

# **Automated Assessment of the Ulcerative Colitis Endoscopic Index of Severity Using a Multi-task Deep Learning Model**

Bing Lv; Qiang Zheng; Xinxin Li; Tao Tao; Jingjing Wu; Yanting Shi\*

\*Corresponding author. E-mail: yantingshi@hotmail.com

## **Supplementary Methods**

### **Model Implementation and Training Details**

UC-MTLNet was implemented in PyTorch (version 2.0.1+cuda11.8) using the timm library (version 1.0.22), with CoAtNet-3 (coatnet\_3\_rw\_224) as the backbone initialized with ImageNet-pretrained weights. All images were resized to an input resolution of  $224 \times 224$  pixels. The model was optimized using AdamW with an initial learning rate of  $8 \times 10^{-5}$ , weight decay of  $1 \times 10^{-4}$ , and a batch size of 24. Training was performed for up to 100 epochs, with early stopping if the validation loss did not improve for 10 consecutive epochs. A linear learning-rate warm-up was applied for the first 3 epochs, followed by cosine annealing with a minimum learning rate of  $8 \times 10^{-7}$ . Gradient clipping was applied with a maximum norm of 1.0.

Class-weighted focal loss was used for each UCEIS descriptor, with a focusing parameter  $\gamma$  of 2 and label smoothing of 0.1. The three descriptor losses were combined using a dynamic task-weighting strategy. The best model for each fold was selected based on the lowest validation total loss.

During training, data augmentation included resizing to  $256 \times 256$  pixels, random cropping to  $224 \times 224$  pixels, random horizontal flipping, random vertical flipping, random rotation, random affine transformation, color jittering, random grayscale conversion, Gaussian blur, and random erasing.

Validation and test images were resized to  $224 \times 224$  pixels without augmentation. All images were normalized using ImageNet mean and standard deviation values.

To facilitate reproducibility, the source code, including the implementation of the dynamic task-weighting strategy and the complete data augmentation pipeline, has been made publicly available at <https://github.com/zibosyt/ulcerative-colitis-classification>. The relevant implementation can be found in the training and dataset-preprocessing scripts.

### **Bootstrap Confidence Interval Estimation**

The same patient-level cluster bootstrap framework was used both to estimate 95% confidence intervals for individual performance metrics and to assess differences in performance between UC-MTLNet and each comparator (single-task models or endoscopists), each with 5,000 iterations. In each iteration, patients were resampled with replacement at the unit of the whole patient (i.e., all endoscopic images from a resampled patient were carried into the resampled dataset together, preserving the within-patient image correlation structure), with the number of resampled patients equal to the original sample size. This procedure was applied consistently for both image-level and patient-level analyses.

For each bootstrap sample, performance metrics were calculated separately for the five independently trained models obtained from five-fold cross-validation and averaged. For comparisons with single-task models, metrics were calculated for the corresponding single-task models and averaged across their five independently trained folds; for comparisons with endoscopists, metrics were calculated separately for each of the five endoscopists and averaged to obtain reader performance. Paired differences (UC-MTLNet minus the corresponding comparator) were computed for each bootstrap sample.

For individual performance metrics, 95% confidence intervals were derived using the percentile method (2.5th and 97.5th percentiles of the bootstrap distribution of the metric). For between-group comparisons, 95% confidence intervals for paired differences were likewise derived using the percentile method (2.5th and 97.5th percentiles of the bootstrap distribution of paired differences). Two-sided P values were calculated by centering the bootstrap distribution at zero and determining the proportion of shifted values whose absolute magnitude equaled or exceeded the absolute observed difference; the minimum reportable P value was 0.0002 (1/5,000). P values were adjusted for multiple comparisons using Bonferroni correction, with the correction family for each comparator and metric comprising the four evaluation settings (image and patient levels in the internal and external test sets). All bootstrap analyses used a fixed random seed (42) to ensure reproducibility.

**Supplementary Table 1.** Ulcerative Colitis Endoscopic Index of Severity descriptors and definitions.

| Descriptor          | Score | Definition                                                                                                                            |
|---------------------|-------|---------------------------------------------------------------------------------------------------------------------------------------|
| Vascular pattern    | 0     | Normal vascular pattern with arborisation of capillaries clearly defined                                                              |
|                     | 1     | Patchy obliteration of vascular pattern                                                                                               |
|                     | 2     | Complete obliteration of vascular pattern                                                                                             |
| Bleeding            | 0     | No visible blood                                                                                                                      |
|                     | 1     | Some spots or streaks of coagulated blood on the surface of the mucosa, which can be washed away                                      |
|                     | 2     | Some free liquid blood in the lumen                                                                                                   |
|                     | 3     | Frank blood in the lumen ahead of endoscope or visible oozing from mucosa after washing intraluminal blood                            |
| Erosions and ulcers | 0     | Normal mucosa, no visible erosions or ulcers                                                                                          |
|                     | 1     | Tiny ( $\leq 5$ mm) defects in the mucosa, of a white or yellow colour with a flat edge                                               |
|                     | 2     | Larger ( $>5$ mm) defects in the mucosa, which are discrete fibrin-covered ulcers in comparison with erosions, but remain superficial |
|                     | 3     | Deep ulcers, with raised edges.                                                                                                       |

**Supplementary Table 2.** Distribution of endoscopic images for UCEIS components.

| UCEIS component     | Category | Development set<br>n (%) (n = 5,833) | Internal test set<br>n (%) (n = 2,546) | External test set<br>n (%) (n = 1,890) |
|---------------------|----------|--------------------------------------|----------------------------------------|----------------------------------------|
| Vascular pattern    | 0        | 2,349(40.3)                          | 1,049(41.2)                            | 706(37.4)                              |
|                     | 1        | 1,435(24.6)                          | 634(24.9)                              | 495(26.2)                              |
|                     | 2        | 2,049(35.1)                          | 863(33.9)                              | 689(36.5)                              |
| Bleeding            | 0        | 3,374(57.8)                          | 1,528(60.0)                            | 1,087(57.5)                            |
|                     | 1        | 1,383(23.7)                          | 588(23.1)                              | 486(25.7)                              |
|                     | 2        | 523(9.0)                             | 222(8.7)                               | 166(8.8)                               |
|                     | 3        | 553(9.5)                             | 208(8.2)                               | 151(8.0)                               |
| Erosions and ulcers | 0        | 3,631(62.2)                          | 1,654(65.0)                            | 1,168(61.8)                            |
|                     | 1        | 1,181(20.2)                          | 484(19.0)                              | 366(19.4)                              |
|                     | 2        | 359(6.2)                             | 160(6.3)                               | 169(8.9)                               |
|                     | 3        | 662(11.3)                            | 248(9.7)                               | 187(9.9)                               |

Data are presented as number of images (%). UCEIS, Ulcerative Colitis Endoscopic Index of Severity.

**Supplementary Table 3.** Initial interobserver agreement among the three reference annotators before consensus discussion.

| Variable            | Complete agreement, n (%) | Two-annotator agreement, n (%) | Three-annotator disagreement, n (%) | Mean pairwise QWK $\pm$ SD | Fleiss' kappa |
|---------------------|---------------------------|--------------------------------|-------------------------------------|----------------------------|---------------|
| Vascular pattern    | 8,330 (81.12)             | 1,939 (18.88)                  | 0 (0.00)                            | 0.9071 $\pm$ 0.0120        | 0.8071        |
| Bleeding            | 7,910 (77.03)             | 2,343 (22.82)                  | 16 (0.16)                           | 0.8703 $\pm$ 0.0137        | 0.7208        |
| Erosions and ulcers | 8,002 (77.92)             | 2,246 (21.87)                  | 21 (0.20)                           | 0.9006 $\pm$ 0.0125        | 0.7295        |
| Total UCEIS         | 6,037 (58.79)             | 3,965 (38.61)                  | 267 (2.60)                          | 0.9451 $\pm$ 0.0054        | 0.6384        |

Agreement was calculated using the independent scores assigned by the three reference annotators before consensus discussion. Complete agreement indicates that all three annotators assigned the same score. Two-annotator agreement indicates that two of the three annotators assigned the same score and the third differed. Three-annotator disagreement indicates that all three annotators assigned different scores.

**Supplementary Table 4.** Distribution of disagreement span among the three reference annotators before consensus discussion.

| Variable            | Span = 0, n (%) | Span = 1, n (%) | Span = 2, n (%) | Span = 3, n (%) | Span = 4, n (%) | Span $\geq$ 5, n (%) | Mean span $\pm$ SD  |
|---------------------|-----------------|-----------------|-----------------|-----------------|-----------------|----------------------|---------------------|
| Vascular pattern    | 8,330 (81.12)   | 1,939 (18.88)   | 0 (0.00)        | NA              | NA              | NA                   | 0.1888 $\pm$ 0.3914 |
| Bleeding            | 7,910 (77.03)   | 2,292 (22.32)   | 67 (0.65)       | 0 (0.00)        | NA              | NA                   | 0.2362 $\pm$ 0.4399 |
| Erosions and ulcers | 8,002 (77.92)   | 2,220 (21.62)   | 47 (0.46)       | 0 (0.00)        | NA              | NA                   | 0.2253 $\pm$ 0.4286 |
| Total UCEIS         | 6,037 (58.79)   | 3,130 (30.48)   | 832 (8.10)      | 255 (2.48)      | 15 (0.15)       | 0 (0.00)             | 0.5472 $\pm$ 0.7592 |

The disagreement span was defined as the maximum minus minimum score among the three reference annotators for each image before consensus discussion. A span of 0 indicates complete agreement. Larger spans indicate greater disagreement in assigned scores. NA indicates that the span was not applicable because the score range of the corresponding descriptor did not allow that value.

**Supplementary Table 5.** Patient-level performance of UC-MTLNet using image-wise maximum aggregation

| Test Set/Component       | Accuracy (%; 95% CI) | Weighted F1 score(95% CI) | QWK (95% CI)           |
|--------------------------|----------------------|---------------------------|------------------------|
| <b>Internal test set</b> |                      |                           |                        |
| Vascular pattern         | 84.27 (80.57-88.18)  | 0.8438 (0.8060-0.8828)    | 0.8785 (0.8428-0.9099) |
| Bleeding                 | 81.17 (76.72-85.62)  | 0.8146 (0.7721-0.8583)    | 0.9115 (0.8843-0.9340) |
| Erosions and ulcers      | 79.03 (74.12-84.06)  | 0.7965 (0.7510-0.8452)    | 0.8801 (0.8338-0.9182) |
| Total UCEIS              | 62.72 (57.58-67.88)  | 0.6265 (0.5714-0.6812)    | 0.9602 (0.9476-0.9705) |
| <b>External test set</b> |                      |                           |                        |
| Vascular pattern         | 85.00 (80.35-89.82)  | 0.8496 (0.8016-0.8982)    | 0.8809 (0.8390-0.9190) |
| Bleeding                 | 87.27 (83.86-90.91)  | 0.8754 (0.8421-0.9104)    | 0.9376 (0.9131-0.9575) |
| Erosions and ulcers      | 77.95 (72.59-83.64)  | 0.7823 (0.7297-0.8385)    | 0.8948 (0.8549-0.9276) |
| Total UCEIS              | 64.09 (58.40-69.68)  | 0.6491 (0.5914-0.7064)    | 0.9644 (0.9519-0.9742) |

Under image-wise maximum aggregation, total UCEIS was first calculated for each image, and the image with the highest total UCEIS within each patient was selected to derive the patient-level score. Descriptor scores in this table correspond to the vascular pattern, bleeding, and erosions and ulcers scores from the selected image. Data are presented as point estimates with 95% confidence intervals. CI, confidence interval; IMA, image-wise maximum aggregation; QWK, quadratic weighted kappa; UCEIS, Ulcerative Colitis Endoscopic Index of Severity.

**Supplementary Table 6.** Performance of UC-MTLNet for endoscopic remission identification.

| Level         | Test set          | Accuracy, %<br>(95% CI) | Sensitivity<br>(95% CI)   | Specificity<br>(95% CI)   | Cohen's kappa<br>(95% CI) | PPV                       | NPV                       |
|---------------|-------------------|-------------------------|---------------------------|---------------------------|---------------------------|---------------------------|---------------------------|
| Image-Level   | Internal test set | 95.66<br>(94.30-96.88)  | 0.9573<br>(0.9374-0.9725) | 0.9561<br>(0.9362-0.9729) | 0.9112<br>(0.8817-0.9355) | 0.9408<br>(0.9075-0.9648) | 0.9689<br>(0.9509-0.9821) |
| Image-Level   | External test set | 94.25<br>(92.20-96.12)  | 0.9595<br>(0.9413-0.9749) | 0.9324<br>(0.9010-0.9593) | 0.8790<br>(0.8312-0.9187) | 0.9058<br>(0.8045-0.9814) | 0.9329<br>(0.8918-0.9665) |
| Patient-Level | Internal test set | 92.62<br>(89.13-95.73)  | 0.7923<br>(0.7000-0.8778) | 0.9714<br>(0.9415-0.9948) | 0.7960<br>(0.7047-0.8730) | 0.9058<br>(0.8045-0.9814) | 0.9329<br>(0.8918-0.9665) |
| Patient-Level | External test set | 91.36<br>(86.14-95.91)  | 0.6909<br>(0.5360-0.8353) | 0.9879<br>(0.9619-1.0000) | 0.7454<br>(0.5982-0.8673) | 0.9495<br>(0.8340-1.0000) | 0.9059<br>(0.8411-0.9582) |

Endoscopic remission was defined as total UCEIS = 0, and active disease was defined as total UCEIS  $\geq 1$ . Data are presented as point estimates with 95% confidence intervals. CI, confidence interval; NPV, negative predictive value; PPV, positive predictive value; UCEIS, Ulcerative Colitis Endoscopic Index of Severity.

**Supplementary Table 7.** Performance of UC-MTLNet for UCEIS severity stratification.

| Level         | Test set          | Accuracy, % (95% CI) | Weighted F1 score (95% CI) | QWK (95% CI)           |
|---------------|-------------------|----------------------|----------------------------|------------------------|
| Image-Level   | Internal test set | 89.21 (87.38-90.97)  | 0.8917 (0.8732-0.9094)     | 0.9452 (0.9297-0.9560) |
| Image-Level   | External test set | 86.74 (84.11-89.19)  | 0.8660 (0.8393-0.8911)     | 0.9340 (0.9092-0.9495) |
| Patient-Level | Internal test set | 80.19 (76.25-84.38)  | 0.8008 (0.7611-0.8434)     | 0.8963 (0.8655-0.9217) |
| Patient-Level | External test set | 77.95 (72.88-83.27)  | 0.7793 (0.7287-0.8321)     | 0.8862 (0.8477-0.9162) |

UCEIS severity was stratified into four categories: remission, UCEIS 0; mild disease, UCEIS 1–3; moderate disease, UCEIS 4–6; and severe disease, UCEIS 7–8. Data are presented as point estimates with 95% confidence intervals. CI, confidence interval; QWK, quadratic weighted kappa; UCEIS, Ulcerative Colitis Endoscopic Index of Severity.

**Supplementary Table 8.** Performance of single-task models for UCEIS prediction.

| Level                | Test Set/Component       | Accuracy (% , 95% CI) | Weighted F1 score(95% CI) | QWK (95% CI)           |
|----------------------|--------------------------|-----------------------|---------------------------|------------------------|
| <b>Image-Level</b>   | <b>Internal test set</b> |                       |                           |                        |
|                      | Vascular pattern         | 90.19 (88.34-91.90)   | 0.9037 (0.8853-0.9208)    | 0.9313 (0.9126-0.9455) |
|                      | Bleeding                 | 87.45 (84.48-90.23)   | 0.8770 (0.8478-0.9048)    | 0.9221 (0.8968-0.9396) |
|                      | Erosions and ulcers      | 88.11 (85.54-90.58)   | 0.8883 (0.8647-0.9109)    | 0.9331 (0.9112-0.9460) |
|                      | Total UCEIS              | 74.60 (71.04-78.14)   | 0.7489 (0.7127-0.7852)    | 0.9724 (0.9639-0.9777) |
|                      | <b>External test set</b> |                       |                           |                        |
|                      | Vascular pattern         | 88.20 (85.55-90.55)   | 0.8832 (0.8569-0.9072)    | 0.9186 (0.8915-0.9381) |
|                      | Bleeding                 | 87.68 (85.04-90.21)   | 0.8787 (0.8524-0.9039)    | 0.9159 (0.8890-0.9320) |
|                      | Erosions and ulcers      | 89.19 (86.80-91.52)   | 0.8951 (0.8721-0.9178)    | 0.9404 (0.9145-0.9551) |
|                      | Total UCEIS              | 72.38 (68.03-76.73)   | 0.7179 (0.6709-0.7646)    | 0.9661 (0.9547-0.9729) |
| <b>Patient-Level</b> | <b>Internal test set</b> |                       |                           |                        |
|                      | Vascular pattern         | 86.21 (82.50-90.16)   | 0.8664 (0.8320-0.9040)    | 0.8910 (0.8593-0.9211) |
|                      | Bleeding                 | 85.05 (81.59-88.53)   | 0.8503 (0.8152-0.8854)    | 0.9291 (0.9089-0.9469) |
|                      | Erosions and ulcers      | 84.85 (81.19-88.75)   | 0.8491 (0.8126-0.8880)    | 0.9314 (0.9060-0.9517) |
|                      | Total UCEIS              | 70.10 (65.88-74.29)   | 0.7016 (0.6570-0.7478)    | 0.9690 (0.9610-0.9758) |
|                      | <b>External test set</b> |                       |                           |                        |
|                      | Vascular pattern         | 88.64 (85.36-92.08)   | 0.8856 (0.8519-0.9202)    | 0.9123 (0.8851-0.9372) |
|                      | Bleeding                 | 88.18 (85.00-91.64)   | 0.8844 (0.8539-0.9173)    | 0.9383 (0.9142-0.9607) |
|                      | Erosions and ulcers      | 83.64 (79.64-87.72)   | 0.8380 (0.7996-0.8781)    | 0.9270 (0.9012-0.9489) |
|                      | Total UCEIS              | 71.14 (66.79-75.64)   | 0.7092 (0.6648-0.7549)    | 0.9643 (0.9530-0.9741) |

Abbreviations: CI, confidence interval; QWK, quadratic weighted kappa; UCEIS, Ulcerative Colitis Endoscopic Index of Severity.

**Supplementary Table 9.** Performance comparison between UC-MTLNet and single-task models for descriptor-level and total UCEIS prediction.

| <b>Metric</b>                            | <b>Image-level<br/>Internal</b> | <b>Image-level<br/>External</b> | <b>Patient-level<br/>Internal</b> | <b>Patient-level<br/>External</b> |
|------------------------------------------|---------------------------------|---------------------------------|-----------------------------------|-----------------------------------|
| <b>UCEIS, QWK</b>                        |                                 |                                 |                                   |                                   |
| UC-MTLNet                                | 0.9690                          | 0.9646                          | 0.9729                            | 0.9783                            |
| Single-task                              | 0.9724                          | 0.9661                          | 0.9690                            | 0.9643                            |
| Difference (95% CI)                      | -0.0035<br>(-0.0066, -0.0009)   | -0.0015<br>(-0.0089, 0.0036)    | 0.0039<br>(-0.0034, 0.0116)       | 0.0140<br>(0.0042, 0.0262)        |
| P value                                  | 0.02 / 0.08                     | 0.627 / 1.00                    | 0.2922 / 1.00                     | 0.0156 / 0.0624                   |
| <b>UCEIS, Accuracy (%)</b>               |                                 |                                 |                                   |                                   |
| UC-MTLNet                                | 74.31                           | 73.1                            | 69.51                             | 71.59                             |
| Single-task                              | 74.60                           | 72.38                           | 69.51                             | 71.14                             |
| Difference (95% CI)                      | -0.29<br>(-2.28, 1.84)          | 0.72<br>(-2.27, 3.41)           | 0.00<br>(-6.02, 6.02)             | 0.45<br>(-5.91, 6.82)             |
| P value                                  | 0.823 / 1.00                    | 0.634 / 1.00                    | 1.00 / 1.00                       | 0.889 / 1.00                      |
| <b>Vascular, QWK</b>                     |                                 |                                 |                                   |                                   |
| UC-MTLNet                                | 0.9334                          | 0.9193                          | 0.8913                            | 0.9127                            |
| Single-task                              | 0.9313                          | 0.9186                          | 0.8910                            | 0.9123                            |
| Difference (95% CI)                      | 0.0021<br>(-0.0076, 0.0128)     | 0.0007<br>(-0.0133, 0.0132)     | 0.0003<br>(-0.0332, 0.0337)       | 0.0005<br>(-0.034, 0.0336)        |
| P value                                  | 0.689/1.00                      | 0.912/1.00                      | 0.984/1.00                        | 0.974/1.00                        |
| <b>Vascular, Accuracy (%)</b>            |                                 |                                 |                                   |                                   |
| UC-MTLNet                                | 90.4                            | 88.3                            | 85.63                             | 88.86                             |
| Single-task                              | 90.19                           | 88.2                            | 86.21                             | 88.64                             |
| Difference (95% CI)                      | 0.21 (-1.08, 1.64)              | 0.10 (-1.82, 1.85)              | -0.58 (-4.47, 3.50)               | 0.23 (-3.64, 3.86)                |
| P value                                  | 0.762/1.00                      | 0.921/1.00                      | 0.772/1.00                        | 0.902/1.00                        |
| <b>Bleeding, QWK</b>                     |                                 |                                 |                                   |                                   |
| UC-MTLNet                                | 0.9077                          | 0.9142                          | 0.9363                            | 0.9508                            |
| Single-task                              | 0.9221                          | 0.9159                          | 0.9291                            | 0.9383                            |
| Difference (95% CI)                      | -0.0144<br>(-0.027, -0.0043)    | -0.0017<br>(-0.0127, 0.0065)    | 0.0072<br>(-0.0107, 0.0252)       | 0.0125<br>(-0.0094, 0.0391)       |
| P value                                  | 0.0144/0.0576                   | 0.709/1.00                      | 0.410/1.00                        | 0.303/1.00                        |
| <b>Bleeding, Accuracy (%)</b>            |                                 |                                 |                                   |                                   |
| UC-MTLNet                                | 86.1                            | 87.16                           | 86.02                             | 89.09                             |
| Single-task                              | 87.45                           | 87.68                           | 85.05                             | 88.18                             |
| Difference (95% CI)                      | -1.36 (-3.03, 0.52)             | -0.52 (-1.97, 0.79)             | 0.97 (-3.30, 5.05)                | 0.91 (-2.73, 4.77)                |
| P value                                  | 0.127/0.510                     | 0.465/1.00                      | 0.645/1.00                        | 0.630/1.00                        |
| <b>Erosions and ulcers, QWK</b>          |                                 |                                 |                                   |                                   |
| UC-MTLNet                                | 0.9337                          | 0.9313                          | 0.9306                            | 0.9451                            |
| Single-task                              | 0.9331                          | 0.9404                          | 0.9314                            | 0.9270                            |
| Difference (95% CI)                      | 0.0006<br>(-0.0095, 0.0085)     | -0.009<br>(-0.0218, 0.0004)     | -0.0008<br>(-0.0164, 0.0154)      | 0.0182<br>(-0.001, 0.0413)        |
| P value                                  | 0.893/1.00                      | 0.097/0.389                     | 0.925/1.00                        | 0.093/0.373                       |
| <b>Erosions and ulcers, Accuracy (%)</b> |                                 |                                 |                                   |                                   |
| UC-MTLNet                                | 88.44                           | 88.04                           | 84.85                             | 86.36                             |
| Single-task                              | 88.11                           | 89.19                           | 84.85                             | 83.64                             |
| Difference (95% CI)                      | 0.33 (-0.99, 1.73)              | -1.14 (-2.91, 0.53)             | 0.00 (-3.30, 3.30)                | 2.73 (-1.59, 7.27)                |
| P value                                  | 0.642/1.00                      | 0.191/0.763                     | 1.00/1.00                         | 0.227/0.907                       |

Differences were calculated as UC-MTLNet minus the corresponding single-task model. Paired patient-level cluster bootstrap with 5,000 resampling iterations was used to estimate differences and 95% confidence intervals. P values are presented as uncorrected / Bonferroni-corrected values. Statistical significance was determined using Bonferroni-corrected P values. CI, confidence interval; QWK, quadratic weighted kappa; UCEIS, Ulcerative Colitis Endoscopic Index of Severity.

**Supplementary Table 10.** Inter-rater agreement among five endoscopists for total UCEIS prediction and severity stratification.

| Level         | Test set          | Task                    | Pairwise QWK mean $\pm$ SD | Range           | Fleiss' kappa |
|---------------|-------------------|-------------------------|----------------------------|-----------------|---------------|
| Image-Level   | Internal test set | Total UCEIS             | 0.9027 $\pm$ 0.0379        | 0.8518 - 0.9745 | 0.5084        |
| Image-Level   | External test set | Total UCEIS             | 0.9311 $\pm$ 0.0183        | 0.9063 - 0.9572 | 0.4744        |
| Image-Level   | Internal test set | Severity stratification | 0.8558 $\pm$ 0.0511        | 0.7931 - 0.9424 | 0.6760        |
| Image-Level   | External test set | Severity stratification | 0.8467 $\pm$ 0.0322        | 0.7940 - 0.8918 | 0.6359        |
| Patient-Level | Internal test set | Total UCEIS             | 0.9557 $\pm$ 0.0167        | 0.9367 - 0.9863 | 0.5846        |
| Patient-Level | External test set | Total UCEIS             | 0.9348 $\pm$ 0.0167        | 0.8996 - 0.9583 | 0.5312        |
| Patient-Level | Internal test set | Severity stratification | 0.8816 $\pm$ 0.0466        | 0.8191 - 0.9776 | 0.7156        |
| Patient-Level | External test set | Severity stratification | 0.8413 $\pm$ 0.0335        | 0.7764 - 0.8998 | 0.6587        |

UCEIS severity stratification included four categories: remission, UCEIS 0; mild disease, UCEIS 1–3; moderate disease, UCEIS 4–6; and severe disease, UCEIS 7–8. QWK, quadratic weighted kappa; SD, standard deviation; UCEIS, Ulcerative Colitis Endoscopic Index of Severity.

**Supplementary Table 11.** Performance of endoscopists for UCEIS prediction.

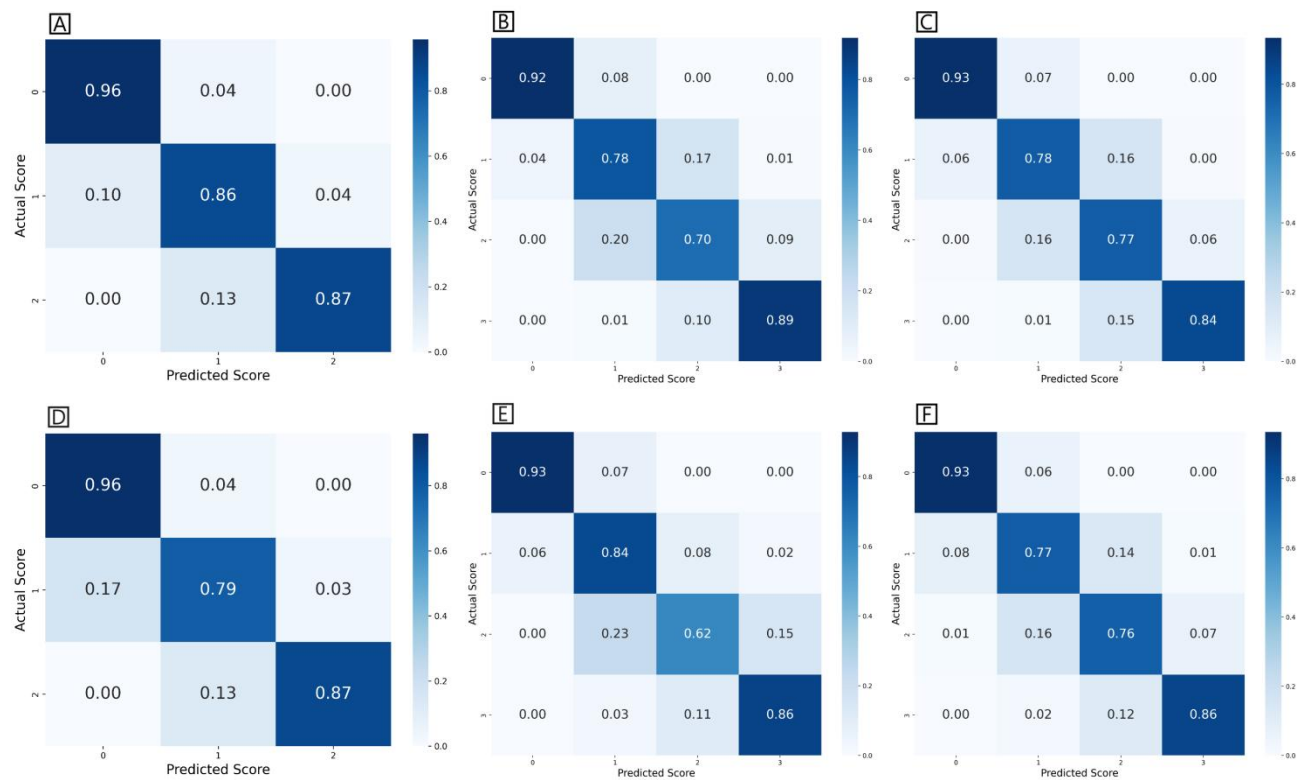
| Level         | Test Set/Component       | Accuracy (%; 95% CI) | Weighted F1 score(95% CI) | QWK (95% CI)           |
|---------------|--------------------------|----------------------|---------------------------|------------------------|
| Image-Level   | <b>Internal test set</b> |                      |                           |                        |
|               | Vascular pattern         | 84.25 (81.42-86.82)  | 0.8454 (0.8172-0.8716)    | 0.8916 (0.8633-0.9133) |
|               | Bleeding                 | 78.50 (74.37-82.53)  | 0.7845 (0.7417-0.8256)    | 0.8429 (0.8052-0.8685) |
|               | Erosions and ulcers      | 78.00 (73.56-82.30)  | 0.7922 (0.7485-0.8337)    | 0.8741 (0.8336-0.8993) |
|               | Total UCEIS              | 59.58 (54.80-64.49)  | 0.6122 (0.5635-0.6618)    | 0.9307 (0.9103-0.9441) |
|               | <b>External test set</b> |                      |                           |                        |
|               | Vascular pattern         | 83.66 (80.76-86.54)  | 0.8392 (0.8100-0.8679)    | 0.8851 (0.8547-0.9080) |
|               | Bleeding                 | 78.73 (75.58-81.91)  | 0.7928 (0.7609-0.8244)    | 0.8484 (0.7821-0.8838) |
|               | Erosions and ulcers      | 82.63 (79.05-86.10)  | 0.8344 (0.7989-0.8682)    | 0.8777 (0.8236-0.9109) |
|               | Total UCEIS              | 59.99 (55.78-64.20)  | 0.6111 (0.5668-0.6546)    | 0.9425 (0.9246-0.9536) |
| Patient-Level | <b>Internal test set</b> |                      |                           |                        |
|               | Vascular pattern         | 84.85 (80.90-89.03)  | 0.8465 (0.8055-0.8889)    | 0.8802 (0.8465-0.9118) |
|               | Bleeding                 | 78.45 (74.19-82.99)  | 0.7866 (0.7442-0.8313)    | 0.8997 (0.8739-0.9223) |
|               | Erosions and ulcers      | 80.97 (76.39-85.76)  | 0.8111 (0.7662-0.8582)    | 0.9190 (0.8928-0.9409) |
|               | Total UCEIS              | 58.06 (52.73-63.48)  | 0.5852 (0.5314-0.6398)    | 0.9514 (0.9393-0.9617) |
|               | <b>External test set</b> |                      |                           |                        |
|               | Vascular pattern         | 87.50 (83.33-91.93)  | 0.8744 (0.8327-0.9189)    | 0.9020 (0.8677-0.9350) |
|               | Bleeding                 | 78.41 (74.10-82.86)  | 0.7829 (0.7393-0.8277)    | 0.8456 (0.7942-0.8933) |
|               | Erosions and ulcers      | 82.27 (77.74-87.04)  | 0.8215 (0.7743-0.8707)    | 0.9001 (0.8659-0.9313) |
|               | Total UCEIS              | 55.00 (50.00-60.01)  | 0.5644 (0.5123-0.6183)    | 0.9544 (0.9435-0.9632) |

Abbreviations: CI, confidence interval; QWK, quadratic weighted kappa; UCEIS, Ulcerative Colitis Endoscopic Index of Severity.

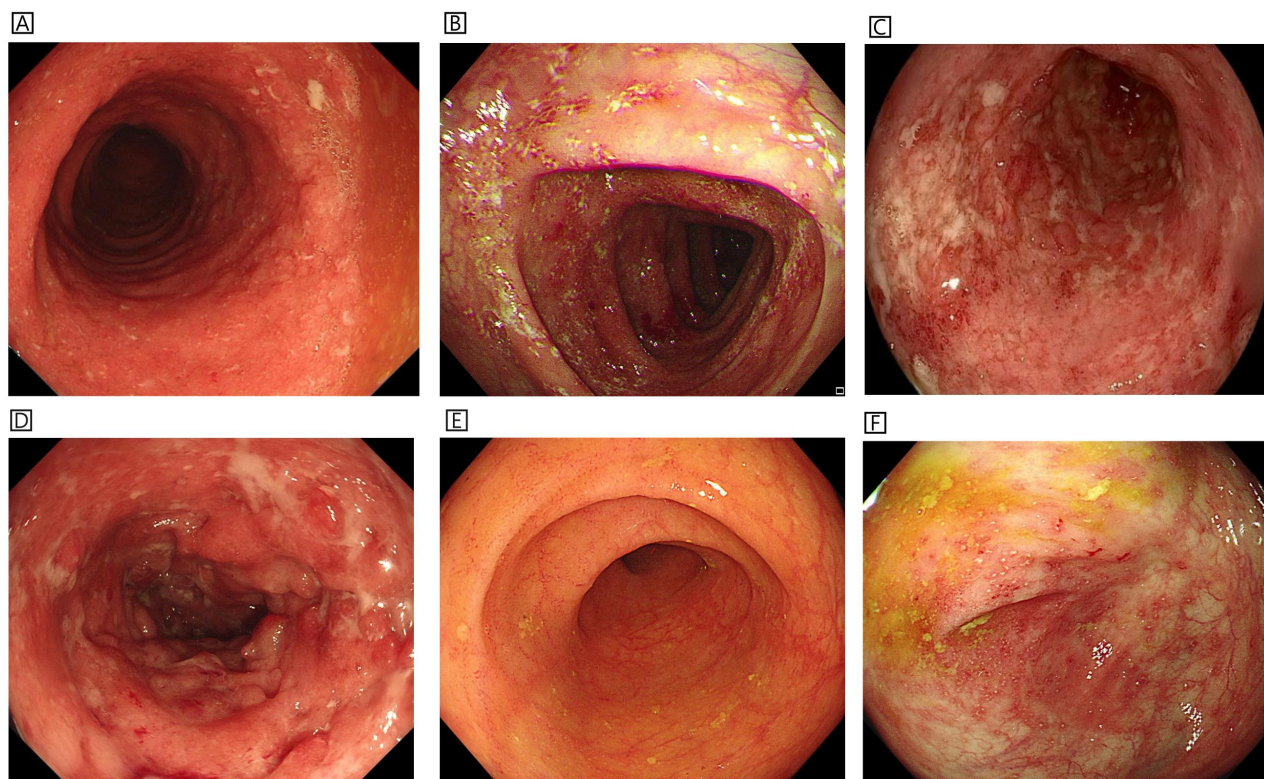
**Supplementary Table 12.** Exploratory analysis of the association between image number and patient-level total UCEIS prediction performance.

| Test Set/Fold            | r (image no. vs. correctness) | r (image no. vs. absolute error) | r (image no. vs. prediction bias) | QWK    | No. of patients |
|--------------------------|-------------------------------|----------------------------------|-----------------------------------|--------|-----------------|
| <b>Internal test set</b> |                               |                                  |                                   |        |                 |
| Fold 0                   | 0.0249                        | -0.0407                          | 0.0092                            | 0.9777 | 103             |
| Fold 1                   | -0.0739                       | 0.0545                           | -0.0172                           | 0.967  | 103             |
| Fold 2                   | 0.0127                        | -0.0083                          | 0.074                             | 0.9745 | 103             |
| Fold 3                   | 0.0077                        | -0.0518                          | 0.0455                            | 0.9703 | 103             |
| Fold 4                   | -0.0372                       | 0.0225                           | -0.0415                           | 0.9752 | 103             |
| <b>External test set</b> |                               |                                  |                                   |        |                 |
| Fold 0                   | 0.0406                        | 0.0107                           | 0.0999                            | 0.979  | 88              |
| Fold 1                   | -0.0441                       | 0.0291                           | 0.0509                            | 0.9709 | 88              |
| Fold 2                   | -0.029                        | 0.0248                           | 0.0175                            | 0.9824 | 88              |
| Fold 3                   | 0.0712                        | -0.0691                          | 0.0192                            | 0.9787 | 88              |
| Fold 4                   | -0.0582                       | 0.0394                           | 0.0909                            | 0.9805 | 88              |

r, Pearson correlation coefficient between the number of images per patient and the specified outcome, calculated separately within the internal test set (n = 103) and the external test set (n = 88) for each independently trained cross-validation model. None of the 30 correlations (15 per test set) reached statistical significance (critical  $|r| \approx 0.195$  for the internal test set and  $\approx 0.211$  for the external test set, at  $\alpha = 0.05$ ).



**Supplementary Figure 1.** Confusion matrices of UC-MTLNet at the image level for the three UCEIS subscores. (A) Vascular pattern, internal test set. (B) Bleeding, internal test set. (C) Erosions and ulcers, internal test set. (D) Vascular pattern, external test set. (E) Bleeding, external test set. (F) Erosions and ulcers, external test set.



**Supplementary Figure 2.** Representative cases illustrating sources of image-level scoring disagreement, selected from misclassified images. (A, B) Disagreement in bleeding scoring (score 0 vs. 1) due to isolated or scattered washable blood traces. (C, D) Disagreement in erosions and ulcers scoring (score 2 vs. 3) due to difficulty judging lesion depth from a single image. (E) Disagreement in vascular pattern scoring (score 0 vs. 1) due to subtle vascular blurring. (F) Combined disagreement in bleeding and erosions and ulcers scoring due to ambiguous fibrin-covered lesions.