Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multicenter Validation of a ResNet50-Based Deep Convolutional Network for Endoscopic Grading of Ulcerative Colitis Severity

Mingrui Qi
The First Affiliated Hospital of Jinzhou Medical University
Cong Dai
The First Affiliated Hospital of China Medical University
Chuqiao Li
The First Affiliated Hospital of Jinzhou Medical University
Wei Niu
The First Affiliated Hospital of Jinzhou Medical University
Baohai Liu
The First Affiliated Hospital of Jinzhou Medical University
Xiaofei Li
The First Affiliated Hospital of Jinzhou Medical University
Quanlin Li
Fudan University Zhongshan Hospital
Lianyi Guo
angle_gly@163.com

The First Affiliated Hospital of Jinzhou Medical University

Research Article

Keywords: Artificial Intelligence, Ulcerative Colitis, Endoscopic Disease Severity, Colonoscopy Images, Deep Convolutional Neural Networks (DCNN)

Posted Date: September 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10802303/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Objective

To tackle inconsistent endoscopic severity scores in ulcerative colitis (UC) caused by subjective clinician interpretation, this study developed a deep convolutional neural network (DCNN)-based system. Unlike many studies relying on single-center or public data, this model was externally validated using multi-institutional data to enhance generalizability.

Methods

Colonoscopy images from patients with UC were retrospectively collected from one hospital (Dec 2019–May 2024) for training and testing. An AI system based on ResNet50 was developed to assess endoscopic severity. Additional images from two other tertiary hospitals (May–Dec 2024) were used for external validation. Performance was measured by sensitivity, specificity, accuracy, PPV, NPV, AUC, and weighted Kappa.

Results

The optimized model achieved sensitivities of 91.8% (normal), 81.8% (mild), 78.8% (moderate), and 64.8% (severe), with a mean AUC of 0.899. External validation on 1,732 images showed improved sensitivities: 96.0% (normal), 84.2% (mild), 79.1% (moderate), and 92.1% (severe), with specificities > 94% across categories and a mean AUC of 0.885. AI-endoscopist agreement (ICC = 0.840, QWK = 0.840) was comparable to inter-expert consistency (ICC = 0.880, QWK = 0.880).

Conclusions

The ResNet50-based AI system demonstrated high diagnostic accuracy for UC severity evaluation, performing comparably to experienced endoscopists. It offers a reliable and generalizable tool for supporting clinical assessment and treatment planning.

Introduction

Ulcerative colitis (UC), an idiopathic chronic inflammatory bowel disease, manifests as recurrent inflammation of the colorectal mucosa. This condition is characterized by diffuse and continuous superficial inflammation extending proximally from the rectum, accompanied by corresponding histopathological alterations. The disease typically initiates in the distal rectum and progresses contiguously to involve segmental or pancolonic regions[1]. Previous studies have established endoscopic mucosal healing as the paramount therapeutic objective in UC. The assessment of endoscopic disease severity—critical for defining mucosal healing—primarily utilizes two validated

scoring systems: the Mayo Endoscopic Score (MES) and the Ulcerative Colitis Endoscopic Index of Severity (UCEIS). Precise evaluation of endoscopic severity is indispensable for personalizing treatment regimens, monitoring disease trajectory, and predicting clinical outcomes. However, these assessments exhibit inherent subjectivity due to inter-observer variability influenced by clinicians' endoscopic interpretation experience and heterogeneous healthcare settings.

Artificial intelligence, as a branch of computer science, performs complex tasks by simulating the human brain[2]. In medical artificial intelligence (AI), applications are broadly categorized into virtual and physical domains. The physical domain is epitomized by the Da Vinci Surgical System—a robotic platform extensively deployed in complex procedures across general surgery, urology, and gynecology[3]. Machine learning (ML)—a fundamental subset of artificial intelligence within the virtual domain—progressively enhances task-specific performance through iterative algorithmic training. Core methodologies include deep learning (DL), supervised learning (SL), support vector machines (SVMs), and convolutional neural networks (CNNs), with the latter constituting a specialized DL architecture particularly effective in visual pattern recognition[4]. Artificial intelligence (AI) applications are currently advancing at an unprecedented pace within gastrointestinal endoscopy. By facilitating real-time analysis of large-scale endoscopic image datasets, AI systems reduce inter-observer variability and enhance diagnostic objectivity during endoscopic examinations.

Artificial intelligence (AI) has made significant strides across various medical disciplines, showcasing a wide array of diagnostic and therapeutic utilities in dermatology[5], radiology[6], ophthalmology[7], stomatology[8], bioinformatics[9], and pathology[10]. These technological advancements are reshaping clinical paradigms by amalgamating patient screening, diagnostic categorization, and prognostic forecasting. In gastroenterology, initial AI applications were primarily centered on the detection of colorectal polyps and aiding in capsule endoscopy. Studies show that models utilizing convolutional neural networks (CNN) can achieve diagnostic accuracy comparable to expert endoscopists.[11] Recent advancements in artificial intelligence have propelled the automated evaluation of UC endoscopic severity into the spotlight of current research. Noteworthy investigations have demonstrated the superiority of the Convolutional Neural Network (CNN)-based Ulcerative Colitis Endoscopic Severity Index (UCEIS) assessment over the conventional Mayo Endoscopic Score (MES) in terms of grading accuracy. Moreover, the CNN-based approach exhibits higher inter-rater reliability. Convolutional neural network (CNN) methods have demonstrated high inter-rater reliability and improved diagnostic accuracy in evaluating the severity of UC, aligning closely with the primary therapeutic objective of achieving favorable endoscopic and histological outcomes in UC management. However, existing research faces limitations, such as reliance on data from single medical institutions or open-source datasets, hindering the universal applicability of AI in identifying UC severity through endoscopy. Moreover, current studies predominantly concentrate on CNNs and tend to confine the use of UC Endoscopic Index of Severity (UCEIS) scores to experimental contexts. Advances in technology have brought deep convolutional neural network (DCNN) models to the forefront, with studies demonstrating the superior image recognition and autonomous learning capabilities of ResNet models within DCNNs[12]. In this

investigation, we conducted architecture selection for the ResNet50 model and integrated it with UCEIS scores to create a DCNN model trained on white light endoscopic images of patients with UC. Following rigorous internal validation using validation and test cohorts, the model underwent external validation to assess endoscopic severity. The primary objective of this research is to evaluate the diagnostic performance of the model compared to that of practicing endoscopists and to establish an objective framework for consistent real-time analysis of UC severity in clinical settings.

Methods

1. Data collection

A retrospective analysis was conducted on 5,000 colonoscopic images of patients with ulcerative colitis (UC) who underwent colonoscopy at the Department of Gastroenterology, First Affiliated Hospital of Jinzhou Medical University between December 1, 2019, and May 1, 2024. These images were divided into internal training and internal test sets. Additionally, colonoscopic images of 1,732 UC patients from two other tertiary hospitals in Liaoning Province, captured between May 2024 and December 2024, were incorporated into the external test set. All procedures were conducted using Olympus colonoscopy equipment (Olympus, Tokyo, Japan). Exclusion criteria comprised: (1) Lesion characteristics exclusion criteria: a. Non-UC characteristic changes (e.g., isolated proctitis, typical longitudinal ulcers in Crohn's disease, microscopic colitis, ischemic colitis, radiation colitis, eosinophilic gastroenteritis). b. Coexistence with other intestinal diseases (e.g., confirmed microscopic colitis, active radiation enteritis). c. Interference from tumorous lesions (e.g., advanced colorectal cancer, invasive lymphoma lesions). (2) Image quality exclusion criteria: a. Inadequate bowel preparation (Boston score ≤ 5). b. Incomplete documentation of critical anatomical sites (e.g., missing ileocecal valve/rectoanal junction). c. Blurred images or artifacts affecting interpretation.The images were evenly distributed across the affected intestinal segments (ileocecal region, ascending colon, transverse colon, descending colon, sigmoid colon, and rectum). Ten colonoscopic images of the affected segments were chosen as candidates and subjected to further screening. Subsequently, the images underwent processing using Adobe Photoshop (PS) software, including adjustments for orientation (left-right, up-down), contrast enhancement, and image sharpening.

Images were annotated based on a validated UCEIS score, which evaluates vascular texture, intestinal bleeding, erosion, and ulceration. Specific score details are outlined in Supplementary Table 1, with illustrative examples provided in Supplementary Fig. 1. In this study, a standardized scoring integration method was utilized to aggregate scores from each dimension of the UCEIS[13] into a four-category grading system: Grade (normal, no active mucosal lesions), Grade 1 (mild), Grade 2 (moderate), and Grade 3 (severe). During data preprocessing, images were organized using a structured file management system: the internal training set (n = 4,000) employed a batch storage approach (each batch comprising 100 images as an independent data unit), while the internal validation set (n = 1,000) and the external multicenter validation set (n = 1732) were stored separately.The labeling process by experts adhered strictly to the principle of blind evaluation and involves three experienced endoscopic specialists (each

with ≥ 10 years of experience). Prior to assessment, all images are desensitized to remove potential biases, such as patient identification, examination timing, and endoscopic equipment settings. Throughout the scoring phase, experts do not have access to the patient's clinical history, laboratory results, or treatment records, past or present. A three-tier review system was established to address scoring discrepancies, defined as differences of ≥ 1 point between individual scores or inconsistencies in total scores. Initially, two conflicting experts, BH Liu and W Niu, conduct a reevaluation. If a consensus is not reached, the final decision is made by the senior chief physician, LY Guo, who has 25 years of experience. The ultimate score is determined through a majority vote.

2. Construction of DCNN models

In the realm of deep learning, convolutional neural networks (CNNs) have proven highly effective in tasks like image recognition and classification due to their strong feature extraction abilities. To aid in diagnosing ulcerative colitis severity via endoscopy, we conducted a comparative analysis of models, ultimately favoring deep convolutional neural networks (DCNNs) and selecting the ResNet (residual network) model. ResNet introduces an innovative architecture to deep neural networks. Unlike traditional deep neural networks, which often face challenges like gradient vanishing or explosion with increased depth, ResNet mitigates these issues by incorporating residual blocks. These blocks include shortcut connections between layers to enable direct gradient flow during training, addressing gradient degradation in deep networks and allowing for the creation of deeper models. ResNet50, a specific implementation of the ResNet architecture, consists of 50 layers. This intermediate depth balances robust feature learning capabilities with manageable computational complexity, avoiding overfitting risks associated with excessively deep networks.

The ResNet50 model provides a distinct benefit through its capability for pre-training on large image datasets and subsequent transfer of pre-trained weights into the model. Fine-tuning the fully connected layers of the model with ulcerative colitis endoscopic image data, a form of transfer learning, markedly decreases training duration, improves training effectiveness and precision, and addresses the risk of overfitting. This method streamlines the process of assisting in diagnosing the severity of ulcerative colitis during endoscopy within a practical timeframe and resource framework.

The ResNet50 architecture is composed of convolutional layers, residual blocks, pooling layers, and a fully connected layer, with the backbone structure functioning as the primary feature extraction element. The model initiates with a 7×7 convolutional layer (strided by 2) followed by downsampling facilitated by a 3×3 maximum pooling layer (strided by 2). Residual blocks, a pivotal feature of ResNet50, address gradient vanishing problems and enhance model expressiveness through shortcut connections. Each residual block typically adopts a Bottleneck design, integrating 1×1, 3×3, and 1×1 convolutional layers with batch normalization and ReLU activation functions. Initial downsampling in pooling layers is achieved using a maximum pooling layer, while global average pooling is utilized for final feature compression and classification. The fully connected layer usually comprises a single layer responsible

for transforming abstract features into the classification space. The schematic illustration of the model structure is presented in Fig. 1.

3.Data processing

We collected 5,000 colonoscopy images from the Department of Gastroenterology at the First Affiliated Hospital of Jinzhou Medical University. These images were then divided into an internal training set and an internal validation set at a ratio of 4:1. Additionally, colonoscopy images from UC patients from two other tertiary hospitals in Liaoning Province were gathered and included in the external validation set. All collected colonoscopy images originated from clinical examinations and contained accompanying text information such as time, scope number, and model. To standardize the images for easy identification, each image underwent cropping using image processing software (e.g., Adobe Photoshop) to retain only the portion visible under the microscope.

Following image processing, the pixel dimensions of each image range from 400*600 to 1980*1020. To enhance the model's sensitivity to image recognition, no specific alterations are made to the image pixels. For improved recognition accuracy, internal training set images undergo enhancement, sharpening, rotation, and flipping while preserving image integrity. These processed images are organized into new folders for subsequent operations, as depicted in Fig. 2 for the internal training set. To ensure even distribution of images in the training set and mitigate potential biases from clinical indicators and treatment variables, a random sampling method and double-blind approach are employed. Expert-labeled images of varying severity in patients are segregated into four severity-marked folders. The model is then trained automatically on these images. Upon completion of data collection, images from the external validation set undergo software processing to retain only the endoscopic lesion area, discarding extraneous text such as imaging details. Subsequently, these processed images are categorized into respective folders using the same methodology.

The colonoscopy images of patients with UC obtained from our hospital's Department of Gastroenterology were expertly annotated for internal model validation. Subsequently, colonoscopy images of UC patients from two other tertiary hospitals in Liaoning Province were also expertly annotated for external validation of the model.

4.Software hardware and model training and evaluation

The study utilized the AutoDL computing cloud platform (www.autodl.com) with Python version 3.12 (Ubuntu 22.04). The hardware configuration included an NVIDIA RTX 4090 GPU (24GB), 16 vCPU Intel(R) Xeon(R) Platinum 8352V CPU@2.10GHz, 120GB of total memory, and a 50GB data disk. The research focused on developing, training, and evaluating an endoscopic image scoring model for patients with ulcerative colitis. Optimization of gradient direction and magnitude was achieved through dynamic adjustment of the learning rate.

The RMSProp optimizer is selected to optimize the model training process due to its ability to dynamically adjust the learning rate based on the training progress. This optimizer accurately optimizes the gradient direction and magnitude, effectively preventing issues such as vanishing or exploding gradients. Consequently, it ensures stable and efficient convergence to the global optimal solution during training. The loss function employed is a cross-entropy loss function that incorporates log_softmax, known for its superior performance in multi-class classification tasks by effectively quantifying the disparity between model predictions and true labels. To mitigate overfitting and enhance model generalization, a label smoothing strategy is introduced. This strategy diminishes the model's reliance on specific sample labels by smoothing the true labels, thereby enhancing the model's performance on unseen data. Additionally, a warm-up strategy is implemented in the learning rate adjustment process. Initially, a small learning rate is set to maintain convergence stability and prevent abrupt oscillations or convergence failures caused by excessively high learning rates. As training progresses to the fifth epoch, the learning rate is gradually increased to expedite model convergence and enhance training efficiency.

The ResNet50 model was utilized to assess endoscopic severity in a retrospective analysis of colonoscopic images from patients with ulcerative colitis in our gastroenterology department. The model's predictions, denoted as "Prediction," along with their corresponding confidence intervals, labeled as "Confidence Interval," were recorded. Subsequently, statistical analyses were performed on the AI model's predictions, including the construction of a Confusion Matrix. The diagnostic outcomes of the AI model and those of endoscopic experts regarding the severity grading of colonoscopic images in patients with ulcerative colitis were statistically compared for both the internal and external test sets. The agreement between the AI model and endoscopic experts was quantified to ascertain any significant disparities in diagnoses.

5.Statistical analysis

We utilized RStudio (R version 4.4) and GraphPad Prism 8 for statistical analysis to assess AI model performance metrics: sensitivity, specificity, accuracy, positive predictive value (PPV), negative predictive value (NPV), and area under the curve (AUC). We also evaluated the intraclass correlation coefficient (ICC) between AI predictions and expert endoscopic assessments, along with the consistency of endoscopic severity of UC using the Quadratic Weighted Kappa Coefficient (QWK). A 95% confidence interval (CI) was determined for these metrics.

Results

1. Data collection

Retrospectively collected colonoscopy images from 6,178 UC patients at the Department of Gastroenterology, First Affiliated Hospital of Jinzhou Medical University underwent screening for duplicates and inconsistent expert ratings, resulting in a final internal dataset of 5,000 images. This dataset was partitioned into a internal training set (4,000 images) and a internal test set (1,000 images)

following a 4:1 ratio. Additionally, a total of 1732 colonoscopic images from 132 patients with UC were gathered from two tertiary hospitals in Liaoning Province to establish an external test set. This dataset was employed to evaluate the diagnostic accuracy of artificial intelligence (AI) algorithms in recognizing the endoscopic severity of UC. Comprehensive baseline patient information for this external validation set is presented in Table 1 external.

Table 1
Clinical characteristics of UC patients in the internal validation set and external validation set

Patients with UC	**Internal(n = 200)**	**External(n = 132)**
Male, n(%)	128(64.0%)	78(59.1%)
Female, n(%)	72(36.0%)	54(40.9%)
Montreal classification, n (%)		
E1	24(12.0%)	20(15.2%)
E2	32(16.0%)	38(28.8%)
E3	144(72.0%)	74(56.1%)
Medical history (years), n (%)		
0–5	160(80.0%)	89(67.4%)
5–10	32(16.0%)	14(10.6%)
> 10	8(4.0%)	29(22.0%)
White blood cell count (*10^9/L) (mean ± SD)	8.81 ± 3.41	8.97 ± 2.45
Red blood cell count (*10^9/L) (mean ± SD)	4.48 ± 0.44	4.13 ± 1.27
Neutrophil count (*10^9/L) (mean ± SD)	5.88 ± 2.80	6.06 ± 2.91
Hemoglobin (g/L) (mean ± SD)	118.54 ± 23.13	112.63 ± 25.14
Erythrocyte sedimentation rate (mm/h) (mean ± SD)	41.27 ± 28.17	43.82 ± 26.42
Albumin (g/L) (mean ± SD)	36.46 ± 5.93	37.42 ± 5.89
CRP (mg/L) (mean ± SD)	31.44 ± 15.10	33.58 ± 16.37

Note: Montreal classification of UC: E1 is ulcerative proctitis, the lesion scope is limited to the rectum;E2 is left half ulcerative colitis, the lesion scope accumulates to the left half colon; the left half colon is involved, the lesion scope starts from the rectum and extends upward, but does not exceed the splenic flexure of the colon;E3 inflammation extends beyond the splenic flexure of the colon and even extends to the whole colon, which means that lesions occur in the whole colon region from the rectum to the ascending colon, transverse colon and descending colon.

2. Artificial intelligence model establishment and testing

2.1 preliminary tests

The AI model effectively predicted the endoscopic severity of UC, classifying grade 0 (normal) with sensitivity, specificity, PPV, NPV, and F1 Score of 81.5%, 92.5%, 74.9%, 94.8%, and 78.0%, respectively. For grade 1 (mild), these metrics were 60.3%, 83.8%, 34.0%, 93.9%, and 43.5%. In predicting grade 2 (moderate), the sensitivity, specificity, PPV, NPV, and F1 Score were 68.4%, 90.7%, 91.1%, 67.3%, and 78.1%. For grade 3 (severe), the model achieved sensitivity, specificity, PPV, NPV, and F1 Score of 91.4%, 95.8%, 65.5%, 99.2%, and 76.3%. Refer to Table 2 for detailed values and confidence intervals, Fig. 3(a) for the confusion matrix, and Fig. 4(a) for model loss rate accuracy. The present study evaluated the diagnostic performance of a predictive model for assessing the endoscopic severity of ulcerative colitis. While the model demonstrated some capacity for diagnosis, it did not meet the anticipated level of diagnostic accuracy. Consequently, the model architecture was optimized, and further internal training and validation were conducted to address the identified limitations.

Table 2
Sensitivity, specificity, accuracy, positive predictive value (PPV), Negative predictive value (NPV),and F1score of the internal validation model before optimization, the optimized internal validation model and external validation model.

validation	Inflammatory severity grading of ulcerative colitis (UC)	Sensitivity, % (95% CI)	Specificity, % (95% CI)	PPV, % (95%CI)	NPV, % (95%CI)	F1score, % (95%CI)
internal validation before optimization	0	81.5% (95%CI 75.6, 86.4)	92.5% (95% CI 90.4, 94.2)	74.9% (95% CI 68.8, 80.3)	94.8% (95% CI 92.9, 96.2)	78.0% (95%CI 73.8, 81.9)
	1	60.3% (95% CI 51.0, 69.1)	83.8% (95% CI 81.2, 86.2)	34.0% (95% CI 27.7, 40.7)	93.9% (95% CI 92.0, 95.5)	43.5% (95%CI 36.7, 49.9)
	2	68.4% (95% CI 64.4, 72.1)	90.7% (95% CI 87.5, 93.3)	91.1% (95% CI 88.0, 93.6)	67.3% (95% CI 63.3, 71.2)	78.1% (95% CI 75.2, 80.6)
	3	91.4% (95% CI 83.0, 96.5)	95.8% (95% CI 94.2, 97.0)	65.5% (95% CI 56.0, 74.2)	99.2% (95% CI 98.4, 99.7)	76.3% (95% CI 69.0, 82.3)
optimized internal validation	0	91.8% (95% CI 87.1, 95.3)	94.0% (95% CI 92.2, 95.6)	78.9% (95% CI 73.1, 84.1)	97.9% (95% CI: 96.7, 98.8)	84.9% (95% CI 81.1, 88.4)
	1	81.0% (95% CI 75.4, 85.9)	93.8% (95% CI 91.8, 95.4)	79.7% (95%CI 74.0, 84.6)	94.2% (95% CI 92.3, 95.8)	80.3% (95% CI 76.1, 84.2)
	2	78.8% (95% CI 74.8, 82.4)	88.6% (95% CI 85.6, 91.2)	86.1% (95% CI 82.5, 89.2)	82.4% (95% CI 79.0, 85.4)	82.3% (95% CI 79.4, 85.1)
	3	64.0% (95% CI 53.8, 73.4)	95.6% (95% CI 94.0, 96.8)	61.5% (95% CI 51.5, 70.9)	96.0% (95% CI 94.5, 97.2)	62.7% (95% CI 54.0, 70.1)

validation	Inflammatory severity grading of ulcerative colitis (UC)	Sensitivity, % (95% CI)	Specificity, % (95% CI)	PPV, % (95%CI)	NPV, % (95%CI)	F1score, % (95%CI)
external validation	0	96.0% (95% CI 93.1, 97.9)	95.2% (95% CI 94.0, 96.3)	80.9% (95% CI 76.4, 84.0)	99.1% (95% CI 98.5, 99.5)	99.1% (95% CI 98.5, 99.5)
	1	84.2% (95% CI 81.6, 86.6)	95.5% (95% CI 94.0, 96.8)	94.6% (95% CI 92.8, 96.1)	86.6% (95% CI 84.4, 88.7)	89.1% (95% CI 87.4, 90.7)
	2	79.1% (95% CI 74.4, 83.2)	94.2% (95% CI 92.9 95.4)	77.3% (95% CI 72.5, 81.5)	94.8% (95% CI 93.5, 95.9)	78.2% (95% CI 74.7, 81.5)
	3	92.1% (95% CI 88.0, 95.1)	96.8% (95% CI 95.7, 97.6)	82.9% (95% CI 77.9, 87.1)	98.6% (95% CI 97.9, 99.2)	87.2% (95% CI 84.0, 90.1)

2.2 Model optimization and Performance

The optimized model architecture was reassessed for predicting endoscopic severity. For classifying UC severity as grade 0 (normal), the AI model achieved a sensitivity of 91.8%, specificity of 94.0%, PPV of 78.9%, NPV of 97.9%, and an F1 Score of 84.9%. For grade 1 (mild), these metrics were 81.0%, 93.8%, 79.7%, 94.2%, and 80.3%, respectively. For grade 2 (moderate), the sensitivity was 78.8%, specificity 88.6%, PPV 86.1%, NPV 82.4%, and F1 Score 82.3%. In predicting grade 3 (severe), the values were 64.0% sensitivity, 95.6% specificity, 61.5% PPV, 96.0% NPV, and 62.7% F1 Score (refer to Table 3 for specific values and confidence intervals, Fig. 3(b) for the confusion matrix, and Fig. 4(b) for model loss rate accuracy). The AUC values for differentiating between grade 0 (normal) and grade 1 (mild), grade 0 (normal) and grade 2 (moderate), grade 0 (normal) and grade 3 (severe), grade 1 (mild) and grade 2 (moderate), grade 1 (mild) and grade 3 (severe), and grade 2 (moderate) and grade 3 (severe) were 0.625, 0.975, 1, 0.856, 1, and 0.925, respectively, with an average AUC of 0.899 (see Fig. 5(a) for details).

Table 3
Benchmarking comparison of AI models for endoscopic UC severity assessment

Study (Year)	Model Architecture	Data Scale	Key Performance Metrics	Validation Method
Our Study	ResNet50 (DCNN)	5,000 internal 1,732 external	AUC 0.899 (internal)/0.885 (external) Sensitivity for severe cases 92.1% QWK = 0.880 (vs. experts)	Multi-center external validation
Alyamani (2024)	MobileNet-V2	Single-center 3,800 images	AUC 0.91 Sensitivity for severe cases 93.8%	Single-center internal validation
Ozawa (2021)	EfficientNet-B4	Single-center 2,148 images	AUC 0.94 for severe lesions	Temporal external validation
Bhambhvani (2020)	ResNeXt-101	Single-center 777 images	Mayo score accuracy 82%	Single-center split validation
Stidham (2019)	Inception-V3	Single-center 5,466 images	UCEIS sub-item AUC 0.89–0.96	Single-center validation

3. External validation of the model

The AI model's performance in predicting UC endoscopic severity yielded the following metrics: for grade 0 (normal), sensitivity, specificity, PPV, NPV, and F1-score were 96.0%, 95.2%, 80.9%, and 99.1%, respectively; for grade 1 (mild), these values were 84.2%, 95.5%, 94.6%, 86.6%, and 89.1%, for grade 2 (moderate), these values were 79.1%, 94.2%, 77.3%, 94.8%, and 78.2%, for grade 3 (severe), these values were 92.1%, 96.8%, 82.9%, 98.6%, and 87.2%. Refer to Table 4 for detailed values and confidence intervals. The AUC values for distinguishing between severity levels were as follows: grade 0 (normal) vs. grade 1 (mild), 0.921; grade 0 (normal) vs. grade 2 (moderate), 0.996; grade 0 (normal) vs. grade 3 (severe), 0.964; grade 1 (mild) vs. grade 2 (moderate), 0.852; grade 1 (mild) vs. grade 3 (severe), 0.843; and grade 2 (moderate) vs. grade 3 (severe), 0.719, with a mean AUC of 0.885. See Fig. 5(b) for further details.

4. Consistency evaluation of endoscopic severity grading of UC between artificial intelligence model and endoscopic experts

In the retrospective internal test set, the AI model's prediction calibration curve against the endoscopic expert score was generated using 100 bootstrap repetitions, yielding a mean absolute error of 0.045 (see Fig. 6(a)). For the external test set, the calibration curve exhibited a mean absolute error of 0.023 (see

Fig. 6(b)).Scatter plots of expert and model scores were generated for both the internal and external test sets. Scatter plots revealed the distribution of model scores for each given expert UCEIS score, indicating a non-linear relationship, indicating a non-linear relationship. Additionally, for certain UCEIS values, points varied with image count, suggesting diverse associations between UCEIS and ResNet50 across different image numbers. The plots shared consistent layout and data distribution, underscoring the model's stability and reliability. (Refer to Fig. 6(c) and 6(d) for details). We compared the AI model output scores with the endoscopic expert scores to assess the consistency of the AI model output scores with the endoscopic expert scores. The ICC between AI and endoscopist was 0.840 (95% CI 0.820–0.858) in the internal test set and 0.880 (95% CI 0.869–0.890) in the external test set. The QWK for endoscopic severity of UC as judged by AI and endoscopic experts was 0.840 (95% CI 0.812–0.865) in the internal test set and 0.880 (95% CI 0.860–0.900) in the external test set.

Discussion

Artificial intelligence has been widely used in the medical field recently. AI in medicine is broadly divided into virtual and physical domains. Machine learning (ML) is an important branch of artificial intelligence. Deep learning (DL), a particularly successful subset of ML, encompasses architectures like convolutional neural networks (CNNs). Other important ML approaches include supervised learning, semi-supervised learning, support vector machines, and random forests[4]. At present, a convolutional neural network is a deep neural network that has been successfully trained widely. Deep learning has great advantages in target recognition, classification, detection and segmentation, and is widely used in medicine, among which convolutional neural network has become its core[14]. CNNs take raw image data as input, automatically learn hierarchical feature representations from large training datasets, and output predictions or feature maps. Convolutional neural networks are widely used in dermatology [15, 16], imaging[17, 18], ophthalmology, stomatology, bioinformatics[19] and pathology[20]. With the breakthrough of computing power and the evolution of big data technology, the structural optimization of convolutional neural networks has become an important evolution direction of artificial intelligence development. In this context, Deep Convolutional Neural Network (DCNN) architectures have matured, which effectively approaches the hierarchical information processing mechanism of biological nervous system through key technical paths such as multi-level network topology construction, mass labeled data driving, network connection paradigm innovation and parameter optimization strategy. This bionic computing framework significantly improves the brain-like features of the model and realizes the ability transition from primary feature extraction to multi-level feature representation in the visual information processing dimension[21].

In recent years, the application of artificial intelligence to the field of gastroenterology has been a hot research direction, whether it is to evaluate the intestinal preparation of patients or to identify polyps and adenomas in human-computer confrontation. UC is a type of inflammatory bowel disease, a chronic, specific inflammation of the colonic mucosa that can affect any segment of the intestine from the ileocecum to the rectum, and is affected by geographic, age, gender, genetic, and environmental

factors[1]. Ulcerative colitis has become a global human health challenge over the past 20 years, and recent epidemiological studies indicate that UC is rapidly increasing in incidence in South America, Europe, Asia and Africa. The recently updated Inflammatory Bowel Disease Research Organisation (IOIBD) Target Selection for Inflammatory Bowel Disease Treatment-II (STRIDE-II, 2021) in the International Inflammatory Bowel Disease Research Organisation emphasises that mucosal healing is associated with improved long-term outcomes in UC patients[21] and that the pursuit of mucosal healing is our current treatment goal.

In patients with ulcerative colitis, endoscopy is essential to assess disease activity and to provide appropriate treatment in a timely manner[22]. At present, there are many endoscopic scoring systems for evaluating the severity of ulcerative colitis. Mayo endoscopic scoring is widely used in clinical trials and clinical practice, but the evaluation results of each scoring system vary greatly. UCEIS, as a new scoring system, scores based on three endoscopic characteristics of vascular texture, hemorrhage and erosion, and each characteristic score is 0–3. The sum of the scores obtained divides the disease state into: Normal 0 points, mild activity 1–3 points, moderate activity 4–6 points, severe activity 7–8 points. In order to prove that the score is easier to apply clinically, Ikeya[23] selected 53 patients with moderate to severe UC who received tamoxix induction therapy. Through colonoscopy, UCEIS and Mayo scores were used to evaluate the endoscopic severity. It was found that UCEIS score can more accurately reflect the endoscopic reality of severe UC patients, and UCEIS score can better reflect the wide distribution of UC severity. In addition, based on current AI research in UC, studies utilizing AI for UC severity prediction have shown that UCEIS scores yield high sensitivity and specificity[24].

Previous studies exclusively utilized convolutional neural network (CNN) architectures, neglecting the newer deep convolutional neural network (DCNN) architectures. Alyamani et al. [25] specifically employed pre-trained MobileNet-V2 models within DCNN architectures. In contrast, we enhanced model performance substantially by incorporating transfer learning, dynamic learning rate adjustment (utilizing RMSProp with a warm-up strategy), and label smoothing techniques.

In contrast to the study by Bhambhvani et al.[26], which utilized the ResNeXt-101 model architecture with a limited dataset of 777 images from a single center for training and validation, our research introduces a novel approach. By meticulously curating lesion image data from standard white light endoscopes, we have effectively engineered a deep convolutional neural network based on the ResNet50 architecture. This has led to the creation of an intelligent diagnostic system for evaluating the severity of endoscopic manifestations in ulcerative colitis patients. Furthermore, we have validated the model's capacity for generalization using data from multiple centers.

Deasi's study[27] disaggregates the three constituent elements of the UCEIS score to derive predictive scores for vascular texture, hemorrhage, erosion, and ulcer individually. In contrast, our investigation centers on the UCEIS score with the primary objective of categorizing current severity levels as normal, mild, moderate, or severe. The hierarchical assessment framework employed in our research, comprising four severity levels (normal, mild, moderate, severe), aligns more closely with clinical standards, such as

the UCEIS score.The experiment comprises three stages: training and internal validation prior to model optimization, internal validation post model optimization, and external validation of the model. The data obtained indicate that the model before optimization effectively predicts severe UC (sensitivity 91.4%, specificity 95.8%) but exhibits limited capability in distinguishing between mild and moderate UC (mild F1 Score 43.5%), falling short of the desired model performance. Through the incorporation of convolution layers, adjustment of learning parameters, and data augmentation, the optimized model demonstrates enhanced overall accuracy, increasing from 72.1% to 80.4%. Specifically, the sensitivity for normal cases improves by 10.3%, while for mild cases, it increases by 20.7%. However, the sensitivity for severe cases decreases to 64.0% due to balancing adjustments. This trade-off may be attributed to the optimization goal of enhancing the model's overall performance rather than solely focusing on increasing sensitivity within a specific category. Notably, during the optimization process, some sensitivity in severe cases may be sacrificed to enhance the model's ability to differentiate between mild and moderate cases. Analysis of the ROC curve reveals that the optimized model effectively discriminates between moderate and severe cases (AUC ≥ 0.925) and moderately discriminates between normal and mild cases (AUC 0.625). The observed blurring of boundaries between mild and moderate endoscopic severity in UC patients may account for the latter finding, yet overall, the model meets the anticipated performance levels.

The model underwent validation using an external dataset, achieving an overall accuracy of 86.4% across 1732 validation images. High reliability was maintained for severe predictions (sensitivity 92.1%, negative predictive value 98.6%), while normal recognition accuracy was also notable (sensitivity 96.0%, specificity 95.2%). However, challenges in distinguishing moderate from severe cases were observed, potentially attributed to feature learning bias stemming from variations in sample size distribution between these categories in the external validation set. Notably, the model exhibited consistent discriminatory performance in moderate prediction (F1 Score 78.2%), with an overall accuracy of 86.4%. Noteworthy is the model's high sensitivity for normal and severe cases (96.%/92.1%) and its advantage in cross-center calibration (mean absolute error = 0.023), indicating robust cross-agency generalization capabilities.

Finally, the model calibration and clinical consistency were analyzed. By constructing calibration curves for internal and external tests of the model, it is found that the predicted probability of the model is in good agreement with the actual probability. Notably, the prediction bias for the low probability interval was more pronounced in the internal data, possibly due to partial similarity in mild and moderate lesion characteristics. The scatter plot of consistency between model and expert diagnosis shows that model scores and expert UCEIS scores are highly concentrated near y = x line, and visual consistency is significant. Quantitative correlation analysis was performed by calculating intra-class correlation coefficients and quadratic weighted kappa coefficients between internal and external validation model predictions and expert predictions, and the results met the high consistency criteria of medical diagnostic tools, proving that the model had clinical equivalence with expert judgments in disease classification.

To contextualize our findings, we compared our model's performance with that of existing AI models existing application models utilized for assessing the endoscopic severity of ulcerative colitis in the current market. Table 3 illustrates that while certain models may exhibit marginally superior sensitivity at specific severity levels, they are developed and tested on a single-center dataset, potentially constraining their applicability across diverse settings. In contrast, our model, trained on a larger and more varied multicenter dataset, sustained a notable sensitivity of 92.1% for severe cases during external validation and demonstrated enhanced concordance with expert evaluations (QWK = 0.880).

Limitations

The accurate performance of deep convolutional neural network models in diagnostics often relies on dependable reference criteria, such as pathological biopsies, to differentiate misdiagnosed cases (false positives/false negatives), as seen in colon cancer screening. This differentiation can be achieved by combining biopsy validation with radiologist reports. However, discrepancies in the reference standard, as identified by the three professional endoscopists in this study, can challenge the interpretation of conventional indicators (e.g., sensitivity, specificity), necessitating a comprehensive evaluation in conjunction with other methodologies. Following model optimization in this study, there were instances where classification accuracy both decreased and increased. This suggests potential model self-adjustment and uneven data distribution, warranting further examination. Notably, the training set data were sourced solely from a single center. Future studies could benefit from multi-center data collection to form training sets for enhanced model refinement. External validation using multi-center data could extend beyond provincial boundaries to validate the model's generalization performance more effectively.

Conclusion

This study presents an AI-assisted diagnostic model utilizing the ResNet50 deep convolutional neural network to assess UC severity during colonoscopy. The model exhibited strong diagnostic performance in internal and multicenter external validation. Subsequent experiments validated its high consistency with endoscopic experts in accurately identifying UC mucosa severity. These findings suggest the model's potential significance in early diagnosis, disease monitoring, and personalized UC treatment, offering promising advancements in clinical care and broad research opportunities.

Declarations

Funding

This research work was conducted without financial support from any public, private, or not-for-profit funding agency.

Conflicts of interest

There are no conflicts of interest.

Author Contributions

Mingrui Qi, Cong Dai, Chuqiao Li, and Lianyi Guo conceived and designed the study. Wei Niu, Baohai Liu, Xiaofei Li, Quanlin Li, and Lianyi Guo analyzed and interpreted the data. Mingrui Qi, Cong Dai, and Chuqiao Li collected the data and performed the literature search. Mingrui Qi drafted of the article. Cong Dai, Lianyi Guo, and Quanlin Li critically revised the article for important intellectual content. All authors read and approved the final manuscript.

Ethics approval

This study was approved by the Ethics Committee of The First Affiliated Hospital of Jinzhou Medical University (Approval No.: KYLL2024407). The requirement for informed consent was waived by the ethics committee due to the retrospective nature of the study and because all data were de-identified.

References

1. Ordás I, Eckmann L, Talamini M, et al. Ulcerative colitis [J]. The Lancet, 2012, 380(9853): 1606–1619.
2. Deo R C. Machine Learning in Medicine [J]. Circulation, 2015, 132(20): 1920–1930.
3. Terashima M, Tokunaga M, Tanizawa Y, et al. Robotic surgery for gastric cancer [J]. Gastric Cancer, 2015, 18(3): 449–457.
4. Yin Z, Yao C, Zhang L, et al. Application of artificial intelligence in diagnosis and treatment of colorectal cancer: A novel Prospect [J]. Frontiers in Medicine, 2023, 10:120 doi:10.3389/fmed.2023.123456
5. Du-harpur X, Watt F M, Luscombe N M, et al. What is AI? Applications of artificial intelligence to dermatology [J]. British Journal of Dermatology, 2020, 183(3): 423–430.
6. Kelly B S, Judge C, Bollard S M, et al. Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE) [J]. European Radiology, 2022, 32(11): 7998–8007.
7. Li J-P O, Liu H, Ting D S J, et al. Digital technology, tele-medicine and artificial intelligence in ophthalmology: A global perspective [J]. Progress in Retinal and Eye Research, 2021, 820–855. doi:10.1016/j.preteyeres.2021.100955
8. Shan T, Tay F R, Gu L. Application of Artificial Intelligence in Dentistry [J]. Journal of Dental Research, 2020, 100(3): 232–244.
9. Karim M R, Islam T, Shajalal M, et al. Explainable AI for Bioinformatics: Methods, Tools and Applications [J]. Briefings in Bioinformatics, 2023, 24(5):101–120
10. Bera K, Schalper K A, Rimm D L, et al. Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology [J]. Nature Reviews Clinical Oncology, 2019, 16(11): 703–715.

11. Hassan C, Spadaccini M, Mori Y, et al. Real-Time Computer-Aided Detection of Colorectal Neoplasia During Colonoscopy [J]. Annals of Internal Medicine, 2023, 176(9): 1209–1220.
12. He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition [Z]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 770–778
13. Travis S P L, Schnell D, Krzeski P, et al. Developing an instrument to assess the endoscopic severity of ulcerative colitis: the Ulcerative Colitis Endoscopic Index of Severity (UCEIS) [J]. Gut, 2012, 61(4): 535–542.
14. Ng S C, Shi H Y, Hamidi N, et al. Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies [J]. The Lancet, 2017, 390(10114): 2769–2778.
15. Cho S I, Sun S, Mun J H, et al. Dermatologist-level classification of malignant lip diseases using a deep convolutional neural network [J]. British Journal of Dermatology, 2019, 182(6): 1388–1394.
16. Sies K, Winkler J K, Fink C, et al. Past and present of computer-assisted dermoscopic diagnosis: performance of a conventional image analyser versus a convolutional neural network in a prospective data set of 1,981 skin lesions [J]. European Journal of Cancer, 2020, 135: 39–46.
17. Choy G, Khalilzadeh O, Michalskil M, et al. Current Applications and Future Impact of Machine Learning in Radiology [J]. Radiology, 2018, 288(2): 318–328.
18. Xu Y, Hosny A, Zeleznik R, et al. Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging [J]. Clinical Cancer Research, 2019, 25(11): 3266–3275.
19. ElmarakebyL H A, Hwang J, Arafeh R, et al. Biologically informed deep neural network for prostate cancer discovery [J]. Nature, 2021, 598(7880): 348–352.
20. Nam J G, Park S, Park C M, et al. Histopathologic Basis for a Chest CT Deep Learning Survival Prediction Model in Patients with Lung Adenocarcinoma [J]. Radiology, 2022, 305(2): 441–451.
21. Shah S C, Cokombel J-F, Sands B E, et al. Mucosal Healing Is Associated With Improved Long-term Outcomes of Patients With Ulcerative Colitis: A Systematic Review and Meta-analysis [J]. Clinical Gastroenterology and Hepatology, 2016, 14(9): 1245–1255.e8.
22. D'haens G, Sandborn W J, Feagan B G, et al. A Review of Activity Indices and Efficacy End Points for Clinical Trials of Medical Therapy in Adults With Ulcerative Colitis [J]. Gastroenterology, 2007, 132(2): 763–786.
23. Ikeya K, Hanai H, Sugimoto K, et al. The Ulcerative Colitis Endoscopic Index of Severity More Accurately Reflects Clinical Outcomes and Long-term Prognosis than the Mayo Endoscopic Score [J]. Journal of Crohn's and Colitis, 2016, 10(3): 286–295.
24. Jahagirdar V, Bapaye J, Chandan S, et al. Diagnostic accuracy of convolutional neural network–based machine learning algorithms in endoscopic severity prediction of ulcerative colitis: a systematic review and meta-analysis [J]. Gastrointestinal Endoscopy, 2023, 98(2): 145–154.e8.
25. Alyamani H J. Enhancing Ulcerative Colitis Diagnosis: A Multi-Level Classification Approach with Deep Learning [J]. Computer Modeling in Engineering & Sciences, 2024, 140(1): 1129–1142.

26. Bhambhvani H P, Zamora A. Deep learning enabled classification of Mayo endoscopic subscore in patients with ulcerative colitis [J]. European Journal of Gastroenterology & Hepatology, 2020, 33(5): 645–649.
27. Desai. 62nd Annual Conference of the Indian Society of Gastroenterology, February 10th – 13th, 2022, Pune [J]. Indian Journal of Gastroenterology, 2022, 41(S1): 171–178.

Table 4

Table 4 is not available with this version

Figures

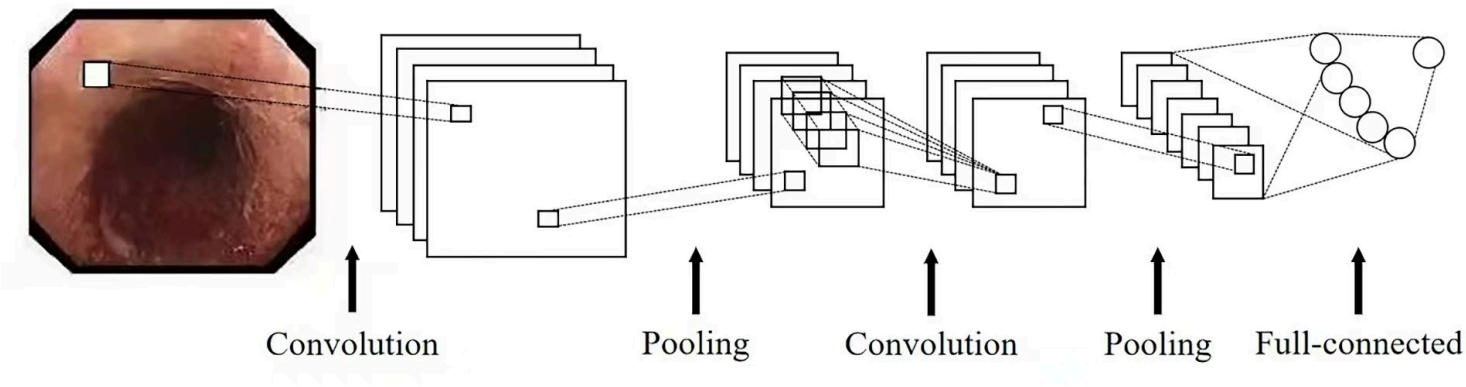


Figure 1

Model Backbone Structure of Deep Convolutional Neural Network for Identifying Endoscopic Severity in Ulcerative Colitis (UC).

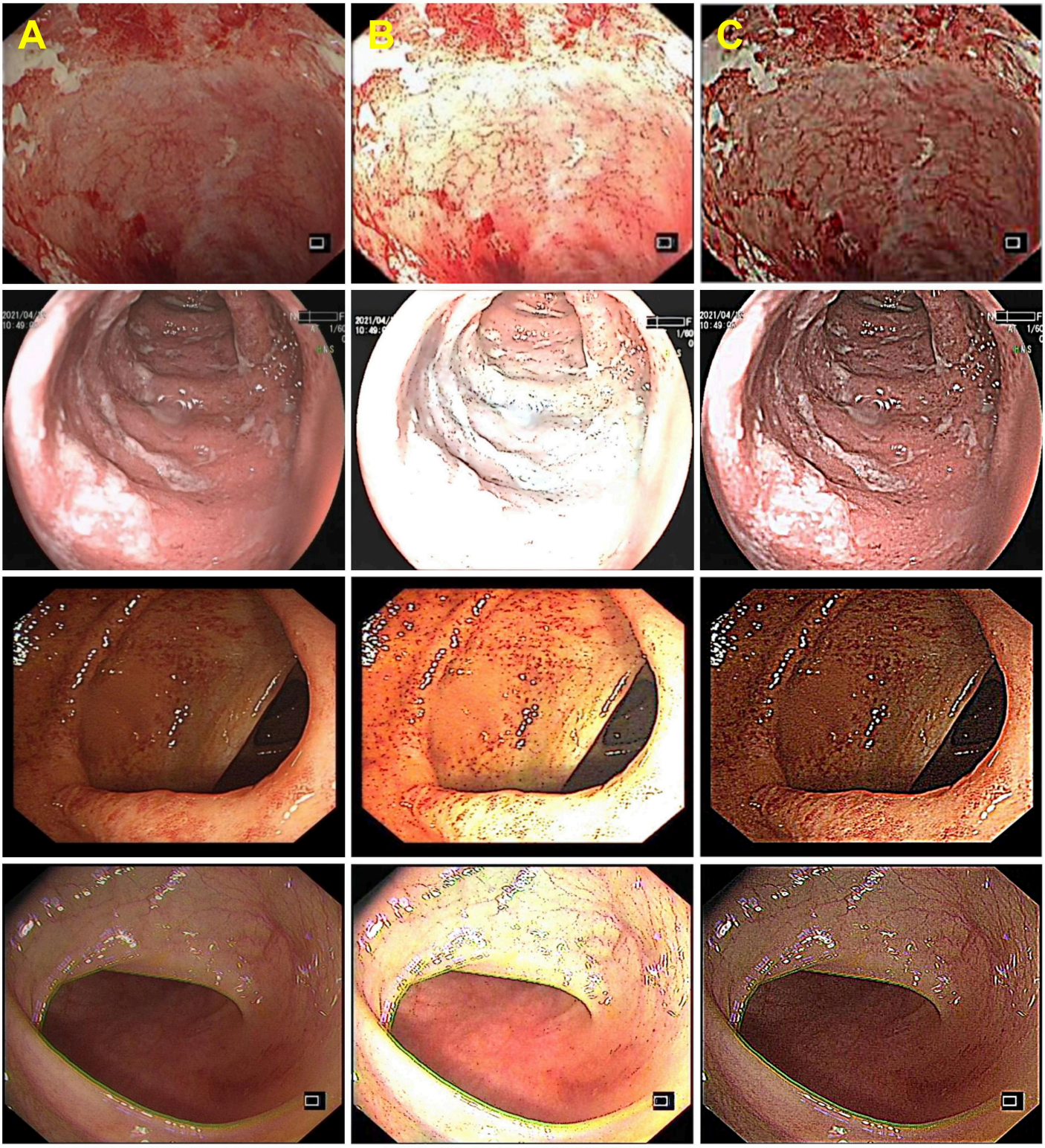


Figure 2

Image Processing Steps for the Internal Training Set. Panel A represents Raw colonoscopy images, B represents Image Enhancement and Augmentationand C represents Image Sharpening via Unsharp Masking.

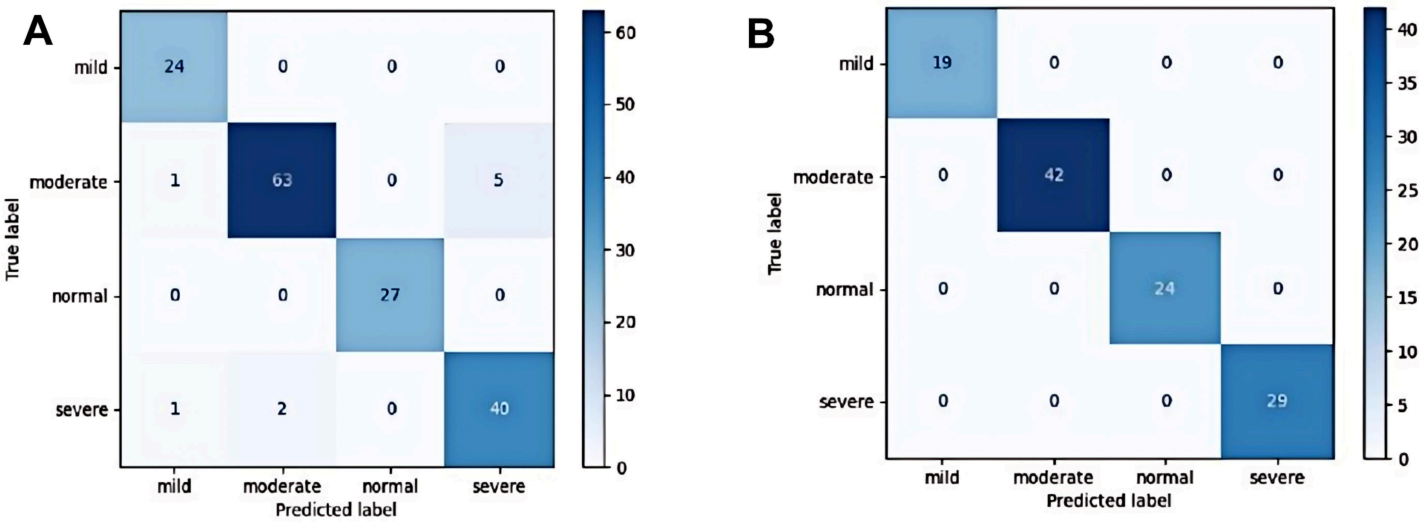


Figure 3

Confusion matrix of the dataset. Panel A represents Pre-optimization model confusion matrix, B represents Optimized Model Confusion Matrix.

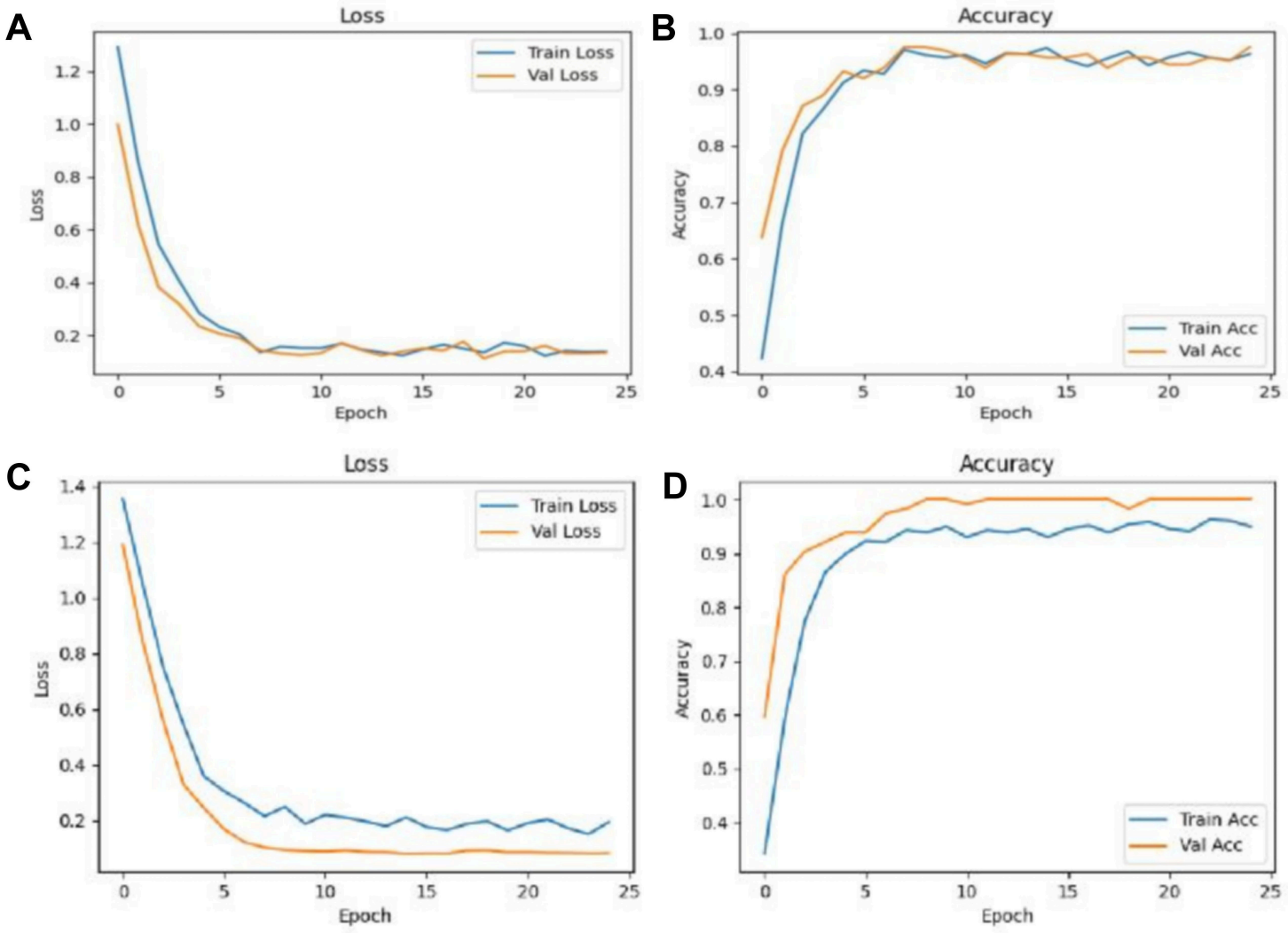


Figure 4

loss rate and accuracy. Panel A represents Pre-optimization model prediction loss rate, B represents Pre-optimization model accuracy, C represents Optimized Model loss rate, D represents Optimized Model accuracy.

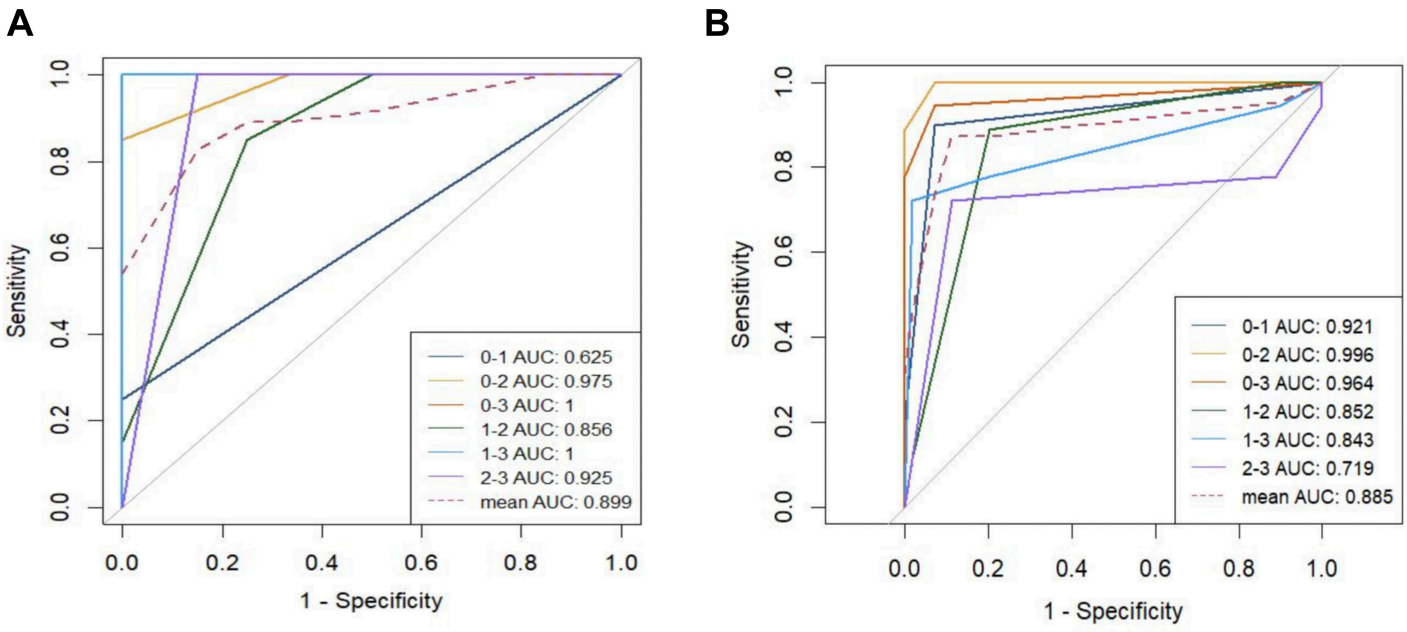


Figure 5

ROC curve for identifying endoscopic severity of UC. Panel A represents after optimization of the model for the image test set ROC curve, B represents external test set test model ROC curve.

Note: 0-normal; 1-mild; 2-moderate; 3-severe; ROC, subject operating characteristic curve

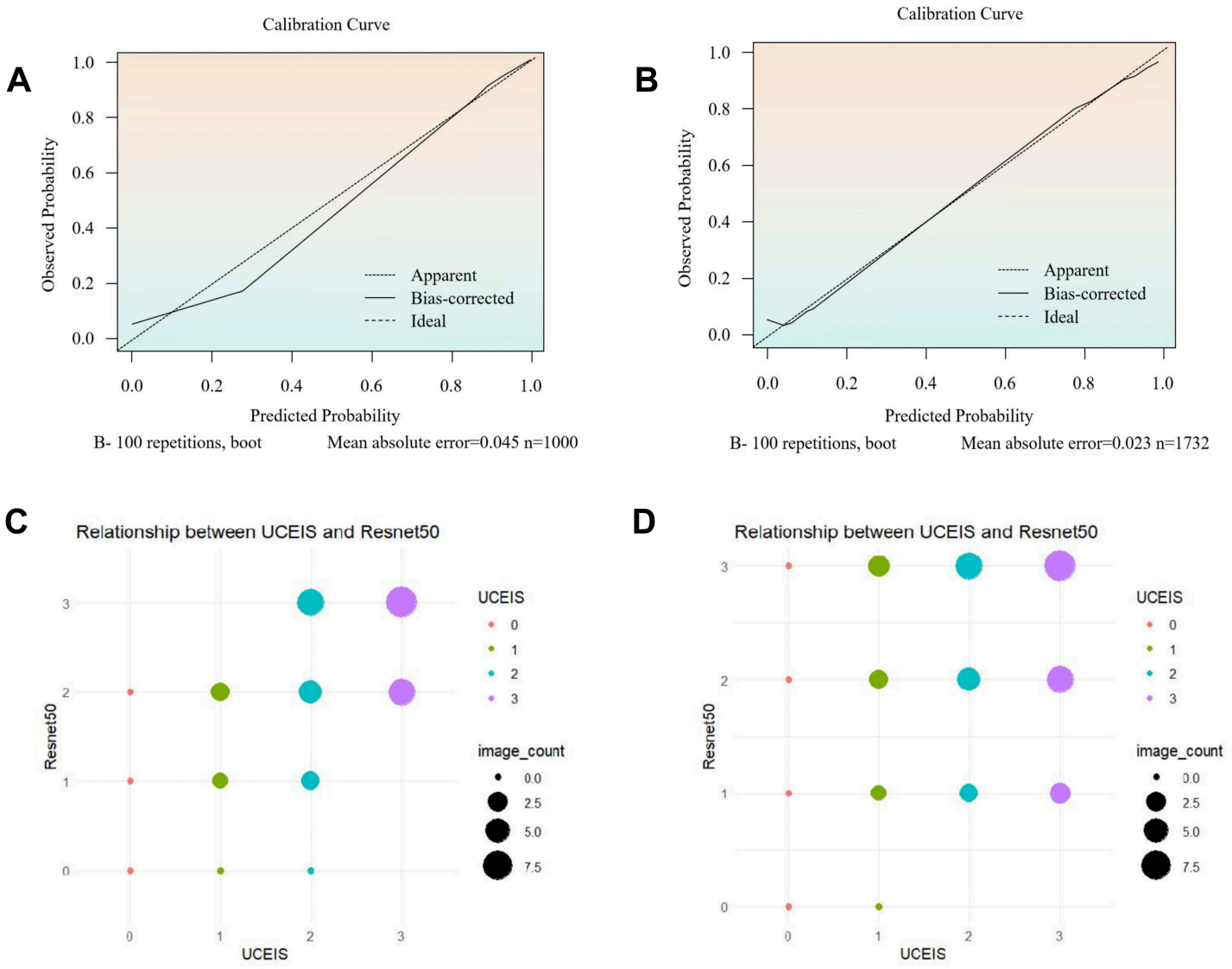


Figure 6

Internal Validation Calibration Curve And Correlation evaluation of UCEIS scores between artificial intelligence models (ResNet50) and experts. Panel A represents Internal Validation Calibration Curve, B represents External Correlation evaluation, C represents External Verification Calibration Curve and D represents External Correlation evaluation.

Note1: B = 100 repetitions, boot, constructed by bootstrap; Mean absolute error; n, number of samples

Note2: Each circle indicates the number of pictures per score; ResNet50, artificial intelligence model; UCEIS in figure, expert rating

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarytable.docx
- SupplementaryFigure.jpg