Supplementary Table 1. Five-fold Cross-validation Performance In The Development Cohort (*n* = 750)

Metric	Value
Four-class MES prediction	
Macro-averaged AUROC (95% CI)	0.96 (0.95–0.97)
Quadratic weighted κ (95% CI)	0.92 (0.90–0.94)
Macro-*F*1 score	0.85
Overall accuracy, %	84
MES 0 versus MES ≥1; complete endoscopic remission	
AUROC (95% CI)	0.98 (0.97–0.99)
Brier score	0.06
MES ≤1 versus MES ≥2; clinically actionable threshold	
AUROC (95% CI)	0.97 (0.96–0.98)
Brier score	0.07

Abbreviations; AUROC, area under the receiver operating characteristic curve; CI, confidence interval; MES, Mayo Endoscopic Subscore.
Note; Performance metrics were calculated from pooled out-of-fold patient-level predictions. The 95% confidence intervals were estimated using 2,000 patient-level bootstrap resamples.

Supplementary Table 2. Development-cohort Performance For The MES ≤1 Versus MES ≥2 Endpoint

Metric	Value
AUROC (95% CI)	0.97 (0.96–0.98)
Sensitivity, % (95% CI)	88 (85–91)
Specificity, % (95% CI)	94 (92–96)
PPV, % (95% CI)	91 (88–94)
NPV, % (95% CI)	92 (89–95)
Brier score	0.07

Abbreviations; *AUROC, area under the receiver operating characteristic curve; CI, confidence interval; MES, Mayo Endoscopic Subscore; NPV, negative predictive value; PPV, positive predictive value.*
Note; AUROC and the Brier score were calculated using the endpoint-specific predicted probabilities. Sensitivity, specificity, PPV, and NPV were calculated by dichotomizing the four-class predicted MES categories according to the prespecified endpoint definition. The 95% confidence intervals were estimated using 2,000 patient-level bootstrap resamples.

Supplementary Table 3. Stratified Performance In The Temporal Validation Cohort

Stratification variable	Stratum	*n*	AUROC for MES ≤1 versus MES ≥2 (95% CI)	Quadratic weighted κ for four-class MES prediction (95% CI)
Bowel preparation quality	BBPS 7–9	175	0.96 (0.94–0.98)	0.91 (0.87–0.94)
	BBPS 6	75	0.94 (0.90–0.97)	0.88 (0.83–0.92)
Video duration	Above median (>8.5 min)	125	0.96 (0.93–0.98)	0.90 (0.86–0.94)
	At or below median (≤8.5 min)	125	0.95 (0.92–0.97)	0.89 (0.85–0.93)
Artifact burden	Low (<25% of frames removed)	120	0.97 (0.94–0.99)	0.92 (0.88–0.95)
	High (≥25% of frames removed)	130	0.94 (0.90–0.97)	0.87 (0.82–0.91)
Disease extent	Proctitis (E1)	68	0.95 (0.91–0.98)	0.88 (0.83–0.92)
	Left-sided colitis (E2)	102	0.96 (0.93–0.98)	0.90 (0.86–0.94)
	Extensive colitis (E3)	80	0.95 (0.92–0.97)	0.89 (0.85–0.93)

Abbreviations; *AUROC, area under the receiver operating characteristic curve; BBPS, Boston Bowel Preparation Scale; CI, confidence interval; MES, Mayo Endoscopic Subscore.*
Note; Low and high artifact burden were defined according to the proportion of frames removed during quality filtering. AUROC and quadratic weighted κ were calculated separately within each stratum. The 95% confidence intervals were estimated using 2,000 patient-level bootstrap resamples. These prespecified subgroup analyses should be interpreted as exploratory.