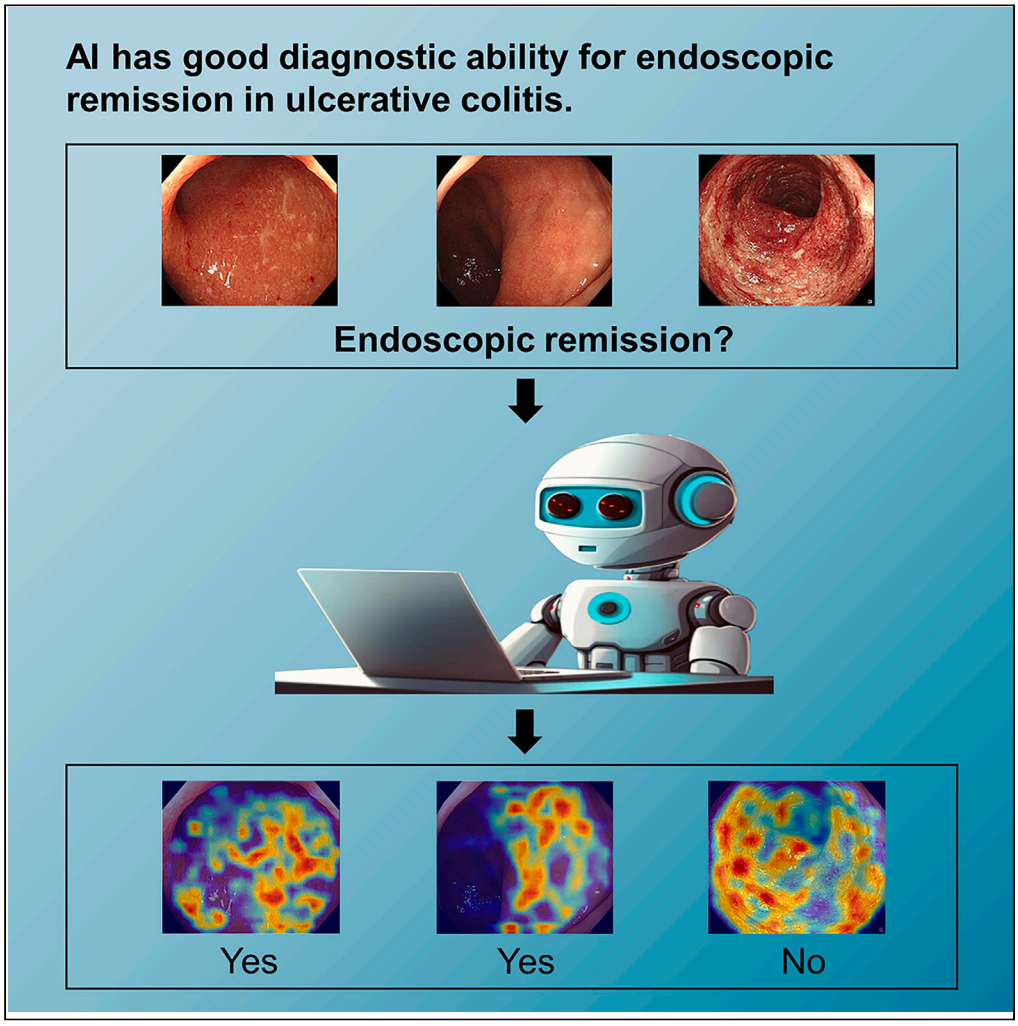
iScience

CellPress
OPEN ACCESS

Article

A systematic review and meta-analysis of artificial intelligence-diagnosed endoscopic remission in ulcerative colitis



Bing Lv, Lihong Ma, Yanping Shi, Tao Tao, Yanting Shi

yantingshi@hotmail.com

Highlights

AI has good diagnostic ability for endoscopic remission in UC

AI performed well for four commonly used endoscopic remission criteria

Provide an evidence-based medical rationale for AI diagnosing endoscopic remission

Lv et al., iScience 26, 108120
November 17, 2023
https://doi.org/10.1016/j.isci.2023.108120

iScience

CellPress
OPEN ACCESS

Article

A systematic review and meta-analysis of artificial intelligence-diagnosed endoscopic remission in ulcerative colitis

Bing Lv,[1] Lihong Ma,[2] Yanping Shi,[3] Tao Tao,[2] and Yanting Shi[2,4,*]

SUMMARY

Endoscopic remission is an important therapeutic goal in ulcerative colitis (UC). The Ulcerative Colitis Endoscopic Index of Severity (UCEIS) and Mayo Endoscopic Score (MES) are the commonly used endoscopic scoring criteria. This systematic review and meta-analysis aimed to evaluate the accuracy of artificial intelligence (AI) in diagnosing endoscopic remission in UC. We also performed a meta-analysis of each of the four endoscopic remission criteria (UCEIS = 0, MES = 0, UCEIS = <1, MES = <1). Eighteen studies involving 13,687 patients were included. The combined sensitivity and specificity of AI for diagnosing endoscopic remission in UC was 87% (95% confidence interval [CI]:81–92%) and 92% (95% CI: 89–94%), respectively. The area under the curve (AUC) was 0.96 (95% CI: 0.94–0.97). The results showed that the AI model performed well regardless of which criteria were used to define endoscopic remission of UC.

INTRODUCTION

Ulcerative colitis (UC) is a chronic inflammatory disease of the colonic mucosa.[1] Clinical presentation includes diarrhea, mucopurulent and bloody stools, and abdominal pain.[2,3] UC has traditionally been considered a disease of Western countries; the incidence and prevalence of UC are the highest in North America and Northern Europe, with an incidence of 9–20 cases per 100,000 person-years and a prevalence of 156–291 cases per 100,000 people.[4] Recent epidemiological studies suggest that the incidence is rapidly increasing in South America, Eastern Europe, Asia, and Africa.[5,6] Repeated, persistent UC attacks severely affect the health and quality of life of patients[7,8] and impose a considerable medical and economic burden on society.

UC can only be remitted, but not completely cured.[9,10] The International Organization for the Study of Inflammatory Bowel Disease has identified endoscopic healing as the preferred long-term treatment goal for UC and recommended endoscopic remission as an important therapeutic goal.[11] Among the many endoscopic scoring systems used to assess the activity of UC, the Mayo Endoscopic Score (MES)[12] and the Ulcerative Colitis Endoscopic Index of Severity (UCEIS)[13] are widely used. The MES ranges from 0 to 3 and is used to evaluate erythema, mucosal fragility, vascular patterns, bleeding, and ulceration.[12] The UCEIS grade ranges from 0 to 8 based on three dimensions: vascular pattern, bleeding, and erosions/ulcerations.[13] There is no clear definition of endoscopic remission in UC.[14] In an international consensus, experts voted on this topic, and the top four were UCEIS = 0, MES = 0, UCEIS = <1, and MES = <1.[14] This study used these four scores as criteria for endoscopic remission in UC.

Currently, the evaluation of endoscopic images relies primarily on the manual judgment of the endoscopist, which depends heavily on the level of experience of the endoscopist and can be affected by fatigue and stress. In recent years, artificial intelligence (AI), particularly deep learning (DL),[15,16] has achieved remarkable results in aiding diagnoses. It provides accurate and objective diagnostic results, reduces endoscopist workload, and improves efficiency.[17–19]

As summarized and analyzed in several review articles, AI has demonstrated great potential in assessing and managing inflammatory bowel diseases (IBD), including UC and Crohn's disease. Takenaka et al.[20–22] detailed recent advances and relevant evidence on using AI for endoscopy in IBD and discussed how AI can improve clinical practice in IBD. Tontini et al.[23] described AI's emerging applications and future potential for endoscopy in IBD. Yang et al.[24] quantified the accuracy of AI in predicting the endoscopic severity of UC using a median-taking approach.

Jahagirdar et al.[25] conducted a meta-analysis of convolutional neural networks to the diagnose endoscopic severity of UC. A total of 12 studies were included, and the pooled sensitivity of the AI was 83.9%, specificity was 92.3% and accuracy was 91.2%. This article analyzed the overall performance of the AI but did not analyze it further for specific endoscopic scores. Our study provides a comprehensive quantitative

[1]School of Computer Science and Technology, Shandong University of Technology, NO.266, Xincunxi Road, Zibo, Shandong 255000, China
[2]Department of Gastroenterology, Zibo Central Hospital, No.10 Shanghai Road, Zibo, Shandong 255000, China
[3]Department of Pediatrics, Zhoucun Maternal and Child Health Care Hospital, No.72 Mianhuashi Street, Zibo, Shandong 255000, China
[4]Lead contact
*Correspondence: yantingshi@hotmail.com
https://doi.org/10.1016/j.isci.2023.108120

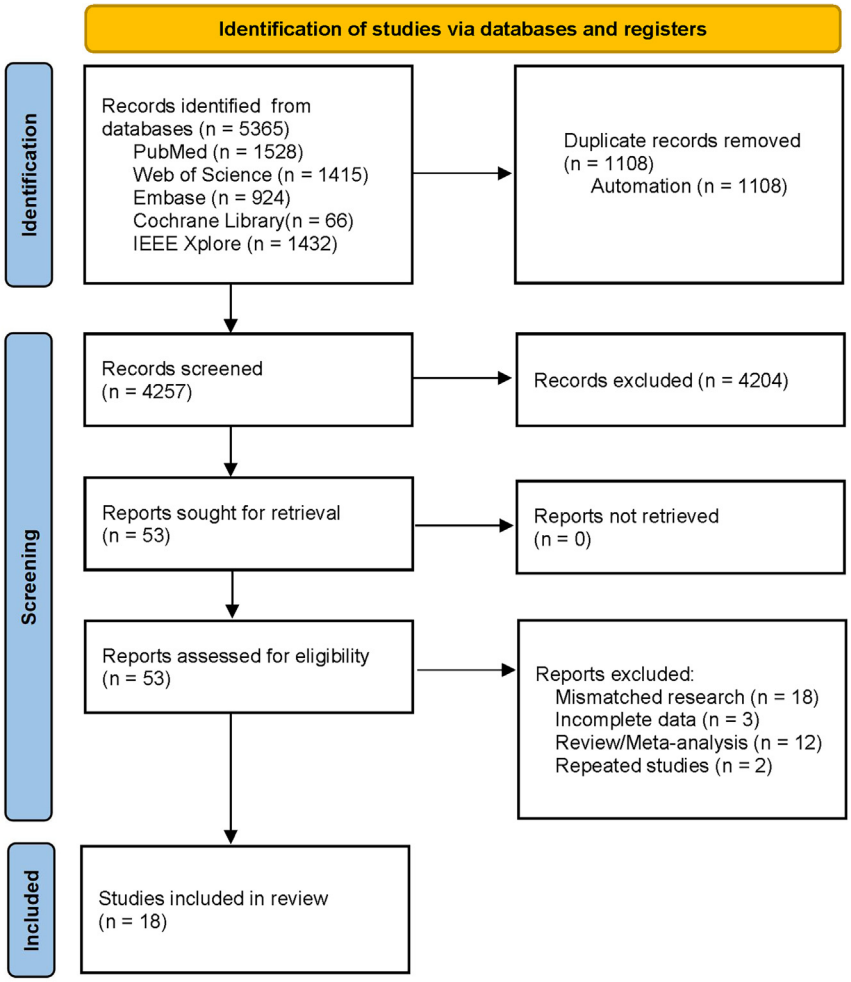


Figure 1. PRISMA Flow diagram for study selection

meta-analysis of AI performance from the perspective of endoscopic remission to provide an evidence-based medical rationale for its clinical application. The primary outcome of this study was the overall performance of the AI model in diagnosing endoscopic remission. The secondary outcome was the ability of AI to diagnose endoscopic remission under different criteria (UCEIS = 0, MES = 0, UCEIS = <1, MES = <1).

RESULT

Included studies and quality assessment

We searched five databases according to the search strategy and retrieved 5,365 articles. EndNote software was used to remove 1,108 duplicates automatically. We excluded 4,204 articles after reviewing their titles and abstracts. The remaining 53 articles were read in detail, of which 18 were included.[26–43] The literature screening process is illustrated in Figure 1. Details of the included studies are presented in Table 1. Three studies by Takenaka et al.[28,44,45] used AI for the endoscopic scoring of UC, and we included only the studies with the largest sample size.[28]

The 18 studies included 13,687 patients. Twelve studies were retrospective and six were prospective[28,30,31,34,35,40]; fourteen studies were single-center and four were multicenter[30,31,33,40]; and nine studies were from Europe and America and nine were from Asia.[26,28,29,33,35,38,39,42,43] All the studies used images obtained from common white-light imaging and used DL algorithms. It should be noted that of the six prospective studies, only Takenaka 2020[28] explicitly stated that the AI model was involved in the actual diagnostic process of the patient, and the other five studies only prospectively collected data for AI testing.

The assessment of the risk of bias in the included studies is shown in Figure 2. One study[32] explored the ability of DL models to automatically grade each MES, but images with MES = 0 were not included in the dataset. This article is valuable for studying AI in assessing the endoscopic severity of UC. However, regarding endoscopic remission, it was considered to have a high risk of bias in the index test section. We will analyze this study separately in the sensitivity analysis. Three studies were considered to have an unknown risk of bias in the patient selection section: two study[29,42] did not describe the patient selection process and one study[39] did not describe the quality of the endoscopic images.

Table 1. Details of the included studies

Study	Country/ Region	Study Center	Study design	Algorithm	Patients(n)	Train set(n)	Test set	Standard Reference	Sensitivity (%)	Specificity (%)
Ozawa 2018[26]	Japan	Single	Retrospective	GoogLeNet	558	26,304 images	image	**MES=0**	**72.54**	**78.42**
								MES<=1	98.42	0.62.9
Stidham 2019[27]	USA	Single	Retrospective	InceptionV3	3,112	14,862 images	frame	**MES=0**	**75.3**	**92.89**
								MES<=1	93.39	87.07
Takenaka 2020[28]	Japan	Single	Prospective	InceptionV3	2,887	40,758 images	image	**UCEIS=0**	**96**	**83**
Huang 2021[29]	Taiwan	Single	Retrospective	Inceptionv3 +SVM +k-NN	54	600 images	image	**MES<=1**	**96.33**	**89.23**
Gottlieb 2021[30]	USA	Multi	Prospective	RNN	249	661 videos	video	**MES=0**	**87.5**	**96.61**
								MES<=1	80	90.48
								UCEIS = 0	66.67	98.44
								UCEIS<=1	69.23	85.19
Yao 2021[31]	USA	Multi	Prospective	InceptionV3	175	51 videos	video	**MES<=1**	**72.09**	**86.88**
Bhambhvani 2021[32]	USA	Single	Retrospective	ResNeXt101	777	698 images	image	**MES<=1**	**66.67**	**91.38**
Gutierrez Becker 2021[33]	Japan	Multi	Retrospective	ResNet50	1105	1338 videos	image	**MES<=1**	**73**	**97.7**
Patel 2022[34]	UK	Single	Prospective	ResNet34	73	38,124 images	image	**UCEIS=0**	**73**	**93**
								UCEIS<=1	99.69	79.63
Byrne 2022[35]	India	Single	Prospective	EfficientNetB3	234	107 videos	image	**MES=0**	**85.71**	**94.6**
								MES<=1	91.29	96.7
								UCEIS = 0	88.18	93.89
								UCEIS<=1	86.14	97.1
Sutton 2022[36]	Canada	Single	Retrospective	DenseNet121	840	669 images	image	**MES=0**	**79**	**91**
Luo 2022[37]	China	Single	Retrospective	Improved DenseNet201	1317	7942 images	image	**MES=0**	**98**	**97.4**
Polat 2022[38]	Turkey	Single	Retrospective	DenseNet121	462	7904 images	image	**MES<=1**	**97.4**	**87.6**
Fan 2023[39]	China	Single	Retrospective	ResNet50	350	11,303 images	image	**MES=0**	**87.5**	**96.68**
								UCEIS = 0	87.4	96.62
Iacucci 2023[40]	UK	Multi	Prospective	Improved VGG16	331	543 images	image	**UCEIS<=1**	**80**	**78**
Lo 2022[41]	Denmark	Single	Retrospective	EfficientNetB2	467	1261 images	image	**MES=0**	**94**	**94**
								MES<=1	93	96
Kadota 2022[42]	Japan	Single	Retrospective	Custom Networks	388	8212 images	image	**MES=0**	**84**	**75**
								MES<=1	95.7	72
Wang 2023[43]	China	Single	Retrospective	CB-HRNet	308	4787 images	image	**MES=0**	**92.25**	**93.2**
								MES<=1	92.87	95.41

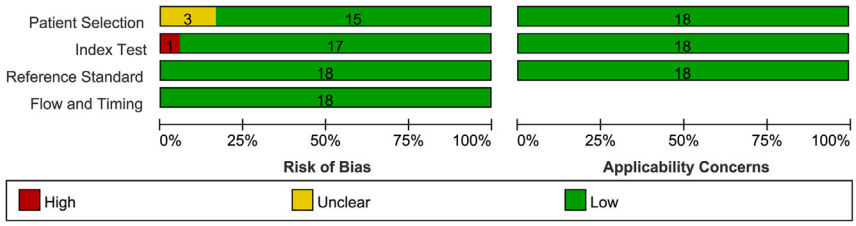


Figure 2. Summary of risk of bias and applicability of concerns graph

Study characteristics and data extraction

We extracted 31 datasets from 18 studies, as detailed in Table 1. One set of data from each study was selected for meta-analysis, as described in section 2.3, to form the main results, and these data are shown in bold in Table 1. Of the 31 datasets, 5, 9, 4, and 13 had UCEIS = 0, MES = 0, UCEIS = < 1, and MES = < 1, respectively, as the criterion. We performed a separate meta-analysis for each criterion to obtain secondary outcomes.

Ozawa et al.[26] constructed an aided diagnostic system based on GoogleNet[46] to identify normal mucosa (MES = 0) and mucosal healing (MES = < 1). A total of 558 patients with UC and 26,304 colonoscopy images were included in this study. The AUC for the systematic identification of MES = 0 and MES = < 1 reached 0.86 and 0.98, respectively.

Stidham et al.[27] used the InceptionV3[47] network to identify endoscopic remission in patients with UC (MES = < 1). The authors tested image and video test sets. A total of 3,112 patients, 16,514 endoscopic images, and 30 endoscopic videos were included in the study. The positive predictive value of the model on the video test set was 68% (95% CI, 67–69%), the negative predictive value was 98% (95% CI, 97–99%), and AUC was 0.966 (95% CI, 0.963–0.969).

Takenaka et al.[28] developed a deep neural network for evaluating endoscopic images of patients with UC. The model identified endoscopic (UCEIS = 0) and histological remission with an accuracy of 90.1% and 92.9%, respectively. This study included 2,887 patients, 44,945 images, and 10,989 biopsies.

Huang et al.[29] combined the deep neural network, support vector machine (SVM), and k-nearest neighbor (k-NN) techniques to diagnose mucosal healing (MES = < 1) with an accuracy, sensitivity, and specificity of 94.5%, 89.2%, and 96.3%, respectively. The model was trained and tested using 856 endoscopic images of 54 patients with UC.

Gottlieb et al.[30] used recurrent neural networks to evaluate the endoscopic videos of patients with UC and provided the MES and UCEIS scores. The study included 249 patients from 14 countries. We calculated the sensitivity and specificity of the model to identify MES = 0, MES = < 0, UCEIS = 0, and UCEIS = < 0 according to the confusion matrix.

Yao et al.[31] developed a fully automated video analysis system based on the InceptionV3 network, using the MES score to grade UC. The authors tested the model using both internal and external test sets. On an external test set containing 264 videos, the system achieved 83.7% accuracy in distinguishing MES = < 1 vs. MES = > 2.

Bhambhvani et al.[32] used the ResNeXt[48] network to distinguish MES 1–3 and trained and tested the model using 777 UC images from the public dataset HyperKvasir.[49] The final model identified MES = 1 with an AUC of 0.89 and an overall accuracy of 77.2%.

Becker et al.[33] developed a DL-based system for endoscopic scoring. The study involved a total of 1,105 patients from 28 countries. The system achieved an AUC of 0.85 $\pm$ 0.0273 for identifying MES = > 2 on the HyperKvasir dataset.

Patel et al.[34] used the ResNet[50] network to assess UC severity. Videos from 73 patients were used to train and test. The model achieved 90% accuracy in identifying normal mucosa vs. active inflammation (UCEIS = 0 vs. UCEIS = > 1) and 98% accuracy in identifying mild inflammation vs. moderate-severe inflammation (UCEIS = < 3 vs. UCEIS = > 4).

Byrne et al.[35] built a DL model to automatically predict the MES and UCEIS scores for UC using EfficientNet[51] as the backbone network. A total of 234 videos were collected to train and test the proposed model. Four datasets were included in the meta-analysis.

Sutton et al.[36] compared the accuracy of four mainstream DL networks to distinguish MES = < 1 vs. MES = > 2. The models were trained and tested using 840 endoscopic images of UC obtained from Hyper-Kvasir. DenseNet[52] performed the best, with sensitivity, specificity, and accuracy of 79%, 91%, and 87.5%, respectively.

Luo et al.[37] proposed a DL network called "Efficient Attention Mechanism Network" for identifying endoscopic remission in UC (MES = 0). A total of 14,306 endoscopic images from 1,317 patients with UC were collected to train and test the network. We extracted the test results of the model on a dataset with a larger sample size, and the model achieved an accuracy of 97.6% and an AUC of 0.975.

Polat et al.[38] created an open-source dataset named LIMUC, which comprises 11,276 endoscopic images from 564 patients with UC. At the same time, the authors proposed a regression-based DL method and evaluated it using the LIMUC dataset. When identifying endoscopic remission (MES = < 1), the sensitivity, specificity, and accuracy of this method reached 97.4%, 87.6%, and 95.7%, respectively.

Fan et al.[39] developed an automated scoring system for UC based on DL technology using 5,875 endoscopic images and 20 videos from 332 patients with UC to train and test the model. The model achieved an accuracy of 86.54% for the MES. For the UCEIS, 90.7% accuracy was achieved in identifying the vascular morphology, 84.6% in identifying erosions and ulcers, and 77.7% in identifying bleeding.

Iacucci et al.[40] developed a computer-aided diagnostic system based on the VGG16[53] network to assess UC severity. A total of 331 patients were included in the study. The system assessed UC endoscopic activity (UCEIS >1) with a sensitivity of 78% and specificity of 80%.

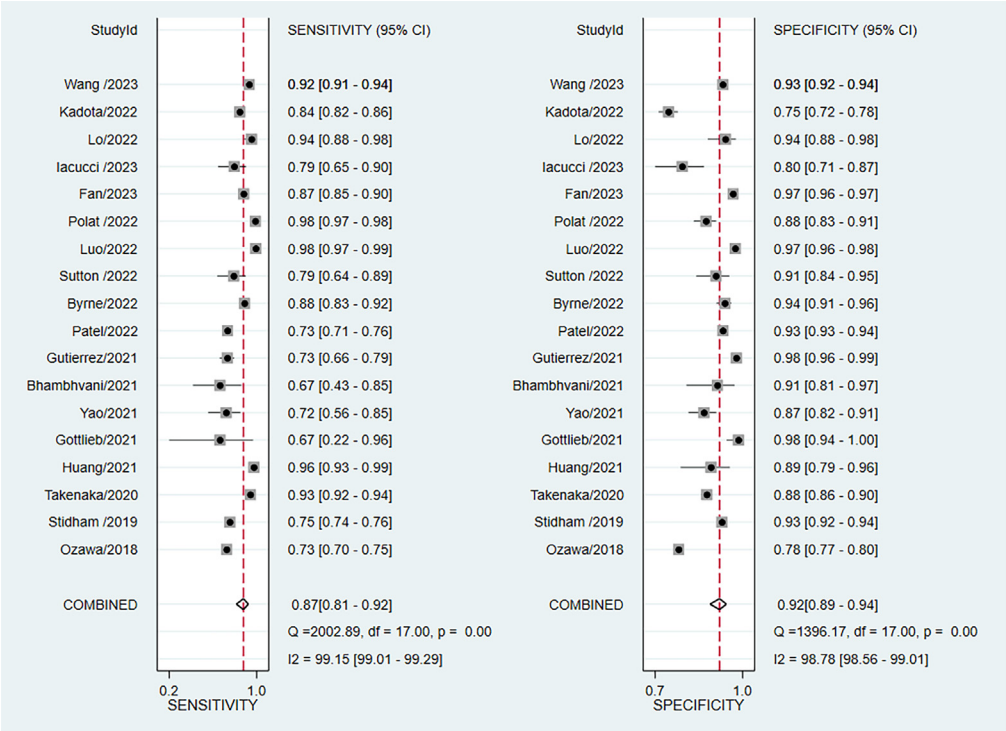


Figure 3. Forest plot of sensitivity and specificity of AI

Lo et al.[41] compared the accuracy of five DL algorithms for identifying the MES for UC, collecting 1484 images from 467 patients to train the test model. The best-performing EfficientNetB2 model achieved 94% accuracy in distinguishing MES 0 from MES 1–3 and 93% accuracy in distinguishing MES 0–1 from MES 2–3.

Bhambhvani et al.[42] proposed a multi-task learning method by combining learning to rank[54] with regression. The method not only realized automatic scoring of endoscopic severity of UC but also greatly reduced the annotation cost of images. The study used 10,265 endoscopic images from 388 patients to train and test the model. The model distinguishes between MES 0 and MES 1–3 with an F1-Score of 0.85.

Wang et al.[43] created an open-source dataset of UC endoscopic images comprising 7,978 images from 308 patients (excluding augmented images). Additionally, the authors proposed a new DL network named CB-HRNet. CB-HRNet achieves an accuracy of 93.73% in distinguishing MES 0 from MES 1–3 and 95.73% in distinguishing MES 0–1 from MES 2–3 with an accuracy of 95.07%.

Primary outcome

AI diagnosed endoscopic remission in UC with a combined sensitivity of 87% (95% CI: 81–92%, I^2 = 99.15), specificity of 92% (95% CI: 89–94%, I^2 = 98.78) (Figure 3), positive likelihood ratio of 11.03(95% CI: 7.72–15.76), negative likelihood ratio of 0.14 (95% CI: 0.09–0.21) (Figure S1), diagnostic score of 4.37 (95% CI: 3.74–5.01), and diagnostic odds ratio of 79.42 (95% CI: 42.14–149.71) (Figure S2). These metrics indicate that AI can be used to identify endoscopic remission effectively. The SROC curve was plotted (Figure 4) with an AUC of 0.96 (95% CI: 0.94–0.97), indicating the high accuracy of AI in identifying endoscopic remission in UC.

We evaluated the clinical utility of AI for diagnosing endoscopic remission using the Fagan plot (Figure 5). When the pre-test probability was set at 50%, the probability that the patient was in endoscopic remission was 92% if AI diagnosed the result as endoscopic remission. The probability that the patient was in endoscopic remission was 12% if AI diagnosed the result as endoscopic activity. This indicates the good diagnostic value of AI in clinical applications.

Subgroup analysis and meta-regression

Although AI performed well in diagnosing endoscopic remission in UC, I^2 showed a high degree of heterogeneity among studies. We performed a subgroup analysis and meta-regression according to the study region (Euro-America or Asia), type (prospective or retrospective), center (multi or single), number of patients (>400 or <400), and endoscopic scoring criteria (MES or UCEIS) to explore possible sources of heterogeneity. The results are summarized in Table 2.

The effects of study type, criteria, and center on sensitivity were statistically significant ($p < 0.05$). The study region significantly affected the sensitivity ($p < 0.001$). The study type, region, endoscopic scoring criteria, and the number of patients significantly affected the specificity

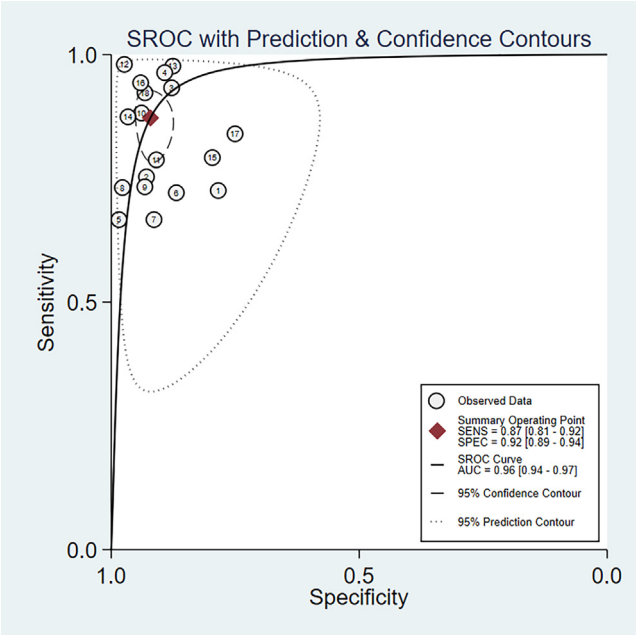


Figure 4. Summary receiver operating characteristic curves
Each circle indicates an individual study, red diamond represents summary sensitivity and specificity.

($p < 0.001$). The effects of the study center on specificity were statistically significant ($p < 0.05$). The combined sensitivity of the Asian studies was significantly higher than that of the Euro-American studies. The combined sensitivity of the single-center studies was significantly higher than that of the multicenter studies.

Sensitivity analysis and publication bias

The study by Bhambhvani et al.[32] was considered to have a high risk of bias at the time of quality assessment. We removed it, and the combined sensitivity was 87% (95% CI: 80–92%), specificity was 93% (95% CI: 90–95%), and the AUC was 0.96 (95% CI: 0.94–0.97). The study did not significantly affect the results of the meta-analysis.

After removing the study by Luo et al.,[37] the combined sensitivity was 85% (95% CI: 78–90%), and the AUC was 0.96 (95% CI: 0.93–0.97). This study had the greatest impact on the combined results but did not change it significantly. This suggests that the results of this meta-analysis are stable and not overly dependent on data from a single study. Publication bias was assessed using the Deeks'funnel plot (Figure 6) and the graph was roughly symmetrical ($p = 0.45$), indicating no publication bias.

Secondary outcome

We performed a separate meta-analysis of the data corresponding to each endoscopic remission criterion. The results are summarized in Table 3. The AUCs of the four meta-analyses ranged from 0.94 to 0.97, indicating that AI performed well regardless of the criteria used to define endoscopic remission in UC. It is important to note that the meta-analysis results must be interpreted with caution because of the small number of studies corresponding to each criterion.

DISCUSSION

AI technology is increasingly used in clinical practice to enhance diagnostic accuracy, stability, and efficiency. We performed a systematic evaluation and meta-analysis of the accuracy of AI in diagnosing UC endoscopic remission. Eighteen publications, 31 datasets, and 13,687 patients were included in this study. The combined results demonstrated the good diagnostic value of AI. High heterogeneity was observed among the studies, and the study region, type, and endoscopic scoring criteria were identified as possible sources of heterogeneity by subgroup analysis and meta-regression. We also performed a meta-analysis for each of the four criteria for endoscopic remission, and AI showed good performance with no significant differences in the results.

Endoscopic remission in UC is used to assess treatment effectiveness and guide the development of subsequent treatment plans. It is important to note that this meta-analysis defined endoscopic remission in terms of endoscopic scores but did not distinguish whether the endoscopic score was pre- or post-treatment. We assumed that the same endoscopic score was consistent with the characteristics of endoscopic imaging both before and after treatment. The effect of this assumption on the results of this meta-analysis requires further evaluation.

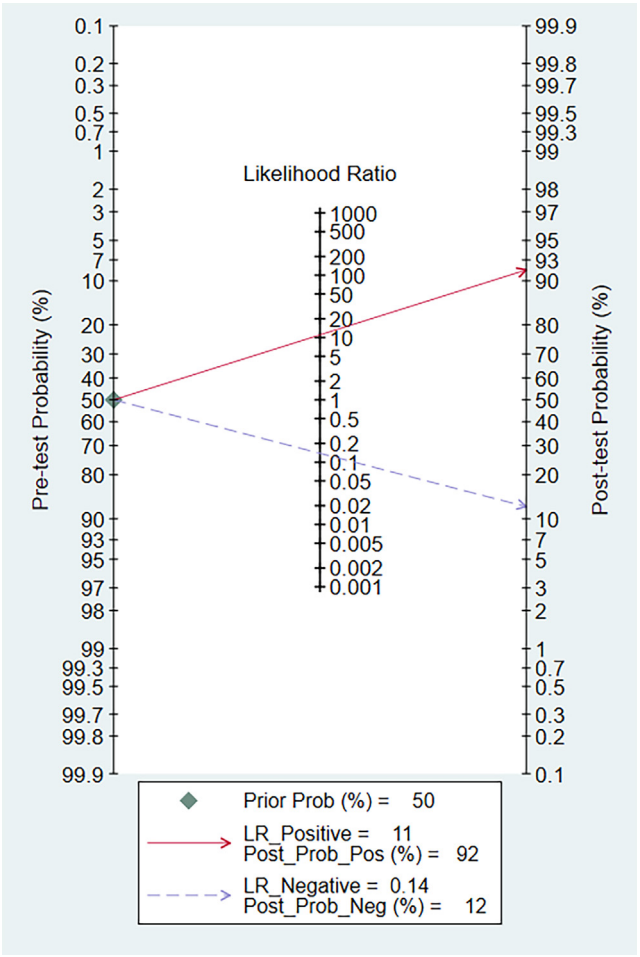


Figure 5. Fagan nomogram of the accuracy of AI

DL techniques can automatically extract features for image recognition. With traditional machine learning, experienced experts must manually extract image features and select a suitable classifier (e.g., random forest or SVM) for statistical analysis. DL is considered to have advantages over traditional machine learning for analyzing big data.[55,56] All 18 included studies used DL techniques to evaluate the images. One study[29] compared the performance of DL techniques (InceptionV3) with that of traditional machine learning techniques (SVM and k-NN) in identifying mucosal healing in UC. SVM performed slightly better than InceptionV3 in this study (AUC, 0.9252 vs. 0.9074), possibly because of the small sample size.

Table 2. Subgroup analyses and meta-regression results

Parameter	Category	Studies(n)	Sensitivity(95%CI)	p value	Specificity(95%CI)	p value
Study Region	Euro-America	9	0.83(0.74–0.92)	<0.001	0.93(0.89–0.97)	<0.001
	Asia	9	0.90(0.84–0.96)		0.92(0.89–0.96)	
Score Criteria	UCEIS	6	085(0.74–0.95)	0.02	0.93(0.89–0.97)	<0.001
	MES	12	0.88(0.82–0.94)		0.93(0.89–0.97)	
Study Type	Prospective	6	0.82(0.71–0.94)	0.01	0.93(0.89–0.98)	<0.001
	Retrospective	12	0.89(0.84–0.94)		0.92(0.89–0.96)	
Study Center	Multi	4	0.75(0.56–0.93)	0.01	0.96(0.93–0.99)	0.01
	Single	14	0.89(0.85–0.94)		0.92(0.88–0.95)	
Patient (n)	>400	9	0.89(0.82–0.95)	0.08	0.93(0.89–0.96)	<0.001
	<400	9	0.86(0.77–0.94)		0.93(0.89–0.97)	

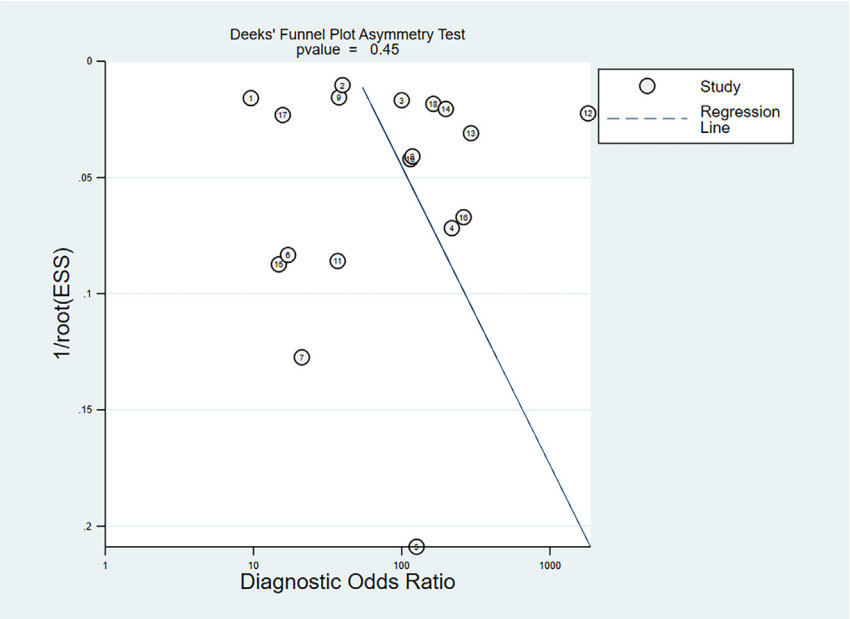


Figure 6. Deeks' funnel plot asymmetry test for publication

In addition to the automated determination of endoscopic remission of UC, studies have been conducted using artificial intelligence to read endoscopic images to determine whether UC is histologic remission.[28,35,44,45] Compared with traditional biopsy techniques, this technique will reduce the time cost and bleeding risk effectively. It will be an important research direction for AI for UC diagnosis.

Although the effectiveness of AI in diagnosing endoscopic remission in UC has been initially validated, some aspects require improvement.

(1) The training and test data used by AI models are based on manual annotation by endoscopists; as the endoscopy scores differ between endoscopists, this may affect the accuracy of AI models and lead to misdiagnoses and missed diagnoses. AI models must be rigorously evaluated and validated using different datasets to ensure their robustness and accuracy.

(2) UC is a relatively niche disease that makes it difficult to provide sufficient training data for AI models, which may cause overfitting or underfitting. HyperKvasir is a publicly available endoscopic image dataset containing 851 UC images labeled with the MES. The public dataset LIMUC[38] comprises 11,276 endoscopic images from 564 patients with UC. The public dataset TMC-UCM[43] comprises 7,978 images from 308 patients with UC. Several studies have used these datasets for AI model training and testing. More publicly available datasets can help researchers reduce time and financial costs, and improve the reproducibility of their studies.

(3) AI, especially DL models, is often considered a black box, mainly because of the complex structure of DL models and the difficulty in explaining the internal operating mechanisms. All 16 included studies used DL techniques, and only two studies explored the explainability of the models.[36,37] Medical image analysis requires rigorous scientific explanations. DL explainability is a key factor in user trust and is an important research direction for DL-aided diagnostic techniques.

In conclusion, this systematic review provides a comprehensive description and analysis of the current AI-assisted diagnosis of endoscopic remission in UC. The results showed that AI has good diagnostic ability and high clinical application value. However, more data are required to train and test the AI models to improve their reliability and generalizability.

Limitations of the study

This systematic review had some limitations. (1) There was a high degree of heterogeneity among the studies. The heterogeneity caused by different AI models, endoscope types, and differences owing to manually labeled data requires further investigation. (2) The sample size of some studies was small and may not have been representative. (3) When a separate meta-analysis was performed for each criterion, the number of studies corresponding to each criterion was small. For example, only four datasets were available for UCEIS = <1, and the analysis results may not be representative. (4) In a real environment, the accuracy of an AI model may be reduced. The actual endoscopic environment is considerably more complex than the experimental environment. Different light intensities, imaging angles, and interference from other diseases can affect the judgment of an AI model. (5) Only English-language literature was included, which may have inserted a language bias.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

Table 3. Meta-analysis results corresponding to different endoscopic remission criteria

Remission criteria	Studies(n)	Sensitivity(95%CI)	Specificity(95%CI)	DOR (95%CI)	AUC (95%CI)
UCEIS = 0	5	0.85(0.76–0.91)	0.95(0.91–0.97)	99.96(54.41–183.68)	0.96(0.94–0.98)
MES = 0	9	0.89(0.82–0.94)	0.93(0.88–0.96)	109.73(39.22–307.02)	0.97(0.95–0.98)
UCEIS<=1	4	0.92(0.64–0.99)	0.88(0.76–0.94)	87.97(13.87–558.00)	0.94(0.92–0.96)
MES<=1	13	0.01(0.84–0.94)	0.91(0.86–0.94)	89.15(49.37–160.97)	0.96(0.94–0.97)

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.108120.

ACKNOWLEDGMENTS

We thank Tao Tao for his helpful advice on our article. This study did not receive any specific grants from funding agencies in the public, commercial, or non-profit sectors.

AUTHOR CONTRIBUTIONS

Conceptualization, Y.S. and B.L.; methodology, B.L. and L.M.; investigation, L.M. and Y.P.S.; formal analysis, Y.S. and B.L., writing-original draft, B.L.; writing-review & editing, Y.S. and Y.P.S.; supervision, T.T.

DECLARATION OF INTERESTS

The authors declare that this study was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

Received: July 12, 2023
Revised: September 8, 2023
Accepted: September 29, 2023
Published: October 5, 2023

REFERENCES

1. Ordás, I., Eckmann, L., Talamini, M., Baumgart, D.C., and Sandborn, W.J. (2012). Ulcerative colitis. Lancet *380*, 1606–1619. https://doi.org/10.1016/S0140-6736(12)60150-0.
2. Swidsinski, A., Ladhoff, A., Pernthaler, A., Swidsinski, S., Loening-Baucke, V., Ortner, M., Weber, J., Hoffmann, U., Schreiber, S., Dietel, M., and Lochs, H. (2002). Mucosal flora in inflammatory bowel disease. Gastroenterology *122*, 44–54. https://doi.org/10.1053/gast.2002.30294.
3. Riley, S.A., Mani, V., Goodman, M.J., and Lucas, S. (1990). Why do patients with ulcerative colitis relapse? Gut *31*, 179–183. https://doi.org/10.1136/gut.31.2.179.
4. Kaplan, G.G. (2015). The global burden of IBD: from 2015 to 2025. Nat. Rev. Gastroenterol. Hepatol. *12*, 720–727. https://doi.org/10.1038/nrgastro.2015.150.
5. Ng, S.C., Shi, H.Y., Hamidi, N., Underwood, F.E., Tang, W., Benchimol, E.I., Panaccione, R., Ghosh, S., Wu, J.C.Y., Chan, F.K.L., et al. (2017). Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. Lancet *390*, 2769–2778. https://doi.org/10.1016/S0140-6736(17)32448-0.
6. Kaplan, G.G., and Ng, S.C. (2017). Understanding and Preventing the Global Increase of Inflammatory Bowel Disease. Gastroenterology *152*, 313–321.e2. https://doi.org/10.1053/j.gastro.2016.10.020.
7. Conrad, K., Roggenbuck, D., and Laass, M.W. (2014). Diagnosis and classification of ulcerative colitis. Autoimmun. Rev. *13*, 463–466. https://doi.org/10.1016/j.autrev.2014.01.028.
8. Sandborn, W.J., Feagan, B.G., Marano, C., Zhang, H., Strauss, R., Johanns, J., Adedokun,

O.J., Guzzo, C., Colombel, J.-F., Reinisch, W., et al. (2014). Subcutaneous golimumab maintains clinical response in patients with moderate-to-severe ulcerative colitis. Gastroenterology *146*, 96–109.e1. https://doi.org/10.1053/j.gastro.2013.06.010.

9. Inflammatory Bowel Disease Group Chinese Society of Gastroenterology Chinese Medical Association (2021). Chinese consensus on diagnosis and treatment in inflammatory bowel disease (2018, Beijing). J. Dig. Dis. *22*, 298–317. https://doi.org/10.1111/1751-2980.12994.
10. Ott, S.J., Kühbacher, T., Musfeldt, M., Rosenstiel, P., Hellmig, S., Rehman, A., Drews, O., Weichert, W., Timmis, K.N., and Schreiber, S. (2008). Fungi and inflammatory bowel diseases: Alterations of composition and diversity. Scand. J. Gastroenterol. *43*, 831–841. https://doi.org/10.1080/00365520801935434.
11. Peyrin-Biroulet, L., Sandborn, W., Sands, B.E., Reinisch, W., Bemelman, W., Bryant, R.V., D'Haens, G., Dotan, I., Dubinsky, M., Feagan, B., et al. (2015). Selecting Therapeutic Targets in Inflammatory Bowel Disease (STRIDE): Determining Therapeutic Goals for Treat-to-Target. Am. J. Gastroenterol. *110*, 1324–1338. https://doi.org/10.1038/ajg.2015.233.
12. Schroeder, K.W., Tremaine, W.J., and Ilstrup, D.M. (1987). Coated oral 5-aminosalicylic acid therapy for mildly to moderately active ulcerative colitis. A randomized study. N. Engl. J. Med. *317*, 1625–1629. https://doi.org/10.1056/NEJM198712243172603.
13. Travis, S.P.L., Schnell, D., Krzeski, P., Abreu, M.T., Altman, D.G., Colombel, J.-F., Feagan, B.G., Hanauer, S.B., Lémann, M., Lichtenstein, G.R., et al. (2012). Developing an instrument to assess the endoscopic severity of ulcerative colitis: the Ulcerative Colitis Endoscopic Index of Severity (UCEIS). Gut *61*, 535–542. https://doi.org/10.1136/gutjnl-2011-300486.
14. Vuitton, L., Peyrin-Biroulet, L., Colombel, J.F., Pariente, B., Pineton De Chambrun, G., Walsh, A.J., Panes, J., Travis, S.P.L., Mary, J.Y., and Marteau, P. (2017). Defining endoscopic response and remission in ulcerative colitis clinical trials: an international consensus. Aliment. Pharmacol. Ther. *45*, 801–813. https://doi.org/10.1111/apt.13948.
15. LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444. https://doi.org/10.1038/nature14539.
16. Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J., and Socher, R. (2021). Deep learning-enabled medical computer vision. NPJ Digit. Med. *4*, 5. https://doi.org/10.1038/s41746-020-00376-2.
17. Soffer, S., Lahat, A., and Klang, E. (2021). Artificial intelligence in colonoscopy. Lancet. Gastroenterol. Hepatol. *6*, 984. https://doi.org/10.1016/S2468-1253(21)00349-6.
18. Chudzik, P., Majumdar, S., Calivá, F., Al-Diri, B., and Hunter, A. (2018). Microaneurysm detection using fully convolutional neural networks. Comput. Methods Progr. Biomed. *158*, 185–192. https://doi.org/10.1016/j.cmpb.2018.02.016.
19. Quellec, G., Charrière, K., Boudi, Y., Cochener, B., and Lamard, M. (2017). Deep image mining for diabetic retinopathy screening. Med. Image Anal. *39*, 178–193. https://doi.org/10.1016/j.media.2017.04.012.
20. Stidham, R.W., and Takenaka, K. (2022). Artificial Intelligence for Disease Assessment in IBD: How Will it Change Our Practice? Gastroenterology *162*, 1493–1506. https://doi.org/10.1053/j.gastro.2021.12.238.
21. Kawamoto, A., Takenaka, K., Okamoto, R., Watanabe, M., and Ohtsuka, K. (2022). Systematic review of artificial intelligence-based image diagnosis for inflammatory bowel disease. Dig. Endosc. *34*, 1311–1319. https://doi.org/10.1111/den.14334.
22. Takenaka, K., Kawamoto, A., Okamoto, R., Watanabe, M., and Ohtsuka, K. (2022). Artificial intelligence for endoscopy in inflammatory bowel disease. Int. Res. *20*, 165–170. https://doi.org/10.5217/ir.2021.00079.
23. Tontini, G.E., Rimondi, A., Vernero, M., Neumann, H., Vecchi, M., Bezzio, C., and Cavallaro, F. (2021). Artificial intelligence in gastrointestinal endoscopy for inflammatory bowel disease: a systematic review and new horizons. Therap. Adv. Gastroenterol. *14*. 17562848211017730. https://doi.org/10.1177/17562848211017730.
24. Yang, L.S., Perry, E., Shan, L., Wilding, H., Connell, W., Thompson, A.J., Taylor, A.C.F., Desmond, P.V., and Holt, B.A. (2022). Clinical application and diagnostic accuracy of artificial intelligence in colonoscopy for inflammatory bowel disease: systematic review. Endosc. Int. Open *10*, E1004–E1013. https://doi.org/10.1055/a-1846-0642.
25. Jahagirdar, V., Bapaye, J., Chandan, S., Ponnada, S., Kochhar, G.S., Navaneethan, U., and Mohan, B.P. (2023). Diagnostic accuracy of convolutional neural network–based machine learning algorithms in endoscopic severity prediction of ulcerative colitis: a systematic review and meta-analysis. Gastrointest. Endosc. *98*, 145–154.e8. https://doi.org/10.1016/j.gie.2023.04.2074.
26. Ozawa, T., Ishihara, S., Fujishiro, M., Saito, H., Kumagai, Y., Shichijo, S., Aoyama, K., and Tada, T. (2019). Novel computer-assisted diagnosis system for endoscopic disease activity in patients with ulcerative colitis. Gastrointest. Endosc. *89*, 416–421.e1. https://doi.org/10.1016/j.gie.2018.10.020.
27. Stidham, R.W., Liu, W., Bishu, S., Rice, M.D., Higgins, P.D.R., Zhu, J., Nallamothu, B.K., and Waljee, A.K. (2019). Performance of a Deep Learning Model vs Human Reviewers in Grading Endoscopic Disease Severity of Patients With Ulcerative Colitis. JAMA Netw. Open *2*, e193963. https://doi.org/10.1001/jamanetworkopen.2019.3963.
28. Takenaka, K., Ohtsuka, K., Fujii, T., Negi, M., Suzuki, K., Shimizu, H., Oshima, S., Akiyama, S., Motobayashi, M., Nagahori, M., et al. (2020). Development and Validation of a Deep Neural Network for Accurate Evaluation of Endoscopic Images From Patients With Ulcerative Colitis. Gastroenterology *158*, 2150–2157. https://doi.org/10.1053/j.gastro.2020.02.012.
29. Huang, T.-Y., Zhan, S.-Q., Chen, P.-J., Yang, C.-W., and Lu, H.H.-S. (2021). Accurate diagnosis of endoscopic mucosal healing in ulcerative colitis using deep learning and machine learning. J. Chin. Med. Assoc. *84*, 678–681. https://doi.org/10.1097/JCMA.0000000000000559.
30. Gottlieb, K., Requa, J., Karnes, W., Chandra Gudivada, R., Shen, J., Rael, E., Arora, V., Dao, T., Ninh, A., and McGill, J. (2021). Central Reading of Ulcerative Colitis Clinical Trial Videos Using Neural Networks. Gastroenterology *160*, 710–719.e2. https://doi.org/10.1053/j.gastro.2020.10.024.
31. Yao, H., Najarian, K., Gryak, J., Bishu, S., Rice, M.D., Waljee, A.K., Wilkins, H.J., and Stidham, R.W. (2021). Fully automated endoscopic disease activity assessment in ulcerative colitis. Gastrointest. Endosc. *93*, 728–736.e1. https://doi.org/10.1016/j.gie.2020.08.011.
32. Bhambhvani, H.P., and Zamora, A. (2021). Deep learning enabled classification of Mayo endoscopic subscore in patients with ulcerative colitis. Eur. J. Gastroenterol. Hepatol. *33*, 645–649. https://doi.org/10.1097/MEG.0000000000001952.
33. Gutierrez Becker, B., Arcadu, F., Thalhammer, A., Gamez Serna, C., Feehan, O., Drawnel, F., Oh, Y.S., and Prunotto, M. (2021). Training and deploying a deep learning model for endoscopic severity grading in ulcerative colitis using multicenter clinical trial data. Ther. Adv. Gastrointest. Endosc. *14*. 2631774521990623. https://doi.org/10.1177/2631774521990623.
34. Patel, M., Gulati, S., Iqbal, F., and Hayee, B. (2022). Rapid development of accurate artificial intelligence scoring for colitis disease activity using applied data science techniques. Endosc. Int. Open *10*, E539–E543. https://doi.org/10.1055/a-1790-6201.
35. Byrne, M.F., Panaccione, R., East, J.E., Iacucci, M., Parsa, N., Kalapala, R., Reddy, D.N., Ramesh Rughwani, H., Singh, A.P., Berry, S.K., et al. (2023). Application of Deep Learning Models to Improve Ulcerative Colitis Endoscopic Disease Activity Scoring Under Multiple Scoring Systems. J. Crohns Colitis *17*, 463–471. https://doi.org/10.1093/ecco-jcc/jjac152.
36. Sutton, R.T., Zai Ane, O.R., Goebel, R., and Baumgart, D.C. (2022). Artificial intelligence enabled automated diagnosis and grading of ulcerative colitis endoscopy images. Sci. Rep. *12*, 2748. https://doi.org/10.1038/s41598-022-06726-2.
37. Luo, X., Zhang, J., Li, Z., and Yang, R. (2022). Diagnosis of ulcerative colitis from endoscopic images based on deep learning. Biomed. Signal Process Control *73*, 103443. https://doi.org/10.1016/j.bspc.2021.103443.
38. Polat, G., Kani, H.T., Ergenc, I., Ozen Alahdab, Y., Temizel, A., and Atug, O. (2023). Improving the Computer-Aided Estimation of Ulcerative Colitis Severity According to Mayo Endoscopic Score by Using Regression-Based Deep Learning. Inflamm. Bowel Dis. *29*, 1431–1439. https://doi.org/10.1093/ibd/izac226.
39. Fan, Y., Mu, R., Xu, H., Xie, C., Zhang, Y., Liu, L., Wang, L., Shi, H., Hu, Y., Ren, J., et al. (2023). A novel deep learning-based computer-aided diagnosis system for predicting inflammatory activity in ulcerative colitis. Gastrointest. Endosc. *97*, 335–346. https://doi.org/10.1016/j.gie.2022.08.015.
40. Iacucci, M., Parigi, T.L., Del Amor, R., Meseguer, P., Mandelli, G., Bozzola, A., Bazarova, A., Bhandari, P., Bisschops, R., Danese, S., et al. (2023). Artificial Intelligence Enabled Histological Prediction of Remission or Activity and Clinical Outcomes in Ulcerative Colitis. Gastroenterology *164*, 1180–1188.e2. https://doi.org/10.1053/j.gastro.2023.02.031.
41. Lo, B., Liu, Z., Bendtsen, F., Igel, C., Vind, I., and Burisch, J. (2022). High Accuracy in Classifying Endoscopic Severity in Ulcerative Colitis Using Convolutional Neural Network. Am. J. Gastroenterol. *117*, 1648–1654. https://doi.org/10.14309/ajg.0000000000001904.
42. Kadota, T., Abe, K., Bise, R., Kawamura, T., Sakiyama, N., Tanaka, K., and Uchida, S.

(2022). Automatic Estimation of Ulcerative Colitis Severity by Learning to Rank With Calibration. IEEE Access *10*, 25688–25695. https://doi.org/10.1109/ACCESS.2022.3155769.

43. Wang, G., Zhang, S., Li, J., Zhao, K., Ding, Q., Tian, D., Li, R., Zou, F., and Yu, Q. (2023). CB-HRNet: A Class-Balanced High-Resolution Network for the evaluation of endoscopic activity in patients with ulcerative colitis. Clin. Transl. Sci. *16*, 1421–1430. https://doi.org/10.1111/cts.13542.
44. Takenaka, K., Fujii, T., Kawamoto, A., Suzuki, K., Shimizu, H., Maeyashiki, C., Yamaji, O., Motobayashi, M., Igarashi, A., Hanazawa, R., et al. (2022). Deep neural network for video colonoscopy of ulcerative colitis: a cross-sectional study. Lancet. Gastroenterol. Hepatol. *7*, 230–237. https://doi.org/10.1016/S2468-1253(21)00372-1.
45. Takenaka, K., Ohtsuka, K., Fujii, T., Oshima, S., Okamoto, R., and Watanabe, M. (2021). Deep Neural Network Accurately Predicts Prognosis of Ulcerative Colitis Using Endoscopic Images. Gastroenterology *160*, 2175–2177.e3. https://doi.org/10.1053/j.gastro.2021.01.210.
46. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9. https://doi.org/10.1109/CVPR.2015.7298594.
47. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826. https://doi.org/10.1109/CVPR.2016.308.
48. Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017). Aggregated Residual Transformations for Deep Neural Networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5987–5995. https://doi.org/10.1109/CVPR.2017.634.
49. Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., Pogorelov, K., Lux, M., Nguyen, D.T.D., et al. (2020). HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Sci. Data *7*, 283. https://doi.org/10.1038/s41597-020-00622-y.
50. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE), pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.
51. Tan, M., and Le, Q. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning (PMLR), pp. 6105–6114.
52. Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K.Q. (2017). Densely Connected Convolutional Networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269. https://doi.org/10.1109/CVPR.2017.243.
53. Simonyan, K., and Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. Preprint at arXiv. https://doi.org/10.48550/arXiv.1409.1556.
54. Liu, T.-Y. (2007). Learning to Rank for Information Retrieval. FNT Information Retrieval *3*, 225–331. https://doi.org/10.1561/1500000016.
55. Chauhan, N.K., and Singh, K. (2018). A Review on Conventional Machine Learning vs Deep Learning. In 2018 International Conference on Computing, Power and Communication Technologies (GUCON), pp. 347–352. https://doi.org/10.1109/GUCON.2018.8675097.
56. Wang, P., Fan, E., and Wang, P. (2021). Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. Pattern Recogn. Lett. *141*, 61–67. https://doi.org/10.1016/j.patrec.2020.07.042.
57. Salameh, J.-P., Bossuyt, P.M., McGrath, T.A., Thombs, B.D., Hyde, C.J., Macaskill, P., Deeks, J.J., Leeflang, M., Korevaar, D.A., Whiting, P., et al. (2020). Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. BMJ *370*, m2632. https://doi.org/10.1136/bmj.m2632.
58. Whiting, P.F., Rutjes, A.W.S., Westwood, M.E., Mallett, S., Deeks, J.J., Reitsma, J.B., Leeflang, M.M.G., Sterne, J.A.C., and Bossuyt, P.M.M.; QUADAS-2 Group (2011). QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. Ann. Intern. Med. *155*, 529–536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009.
59. Van Houwelingen, H.C., Zwinderman, K.H., and Stijnen, T. (1993). A bivariate approach to meta-analysis. Stat. Med. *12*, 2273–2284. https://doi.org/10.1002/sim.4780122405.
60. Van Houwelingen, H.C., Arends, L.R., and Stijnen, T. (2002). Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat. Med. *21*, 589–624. https://doi.org/10.1002/sim.1040.
61. Booth, A., Clarke, M., Ghersi, D., Moher, D., Petticrew, M., and Stewart, L. (2011). An international registry of systematic-review protocols. Lancet *377*, 108–109. https://doi.org/10.1016/S0140-6736(10)60903-8.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Literature summary statistics	Science Data Bank	https://doi.org/10.57760/sciencedb.10950
Software and algorithms		
StataSE 16.0	StataCorp LLC	https://www.stata.com
Review Manager 5.4	The Cochrane Collaboration	https://revman.cochrane.org/info
EndNote 20	Clarivate Analytics LLC	https://endnote.com/downloads

RESOURCE AVAILABILITY

Lead contact

For additional information and resources should be directed to the lead contact, Yanting Shi (yantingshi@hotmail.com).

Materials availability

This study is a meta-analysis and did not use or generate any reagents.

Data and code availability

The data used in this meta-analysis came from published studies, and no new data or codes were used. All data are described in the "key resources table" section.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Our study does not use experimental models typical in the life sciences. The participant characteristics and AI algorithmic details of the included studies are shown in Table S1.

METHOD DETAILS

This study was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy Studies (PRISMA-DTA)[57] guidelines. The PRISMA-DTA checklist is presented in Table S2. All the data for this study were collected from the included literature, and ethical approval was not required.

Searching strategy

We searched five databases: PubMed, Web of Science, Embase, IEEE Xplore, and the Cochrane Library, without any time restrictions. The final search was performed on August 15, 2023. Keywords related to DL included "deep learning," "artificial intelligence," "machine learning," "computer-aided," and "natural networks." Keywords related to UC included "inflammatory bowel disease", "ulcerative colitis", "IBD," and "UC." Table S3 provides the detailed search strategy. In addition, we checked the references of related articles to identify more relevant studies.

Eligibility criteria

The inclusion criteria were as follows: (1) prospective or retrospective studies, (2) used AI to analyze colonoscopy images/videos to assess UC severity, (3) use of UCEIS or MES as the scoring criteria, (4) ability to obtain 2×2 tables (true positive, false negative, false positive, and true negative) directly or indirectly, (5) the data in the 2×2 table represented one of the following four classifications: UCEIS 0 vs. UCEIS 1–8, MES 0 vs. MES 1–3, UCEIS 0–1 vs. UCEIS 2–8, MES 0–1 vs. MES 3–4, (6) used publicly available AI algorithms with detailed descriptions, and (7) provided a detailed description of the datasets used for model training and testing.

The exclusion criteria were as follows: (1) articles without experimental data, such as conference abstracts, reviews, and letters, (2) articles without the full text, and (3) articles for which a 2×2 table data could not be extracted.

Two authors (B.L. and L.M.) reviewed the retrieved articles using this strategy, and any disagreements were resolved through discussion.

Data extraction

Two authors (B.L. and L.M.) independently extracted data from eligible articles and resolved conflicts through discussions with Y.S. The following data were extracted: first author, publication year, study site, patient information, endoscopic imaging type, AI algorithm, endoscopy scoring criteria, the training set sample information, the test set sample information, 2×2 tables, sensitivity, and specificity.

Multiple test cohorts in a study were prioritized in the following order: external test cohort, video test cohort, and large sample size cohort.

Multiple endoscopic remission criteria, if included, were all extracted; however, only one test result per study was selected for the primary meta-analysis and prioritized in the following order: UCEIS = 0, MES = 0, UCEIS = < 1, MES = < 1. Others were used for the analysis to obtain secondary results.

QUANTIFICATION AND STATISTICAL ANALYSIS

Quality assessment

The quality of the screened literature was evaluated using the Quality Assessment of Diagnostic Accuracy Studies 2 tool,[58] which consists of four main components: patient selection, index test, reference standard, and process and progress. Quality evaluation was performed independently by two authors (L.M. and T.T.), and in cases of disagreement, the final results were decided by a joint discussion. Charts for quality assessment were drawn using the Review Manager 5.4 (Cochrane Collaboration, Oxford, UK).

Statistical analysis

The meta-analysis was performed using Stata/SE software (version 16.0; Stata, College Station, TX, USA) with the Midas package installed. Midas commands used a bivariate mixed-effects regression model[59,60] to pool the data. We pooled the sensitivity, specificity, positive likelihood ratio (PLR), negative likelihood ratio (NLR), diagnostic odds ratio (DOR), and 95% confidence intervals (CI). Summary receiver operating characteristic (SROC) curves were plotted and the area under the curve (AUC) was calculated.

The I^2 statistic assessed the heterogeneity between studies, and values of $I^2 > 50\%$ indicated substantial heterogeneity. Sources of heterogeneity were explored using subgroup and meta-regression analyses. In each subgroup, for $P < 0.05$, $P < 0.01$, and $P < 0.001$, the differences were considered statistically significant, more significant, and extremely significant, respectively.

Publication bias was assessed using Deeks' funnel plot, which indicated the possibility of publication bias if the funnel plot was asymmetrical. The slope coefficients $P < 0.1$ and $P < 0.05$ indicated asymmetry and significant asymmetry, respectively, in the funnel plot.

The stability and reliability of the meta-analysis results were assessed using sensitivity analysis. Changes in I^2 value were observed by sequentially excluding each study, and a large change indicated heterogeneity in the study.

ADDITIONAL RESOURCES

The study was registered, before its initiation, in PROSPERO[61] on February 06, 2023 (ID: CRD42023391093).