TITLE PAGE

Title

Application of Deep Learning Models to Improve Ulcerative Colitis Endoscopic Disease Activity Scoring Under Multiple Scoring Systems

Short title

Artificial Intelligence in Ulcerative Colitis

Authors

Michael F Byrne[1,2], Remo Panaccione[3], James E East[4], Marietta Iacucci[5], Nasim Parsa[2,6], Rakesh Kalapala[7], Duvvur N Reddy[7], Hardik Ramesh Rughwani[7], Aniruddha P Singh[7], Sameer K Berry[8], Ryan Monsurate[2], Florian Soudan[9], Greta Laage[9], Enrico D Cremonese[9], Ludovic St-Denis[9], Paul Lemaître[9], Shima Nikfal[9], Jerome Asselin[9], Milagros L Henkel[2], Simon P Travis[4]

Affiliations

1 Division of Gastroenterology, Department of Medicine. Vancouver General Hospital, University of British Columbia, Vancouver, British Columbia, Canada.
2 Satisfai Health, Vancouver, British Columbia, Canada.
3 Division of Gastroenterology, University of Calgary, Calgary, Canada.
4 Translational Gastroenterology Unit and Oxford NIHR Biomedical Research Centre, Nuffield Department of Clinical Medicine, Experimental Medicine Division, University of Oxford, John Radcliffe Hospital, Headley Way, Headington, Oxford OX3 9DU, United Kingdom.
5 Institute of Translational Medicine, Division of Gastroenterology, Birmingham, United Kingdom.
6 Division of Gastroenterology, Mayo Clinic, Scottsdale, Arizona, USA.
7 Asian Institute of Gastroenterology (AIG Hospitals), Gachibowli, Hyderabad, India.
8 Division of Gastroenterology & Hepatology, University of Michigan, Ann Arbor, MI, USA.
9 IVADO Labs, Montreal, Canada.

Grant support

None

Abbreviations

AI, artificial intelligence; IBD, inflammatory bowel disease; MES, mayo endoscopic subscore; DL, deep learning; CNN, convolutional neural network; QA, qualification accuracy; QWK, quadratic weighted kappa; UC, ulcerative colitis; UCEIS, ulcerative colitis endoscopic index of severity; GI, gastrointestinal; CR, central reader.

Correspondence

Michael F. Byrne MA MD (Cantab) MRCP FRCPC
Clinical Professor of Medicine
Vancouver General Hospital
University of British Columbia
5153 - 2775 Laurel Street
Vancouver, BC, Canada
Email: michael.byrne@vch.ca
Phone: +1 604 875 5474
Fax: +1 604 628 2419

Disclosures

Dr. Byrne: CEO, Founder, and shareholder in Satisfai Health Inc.
Dr. East: Has served on the clinical advisory board and has share options in Satisfai Health Inc.; and reports speaker fees from Falk.
Dr. Panaccione: Shareholder in Satisfai Health Inc.
Dr. Henkel: Shareholder in Satisfai Health Inc.
Ms. Laage: Employee, IVADO labs.
Mr. St-Denis: Employee, IVADO labs.
Mr. Lemaître: Employee, IVADO labs.
Ms. Nikfal: Employee, IVADO labs.
Mr. Asselin: Employee, IVADO labs.
Mr. Monsurate: Shareholder in Satisfai Health Inc.
Mr. Cremonese: Employee, IVADO labs.
Mr. Soudan: Employee, IVADO labs.
Dr. Travis: Consultant for Satisfai Health Inc.
Dr. Parsa: Shareholder in Satisfai Health Inc.
Dr. Singh has no financial relationships to disclose.
Dr. Kalapala has no financial relationships to disclose.
Dr. Berry has no financial relationships to disclose.
Dr. Reddy has no financial relationships to disclose.
Dr. Singh has no financial relationships to disclose.
Dr. Rughwani has no financial relationships to disclose.
Dr. Iacucci has research grants from Pentax, Olympus, and Fuji.

Author Contributions

M.F. Byrne (Conceptualization: Lead Investigation: Lead; Supervision: Lead; Writing – original draft: Lead; Writing – review & editing: Equal)
R. Panaccione (Conceptualization: Lead; Investigation: Lead; Writing – original draft: Supporting; Writing – review & editing: Equal)
J.E. East (Conceptualization: Lead; Investigation: Lead; Writing – original draft: Supporting; Writing – review & editing: Equal)
M. Iacucci (Conceptualization: Supporting; Data curation: Equal; Writing – review & editing: Equal)
N. Parsa (Writing – original draft: Lead; Writing – review & editing: Equal)
R. Kalapala (Data curation: Equal; Writing – review & editing: Equal)
N.R. Duvvur (Data curation: Equal; Writing – review & editing: Equal)
H. Rughwani (Data curation: Equal; Writing – review & editing: Equal)
A.P. Singh (Data curation: Equal; Writing – review & editing: Equal)
S. Berry (Writing – original draft: Lead; Writing – review & editing: Equal)
R. Monsurate (Conceptualization: Supporting; Writing – original draft: Lead; Writing – review & editing: Equal)
F. Soudan (Conceptualization: Lead; Investigation: LEAD; Methodology: Lead; Supervision: Lead; Writing – original draft: Lead; Writing – review & editing: Equal)
G. Laage (Methodology: Supporting; Writing – original draft: Lead; Writing – review & editing: Equal)
E.D. Cremonese (Conceptualization: Supporting; Writing – review & editing: Equal)
L. St-Denis (Conceptualization: Supporting; Methodology: Supporting; Writing – original draft: Lead; Writing – review & editing: Equal)
P. Lemaître (Methodology: Supporting; Writing – original draft: Lead; Writing – review & editing: Equal)
S. Nikfal (Methodology: Supporting; Writing – original draft: Lead; Writing – review & editing: Equal)
J. Asselin (Methodology: Supporting; Writing – original draft: Lead; Writing – review & editing: Equal)
M.L. Henkel (Data curation: Equal; Writing – original draft: Lead; Writing – review & editing: Equal)
S.P. Travis (Investigation: LEAD; Data curation: Equal; Supervision: Lead; Writing – original draft: Supporting; Writing – review & editing: Equal)

Data availability statement

Data available on request: the data underlying this article will be shared upon request to the corresponding author.

Keywords
Inflammatory Bowel Disease; Deep learning; Mayo Endoscopic Subscore; Ulcerative Colitis Endoscopic Index of Severity.

ABSTRACT

Background & Aims: Lack of clinical validation and inter-observer variability are two limitations of endoscopic assessment and scoring of disease severity in patients with Ulcerative Colitis. We developed a deep learning (DL) model to improve, accelerate and automate UC detection, and predict the Mayo Endoscopic Subscore (MES) and the Ulcerative Colitis Endoscopic Index of Severity (UCEIS).

Methods: A total of 134 prospective videos (1,550,030 frames) were collected and those with poor quality were excluded. The frames were labeled by experts based on MES and UCEIS scores. The scored frames were used to create a preprocessing pipeline and train multiple convolutional neural networks (CNNs) with proprietary algorithms in order to filter, detect and assess all frames. These frames served as the input for the DL model, with the output being continuous scores for MES and UCEIS (and its components). A graphical user interface was developed to support both labeling video sections and displaying the predicted disease severity assessment by the AI from endoscopic recordings.

Results: Mean absolute error (MAE) and mean bias were used to evaluate the distance of continuous model's predictions from ground truth and its possible tendency to over/under-predict were excellent for MES and UCEIS. The quadratic weighted kappa used to compare the inter-rater agreement between experts' labels and the model's predictions showed strong agreement (0.87, 0.88 frame-level, 0.88, 0.90 section-level and 0.90, 0.78 at video-level, for MES and UCEIS, respectively).

Conclusions: We present the first fully automated tool that improves the accuracy of the MES and UCEIS, reduces the time between video collection and review, and improves subsequent quality assurance and scoring.

What You Need to Know:

Background and context: Endoscopic assessment and scoring the disease severity in UC is limited by inter-observer variability and lack of clinical validation.

New findings: We present the first fully automated AI model for UC disease activity scoring under both the MES and UCEIS, at frame, section, and video levels, that is ready for use in clinical practice. Our model improves the accuracy of both scoring systems, reduces the time between video collection and review, and improves subsequent quality assurance and scoring.

Limitations: Limited dataset with imbalanced classes, limited generalizability, difficulty in describing a fair comparison with the literature due to the lack of an open dataset, subjective ground truth for MES and UCEIS resulting in potential bias for the labelers reviewing AI-generated sections with GUI.

Impact: Our results enable the development of a model that can be used to improve the efficiency and accuracy of UC endoscopic assessment and scoring at different stages of the clinical journey such as video quality assurance by physicians and increase the efficiency of central reading in clinical trials.

Introduction

Ulcerative colitis (UC) is a chronic inflammatory disease of the colon and rectum with increasing incidence and prevalence worldwide[1]. Several treatment options are available for UC, based on disease activity, severity, and prior response to medical treatments[2]. In patients with UC, the disease activity and severity can be assessed using inflammatory markers, clinical symptom scores, endoscopic inflammation scores, and histologic scoring systems[3-7]. One of the main goals of therapy in patients with UC is to achieve "mucosal healing" which has been shown to be associated with decreased rates of steroid use, hospitalization, colectomy, and improved quality of life[8]. The status of mucosal inflammation during colonoscopy can be reported with scoring systems such as the Mayo Endoscopic Subscore (MES) and the Ulcerative Colitis Endoscopic Index of Severity (UCEIS)[9-10]. The MES 0-1 has been reported to be associated with improved rates of clinical remission, while the UCEIS score has been shown to be a more accurate reflection of UC severity and clinical remission, and the short and long-term clinical outcomes---clinical remission (UCEIS 0–1), mild (UCEIS 2–4), moderate (UCEIS 5–6) and severe (UCEIS 7–8)[11-14].

While disease severity scoring systems are established, the presence of inter-observer variability and lack of clinical validation remain two important limitations of endoscopic assessment and scoring of disease severity in patients with UC[10]. To overcome these limitations and improve the inter-observer agreement, central reading by clinically blinded off-site experienced endoscopists--Central Readers (CRs)-- has been used as a crucial component in UC clinical trials[15]. Recently, artificial Intelligence (AI) has been utilized to enhance the interpretation of endoscopic images to assess disease severity in patients with UC and to strive at reducing the delay and cost associated with central reading activity[16,17]. Studies have shown encouraging results in the application of deep learning (DL) models in the UC diagnostic paradigm to improve the disease activity scoring, especially in central reading for clinical trials when using the MES[16-20]. Stidham et al built a DL model which successfully distinguished between active disease (UCEIS 2–4) or endoscopic remission (UCEIS 0–1) from colonoscopy videos and was able to identify exact Mayo sub scores with comparable accuracy to three experienced human reviewers[18]. However, the UCEIS scoring system, which focuses on design features that minimize interobserver variability, may potentially allow for training models with superior assessment and scoring ability if all the features of the

UCEIS score are used, rather than just binary distinctions between active disease and endoscopic remission. Therefore, we developed a DL methodology to improve, accelerate and automate UC disease detection. Specifically, we trained several convolutional neural networks models (CNN) to preprocess endoscopic recordings with the final goal of assessing the MES, UCEIS and its three descriptor indices to video sections. Sections are automatically generated in continuous intervals of recording depicting a stable, observable disease state. A user-friendly and interactive Graphical User Interface (GUI) was designed to show the results of the various CNN models to help experts efficiently assign UC disease activity. Our system as described here was originally developed to work on recorded video. However, our models can infer fast enough to output results on the GUI to be considered real time.

Materials and Methods

General

The aim of the models developed is to predict the MES, UCEIS, and its three descriptor indices from the reviewed sections. We describe in detail the different steps of the approach in the next sections.

Data Collection

We used unaltered and de-identified colonoscopy and sigmoidoscopy videos of UC patients provided by the Asian Institute of Gastroenterology (AIG Hospitals). Local institutional review board approval was obtained for this study (AIGAEC-BH&R 08/10/2020-03). We carried out various simulations to estimate the required sample size for our models' accuracy. In brief, we generated 10000 samples of random discrete uniformly distributed UCEIS scores with independent random normally distributed errors to simulate the model's error.

A total of 134 prospective videos were collected with Olympus HDR-60 scopes (190 and 180 Series) between October 2020 and April 2021. These videos were encoded with the YUV420p pixel format, and were also deinterlaced at 25 frames per second, resulting in a total dataset of

1,550,030 frames. We describe the DL workflow below and the details of our dataset in the Results section.

First Step: Video Quality Assessment

The first step of the proposed methodology consisted of a quality assurance control to filter out videos that were deemed of poor quality by a domain expert. The reasons for excluding a video included, but were not limited to, poor bowel preparation, ex-vivo footage, and being out-of-focus. This step was qualitatively applied on whole videos (i.e., not for each frame).

For further clarity, we only removed videos where there was very little or almost no visible mucosa. Most videos were retained and used for training, and had a mix of good and poor quality sequences, thus representing real world scenario. The recordings that were not filtered out were decomposed into frames and sent to the second step of the workflow.

Second Step: Pre-Processing Pipeline Application

The pre-processing pipeline consisted of four sub-steps as described in Figure 1: a blue light identifier algorithm, a scorability assignment model, a biopsy procedure and ex-vivo detector, and a frame-based disease severity assessment model. Each algorithm and model used in the pre-processing pipeline was developed and trained internally on proprietary endoscopy videos and labels.

Blue Light Identifier Algorithm

With the purpose of keeping only white-light frames with normal magnification throughout the process, a heuristic to identify image-enhancement using blue light imaging based on pixel color was applied on all frames.

Scorability Assignment Model

The objective of this sub-step is to distinguish “scorable” from “non-scorable” frames. Scorable frames are defined such that they easily allow the assessment of UC activity by a GI specialist. On the other hand, non-scorable frames are considered as challenging frames for scoring UC due to

either feces and/or water jet presence, visible biopsy tool, and post biopsy bleeding (not to be confused with blood from the disease itself). They can also include ex-vivo and out of focus (shadowed, too close to the mucosa, or blurred) frames. For this task, we employed a CNN which outputs the probability for a frame to be scorable, and we kept the frames that met a given threshold.

Biopsy Procedure and Ex-Vivo Detector

With the aim of providing more contextual information for our workflow, we leveraged a CNN model taking as input non-scorable frames detected 'as is' in the previous sub-step of the pre-processing pipeline. We focused on two specific scenarios--biopsy procedures and ex-vivo footage. Detection of a biopsy procedure was needed to avoid confusion between disease-sate bleeding and fresh blood from biopsies when assessing the UCEIS Bleeding descriptor. Furthermore, frames detected as ex-vivo were essential to define sections for which a review is impossible.

Frame-Based Disease Severity Assessment Model

The goal of this sub-step was to assign MES, UCEIS, and the three descriptors for UCEIS (Erosions and Ulcers, Vascular Pattern, and Bleeding) to scorable frames detected 'as is' by the scorability assignment model with a dedicated CNN. The predicted scorable images were used to create continuous stable disease state sections in the next step of the workflow.

Third Step: Section Generation

To mimic the performance of experts in the reviewing process for UC disease assessment, we broke down endoscopy videos into short sections of continuous frames representing stable disease states in order to score coherent parts of the videos. Decomposition into such sections was motivated by two goals: the need to stabilize the endoscopic video review process by expert readers, and to also develop an efficient in-house labeling system leveraging the expertise of an internal specialist team. This team consisted of one global central reading expert (ST, gold

standard), six GI specialists (silver standard), and 20 GI trainees (bronze standard). ST was the clinician who first described the UCEIS scoring system.

Since endoscopic disease activity (and therefore the MES and the UCEIS scores) can vary widely in a short time window, visual 'noise' is created for readers who are assessing UC severity. It was for this reason that the UCEIS was designed to be scored in the worst affected area during flexible sigmoidoscopy, although no such stipulation is described for the MES. To manage the variation and facilitate assessment, there was a need to create consistent, appropriately long representative sections of stable disease to help readers navigate within the GUI to scroll through the videos quickly to assess UC severity.

We developed an algorithm in which inputs were the frames that were assigned a score in the last step of the preprocessing pipeline, and which had output sections varying from three to twenty seconds. Biopsy procedure, ex-vivo, and non-scorable frames were similarly processed through the algorithm but were not categorized as scorable sections.

Sections were created from all the scored frames, whether they were scorable or non-scorable. Our generated sections from the algorithm were short sections of continuous and mainly scorable frames representing stable disease states as explained in the previous paragraph. Of course, non-scorable frames (outliers) could be integrated in the sections, but they were negligible. The same algorithm was also able to flag non-scorable sequences (biopsy, ex-vivo, etc.). We did not call them "sections" as we did not want to confuse the reader. "Sections" were only scorable sequences for the purposes of our description. In addition, the section creation process was developed as an offline process to be executed on recorded videos. As stated earlier, our models can infer and create sections fast enough to be considered as real-time.

Fourth Step: Graphical User Interface Leverage

The previous steps of the DL approach enabled the development of a GUI displaying videos and their respective created sections for review. The web-based interface was built with sequential ordered steps to optimize the endoscopic videos review workflow, described in Figure 2.

Reviewers can first access the tool whenever endoscopic videos from patients with UC need to be reviewed (step 1 in fig. 2). Frames from videos under review have gone through the automated

preprocessing pipeline and were estimated either as non-scorable or scorable. Scorable frames underwent evaluation by the section creation and refinement algorithms such that newly created and refined sections could be represented along with the non-scorable frames in a timeline under the video in review (step 2 in fig.2). This timeline with markers indicated contextual information according to a color code. For the continuous sections representing stable disease state, gradient colors are used to highlight the severity of the UC: light grey for no disease activity; green for low MES/UCEIS scores; yellow/orange for medium MES/UCEIS scores; and red for high MES/UCEIS scores. Non-scorable, blue light, and out of the body video portions are highlighted in a gradient of grey colors. Biopsy procedures are also highlighted in blue.

The tool featured the presentation - in decreasing order - of the first high disease activity section down to the last one based on MES and UCEIS scores. In this way, the expectation is that clinical readers save time by reviewing only the relevant sections of the video to confirm the score assigned to each section and the whole recording (step 3 in fig.2). If needed, users can tag specific features in the reviewed video such as scope trauma, biopsy blood, and poor bowel preparation, all of which can be leveraged later to optimize the preprocessing pipeline (step 4 in fig.2). Once the user scores a section that is at least 2 points higher than any of the remaining AI scored sections, the video will be assigned the highest UCEIS score and MES score (step 5 in fig.2). (Note that although it is not shown in Figure 2, the highest MES is also displayable by configuration.) This results in a live, simple, interactive, and user-friendly application usable to improve the reviewed workflow, thus speeding up the reading process while improving the accuracy of the UC disease activity assessment.

For the work presented in this paper, the GUI was used to obtain high-quality labels at a section-level. Thus, the tool was used by GI specialists to review each generated section to either confirm or refute the estimated MES, UCEIS, and the three UCEIS descriptors as required. At least two reviewers scored every section used in the training phase.

The utilization of the sections built from raw labels and reviewed by medical experts resulted in two major improvements: a faster review process, and a large quantity of high-quality ground truth labels. We used the latter to train the section-based severity assignment model in the final step of the approach. It was therefore an iterative process (Of note, the GUI can also be used as

a review tool by central readers to automatically characterize UC disease activity in endoscopic videos).

Fifth step: Disease Severity Assessment

Dataset Creation

In the final step of the DL workflow, we trained a CNN referred to as 'Section-based Disease Assessment (SDA)'. The data used to develop the model contained frames from each section that was assigned an MES score, or UCEIS and its three descriptors. The scores of a section were assigned to all of the frames which made up that section. The dataset used in this final step of the workflow was split at a video-level into training, validation, and test sets in a 60%-20%-20% distribution, with no overlap of videos used in these three sets.

Model Generation

We formulated the problem as a regression task. Therefore, we developed a CNN model taking frames from reviewed sections as input, which then outputs continuous scores for MES, UCEIS, and its three descriptors, also at frame-level. The objective was to provide a precise score to help UC disease severity assessment as well as to provide granular reporting of results. In the Results section below, we explain the approach to infer frame-level findings to section-level.

The CNN based model applied is an EfficientNetB3 architecture with weights pre-trained on ImageNet. We appended a global average pooling layer and dense layers to this network. The output layer contained five separate dense layers predicting continuous scores at frame-level according to their respective scale: MES, aggregate UCEIS, and the individual UCEIS descriptors: Vascular Pattern, Bleeding, and Erosions and Ulcers. Thus, for each input frame, the model predicted five scores. The model high-level architecture is shown in Figure 3.

Model Training

Each image was resized to a resolution of 320 x 320 for processing efficiency with no degradation of outputs. Data balancing and standard computer vision data augmentation techniques such as

cropping, resizing, flipping, rotating, and color modification (including contrast, brightness and saturation) were applied to improve the model's ability to generalize.

Various iterations of the architecture were evaluated by varying the random seed and the hyperparameters such as the loss function, dropout rate, learning rate, number of epochs, and optimizer to both ensure high precision results and prevent overfitting. In addition to the hyperparameter search, an architecture search was also performed, exploring a variety of CNN architectures and dense layer configurations. Note that while all models were assessed on a validation set, all results shown in this manuscript are based on a separate hold-out / test set, unseen during training or validation. Model generation and training experiments were performed using TensorFlow.

Throughout the DL workflow, we developed an iterative approach to improve the section definition and thus the model's predictions. With the SDA CNN's outputs, we updated the heuristic that defined section boundaries in videos. These updated section definitions had greater autocorrelation, resulting in more consistent and accurate section level reads. Once these new sections were reviewed, an improved model was trained based on the new sections, allowing for an iterative approach to improving section scores. These iterative refinements permitted a more granular score to support the assessment of UC disease severity.

Performance metrics

We assessed the model performance on MES, UCEIS, and UCEIS descriptors at section and video level. Since we tried to mimic expert scoring behavior by creating coherent sections of videos to be reviewed, we placed an emphasis on section-level results.

In order to obtain section-level scores from frame-level predictions, we computed the 83rd percentile for each score over the frames belonging to a section (this number was chosen based on the results of the validation set). We decided to use the 83rd percentile of each score over the frames to infer from a section-level to a frame-level since it gave us the best results based on several tests. In that way, we were able to remove outliers to infer properly. To infer section-level predictions at video-level, we calculated the maximum of each score over the sections belonging to a video. We assessed the performance with the metrics described below.

We considered the Mean Absolute Error (MAE), a well-known measure of accuracy for regression problems. It is defined as the absolute difference between the model inferences and the ground truth for continuous results. The bias was then used to evaluate the direction of the performance error. Those metrics were assessed mainly to evaluate and compare performance of our different regression model iterations.

Although we formulated the problem as a regression task, we proceeded to analyze results as a classification problem because scoring conventions are on discrete MES and UCEIS values. To do so, continuous scores were rounded in order to be appropriately compared to the ground truth. The Quadratic Weighted Kappa (QWK) was identified as the primary evaluation index as it is particularly suited for classification tasks. This variant of the Weighted Kappa is the reference statistic to find the degree of agreement between two raters, humans or not, thus measuring inter-observer variability, especially with ordinal scale items. This metric strongly penalizes large errors by putting a bigger weight on such errors compared to small errors (i.e., when the predictions are closer to the ground truth). We also used multiple typical classification metrics such as area under the ROC (AUROC), accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Finally, we considered two binary classification tasks: the first one compared MES 0-1 against MES 2-3, and the second task considered UCEIS ≤ 3 against UCEIS >3 to compare with existing results. To evaluate the quality of the GUI, qualitative feedback was collected from four clinical specialists. Statistical analysis on the amount of time spent reviewing videos and their respective sections was performed.

Results

Data Summary

The dataset used in this work contained 134 videos, accounting for 1,550,030 frames. We describe in Table 1 the breakdown of each step and the resulting associated data.

The final dataset, obtained after the fifth step and used to train the SDA model, is partitioned as followed: 126,320 (33%), 33,742 (9%), 84,937 (22%), and 141,433 (36%) for MES score of 0, 1, 2, and 3 respectively; 159,780 (41%), 84,411 (22%), 81,986 (21%), and 60,255 (16%) for Erosions and Ulcers of 0, 1, 2, and 3 respectively; 126,878 (33%), 71,098 (18%), and 188,456 (49%) for

Vascular Pattern of 0, 1, and 2 respectively; 151,993 (40%), 190,624 (49%), 38,349 (10%), and 5,466 (1%) for Bleeding of 0, 1, 2, and 3 respectively (Figure 4).

Model performance compared to expert labels

We provide in Table 2 the performance of the model according to the MAE and bias metrics. Overall, the model produced good performances at both section-level and video-level. The MAE and Bias were relatively low considering the magnitude of the scoring scale, especially for the UCEIS. In fact, at both section-level and video-level, for the MES and UCEIS individual subscores, SDA's predictions were equal or less than a half point away from the true value. The model's predictions for UCEIS are less than a point away from the ground truth.

While the MAE is an appropriate metric to evaluate and compare regression models, the QWK metric is more suitable for classification tasks and to compare our best model to the GI experts' labels. According to the scientific literature, a QWK between 0.61 and 0.8 is considered as substantial, while a QKW above 0.80 is stated as almost perfect agreement. Table 3 demonstrates the interobserver agreement between expert endoscopists and the SDA model at section and at video level using the QWK metric. The model's predictions at section-level were excellent, with a QWK over 0.8, except for the Bleeding descriptor. At video-level, the model's performance was good with a QWK over 0.6 except for the Bleeding descriptor.

Model results were also presented at severity-level for both MES (Supplementary Table 1) and UCEIS (Supplementary Table 2) using classification metrics that included specificity, sensitivity, NPV, PPV, and area under the curve (AUC). The best MES model's performance was for severity-level 0 and 3 with specificity of 94.60% and 87.90%, respectively; sensitivity of 85.71% and 69.14%, respectively; NPV of 92.00% and 87.70%, respectively; and PPV of 90.14% and 69.54%, respectively.

Confusion matrices for MES and UCEIS at section-level are shown in Supplementary Figure 1 and 2. The accuracies were 69.00% and 54.80% for MES and UCEIS, respectively. Additionally, we computed the accuracy at +/- 1 severity-level for UCEIS to determine the degree of disagreement between the model's prediction and the ground truth. The +/- 1 accuracy was 87.4, meaning a low error amplitude.

As described in the Performance Metrics subsection, we also considered two binary classification tasks: MES 0-1 versus MES 2-3, and UCEIS versus UCEIS > 3. The results are presented in Supplementary Table 3 and Supplementary Table 4 for both tasks respectively.

GUI evaluation

In a preliminary review of the system by four clinical experts, the user experience was overwhelmingly positive. Many advantages were expressed regarding the utilization of our tool to assess UC severity from endoscopic videos. Firstly, it allowed the user to focus on relevant sections instead of the whole video, reducing by a third the number of frames that a user needs to review. Indeed, out of the 1,550,030 total number of frames, 386,432 were kept in the created sections reviewed by the GI specialists. It also improved the quality and consistency of scores since users are reviewing all the same segments, allowing for the multiple labels to be used iteratively to re-train and improve the underlying DL models.

In this work, the developed GUI was successfully used by GI specialists in order to perform video quality assessment and section scoring to assess UC disease severity. On average, a section review took 26 seconds, with an average total video review time of 8 minutes. These numbers were obtained when the specialists reviewed all the sections of the videos for the purpose of this work. In research and clinical applications, however, the tool allows a much faster, more streamlined scoring process by expecting that clinical readers evaluate only pertinent sections of a video to confirm the scores assigned to each section.

A short video demonstration of our GUI and AI model in action is attached (video 1)

Discussion

The endoscopic scoring of UC disease activity with MES and UCEIS has been traditionally challenging due to the lack of clinical validation, and also as a result of disagreement on repeated observations. Central readers who are clinically blinded expert endoscopists have been utilized in UC clinical trials in an attempt to standardize the endoscopic assessment of UC[22]. Central reader validation in the endoscopic scoring of UC is inherently limited due to the lack of a true gold standard (i.e., biopsy), and therefore any measures of accuracy may be impacted by the

quality of central readers, algorithm performance, or inherent problems with the MES and UCEIS scoring system. Previous studies have reported the application of DL for the analysis of large endoscopy image datasets in order to improve and standardize UC disease severity grading. Ozawa et al developed a DL based model based on a GoogLeNet architecture to identify MES 0 and mucosal healing (score 0–1) in an independent test set of 3981 images from 114 UC patients, with a reported Area Under the Curve (AUC) of 0.86 and 0.98, respectively[16]. Stidham et al focused on the binary classification task of MES 0-1 and MES 2-3, and reported a sensitivity, specificity, PPV, NPV, accuracy, and QWK of 93%, 87%, 84%, 94%, 90%, and 79%, respectively[18]. We also compared the MES 0-1 scores against MES 2-3 scores and reported better results for all the aforementioned values including sensitivity, specificity, PPV, NPV, accuracy, and QWK of 96%, 91%, 91%, 96%, 94%, and 87%, respectively. Takenaka et al developed a CNN model to differentiate remission (UCEIS 3) from moderate-severe disease (UCEIS 3), and reported excellent reproducibility for their model with sensitivity, specificity, PPV, NPV, and AUROC of 83%, 96%, 87%, 94%, and 0.966, respectively[17]. Our results on a similar task are 93%, 93%, 92%, 94%, 0.936, respectively. Yao et al evaluated their CNN based video analysis model on 264 videos and reported 83.7% accuracy for differentiating remission from active disease, with an AUC of 0.93, average F1 score of 0.77, and a positive level of agreement with gastroenterologist scoring (kappa = 0.84)[19]. Gottlieb et al performed a randomized controlled trial to evaluate their CNN model in the assessment of mucosal inflammation according to MES and UCEIS. Their model's overall performance on the primary objective metric showed almost perfect agreement, with QWK of 0.84 for MES and 0.85 for UCEIS, respectively[20]. We have slightly better results on those same metrics, but at section-level as shown in Table 3 (0.886 for MES and 0.904 for UCEIS, respectively). Their model's accuracy at score-level results on video for MES is 70.2% compared to 69.0% for our work, and on UCEIS their accuracy is 45.5% compared to 54.8% for our work.

While previous AI work in the field can score UC activity at the frame or video level for either UCEIS or MES, we present the first fully automated DL model for scoring disease activity under both the MES and UCEIS scores, at frame, section, and video levels, and with an architecture that can accommodate other scoring systems such as Paddington International virtual Chromoendoscopy Score (PiCaSSO)[23]. We are also the first to describe an AI model that is ready

for clinical evaluation both in terms of robustness but also in relation to usability and fitting in with current workflow. This has been a big challenge for AI tools in endoscopy, namely that they do not hinder the physician, but rather add true assistance and benefit. We have dedicated some of our efforts in building this solution with the practical usability of said tool very much at the forefront of our thinking. Our system improves the accuracy of the MES and UCEIS scores, reduces the time between video collection and review, and improves subsequent quality assurance and scoring. Overall, our model performed well, as MAE and mean Bias at both section-level and video-level were relatively close to the ground truth considering the magnitude of the scoring scale, especially for the UCEIS. In our investigation, the QWK was used to compare the interobserver agreement between central readers' labels and the AI model's predictions. The results were excellent at section-level for both MES and UCEIS, with QWK of 0.886 and 0.904, respectively.

Our study has several limitations. One of the limitations of our study is a limited dataset with imbalanced classes. The generalizability of our model is also limited. However, to improve it, we have started validating our results by testing and training on different datasets with various endoscopy sites, equipment, and recording techniques. Another limitation is the great difficulty in describing a fair comparison with the literature due to the lack of an open dataset. Moreover, the ground truth for MES and UCEIS is subjective, and there is a potential bias for the labelers when reviewing AI-generated sections with our GUI. In addition, the model we describe in this study was designed to work on pre-recorded videos of UC patients. This is therefore an offline tool. However, our results show that the models described can infer quickly enough to easily create a real time application, which we will reveal in the near future.

Overall, our results enable the development of a model that can be used to improve the efficiency and accuracy of endoscopic assessment and scoring of UC at different stages of the clinical journey, whether offline or live. It can be used by physicians at site level for video quality assurance and also by central reading organizations and the pharmaceutical industry to score videos and increase the efficiency of central reading in clinical trials. It will also be usable as a tool for evaluation during live endoscopy where it could serve as an accurate reproducible

measurement of endoscopic disease activity. Finally, there is an opportunity for education at the level of the GI trainees to set up training modules.

Conclusions

In summary, we report a fully automated DL model that improves the accuracy of the MES and UCEIS scores, reduces the time between video collection and review, and improves subsequent quality assurance and scoring. Our model demonstrated relevant feature identification for scoring of disease activity in UC, well aligned with scoring guidelines and performance of the experts. We present work that builds a frame level regression scoring system paired with a clustering algorithm and video level heuristics that scores simultaneously under both scoring modalities. Going forward, we aim to continue developing our detection and scoring systems in order to produce a system that can score at a superhuman level and with greater precision than current scoring modalities. More data in terms of volume and diversity is being collected and analysed to drive towards a final product ready for clinical use. We also are doing more formal evaluation of the usability of the graphic user interface described in this study so that we can have a tool that is truly one that will offer timesaving and better user satisfaction.

Funding

James East and Simon Travis are funded by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre. The views expressed are those of the authors and not necessarily those of the National Health Service, the NIHR or the Department of Health.

References

1. Dignass A, Eliakim R, Magro F, et al. Second European evidence-based consensus on the diagnosis and management of ulcerative colitis Part 1: definitions and diagnosis. J Crohn Colitis 2012;6:965–990.
2. Singh S, Fumery M, Sandborn WJ. Systematic review with network meta-analysis: first-and second-line pharmacotherapy for moderate-severe ulcerative colitis. Aliment Pharmacol Ther 2018;47:162–715.
3. Lewis JD, Chuai S, Nessel L, et al. Use of the noninvasive components of the Mayo score to assess clinical response in ulcerative colitis. Inflamm Bowel Dis. 2008;14:1660-1666.
4. Jones J, Loftus EV Jr, Panaccione R, et al. Relationships between disease activity and serum and fecal biomarkers in patients with Crohn's disease. Clin Gastroenterol Hepatol. 2008; 6:1218-1224.
5. Schoepfer AM, Beglinger C, Straumann A, et al. Fecal calprotectin more accurately reflects endoscopic activity of ulcerative colitis than the Lichtiger Index, C-reactive protein, platelets, hemoglobin, and blood leukocytes. Inflamm Bowel Dis. 2013;19:332-341.
6. Xie T, Zhang T, Ding C, et al. Ulcerative Colitis Endoscopic Index of Severity (UCEIS) vs Mayo Endoscopic Score (MES) in guiding the need for colectomy in patients with acute severe colitis. Gastroenterol Rep (Oxf). 2018;6:38-44.
7. Novak G, Parker CE, Pai RK, et al. Histologic scoring indices for evaluation of disease activity in Crohn's disease. Cochrane Database Syst Rev. 2017;7:CD012351.
8. Neurath MF, Travis SP. Mucosal healing in inflammatory bowel diseases: a systematic review. Gut 2012; 61: pp. 1619-1635.
9. Schroeder KW, Tremaine WJ, Ilstrup DM. Coated oral 5- aminosalicylic acid therapy for mildly to moderately active ulcerative colitis. N Engl J Med 1987;317:1625– 1629.
10. Travis SPL, Schnell D, Krzeski P, et al. Developing an instrument to assess the endoscopic severity of ulcerative colitis: The Ulcerative Colitis Endoscopic Index of Severity of ulcerative colitis: the Ulcerative Colitis Endoscopic Index of Severity (UCEIS). Gut 2012; 61:535–542.
11. Bounuen G, Levesque BG, Pola S, et al. Feasibility of endoscopic assessment and treating to target to achieve mucosal healing in ulcerative colitis. Inflamm Bowel Dis 2014;20:231-239.
12. Mazzuoli S, Guglielmi FW, Antonelli E, et al. Definition and evaluation of mucosal healing in clinical practice. Dig Liv Dis 2013;45: 969-977.
13. Reinink AR, Lee TC, Higgins PD. Endoscopic mucosal healing predicts favorable clinical outcomes in inflammatory bowel disease: a meta analysis. Inflamm Bowel Dis 2016;22:1859-1869.
14. 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 Crohns Colitis. 2016 Mar;10(3):286-295.
15. Gottlieb K, Daperno M, Usiskin K, et al. Endoscopy and central reading in inflammatory bowel disease clinical trials: achievements, challenges, and future developments [published online ahead of print July 22, 2020]. Gut doi:10.1136/gutjnl-2020-320690.
16. Ozawa T, Ishihara S, Fujishiro M, et al. Novel computer-assisted diagnosis system for endoscopic disease activity in patients with ulcerative colitis. Gastrointest Endosc. 2019 Feb;89(2):416-421.

17. Takenaka K, Ohtsuka K, Fujii T, et al. Development and Validation of a Deep Neural Network for Accurate Evaluation of Endoscopic Images From Patients With Ulcerative Colitis. Gastroenterology. 2020 Jun;158(8):2150-2157.
18. 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 NetwOpen 2019; 2: e193963.
19. Yao H, Najarian K, Gryak J, et al. Fully automated endoscopic disease activity assessment in ulcerative colitis. Gastrointest Endosc. 2021 Mar;93(3):728-736.e1.
20. Gottlieb K, Requa J, Karnes W, et al. Central Reading of Ulcerative Colitis Clinical Trial Videos Using Neural Networks. Gastroenterology. 2021 Feb;160(3):710-719.e2.
21. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977 Mar;33(1):159-74. PMID: 843571.
22. Panés J, Feagan BG, Hussain F, et al. Central Endoscopy Reading in Inflammatory Bowel Diseases. J Crohns Colitis. 2016 Sep;10 Suppl 2:S542-7.
23. Iacucci M, Smith SCL, Bazarova A, et al. An International Multicenter Real-Life Prospective Study of Electronic Chromoendoscopy Score PICaSSO in Ulcerative Colitis. Gastroenterology. 2021 Apr;160(5):1558-1569.e8. doi: 10.1053/j.gastro.2020.12.024. Epub 2021 Feb 6. PMID: 33347880.

Tables

Table 1. Quantity of data after each step of the workflow

Step	Sub-step	Data
Video Quality Assessment		- 134 high quality videos - 1,550,030 frames.
Preprocessing Pipeline Application	Blue light identifier	- 1,176,441 white-light frames
	Scorability assignment model	- 582,448 scorable frames - 593,993 non-scorable.
	Biopsy procedure and ex-vivo detector	- 22,543 biopsy procedure frames - 66,910 ex-vivo frames
	Frame-based Disease Severity assessment model	- 582,448 scorable frames predicted
Section Generation		- 2630 scorable sections - 386,432 scorable frames*
Graphical User Interface Leverage		
Disease Severity Assessment (SDA)		- 2630 reviewed sections - 386,432 scorable frames

** Not all scorable frames were used to create sections*

Table 2. MAE and Bias measures of the SDA model

	Section-level		Video-level	
	MAE	Bias	MAE	Bias
Mayo Endoscopic Subscore	0.32	0.05	0.19	0.19
UCEIS	0.65	0.07	0.94	0.44
Erosions and Ulcers	0.36	0.10	0.50	0.12
Vascular Pattern	0.20	-0.01	0.06	0.06
Bleeding	0.24	0.01	0.44	0.06

MES Scale (0-3); UCEIS Scale (0-8); Erosions and Ulcers (0-3), Vascular Pattern Scale (0-2); Bleeding Scale (0-3).Note. We do not provide the results at frame-level because the ground truth is the score at section-level, projected down at frame-level to train the model. There is hence inherent error in the truth at frame-level that would be included in the performance results.

Table 3. The results of inter-observer agreement QWK at section and video-level.

	Section-level	Video-level
Mayo Endoscopic Subscore	0.886	0.821
UCEIS	0.904	0.646
Erosions and Ulcers	0.800	0.600
Vascular Pattern	0.905	0.879
Bleeding	0.754	0.391

Figures

Figure 1. Our fully automated UC detection and scoring decision support methodology.

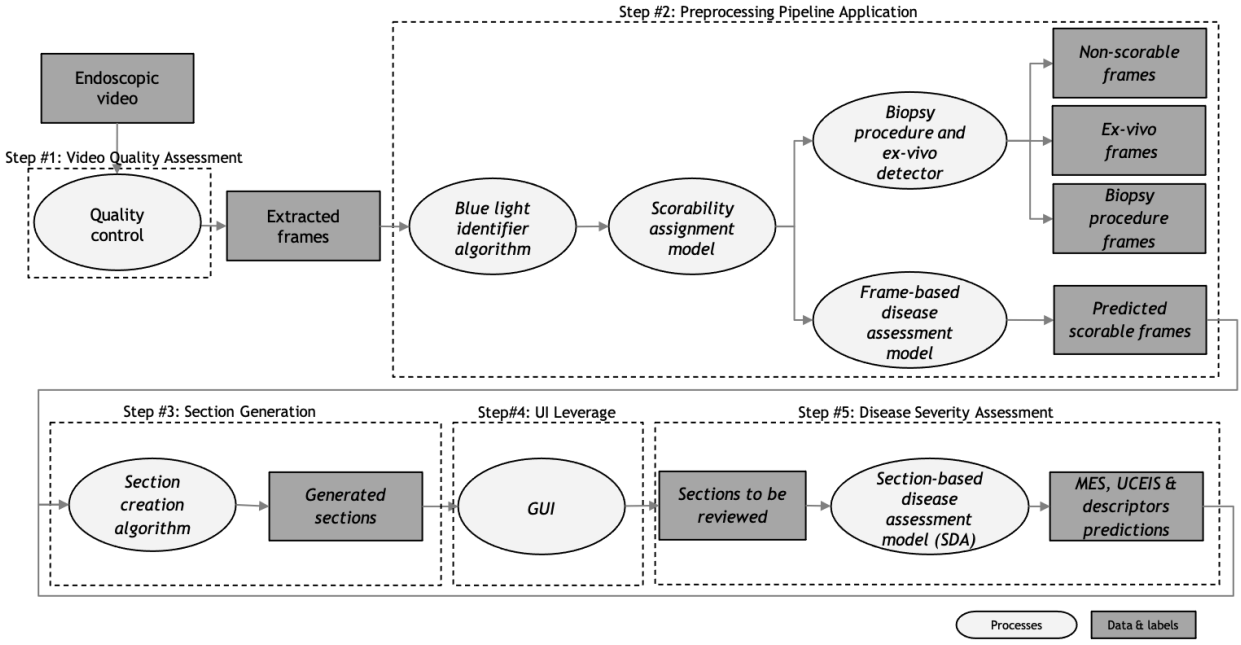


Figure 2. Overview of the GUI.

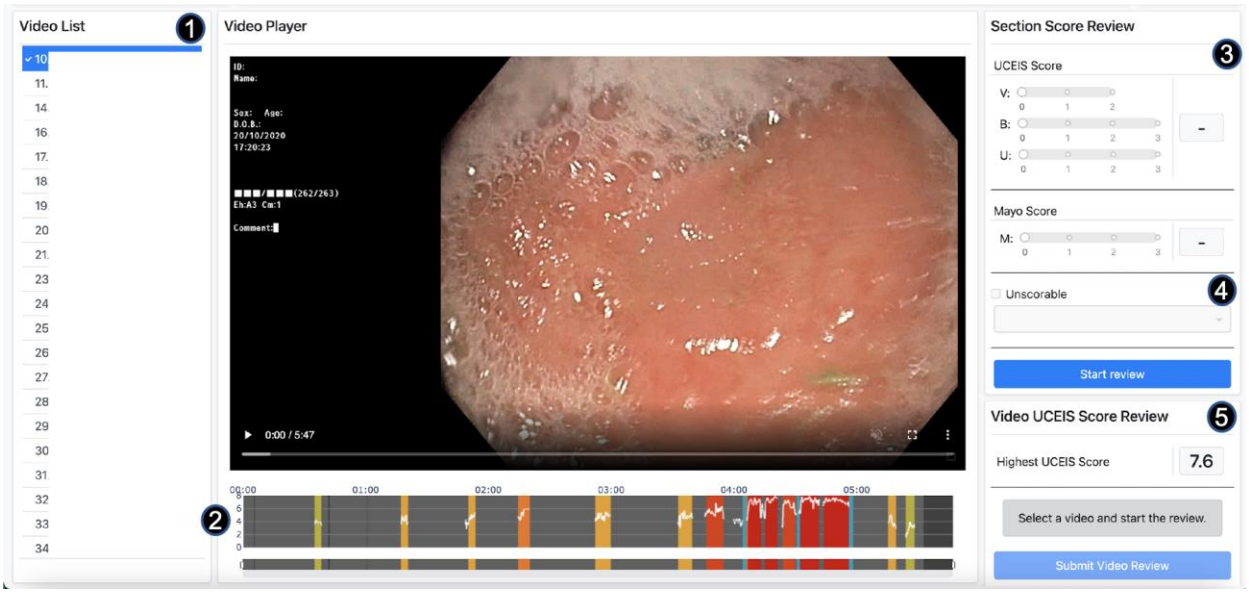


Figure 3. High-level architecture of the SDA CNN model predicting MES, UCEIS, and its descriptors.

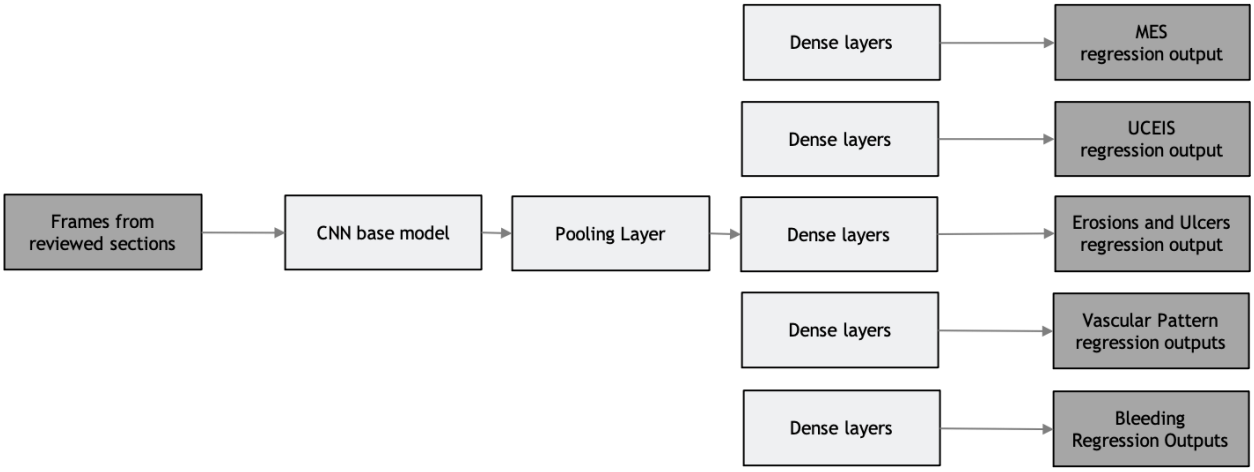
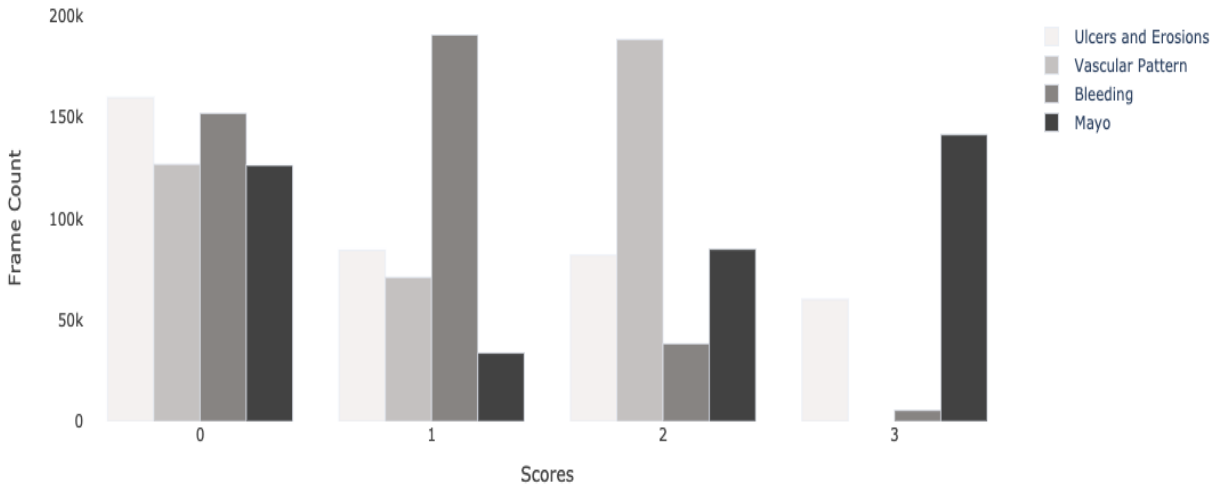


Figure 4. Distribution of MES and UCEIS descriptors for each frame within reviewed sections.



Supplementary Material:

Figure 1. MES Confusion Matrix at section-level

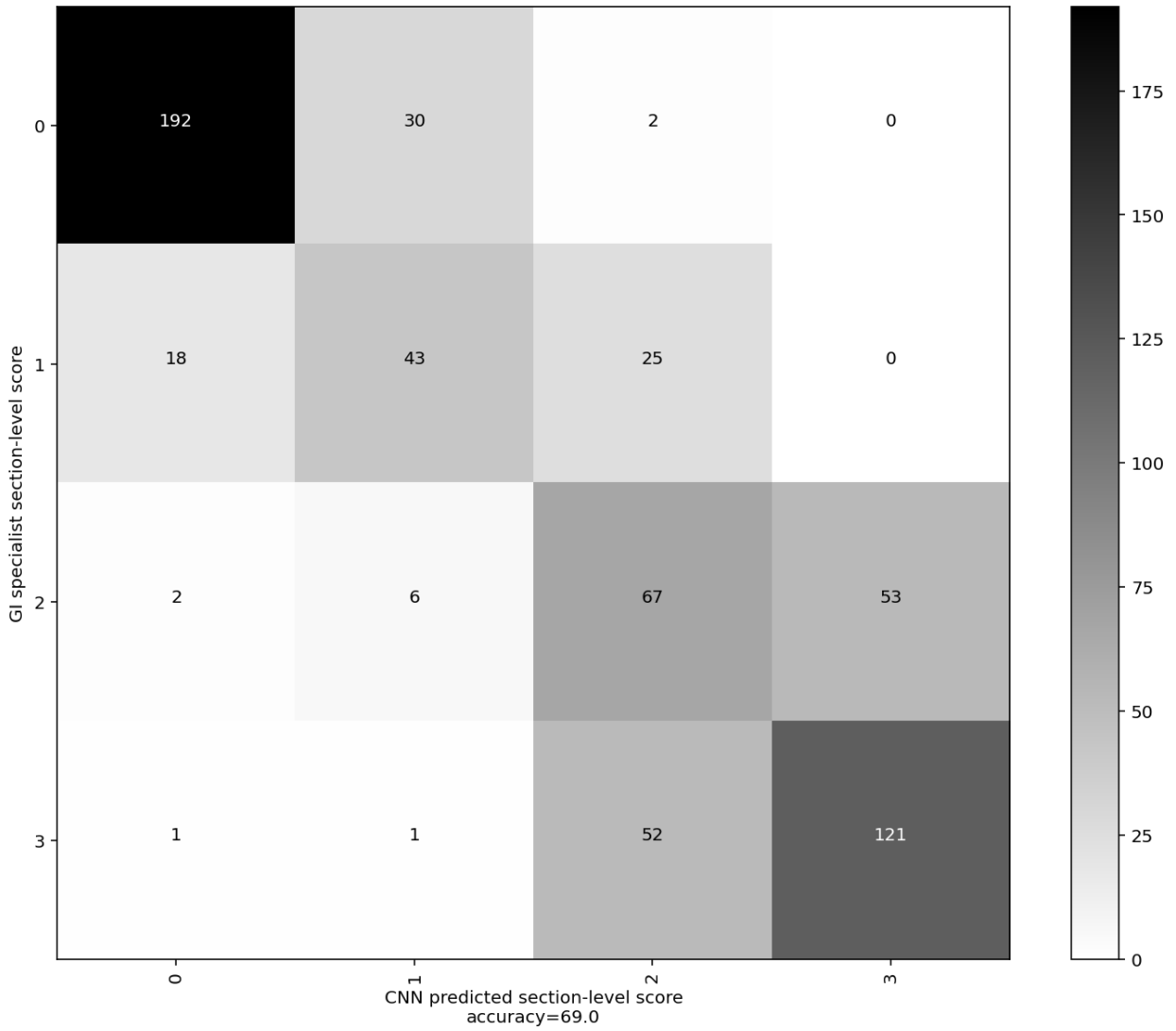


Figure 2. UCEIS Confusion Matrix at section-level

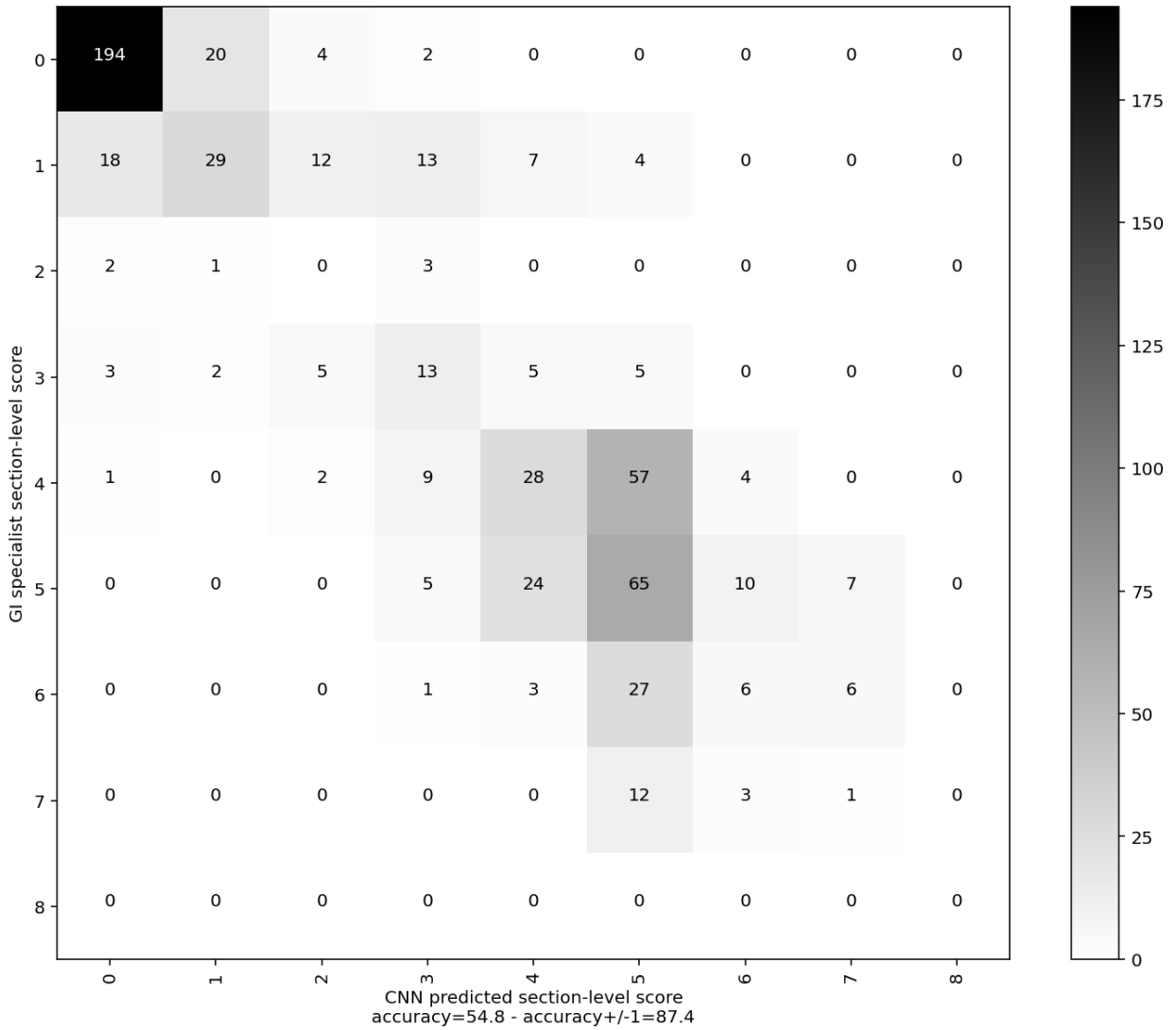


Table 1. MES severity-level results at section-level

Metric/MES	0	1	2	3
Specificity (%)	94.60	92.98	83.71	87.90
Sensitivity (%)	85.71	50.00	54.34	69.14
NPV (%)	92.00	91.93	86.94	87.70
PPV (%)	90.14	53.75	45.89	69.54
AUC	0.902	0.715	0.680	0.785

Table 2. UCEIS severity-level results at section-level

Metric/UCEIS	0	1	2	3	4	5	6	7
Specificity (%)	93.89	95.66	96.21	94.31	92.38	79.08	97.02	97.82
Sensitivity (%)	88.18	34.94	0.00	39.39	27.72	58.56	13.95	6.25
NPV (%)	93.42	90.37	98.98	94.47	86.63	89.62	93.73	97.05
PPV (%)	89.99	55.77	0.00	28.26	41.79	38.24	26.09	7.14
AUC	0.910	0.653	0.481	0.669	0.601	0.688	0.555	0.520

Note. It is possible to have a 0% value for Sensitivity or PPV in case there was no true positives (TPs)

Table 3. Binary classification task #1 - MES 0-1 versus MES 2-3

Metric	Results
Accuracy (%)	94.00%
Specificity (%)	91.29%
Sensitivity (%)	96.70%
NPV (%)	96.59%
PPV (%)	91.56%
AUC	0.941
QWK (%)	87.93%

Table 4. Binary classification task #1 - UCEIS ≤ 3 versus UCEIS >3

Metric	Results
Accuracy (%)	94.00%
Specificity (%)	93.86%
Sensitivity (%)	93.36%
NPV (%)	94.69%
PPV (%)	92.33%
AUC	0.936
QWK (%)	87.12%