This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0). To view a copy of the license, visit **https://creativecommons.org/licenses/by-nc/4.0/**

Annals of Clinical and Analytical Medicine
Original Research

Development and temporal validation of a deep learning model for automated Mayo Endoscopic Subscore assessment in ulcerative colitis

Video-based deep learning for MES assessment

Yavuz Özden[1] Nuh Mehmet Büyükberber[1]

[1]Department of Gastroenterology, Kayseri City Hospital, University of Health Sciences, Kayseri, Türkiye.

Abstract

Aim: The Mayo Endoscopic Subscore (MES) in ulcerative colitis (UC) is limited by interobserver variability, whereas central reading is resource-intensive. We developed and temporally validated a video-based deep learning system for automated MES assessment.

Methods: This retrospective diagnostic accuracy study included 1,000 adults with UC who underwent white-light colonoscopy between November 2020 and November 2025. The model was developed in 750 patients using five-fold patient-level cross-validation and evaluated after model freezing in a temporally separated cohort of 250 patients. Two expert gastroenterologists assigned MES 0 to 3 from full-length videos, with third-reader adjudication of discordant cases. An EfficientNet-B0 model with attention-based multiple instance learning generated predictions. The primary outcome was four-class MES classification; dichotomous endpoints were MES 0 versus MES ≥1 and MES ≤1 versus MES ≥2.

Results: Pre-adjudication interobserver agreement was high (quadratic weighted κ, 0.84; 95% CI, 0.81–0.87). In temporal validation, macro-averaged AUROC was 0.95 (95% CI, 0.93–0.97), overall accuracy was 83.6%, macro-F1 score was 0.84, and quadratic weighted κ was 0.79 (95% CI, 0.68–0.87). AUROCs were 0.97 (95% CI, 0.95–0.98) for MES 0 versus MES ≥1 and 0.96 (95% CI, 0.94–0.97) for MES ≤1 versus MES ≥2. MES 1–2 errors accounted for 58.5% of classification errors. Brier scores were 0.07 and 0.06, respectively, and decision-curve analysis demonstrated a positive net benefit.

Conclusions: The model demonstrated high discrimination, good agreement with the adjudicated reference standard, good calibration, and interpretable errors. Multicenter external validation and prospective studies are required before implementation.

Keywords

colitis, ulcerative, colonoscopy, artificial intelligence, deep learning, image interpretation

DOI: 10.4328/ACAM.50232 Ann Clin Anal Med 2026;17(10):1066-1072
Received: 07.06.2026 Accepted: 06.09.2026 Published Online: 06.09.2026 Printed: 01.10.2026
Corresponding Author: Yavuz Özden, Department of Gastroenterology, Kayseri City Hospital, University of Health Sciences, Kayseri, Türkiye.
Email: yavuzozden@gmail.com **P.:** +90 535 888 36 76
Corresponding Author ORCID: https://orcid.org/0000-0002-5760-3529
Other Authors ORCID: Nuh Mehmet Büyükberber, https://orcid.org/0000-0003-1938-4377

Introduction

The Mayo Endoscopic Subscore (MES) is among the most widely used measures of endoscopic disease severity in ulcerative colitis (UC), informing treat-to-target strategies and serving as a common endpoint in clinical trials.[1,2] Its four-grade structure facilitates routine clinical use, but its visual interpretation is affected by substantial interobserver variability. A recent systematic review and meta-analysis reported a pooled weighted kappa of 0.58 for MES, indicating that disagreement is an inherent limitation of visual severity assessment rather than an occasional finding.[3] This variability may affect treatment decisions, longitudinal monitoring, and comparability across clinical studies.

Central reading has been introduced to standardize endoscopic scoring in research settings. Although central reading can reduce variability, it remains time-consuming, resource-intensive, and dependent on expert human interpretation.[4] Moreover, its scalability in routine practice is limited because large volumes of endoscopic video are generated but are rarely converted into structured and reproducible measures of disease activity.

Deep learning may offer a scalable approach to standardized endoscopic assessment without substantially increasing the burden of human review. Previous studies have shown that convolutional neural networks can classify UC endoscopic severity from still images with performance comparable to that of expert reviewers.[5,6] More recent video-based studies have applied this approach to full-length colonoscopy recordings and demonstrated agreement with expert or central-reader assessments.[4,7,8] Meta-analytic evidence also suggests favorable diagnostic performance of artificial intelligence (AI)-based systems for identifying endoscopic remission across commonly used definitions.[9]

Nevertheless, important methodological limitations remain. Many models have been developed using retrospective, single-center datasets and evaluated using random data splits that may not adequately assess robustness over time or under changing clinical conditions. The use of curated images or selected video segments may exclude ambiguous frames that contribute to disagreement in routine practice, thereby inflating apparent performance and limiting generalizability.[10] Temporal validation, in which a finalized model is evaluated using data collected after model development, remains uncommon but provides a more rigorous assessment of performance stability and reduces the risk of overly optimistic performance estimates.[11]

We therefore developed a deep learning system to assign patient-level MES from full-length routine colonoscopy videos and evaluated its performance in a prespecified temporal validation cohort after model freezing. We hypothesized that model discrimination would remain high over time and that residual errors would cluster in clinically recognizable scenarios, particularly at the boundary between MES 1 and MES 2.

Materials and Methods

Study Design and Setting

This retrospective diagnostic accuracy study was conducted to develop and temporally validate a deep learning model for automated Mayo Endoscopic Subscore (MES) assessment in ulcerative colitis (UC). The study followed a prespecified out-of-time validation framework. This single-center study was conducted at the tertiary-care endoscopy unit of Kayseri City Hospital, Kayseri, Türkiye. The study was reported in accordance with the STARD 2015 and STROBE recommendations.[12,13] A completed STARD 2015 checklist was submitted as a separate supplementary file.

Study Period and Temporal Split

The study period extended from November 2020 to November 2025. Colonoscopies performed between November 2020 and October 2024 formed the development data set (n = 750), whereas examinations performed between November 2024 and November 2025 formed the temporal validation data set (n = 250). Model architecture, hyperparameters, preprocessing rules, and the training pipeline were finalized using only the development data set. The temporal validation data set remained fully held out until model freezing. No patient contributed examinations to both data sets.

Participants

Consecutive adults aged 18 years or older with established UC who underwent colonoscopy during the study period were screened. To prevent within-patient correlation and data leakage, only one colonoscopy per patient was included. Inclusion criteria were established UC, availability of a full-length routine white-light colonoscopy video, and reference-standard MES assigned through the expert labeling workflow. Exclusion criteria were inadequate bowel preparation, defined as a total Boston Bowel Preparation Scale (BBPS) score <6;[14] major video corruption or missing key segments; insufficient mucosal visualization; and dominant alternative colonic pathology, such as cytomegalovirus colitis, ischemic colitis, or malignancy, that invalidated UC activity grading.

Video Acquisition, De-identification, and Frame Sampling

Colonoscopy videos were retrieved from the institutional endoscopy archive and de-identified before analysis. All videos included in the study were acquired using the ELUXEO Lite 6000 endoscopy system, comprising an EP-6000 video processor with an integrated light source and EC-760ZP-V/L video colonoscopes (Fujifilm Corporation, Tokyo, Japan). Video recordings were captured through the institutional endoscopy video recording infrastructure and stored in the hospital Picture Archiving and Communication System. The specific manufacturer and model of the recording and archiving system were not documented in the available institutional technical records; however, all recordings were obtained using the same standardized institutional workflow throughout the study period. Only routine white-light imaging sequences were used.

Frames were extracted at 1 frame per second to reduce redundancy while preserving temporal diversity. A prespecified rule-based quality filter removed noninformative frames, including those with severe motion blur or defocus, marked underexposure or overexposure, and instrument-dominant or content-obscured views in which mucosal assessment was not feasible.

For each video, total duration, extracted frame count, and retained-frame proportion were recorded. Median video duration was 8.2 minutes (interquartile range [IQR], 6.1–11.4) in the development cohort and 8.5 minutes (IQR, 6.3–11.2) in the temporal validation cohort. The median number of extracted frames was 492 (IQR, 366–684), and the median retained-frame proportion was 74% (IQR, 66–81). Patient-level partitioning was performed before frame extraction and preprocessing.

Reference Standard and Outcomes

The reference standard was the patient-level MES assigned from full-length video review according to standard MES definitions. The final MES reflected the most severe endoscopic inflammatory activity observed in the examined colonic mucosa. Two independent board-certified gastroenterologists with expertise in inflammatory

bowel disease and endoscopy reviewed each video and assigned an MES of 0 to 3. Readers were blinded to clinical variables and model outputs. Discordant cases were adjudicated by a third senior gastroenterologist who was blinded to the initial ratings and clinical variables. Interobserver agreement was quantified using quadratic weighted Cohen κ with 95% confidence intervals.

An adjudicated expert MES was selected as the reference standard because MES is the accepted clinical and trial-based measure of endoscopic inflammatory activity in UC and no fully objective gold standard is available.

The primary outcome was four-class MES classification. Prespecified dichotomous endpoints were complete endoscopic remission, defined as MES 0 versus MES ≥1, and clinically actionable endoscopic activity, defined as MES ≤1 versus MES ≥2, consistent with treat-to-target frameworks and guideline recommendations.[2,15]

Index-test predictions and reference-standard assessments were derived from the same colonoscopy video; therefore, there was no time interval or intervening clinical intervention between the index test and the reference standard.

Model Development

The model used frame-level feature extraction followed by patient-level aggregation with attention-based multiple instance learning (MIL). Frames were resized to 224 × 224 pixels and normalized using ImageNet channel statistics. During training only, augmentations included random cropping and resizing, mild brightness and contrast jitter, low-probability Gaussian blur, and JPEG compression.

EfficientNet-B0 pretrained on ImageNet was fine-tuned end-to-end as the convolutional backbone. Frame embeddings were passed to a 2-layer attention-pooling module with a hidden dimension of 256 and 1 attention logit per frame, followed by softmax normalization. The weighted patient-level representation was passed to a 4-output linear classifier producing softmax probabilities for MES 0 to 3. Weighted cross-entropy loss was used to address class imbalance, with weights inversely proportional to development-set class frequencies.

Table 1. Baseline Characteristics and MES Distribution

Characteristic	Development cohort (*n* = 750)	Temporal validation cohort (*n* = 250)
Age, years, median (IQR)	44 (32–58)	45 (33–59)
Female sex, *n* (%)	360 (48)	120 (48)
Disease duration, y, median (IQR)	6.2 (2.8–11.4)	6.5 (3.1–12.0)
BBPS, median (IQR)	7 (6–8)	7 (6–8)
BBPS ≥7, *n* (%)	525 (70)	175 (70)
Video duration, minutes, median (IQR)	8.2 (6.1–11.4)	8.5 (6.3–11.2)
Frames extracted per video, median (IQR)	492 (366–684)	510 (378–672)
Frames retained after filtering, median % (IQR)	74 (66–81)	73 (65–80)
MES 0, *n* (%)	210 (28)	65 (26)
MES 1, *n* (%)	255 (34)	88 (35)
MES 2, *n* (%)	180 (24)	62 (25)
MES 3, *n* (%)	105 (14)	35 (14)

Abbreviations: *BBPS, Boston Bowel Preparation Scale; IQR, interquartile range; MES, Mayo Endoscopic Subscore.*

Optimization was performed using AdamW with an initial learning rate of 1×10^{-4}, weight decay of 1×10^{-5}, and cosine annealing. Five-fold patient-level cross-validation was used for model selection based on the macro-averaged area under the receiver operating characteristic curve (AUROC). Repeated runs used random seeds of 42, 123, and 456; standard deviations were <0.01 for AUROC and <0.02 for quadratic weighted κ. The final model was retrained on the full development data set and frozen before temporal validation.

Training was performed using a single NVIDIA A100 40-GB graphics processing unit (NVIDIA Corporation, Santa Clara, California, USA) and took approximately 6 hours. The mean inference time per video, including frame extraction, filtering, and the model forward pass, was 12.4 seconds.

Temporal Validation

The frozen model was evaluated on the temporally separated validation data set without parameter updates, threshold retuning, or validation-set recalibration.

Ethical Approval

The study protocol was approved by the Kayseri City Hospital Non-Interventional Clinical Research Ethics Committee (Date: 02.12.2025; Decision No: 670).

Statistical Analysis

All analyses were conducted at the patient level. Continuous variables are presented as median and IQR, and categorical variables as number and percentage.

For the temporal validation cohort, 95% confidence intervals for

Table 2. Model Performance for 4-Class MES Prediction and Prespecified Dichotomous Outcomes in the Temporal Validation Cohort

Metric	Value
Four-class MES prediction (*n* = 250)	
Macro-averaged AUROC (95% CI)	0.95 (0.93–0.97)
Quadratic weighted κ (95% CI)	0.90 (0.87–0.93)
Overall accuracy	**84%**
Macro-F1	0.84
MES 0 F1-score	0.89
MES 1 F1-score	0.81
MES 2 F1-score	0.80
MES 3 F1-score	0.85
MES 0 versus ≥1: complete endoscopic remission	
AUROC (95% CI)	0.97 (0.95–0.98)
Sensitivity, % (95% CI)	89 (85–93)
Specificity, % (95% CI)	93 (90–96)
PPV, % (95% CI)	92 (88–95)
NPV, % (95% CI)	91 (87–94)
Brier score	0.07
MES ≤1 versus ≥2: clinically actionable threshold	
AUROC (95% CI)	0.96 (0.94–0.97)
Sensitivity, % (95% CI)	86 (82–90)
Specificity, % (95% CI)	92 (89–95)
PPV, % (95% CI)	89 (85–93)
NPV, % (95% CI)	90 (86–94)
Brier score	0.06

Abbreviations: *AUROC, area under the receiver operating characteristic curve; CI, confidence interval; MES, Mayo Endoscopic Subscore; NPV, negative predictive value; PPV, positive predictive value.*

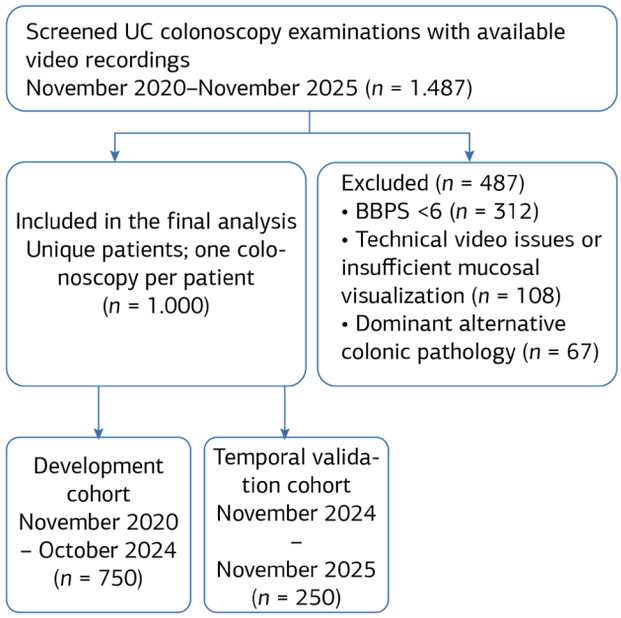


Figure 1. *Study flow diagram.*

Flow diagram of cohort selection from November 2020 through November 2025. Of 1,487 UC colonoscopy examinations with video recordings screened, 487 were excluded because of inadequate bowel preparation (BBPS <6; n=312), technical video issues or insufficient mucosal visualization (n=108), or dominant alternative colonic pathology (n=67). The final analytic cohort comprised 1,000 unique patients, with one colonoscopy per patient, allocated to the development cohort (November 2020–October 2024; n=750) and the temporal validation cohort (November 2024–November 2025; n=250). BBPS, Boston Bowel Preparation Scale; UC, ulcerative colitis.

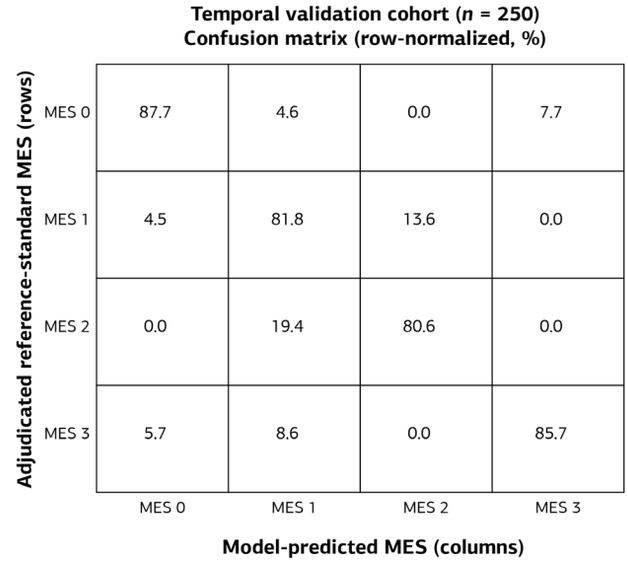


Figure 2. *Confusion matrix for 4-class MES prediction in the temporal validation cohort.*

The confusion matrix compares adjudicated reference-standard MES categories with model-predicted MES categories in the temporal validation cohort. Cell values are row-normalized percentages. Misclassifications were concentrated predominantly between adjacent categories, most commonly MES 1↔2, which accounted for 58% of errors. MES, Mayo Endoscopic Subscore.

macro-averaged one-versus-rest AUROC, dichotomous AUROC, quadratic weighted Cohen κ, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were estimated using nonparametric patient-level bootstrap resampling with 2000 iterations. The 2.5th and 97.5th percentiles of the bootstrap distributions were reported as the 95% confidence intervals. The confidence interval for preadjudication interobserver agreement was calculated using the same patient-level bootstrap procedure. For the development cohort, confidence intervals were calculated from pooled out-of-fold patient-level predictions using 2000 bootstrap iterations.

For four-class classification, the predicted MES category was defined as the category with the highest softmax probability. For the dichotomous endpoints, AUROC and Brier scores were calculated using the corresponding endpoint-specific probabilities, whereas sensitivity, specificity, PPV, and NPV were calculated by dichotomizing the four-class predicted MES categories according to the prespecified endpoint definitions.

Decision-curve analysis for the MES ≤1 versus MES ≥2 endpoint followed the net-benefit framework of Vickers and Elkin across threshold probabilities ranging from 0.10 to 0.90 using a Python implementation.[16] Calibration was assessed using calibration curves and Brier scores. Prespecified sensitivity analyses were performed according to bowel preparation quality, video duration, artifact burden, and disease extent.

All statistical tests were 2-sided, and a *P* value <.05 was considered statistically significant. Performance results were interpreted primarily using point estimates and 95% confidence intervals.

The model generated a definitive prediction for every examination included in the analysis; therefore, there were no indeterminate index-test results. No imputation was performed, and no included patient had missing index-test predictions or reference-standard outcomes.

No formal a priori sample size calculation was performed because this retrospective study included all consecutive eligible patients with complete colonoscopy video recordings during the predefined study period. The sample size was therefore determined by the number of eligible examinations available within the prespecified temporal windows. The temporal validation cohort was defined by calendar time before model evaluation and included 250 unique patients representing all 4 MES categories.

Analyses were performed using Python version 3.10, NumPy version 1.24.3, scikit-learn version 1.2.2, and PyTorch version 2.0.1.

Reporting Guidelines

This study was reported in accordance with the STARD 2015 and STROBE guidelines.

Results

Study Cohort and Baseline Characteristics

During the study period, 1,487 UC colonoscopy examinations with available video recordings were screened. After application of the prespecified exclusion criteria, 1,000 unique patients were included in the final analysis (Figure 1). The development cohort comprised 750 patients examined between November 2020 and October 2024, and the temporally separated validation cohort comprised 250 patients examined between November 2024 and November 2025.

The MES distribution was similar between the cohorts. In the development cohort, MES 0, 1, 2, and 3 accounted for 28%, 34%, 24%, and 14% of cases, respectively. The corresponding proportions in the temporal validation cohort were 26%, 35%, 25%, and 14% (Table 1). Baseline characteristics were also comparable between the cohorts. Median age was 44 years (IQR, 32–58) in the development cohort and 45 years (IQR, 33–59) in the temporal validation cohort. Female patients accounted for 48% of

both cohorts, and the median BBPS score was 7 (IQR, 6–8) in both cohorts. Median video duration was 8.2 minutes (IQR, 6.1–11.4) in the development cohort and 8.5 minutes (IQR, 6.3–11.2) in the temporal validation cohort.

Reference-Standard Labeling and Interobserver Agreement

Two independent expert gastroenterologists assigned an MES score to each video, with third-reader adjudication of discordant cases. Before adjudication, agreement between the two primary readers was high, with a quadratic weighted κ of 0.84 (95% CI, 0.81–0.87). Disagreements were predominantly limited to adjacent categories. Disagreement between MES 1 and MES 2 accounted for 62% of all discordant ratings, whereas disagreements involving a difference of 2 or more MES grades accounted for 8%.

Model Performance in the Development Cohort

In the development cohort, five-fold patient-level cross-validation demonstrated stable performance across folds (Supplementary Table S1). For four-class MES prediction, the cross-validated macro-averaged AUROC was 0.96 (95% CI, 0.95–0.97), the macro-F1 score was 0.85, and the quadratic weighted κ was 0.92 (95% CI, 0.90–0.94).

For the prespecified dichotomous endpoints, the cross-validated AUROC was 0.98 (95% CI, 0.97–0.99) for MES 0 versus MES ≥1 and 0.97 (95% CI, 0.96–0.98) for MES ≤1 versus MES ≥2. After dichotomization of the four-class predictions, sensitivity and specificity for the MES ≤1 versus MES ≥2 endpoint were 88% and 94%, respectively. Detailed performance metrics are provided in Supplementary Table S2.

Temporal Validation Performance

In the temporally separated validation cohort, the frozen model maintained high discrimination. For four-class MES prediction, the macro-averaged AUROC was 0.95 (95% CI, 0.93–0.97), overall accuracy was 83.6%, the macro-F1 score was 0.84, and the quadratic weighted κ was 0.79 (95% CI, 0.68–0.87) (Figure 2 and Table 2).

A total of 41 four-class misclassifications occurred. Errors between MES 1 and MES 2 accounted for 24 of these errors (58.5%), indicating that misclassification was concentrated primarily at this adjacent-category boundary.

For complete endoscopic remission, defined as MES 0 versus MES ≥1, the AUROC was 0.97 (95% CI, 0.95–0.98). For clinically actionable endoscopic activity, defined as MES ≤1 versus MES ≥2, the AUROC was 0.96 (95% CI, 0.94–0.97). After dichotomization of the four-class predictions, sensitivity, specificity, PPV, and NPV were 87.7%, 96.8%, 90.5%, and 95.7%, respectively, for MES 0 versus MES ≥1, and 82.5%, 88.9%, 82.5%, and 88.9%, respectively, for MES ≤1 versus MES ≥2 (Table 2).

Decision-Curve Analysis and Calibration

For the MES ≤1 versus MES ≥2 endpoint, decision-curve analysis demonstrated a positive net benefit across threshold probabilities ranging from 0.10 to 0.80, with the highest net benefit observed between 0.30 and 0.60. In the temporal validation cohort, the Brier score was 0.07 for MES 0 versus MES ≥1 and 0.06 for MES ≤1 versus MES ≥2. Calibration assessment demonstrated good agreement between predicted probabilities and observed outcomes without recalibration in the temporal validation cohort.

Error and Sensitivity Analyses

False-positive predictions for MES ≥2 were associated primarily with adherent mucus mimicking inflammation, mild erythema accompanied by suboptimal luminal distension, and healing ulcers with regenerative changes. False-negative predictions were associated primarily with patchy right-sided inflammation, subtle friability without overt exudate, and limited visualization of the most severely affected colonic segment.

Model performance remained generally consistent across the prespecified strata. In the temporal validation cohort, the AUROC for MES ≤1 versus MES ≥2 ranged from 0.94 to 0.97, while quadratic weighted κ for four-class MES prediction ranged from 0.87 to 0.92 across strata defined by bowel preparation quality, video duration, artifact burden, and disease extent (Supplementary Table S3).

Discussion

This study developed and temporally validated a video-based deep learning system for automated MES grading in UC. The model achieved high multiclass discrimination, good agreement with an adjudicated expert reference standard, and robust performance at clinically relevant dichotomous thresholds. Importantly, performance was maintained in a temporally separated cohort evaluated only after model freezing, providing a more stringent assessment than random data splitting.

Temporal validation remains underused in computer vision studies of inflammatory bowel disease, although it is important for assessing performance stability under changing clinical and technical conditions. Systematic reviews have noted that many AI-based prediction models in this field remain at high risk of bias, partly because of limited external or temporal validation.[11,17] By using a prespecified out-of-time validation cohort, our study addresses this methodological limitation, while recognizing that temporal validation within a single institution does not replace multicenter external validation.

The model's residual error pattern mirrored recognized areas of uncertainty in human MES grading. Expert disagreement was concentrated in adjacent categories, particularly MES 1 and MES 2, and model errors followed a similar pattern. Errors between MES 1 and MES 2 accounted for 24 of the 41 misclassifications (58.5%), supporting the interpretation that most errors occurred at a clinically recognized perceptual boundary. However, a small number of nonadjacent errors, particularly between MES 0 and MES 3, had a disproportionate effect on the quadratic weighted κ and contributed to the lower agreement estimate despite an overall accuracy of 83.6%. The observed visual confounders, including adherent mucus, suboptimal luminal distension, healing ulcers, patchy inflammation, and subtle friability, may guide future model refinement and reader training.

The model also demonstrated good calibration and a positive net benefit across clinically relevant threshold probabilities. These findings support the potential value of probability-based outputs for selecting decision thresholds according to the intended clinical application. However, they should not be interpreted as evidence of improved patient outcomes or real-world clinical effectiveness.

Our findings are consistent with previous evidence showing that deep learning can classify UC endoscopic severity from images and videos with performance comparable to expert assessment.[4-8,18,19] The principal contribution of the present study is its validation framework, which incorporated full-length routine colonoscopy videos, strict patient-level partitioning, temporal validation after model freezing, calibration assessment, characterization of observed error patterns, and prespecified subgroup analyses. These elements address concerns raised in previous literature, including heterogeneous reporting, limited validation, and uncertain

generalizability.[11,17,20,21] The BBPS threshold of ≥6 allowed adequate mucosal assessment while retaining clinically relevant variability in bowel preparation quality.[14]

Limitations

This study has several limitations. First, it was retrospective and conducted at a single center using a single endoscopy platform. Endoscopy systems, video processors, imaging settings, bowel preparation protocols, patient populations, and documentation workflows vary across institutions; therefore, model performance may differ in external settings. Multicenter external validation across different endoscopy platforms and patient populations is essential before clinical deployment.

Second, although the reference standard was based on independent expert review with third-reader adjudication, MES remains inherently subjective. Ambiguity between adjacent categories, particularly MES 1 and MES 2, cannot be fully eliminated. The model may therefore have learned expert scoring conventions rather than a fully objective representation of inflammatory activity. Future studies should evaluate associations between model predictions and histologic remission, objective biomarkers, and longitudinal clinical outcomes.

Third, although most classification errors occurred between adjacent MES categories, several nonadjacent misclassifications were observed and reduced the quadratic weighted agreement. These cases should be examined in future model-development studies to determine whether they resulted from inadequate visualization, video-level aggregation, or limitations in the learned representations.

Fourth, the model assigns a patient-level MES but does not quantify the anatomic extent or spatial distribution of inflammation, which may limit its ability to represent total disease burden. Fifth, clinical impact was not directly evaluated, including effects on reporting time, real-time interobserver agreement, treatment decisions, histologic outcomes, relapse, hospitalization, colectomy, or treatment response. Although the mean inference time of 12.4 seconds per video may be suitable for offline analysis, real-time implementation would require further technical optimization and prospective workflow evaluation. Finally, the subgroup analyses were exploratory and require confirmation in independent cohorts.

Conclusion

In this retrospective temporal validation study, a video-based deep learning model achieved high discrimination, good agreement with an adjudicated expert reference standard, good calibration, and a positive net benefit for automated MES assessment in UC. Errors were concentrated at clinically recognized perceptual boundaries, particularly between MES 1 and MES 2. These findings support the feasibility of standardized video-based endoscopic grading; however, multicenter external validation and prospective clinical impact studies remain essential before implementation in routine practice.

Declarations

Animal and Human Rights Statement

All procedures performed in this study were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Informed Consent

The requirement for informed consent was waived due to the retrospective study design and use of de-identified data.

Data Availability

The data supporting the findings of this study are available from the corresponding author on reasonable request.

Conflict of Interest

The authors declare that they have no conflict of interest.

Funding

None.

Author Contributions (CRediT Taxonomy)

Conceptualization: Y.Ö., N.M.B.
Methodology: Y.Ö., N.M.B.
Software: Y.Ö.
Validation: Y.Ö., N.M.B.
Formal Analysis: Y.Ö.
Investigation: Y.Ö.
Resources: Y.Ö., N.M.B.
Data Curation: Y.Ö.
Writing – Original Draft Preparation: Y.Ö.
Writing – Review & Editing: Y.Ö., N.M.B.
Visualization: Y.Ö.
Supervision: N.M.B.
Project Administration: Y.Ö.

AI Usage Disclosure

Artificial intelligence-assisted tools, including ChatGPT (OpenAI), were used solely for language editing and improving the clarity of the manuscript. These tools were not used to generate or modify study data, perform statistical analyses, create figures, or determine scientific conclusions. All content was reviewed and verified by the authors, who take full responsibility for the final manuscript.

Abbreviations

AI: Artificial intelligence
AUROC: Area under the receiver operating characteristic curve
BBPS: Boston bowel preparation scale
CI: Confidence interval
GPU: Graphics processing unit
IQR: Interquartile range
MES: Mayo endoscopic subscore
MIL: Multiple instance learning
NPV: Negative predictive value
PPV: Positive predictive value
STARD: Standards for reporting diagnostic accuracy studies
STROBE: Strengthening the reporting of observational studies in epidemiology
UC: Ulcerative colitis

References

1. *Turner D, Ricciuto A, Lewis A, et al. STRIDE-II: an update on the Selecting Therapeutic Targets in Inflammatory Bowel Disease (STRIDE) initiative of the International Organization for the Study of IBD (IOIBD): determining therapeutic goals for treat-to-target strategies in IBD. Gastroenterology. 2021;160(5):1570-1583. doi:10.1053/j.gastro.2020.12.031*

2. *Rubin DT, Ananthakrishnan AN, Siegel CA, et al. ACG clinical guideline update: ulcerative colitis in adults. Am J Gastroenterol. 2025;120(6):1187-1224. doi:10.14309/ajg.0000000000003463*

3. *Hashash JG, Ng FY, Farraye FA, et al. Inter- and intraobserver variability on endoscopic scoring systems in Crohn disease and ulcerative colitis: a systematic review and meta-analysis. Inflamm Bowel Dis. 2024;30(11):2217-2226. doi:10.1093/ibd/izae051*

4. *Gottlieb K, Requa J, Karnes W, et al. Central reading of ulcerative colitis clinical trial videos using neural networks. Gastroenterology. 2021;160(3):710-719.e2. doi:10.1053/j.gastro.2020.10.024*

5. *Stidham RW, Liu W, Bishu S, et al. Performance of a deep learning model vs human reviewers in grading endoscopic disease severity of patients with ulcerative colitis. JAMA Netw Open. 2019;2(5). doi:10.1001/jamanetworkopen.2019.3963*

6. *Lo B, Liu Z, Bendtsen F, et al. High accuracy in classifying endoscopic severity in ulcerative colitis using a convolutional neural network. Am J Gastroenterol. 2022;117(10):1648-1654. doi:10.14309/ajg.0000000000001904*

7. *Yao H, Najarian K, Gryak J, et al. Fully automated endoscopic disease activity assessment in ulcerative colitis. Gastrointest Endosc. 2021;93(3):728-736.e1. doi:10.1016/j.gie.2020.08.011*

8. *Takenaka K, Fujii T, Kawamoto A, et al. Deep neural network for video colonoscopy of ulcerative colitis: a cross-sectional study. Lancet Gastroenterol Hepatol. 2022;7(3):230-237. doi:10.1016/s2468-1253(21)00372-1*

9. *Lv B, Ma L, Shi Y, et al. A systematic review and meta-analysis of artificial intelligence-*

diagnosed endoscopic remission in ulcerative colitis. iScience. 2023;26(11):108120. doi:10.1016/j.isci.2023.108120

10. *Ali S, Zhou F, Daul C, et al. Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy. Med Image Anal. 2021;70:102002. doi:10.1016/j.media.2021.102002*

11. *Liu X, Reigle J, Prasath VBS, et al. Artificial intelligence image-based prediction models in IBD exhibit high risk of bias: a systematic review. Comput Biol Med. 2024;171:108093. doi:10.1016/j.compbiomed.2024.108093*

12. *Bossuyt PM, Reitsma JB, Bruns DE, et al; STARD Group. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015;351. doi:10.1148/radiol.2015151516*

13. *von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med. 2007;4(10). doi:10.1371/journal.pmed.0040296*

14. *Calderwood AH, Jacobson BC. Comprehensive validation of the Boston Bowel Preparation Scale. Gastrointest Endosc. 2010;72(4):686-692. doi:10.1016/j.gie.2010.06.068*

15. *Khanna R, Ma C, Jairath V, et al. Endoscopic assessment of inflammatory bowel disease activity in clinical trials. Clin Gastroenterol Hepatol. 2022;20(4):727-736.e2. doi:10.1016/j.cgh.2020.12.017*

16. *Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565-574. doi:10.1177/0272989x06295361*

17. *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. Gastrointest Endosc. 2023;98(2):145-154.e8. doi:10.1016/j.gie.2023.04.2074*

18. *Rimondi A, Gottlieb K, Despott EJ, et al. Can artificial intelligence replace endoscopists when assessing mucosal healing in ulcerative colitis? A systematic review and diagnostic test accuracy meta-analysis. Dig Liver Dis. 2024;56(7):1164-1172. doi:10.1016/j.dld.2023.11.005*

19. *Pal P, Pooja K, Nabi Z, et al. Artificial intelligence in endoscopy related to inflammatory bowel disease: a systematic review. Indian J Gastroenterol. 2024;43(1):172-187. doi:10.1007/s12664-024-01531-3*

20. *Fan Y, Mu R, Xu H, et al. Novel deep learning-based computer-aided diagnosis system for predicting inflammatory activity in ulcerative colitis. Gastrointest Endosc. 2023;97(2):335-346. doi:10.1016/j.gie.2022.08.015*

21. *Stidham RW, Cai L, Cheng S, et al. Using computer vision to improve endoscopic disease quantification in therapeutic clinical trials of ulcerative colitis. Gastroenterology. 2024;166(1):155-167.e2. doi:10.1053/j.gastro.2023.09.049*

How to Cite

Yavuz Özden, Nuh Mehmet Büyükberber. Development and temporal validation of a deep learning model for automated Mayo Endoscopic Subscore assessment in ulcerative colitis. Ann Clin Anal Med 2026;17(10):1066-1072. doi:10.4328/ACAM.50232