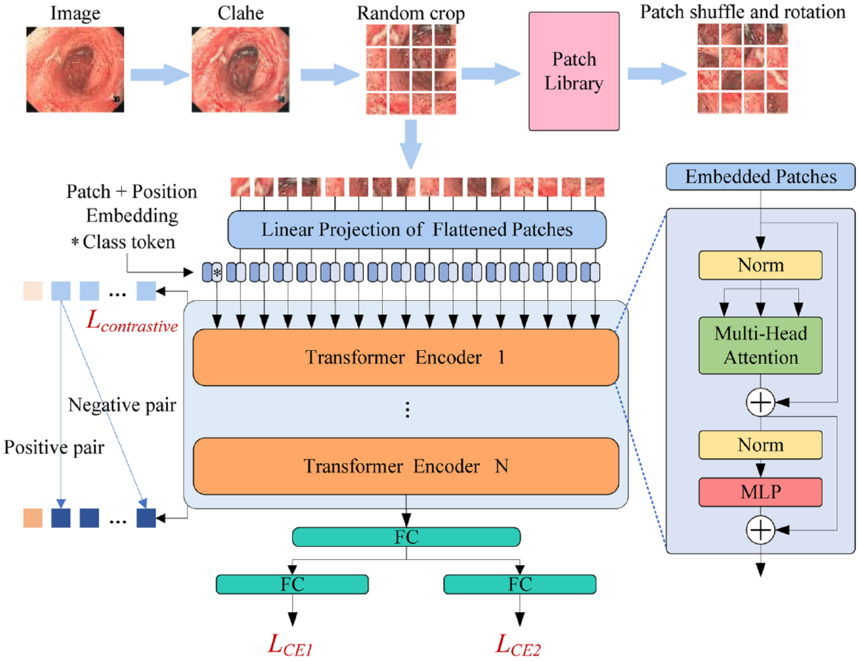

Figure 2. Structure diagram of the developed UC-former.
UC, ulcerative colitis.

Annotation of Mayo endoscopic score

For all collected UC endoscopic images, the Mayo endoscopic score of each image was independently annotated by two endoscopic experts with more than 20 years of working experience. If the labels of the identical image were not consistent between the two experts, the third expert was invited to assist in collectively making a final decision. Mayo endoscopic score annotated by the three endoscopic experts were considered as the Ground Truth labels in this study.

Training process

The structure diagram of the developed UC-former is shown in Figure 2. Firstly, the contrast-limited adaptive histogram equalization (Clahe) algorithm was utilized to highlight the features of each image. Second, we randomly cropped a 224 × 224 area, which was input into the ViT network after data enhancement operations such as random flipping and random brightness jitter. ViT decomposed the input image into a series of patches, linearly embedded each patch, added location information, and then fed the resulting vector sequence to the transformer encoder. In addition, an additional learnable class token was added to the vector sequence for subsequent image classification. The structure of each transformer encoder, consisting of two layerNorm layers, a multihead attention module, and a multilayer perceptron module, is shown on the right side of Figure 2.

Instead of dividing the image into 16 patches in the schematic, we divided each 224 × 224 image into 196 (14 × 14) patches, and the size of each patch was 16 × 16. To compensate for the class imbalances in the dataset while accommodating the input characteristics of ViT, we built a patch library for the category with a small sample size, which contained all the patches for splitting all images of that category in the training set. We then randomly selected patches in the patch library, rotated them, and rearranged them to generate new images. These newly generated images were added to the training set; they recombined the different image patches of that category, which was beneficial to improving the generalization of the deep model.

To improve the training efficiency, in addition to adopting the transfer learning strategy, we employed three loss functions to encourage the deep model to learn as many correlations between two patches as possible to improve its ability to assess the severity of UC. Since the classification of Mayo 1 and Mayo 2 is less consistent with the evaluation of endoscopists,[16] we combined Mayo 0 and Mayo 1 into one category (mild) and Mayo 2 and Mayo 3 into another category (severe). We expected the classification of both mild and severe, which was used to assist the final four-level classification task, to be as accurate as possible, and their loss was applied to update the network parameters. The cross-entropy loss functions between the ground truth labels and the predicted probabilities are given as follows

$$L_{\mathrm{CE1}} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(p_i) + (1-y_i)\log(1-p_i) \qquad (1)$$

$$L_{\mathrm{CE2}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{K} y_{ic} \log(p_{ic}) \qquad (2)$$

where L_{CE1} and L_{CE2} represent the cross-entropy loss of the auxiliary task and that of the primary task, respectively; y_i and y_{ic} represent the true label; p_i and p_{ic} represent the corresponding predicted probability; K represents the number of categories; and N represents the sample size of a batch.

Note that the output of each transformer encoder in ViT, which contains the high-dimensional features of 196 patches, has the same size. Inspired by Gong's work,[23] we hope that the shallow features and deep features in the training process should be as similar as possible for the identical patch and as different as possible for different patches so that the representation of patches can be diversified. The UC-former employs a contrast loss function to achieve the above purpose. Let $q^{[1]}$ and $q^{[G]}$ denote the output patch set of the first layer and last layers, respectively, and let Q denote the number of patches. Next, the contrast loss function can be expressed as follows

$$L_{\mathrm{contrastive}} = -\frac{1}{Q}\sum_{m=1}^{Q} \log \frac{\exp(q_m^{[1]^{\mathrm{T}}} q_m^{[G]})}{\exp(q_m^{[1]^{\mathrm{T}}} q_m^{[G]}) + \exp\left(q_m^{[1]^{\mathrm{T}}} \left(\frac{1}{Q-1} \sum\nolimits_{m\neq n} q_n^{[G]} \right)\right)} \qquad (3)$$

We trained the UC-former by simply minimizing the constructed loss function which was obtained by weighting the above three loss functions as follows

$$L = \alpha L_{\mathrm{CE1}} + \beta L_{\mathrm{CE2}} + \gamma L_{\mathrm{contrastive}} \qquad (4)$$

where $\alpha + \beta + \gamma = 1$. In particular, the values of α, β, and γ were set to 0.5, 0.4, and 0.1, respectively. Note that 10-fold cross-validation was repeated five times on the training set to tune the hyperparameters of the UC-former. This task was carried out based on the PyCharm 2020.1.3 platform, with PyTorch framework version 1.7.0 and Python version 3.8.5. Eight NVIDIA GeForce RTX 2080 Ti graphics cards were used to train the deep network. During the process of model training, the initial learning rate was set to 0.01, which was adjusted by the exponential attenuation method. Stochastic gradient descent optimizer with the momentum of 0.9 and the weight decay of 1×10^{-5} was used to train the deep model for maximally 50 epochs.

Test process

Once the UC-former was trained, 1755 images from the internal test set were utilized to evaluate its performance. According to the standard definition, we calculated the AUC, ACC, sensitivity (SEN), specificity (SPE), positive prediction value (PPV), and negative prediction value (NPV). To better understand the decision-making basis of the network, heatmaps produced by the method in the study by Chefer et al.,[24] which can intuitively show the more important parts related to the decision-making, were given. In addition, the attention maps in the multihead attention modules were extracted to observe patch areas that the model paid more attention to as the network deepened.

The performance of the UC-former was also compared with that of six endoscopists, including three junior endoscopists with approximately 7 years of working experience and three senior endoscopists with more than 15 years of working experience. All endoscopists have experienced professional training in colonoscopy diagnosis and can skillfully operate colonoscopy, judge intestinal mucosal lesions, accurately write endoscopic diagnosis and collect endoscopic images. All six endoscopists independently scored each UC image on the internal test set. Moreover, to evaluate the generalization of the UC-former, external datasets collected from three hospitals were employed.

Statistical analysis

We chose ACC, SEN, SPE, PPV, and NPV to evaluate the classification performance. The

Table 1. Clinical statistical characteristics in the training and test set.

Patients	Training set	Test set
Males, *n* (%)	373 (55.59%)	56 (57.73%)
Age, median	50	47
BBPS (Boston Bowel Preparation Scale) (mean, median, range)	6.75, 7, (6–9)	6.72, 6, (6–9)
Disease duration, years, mean (SD)	2.58 (1.24)	2.58 (1.24)
Diarrhea, *n* (%)	346 (78.64%)	61 (86.11%)
Abdominal pain (%)	322 (73.18%)	56 (77.78%)

SD, standard deviation.

McNemar test was used to assess significant differences among ACC, SEN, and SPE, while the Chi-square test was applied to the PPV and NPV, wherein a *p* value less than 0.05 was considered to be a significant difference. The 95% Wilson confidence interval was applied for each evaluating indicator. All statistical analyses were implemented by IBM SPSS statistical software 25.0.

Results

Patient enrollment

Between January 2018 and December 2021, 15,120 images from 768 patients were obtained from two hospitals. Overall, 13,365 endoscopic images from 671 patients were selected to build the UC-former, and 1755 endoscopic images from 97 patients were selected to evaluate its performance. Moreover, the performance of the UC-former was also compared with that of six endoscopists. Colonoscopy images from 98 patients in three hospitals were employed as the external datasets to evaluate the generalization capability (Supplementary Figure 1). After randomized allocation, the four groups of patient and image characteristics in the training and test sets had similar background data regarding age, sex, bowel preparation, and clinical severity (Table 1).

Performance of UC-former

The average results of 10-fold cross-validation repeated 5 times on the training set is shown in Table 2, with the overall ACC reaching 0.871 (95% CI, 0.865–0.876). In the internal test, we plotted the confusion matrix and the receiver operating characteristic (ROC) curves for Mayo 0, Mayo 1, Mayo 2, and Mayo 3 (Figure 3), where the corresponding AUCs for each level were 0.998 (95% CI, 0.995–1.000), 0.984 (95% CI, 0.978–0.991), 0.973 (95% CI, 0.958–0.989), and 0.990 (95% CI, 0.981–0.998), respectively (Supplementary Table 2).

The comparison of performance between the UC-former and the endoscopists is shown in Table 3 and Figure 4. The classification performance varies for different Mayo endoscopic scores, whether they are UC-formers or endoscopists. Notably, the performance of senior endoscopists was significantly higher than that of junior endoscopists in terms of overall ACC. Note that the overall ACC of the UC-former was 0.908 (95%CI, 0.893–0.920), which is much higher than that of the best senior endoscopist [0.773 (95%CI, 0.753–0.792)] and the best junior endoscopist [0.849 (95%CI, 0.831–0.865)], both with significant differences.

Feature visualization is widely employed to explore the working mechanism and judgment basis of deep networks. The feature maps in the multihead attention module in each transformer encoder are shown in Supplementary Figure 2. For the UC-former, the number of multiheads was 12, and each head was used to extract different correlations between two patches, so for each transformer encoder, there were 12 feature maps. In the shallow layer, the features learned by the UC-former were scattered, while as the network deepened, the features became increasingly focused, and the patches with lesions appeared to be highlighted. Heatmaps generated by the UC-former are shown in Figure 5; they partly explain the classification results achieved by the UC-former.

Multicenter validation

The UC-former showed better generalization performance on the external test sets (Table 4). In the First Affiliated Hospital of Chongqing Medical University, the Sixth Affiliated Hospital of Sun Yat-Sen University, and Tongji Hospital, the overall ACC was 0.824 (95% CI, 0.788–0.854), 0.850 (95% CI, 0.799–0.890), and 0.836 (95% CI, 0.771–0.886), respectively. The SEN and SPE of all categories were higher than 0.8 in

Table 2. Results of 10-fold cross-validation on the training set.

ACC (95% CI)	Level	SEN (95% CI)	SPE (95% CI)	PPV (95% CI)	NPV (95% CI)	AUC (95% CI)
0.871 (0.865–0.876)	Mayo 0	0.973 (0.967–0.978)	0.933 (0.928–0.938)	0.843 (0.832–0.854)	0.989 (0.987–0.991)	0.990 (0.988–0.992)
	Mayo 1	0.886 (0.878–0.894)	0.902 (0.895–0.909)	0.879 (0.870–0.887)	0.908 (0.901–0.915)	0.964 (0.961–0.968)
	Mayo 2	0.933 (0.919–0.944)	0.888 (0.882–0.893)	0.523 (0.504–0.541)	0.990 (0.988–0.992)	0.971 (0.965–0.977)
	Mayo 3	0.954 (0.945–0.962)	0.933 (0.929–0.938)	0.744 (0.728–0.760)	0.990 (0.988–0.992)	0.987 (0.983–0.990)

ACC, accuracy; CI, confidence interval; NPV, negative prediction value; PPV, positive prediction value; SEN, sensitivity; SPE, specificity.

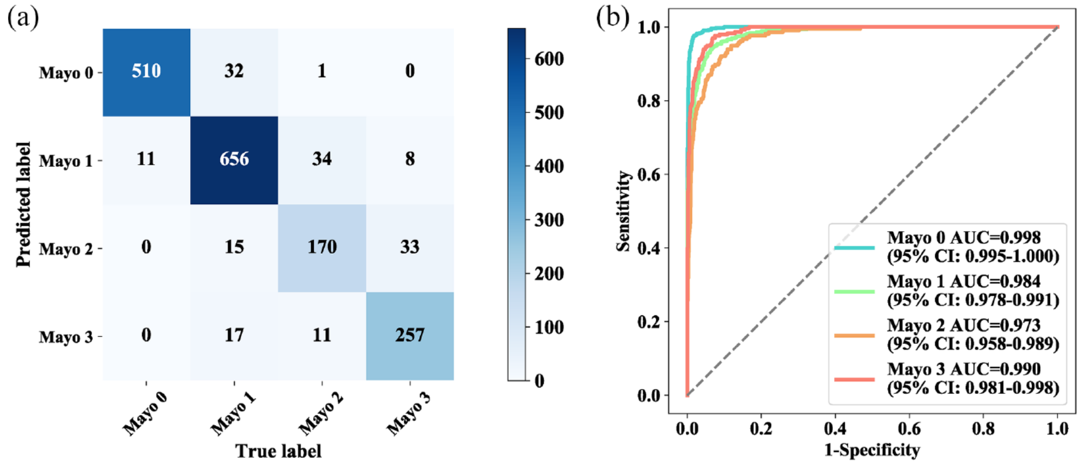

Figure 3. Confusion matrix and ROC curves for UC-former. Confusion matrix (a) and ROC curves (b) for the test results of four groups.

AUC, the area under the receiver operating characteristic curve; ROC, receiver operating characteristic; UC, ulcerative colitis.

Table 3. Comparison of classification performance between UC-former and endoscopists.

	UC-former	Highest junior endoscopist	*p* Value	Highest senior endoscopist	*p* Value
Image analysis					
ACC (95% CI)	0.908 (0.893–0.920)	0.773 (0.753–0.792)	$p<0.001$	0.849 (0.831–0.865)	$p<0.001$
Mayo 0					
SEN (95% CI)	0.977 (0.960–0.987)	0.766 (0.728–0.800)	$p<0.001$	0.766 (0.728–0.800)	$p<0.001$
SPE (95% CI)	0.982 (0.973–0.988)	0.987 (0.979–0.992)	$p=0.016$	0.998 (0.994–1.000)	$p<0.001$
PPV (95% CI)	0.959 (0.938–0.972)	0.961 (0.938–0.976)	$p=0.122$	0.995 (0.982–0.999)	$p<0.001$
NPV (95% CI)	0.990 (0.983–0.994)	0.909 (0.892–0.923)	$p<0.001$	0.910 (0.893–0.924)	$p<0.001$
Mayo 1					
SEN (95% CI)	0.947 (0.928–0.961)	0.749 (0.716–0.779)	$p<0.001$	0.872 (0.846–0.895)	p=0.015
SPE (95% CI)	0.929 (0.912–0.944)	0.863 (0.840–0.882)	$p<0.001$	0.870 (0.848–0.889)	$p<0.001$

(Continued)

Table 3. (Continued)

	UC-former	Highest junior endoscopist	*p* Value	Highest senior endoscopist	*p* Value
PPV (95% CI)	0.903 (0.880–0.922)	0.791 (0.759–0.820)	$p<0.001$	0.823 (0.794–0.849)	$p<0.001$
NPV (95% CI)	0.962 (0.948–0.972)	0.831 (0.808–0.853)	$p<0.001$	0.907 (0.888–0.924)	$p=0.007$
Mayo 2					
SEN (95% CI)	0.940 (0.900–0.964)	0.676 (0.611–0.735)	$p=0.013$	0.898 (0.851–0.932)	$p=0.001$
SPE (95% CI)	0.891 (0.874–0.905)	0.914 (0.898–0.927)	$p<0.001$	0.941 (0.928–0.952)	$p<0.001$
PPV (95% CI)	0.547 (0.496–0.597)	0.523 (0.465–0.581)	$p<0.001$	0.681 (0.624–0.732)	$p=0.014$
NPV (95% CI)	0.991 (0.984–0.995)	0.953 (0.941–0.962)	$p=0.013$	0.985 (0.977–0.990)	$p=0.006$
Mayo 3					
SEN (95% CI)	0.977 (0.952–0.989)	0.916 (0.879–0.943)	$p=0.056$	0.903 (0.864–0.931)	$p=0.148$
SPE (95% CI)	0.926 (0.911–0.938)	0.927 (0.912–0.939)	$p<0.001$	0.975 (0.965–0.982)	$p=0.314$
PPV (95% CI)	0.729 (0.684–0.771)	0.718 (0.671–0.761)	$p<0.001$	0.879 (0.838–0.911)	$p=0.379$
NPV (95% CI)	0.995 (0.989–0.997)	0.982 (0.973–0.988)	$p=0.086$	0.980 (0.971–0.986)	$p=0.164$

ACC, accuracy; CI, confidence interval; NPV, negative prediction value; PPV, positive prediction value; SEN, sensitivity; SPE, specificity; UC, ulcerative colitis.

the three multicenter datasets. Note that the PPV of Mayo 2 was low in the results of the First Affiliated Hospital of Chongqing Medical University, possibly because the number of images belonging to Mayo 2 was too small, accounting for only 7.828% in the entire dataset. The confusion matrix and ROC curves of multicenter validation were shown in Figure 6.

Discussion

Artificial intelligence techniques have been widely utilized in an increasing number of medical fields, such as the accurate diagnosis of pathology, ultrasound, and cardiac imaging.[25,26] In the field of diagnosis of UC, it has been reported that artificial intelligence is introduced to evaluate the histological activity of UC.[12] In the clinical treatment of UC, recovery under endoscopy can be regarded as the gold standard to evaluate the efficiency of clinical treatment. Manual interpretation is limited to experienced clinicians who also exhibit great differences. A few clinical studies have evaluated the severity of UC under endoscopy. A recent study proposed UC-DenseNet can effectively diagnose UC and assist the endoscopist in formulating the treatment strategy.[27] However, it is still necessary to be validated by the multicenter dataset in clinical practice, and compared with endoscopists in order to more effectively demonstrate the efficiency of the deep model in clinical experiments.

In our study, based on retrospective stored images, the developed UC-former demonstrated high ACC, SEN and SPE for the prediction of the Mayo endoscopic score of UC images. As a kind of supervised learning, accurate annotation of Mayo endoscopic score for each UC image is crucial for training an effective deep model. In this study, a more reliable annotation mode was introduced. Two experts were invited to annotate all UC images independently, and the participation of the third expert mainly solved those inconsistent labels for the identical image annotated by the first two experts. Note that it is also a popular annotation mode to extract keywords from diagnostic reports using natural language processing, which could significantly reduce the annotation workload; however, it inevitably introduces annotation noise. In contrast to previous efforts using deep learning to grade UC severity, we adopted a more advanced deep learning architecture, that is, ViT. Furthermore, we designed auxiliary

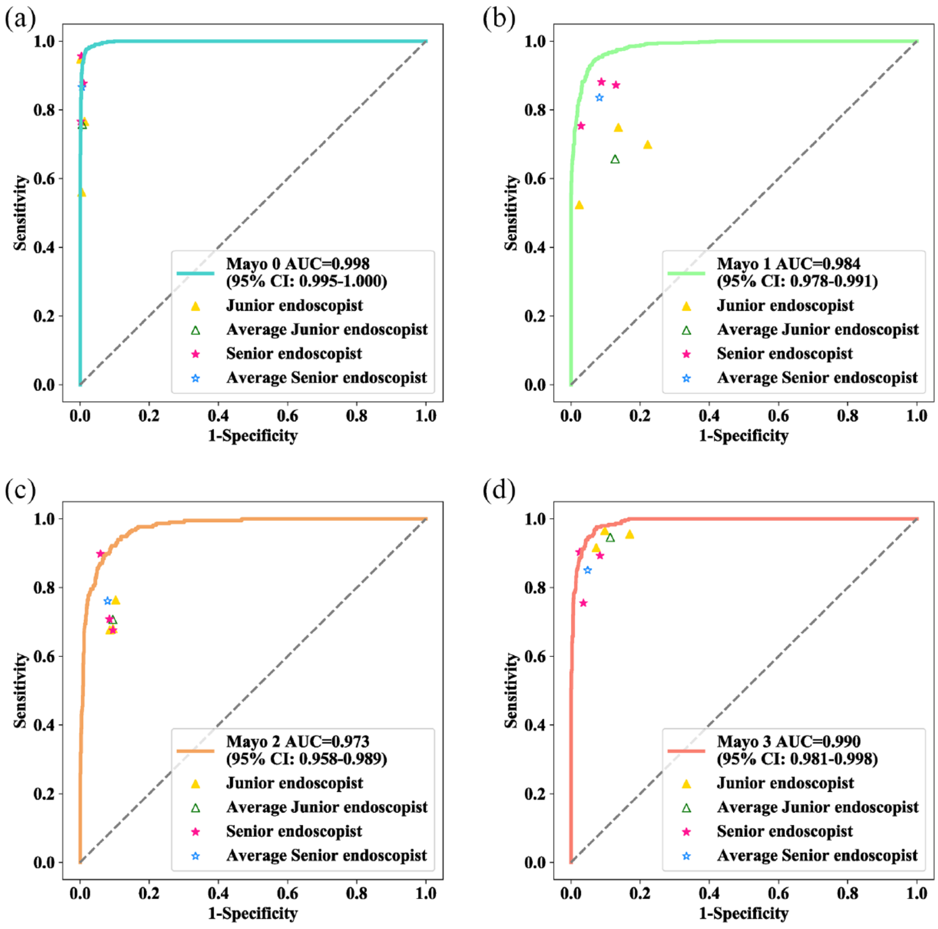

Figure 4. Comparison of performance between UC-former and endoscopists on the internal test set. ROC curve for Mayo 0 (a), Mayo 1 (b), Mayo 2 (c), and Mayo 3 (d). The yellow triangles indicate the diagnostic sensitivities and specificities of the junior endoscopists, the blue triangle indicates the pooled sensitivities and specificities of all junior endoscopists, the pink stars indicate the diagnostic sensitivities and specificities of the senior endoscopists, and the blue star indicates the pooled sensitivities and specificities of all senior endoscopists.
AUC, the area under the receiver operating characteristic curve; ROC, receiver operating characteristic.

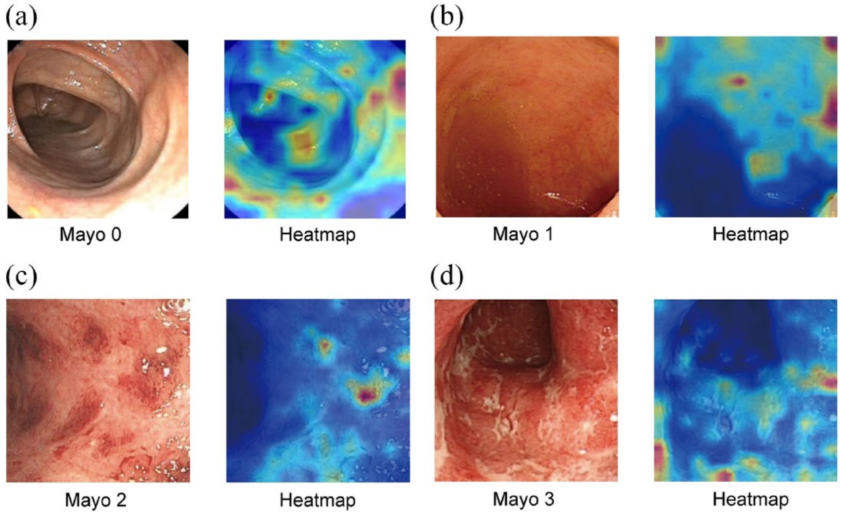

Figure 5. Heatmaps generated by UC-former.
UC, ulcerative colitis (a) The endoscopic image of Mayo 0 and its heatmap, (b) The endoscopic image of Mayo 1 and its heatmap, (c) The endoscopic image of Mayo 2 and its heatmap, (d) The endoscopic image of Mayo 3 and its heatmap.

Table 4. Results of multicenter validation achieved by UC-former.

Hospital	ACC (95% CI)	Level	SEN (95% CI)	SPE (95% CI)	PPV (95% CI)	NPV (95% CI)
The First Affiliated Hospital of Chongqing Medical University		Mayo 0	0.968 (0.909–0.989)	0.950 (0.924–0.967)	0.811 (0.728–0.873)	0.992 (0.978–0.997)
	0.824 (0.788–0.854)	Mayo 1	0.882 (0.838–0.916)	0.843 (0.792–0.883)	0.856 (0.809–0.893)	0.871 (0.822–0.907)
		Mayo 2	0.925 (0.801–0.974)	0.849 (0.814–0.879)	0.343 (0.260–0.436)	0.993 (0.978–0.997)
		Mayo 3	0.948 (0.891–0.976)	0.866 (0.829–0.896)	0.673 (0.597–0.740)	0.983 (0.963–0.992)
The Sixth Affiliated Hospital of Sun Yat-Sen University		Mayo 0	1.000 (0.954–1.000)	0.994 (0.964–0.999)	0.988 (0.933–0.998)	1.000 (0.976–1.000)
	0.850 (0.799–0.890)	Mayo 1	0.983 (0.909–0.997)	0.824 (0.761–0.873)	0.648 (0.544–0.739)	0.993 (0.962–0.999)
		Mayo 2	0.898 (0.782–0.956)	0.859 (0.802–0.902)	0.629 (0.511–0.732)	0.970 (0.931–0.987)
		Mayo 3	0.958 (0.860–0.989)	0.946 (0.904–0.971)	0.821 (0.702–0.900)	0.989 (0.960–0.997)
Tongji Hospital		Mayo 0	1.000 (0.923–1.000)	0.973 (0.925–0.991)	0.939 (0.835–0.979)	1.000 (0.966–1.000)
	0.836 (0.771–0.886)	Mayo 1	0.864 (0.733–0.936)	0.939 (0.880–0.970)	0.844 (0.712–0.923)	0.947 (0.890–0.976)
		Mayo 2	0.939 (0.804–0.983)	0.881 (0.813–0.927)	0.674 (0.530–0.791)	0.982 (0.938–0.995)
		Mayo 3	0.972 (0.858–0.995)	0.911 (0.847–0.949)	0.761 (0.621–0.861)	0.991 (0.952–0.998)

ACC, accuracy; CI, confidence interval; NPV, negative prediction value; PPV, positive prediction value; SEN, sensitivity; SPE, specificity; UC, ulcerative colitis.

branches to enable the model to better learn the subtle features of UC images with different Mayo endoscopic scores. Note that a common problem in clinical practice is that there are fewer positive samples than negative samples, which is manifested in our study, as the number of samples for Mayo 2 was much less than the number of samples for Mayo 0 and Mayo 1. Although we designed a data enhancement method for UC-former to compensate for the class imbalance problem in the dataset, the experimental results show that the classification of Mayo 2 still remains difficult. Notably, there were significant differences in both SEN and SPE between junior endoscopists and senior endoscopists. Furthermore, there were differences between two different senior endoscopists or between two different junior endoscopists because the clinical experience and cognition of each endoscopist were different. In terms of average ACC, the classification performance of senior endoscopists was better than that of junior endoscopists. Compared with endoscopists, the UC-former achieved much higher classification ACC, demonstrating its advantages in the four-level classification of the Mayo endoscopic score. In terms of SEN, SPE, PPV, and NPV, the UC-former still outperformed most endoscopists. In addition, compared with endoscopists, the UC-former made decisions much faster, with an average speed of about 0.017 s per image. Therefore, it is believed that it is feasible to apply this deep model to real-time classification of endoscopic videos in the future. Note that changes in the brightness, color, and contrast of endoscopic images will greatly affect the classification performance of the deep model, which can be demonstrated in the results of multicenter validation. Generally, endoscopic images collected by different medical centers have various colors, brightness, and contrast, so the performance of the UC-former on the multicenter test dataset is usually worse than that on the internal test dataset. Note that data enhancement can increase the diversity of data, thus improving the generalization ability of the deep model. Since the UC-former employed a patch-based data enhancement method, it still achieved excellent performance on the three multicenter external datasets, indicating that it had better generalization ability.

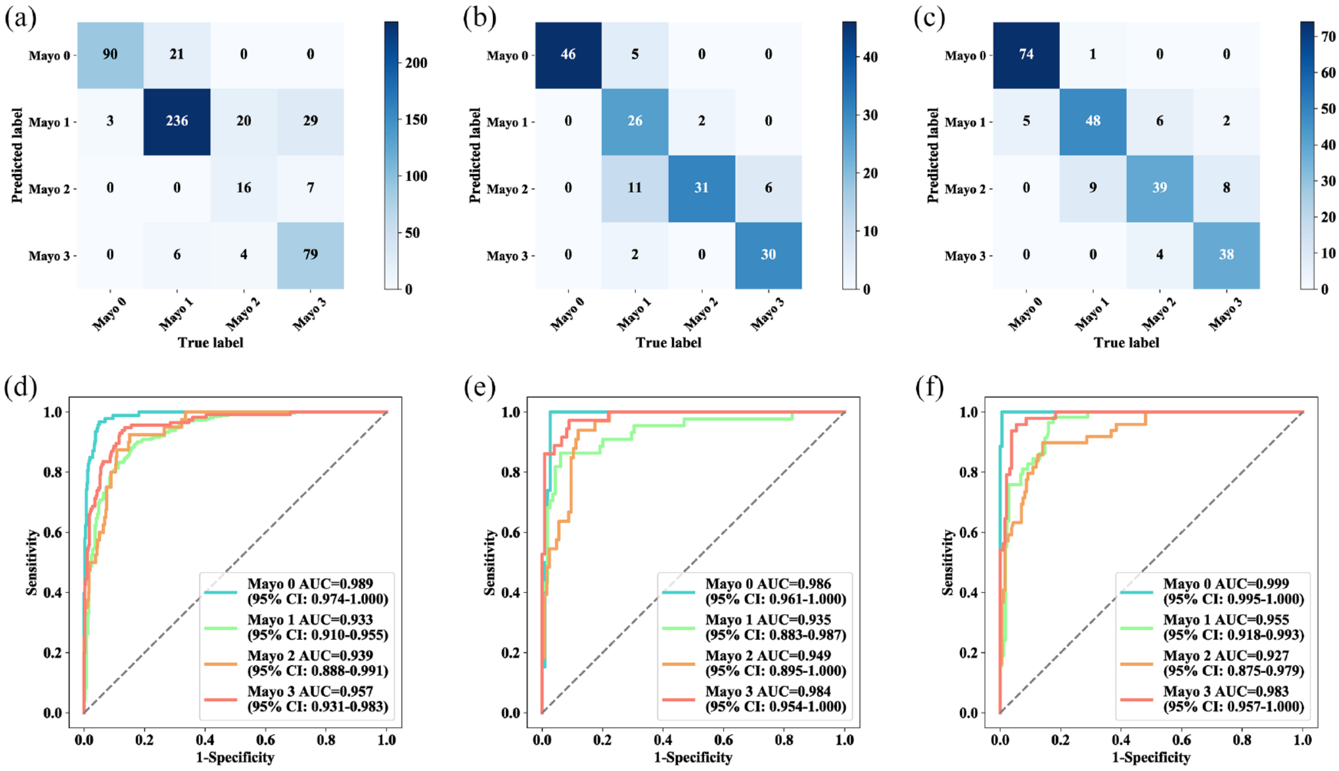

Figure 6. Confusion matrix and ROC curves of multicenter validation. Confusion matrix (a–c) and ROC curves (d–f) for three hospitals. (a and d) The First Affiliated Hospital of Chongqing Medical University. (b and e) Tongji Hospital Affiliated with Huazhong University of Science and Technology (c and f) The Sixth Affiliated Hospital of Sun Yat-Sen University.
AUC, the area under the receiver operating characteristic curve; ROC, receiver operating characteristic.

The heatmaps are used to display which part of the input image has a role in the final classification of the image. Note that most responses of heatmaps generated by UC-former were in the lesion area of UC images, which were relatively significant areas, showing that UC-former learned significant areas and more distinguishing features during the training process. These findings also suggested that UC-former was a reliable and robust deep model in this study.

In this study, we developed and validated a deep learning-based approach to predict the Mayo endoscopic score of UC images. Due to its high ACC, fidelity, and stability, the developed UC-former may assist endoscopists in improving prediction ACC in clinical settings. This approach may lay a foundation for the study of the Mayo endoscopic score predicted by deep learning to evaluate the severity of UC in patients. However, our study still has some limitations. First, we collected data from only two centers to construct the training set, and the data diversity was still insufficient for deep learning. Due to differences in equipment, lighting, and endoscopist manipulation, training data should include as many UC images in various situations as possible to improve the generalization performance of the UC-former. Second, in terms of performance, although we performed both internal validation and multicenter validation, performance validation with larger cohorts or prospective clinical trials is warranted.

In summary, the developed UC-former may be an original and welcome step for the automatic and accurate prediction of the Mayo endoscopic score in UC patients. In the future, in addition to further expanding the dataset, we will continue to improve UC-former and to enhance its feature extraction capability to better extract key features that distinguish UC images with different Mayo endoscopic scores, which may further improve the classification performance of UC-former.

Declarations

Ethics approval and consent to participate

This study was approved by the Ethics Committee of Army Medical Center of PLA (No. 2021-285) on 31 December 2021. Written informed consent was not required for the study on human participants in accordance with the local legislation and institutional requirements.

Consent for publication

Not applicable.

Author contributions

Jing Qi: Conceptualization; Data curation; Writing – original draft; Writing – review & editing.

Guangcong Ruan: Conceptualization; Data curation; Writing – original draft; Writing – review & editing.

Yi Ping: Data curation; Validation; Writing – review & editing.

Zhifeng Xiao: Data curation; Formal analysis; Validation; Writing – review & editing.

Kaijun Liu: Formal analysis; Supervision; Validation.

Yi Cheng: Data curation; Formal analysis; Supervision; Writing – review & editing.

Rongbei Liu: Data curation; Writing – review & editing.

Bingqiang Zhang: Data curation; Formal analysis; Validation.

Min Zhi: Data curation; Formal analysis; Validation.

Junrong Chen: Data curation; Investigation; Validation; Visualization.

Fang Xiao: Data curation; Formal analysis; Supervision; Validation.

Tingting Zhao: Investigation; Methodology; Validation.

Jiaxing Li: Data curation; Validation; Visualization.

Zhou Zhang: Data curation; Formal analysis; Validation.

Yuxin Zou: Methodology; Writing – review & editing.

Qian Cao: Data curation; Methodology; Writing – original draft; Writing – review & editing.

Yongjian Nian: Funding acquisition; Methodology; Supervision; Writing – original draft; Writing – review & editing.

Yanling Wei: Conceptualization; Funding acquisition; Methodology; Writing – original draft; Writing – review & editing.

Acknowledgements

None.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by the Chongqing Technology Innovation and Application Development Project (2022TIAD-KPX0099), Key Project of Chongqing Social Livelihood (cstc2018jscx-mszdX0026), Chongqing Postgraduate Scientific Research Innovation Project (CYS22746), Chongqing Postgraduate Education Teaching Reform Research Project (yjg203140), and Undergraduate Scientific Research Training Program of Army Medical University (2020XBK31).

Competing interests

The authors declare that there is no conflict of interest.

Availability of data and materials

The datasets collected in this study are available from the corresponding author on reasonable request.

ORCID iDs

Yongjian Nian https://orcid.org/0000-0001-6779-0850
Yanling Wei https://orcid.org/0000-0002-3307-3092

Supplemental material

Supplemental material for this article is available online.

References

1. Soffer S, Kopylov U and Klang E. Artificial intelligence for the evaluation of mucosal healing in IBD: the future is here. *Gastroenterology* 2021; 161: 1073–1074.
2. Cosnes J, Gower-Rousseau C, Seksik P, *et al.* Epidemiology and natural history of inflammatory

bowel diseases. *Gastroenterology* 2011; 140: 1785–1794.

3. Ng SC, Shi HY, Hamidi N, *et al.* Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. *Lancet* 2017; 390: 2769–2778.
4. Jurij H and Vipul J. A TIGER among endoscopic indices in inflammatory bowel disease. *J Crohns Colitis* 2022; 16: 519–520.
5. Siow VS, Bhatt R and Mollen KP. Management of acute severe ulcerative colitis in children. *Semin Pediatr Surg* 2017; 26: 367–372.
6. Ben-Horin S, Lahat A, Amitai MM, *et al.* Assessment of small bowel mucosal healing by video capsule endoscopy for the prediction of short-term and long-term risk of Crohn's disease flare: a prospective cohort study. *Lancet Gastroenterol Hepatol* 2019; 4: 519–528.
7. Barash Y, Azaria L, Soffer S, *et al.* Ulcer severity grading in video capsule images of patients with Crohn's disease: an ordinal neural network solution. *Gastrointest Endosc* 2021; 93: 187–192.
8. Pabla BS and Schwartz DA. Assessing severity of disease in patients with Ulcerative Colitis. *Gastroenterol Clin North Am* 2020; 49: 671–688.
9. Mohammed VN, Samaan M, Mosli MH, *et al.* Endoscopic scoring indices for evaluation of disease activity in ulcerative colitis. *Cochrane Database Syst Rev* 2018; 1: CD011450.
10. Eliakim R, Yablecovitch D, Lahat A, *et al.* A novel PillCam Crohn's capsule score (Eliakim score) for quantification of mucosal inflammation in Crohn's disease. *United European Gastroenterol J* 2020; 8: 544–551.
11. Klang E, Barash Y, Margalit RY, *et al.* Deep learning algorithms for automated detection of Crohn's disease ulcers by video capsule endoscopy. *Gastrointest Endosc* 2020; 91: 606–613.e2.
12. Gottlieb K, Requa J, Karnes W, *et al.* Central reading of Ulcerative Colitis clinical trial videos using neural networks. *Gastroenterology* 2021; 160: 710–719.e2.
13. Zhu Y, Wang QC, Xu MD, *et al.* Application of convolutional neural network in the diagnosis of the invasion depth of gastric cancer based on conventional endoscopy. *Gastrointest Endosc* 2019; 89: 806–815.e1.
14. Ding Z, Shi HY, Zhang H, *et al.* Gastroenterologist-level identification of small-bowel diseases and normal variants by capsule endoscopy using a deep-learning model. *Gastroenterology* 2019; 157: 1044–1054.e5.
15. Ozawa T, Ishihara S, Fujishiro M, *et al.* Novel computer-assisted diagnosis system for endoscopic disease activity in patients with ulcerative colitis. *Gastrointest Endosc* 2019; 89: 416–421.e1.
16. Stidham RW, Liu WS, Bishu S, *et al.* Performance of a deep learning model vs human reviewers in grading endoscopic disease severity of patients with Ulcerative Colitis. *JAMA Netw Open* 2019; 2: e193963.
17. Bhambhvani HP and Zamora A. Deep learning enabled classification of Mayo endoscopic subscore in patients with ulcerative colitis. *Eur J Gastroenterol Hepatol* 2021; 33: 645–649.
18. Becker BG, Arcadu F, Thalhammer A, *et al.* Training and deploying a deep learning model for endoscopic severity grading in ulcerative colitis using multicenter clinical trial data. *Ther Adv Gastrointest Endosc* 2021; 14: 1–15.
19. Zamir SW, Arora A, Khan S, *et al.* Restormer: efficient transformer for high-resolution image restoration. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, New Orleans, USA, 2022, pp. 5718–5729.
20. Dosovitskiy A, Beyer L, Kolesnikov A, *et al.* An image is worth 16x16 words: transformers for image recognition at scale. In: *Proceedings of the international conference on learning representations*, Virtual conference, Vienna, Austria, 2021, pp. 1–22.
21. Qi J, Ruan GC, Liu J, *et al.* PHF^3 technique: a pyramid hybrid feature fusion framework for severity classification of Ulcerative Colitis using endoscopic images. *Bioengineering* 2022; 9: 632.
22. Von Elm E, Altman DG, Egger M, *et al.* The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann Intern Med* 2007; 147: 573–577.
23. Gong CY, Wang DL, Li M, *et al.* Vision transformers with patch diversification. *arXiv preprint* 2021; arXiv:2104.12753.
24. Chefer H, Gur S and Wolf L. Transformer interpretability beyond attention visualization. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Virtual conference, Kuala Lumpur, Malaysia, 2021, pp. 782–791.
25. Soffer S, Ben-Cohen A, Shimon O, *et al.* Convolutional neural networks for radiologic images: a radiologist's guide. *Radiology* 2019; 290: 590–606.
26. Soffer S, Klang E, Shimon O, *et al.* Deep learning for wireless capsule endoscopy: a systematic review and meta-analysis. *Gastrointest Endosc* 2020; 92: 831–839.e8.
27. Luo XD, Zhang JH, Li ZG, *et al.* Diagnosis of ulcerative colitis from endoscopic images based on deep learning. *Biomed Signal Process Control* 2022; 73: 103443.