Inflammatory Bowel Diseases, 2023, **29**, 1431–1439
https://doi.org/10.1093/ibd/izac226
Advance access publication 16 November 2022
Original Research Articles - Clinical

Improving the Computer-Aided Estimation of Ulcerative Colitis Severity According to Mayo Endoscopic Score by Using Regression-Based Deep Learning

Gorkem Polat, MSc,[*,†,a] Haluk Tarik Kani, MD,[‡,a] Ilkay Ergenc, MD,[‡,a] Yesim Ozen Alahdab, MD,[‡] Alptekin Temizel, PhD,[*,†,b] and Ozlen Atug, MD[‡,b]

From the [*]Graduate School of Informatics, Middle East Technical University, Ankara, Turkey;
[†]Neuroscience and Neurotechnology Center of Excellence, Middle East Technical University, Ankara, Turkey; and
[‡]Department of Gastroenterology, School of Medicine, Marmara University, Istanbul, Turkey.

[a]These authors contributed equally as co-first authors.
[b]These authors contributed equally as co-senior authors.
Address correspondence to: Haluk Tarik Kani, MD, Marmara Üniversitesi Tıp Fakültesi, Başıbüyük Yolu No: 9 D:2, 34854 Maltepe, İstanbul, Turkey (drhtkani@gmail.com).

Abstract

Background: Assessment of endoscopic activity in ulcerative colitis (UC) is important for treatment decisions and monitoring disease progress. However, substantial inter- and intraobserver variability in grading impairs the assessment. Our aim was to develop a computer-aided diagnosis system using deep learning to reduce subjectivity and improve the reliability of the assessment.

Methods: The cohort comprises 11 276 images from 564 patients who underwent colonoscopy for UC. We propose a regression-based deep learning approach for the endoscopic evaluation of UC according to the Mayo endoscopic score (MES). Five state-of-the-art convolutional neural network (CNN) architectures were used for the performance measurements and comparisons. Ten-fold cross-validation was used to train the models and objectively benchmark them. Model performances were assessed using quadratic weighted kappa and macro F1 scores for full Mayo score classification and kappa statistics and F1 score for remission classification.

Results: Five classification-based CNNs used in the study were in excellent agreement with the expert annotations for all Mayo subscores and remission classification according to the kappa statistics. When the proposed regression-based approach was used, (1) the performance of most of the models statistically significantly increased and (2) the same model trained on different cross-validation folds produced more robust results on the test set in terms of deviation between different folds.

Conclusions: Comprehensive experimental evaluations show that commonly used classification-based CNN architectures have successful performance in evaluating endoscopic disease activity of UC. Integration of domain knowledge into these architectures further increases performance and robustness, accelerating their translation into clinical use.

Key Words: colonoscopy, computer-assisted diagnosis, deep learning, inflammatory bowel diseases, Mayo score, ulcerative colitis

INTRODUCTION

Ulcerative colitis (UC) is a chronic, idiopathic, and remitting disorder that is characterized by persistent inflammation of the colon mucosa. The goal of medical therapy is to induce and maintain clinical remission. Assessing the disease severity plays a key role in patient management, and the Mayo score is the most common score that is used to assess the disease severity of UC.[1] The Mayo endoscopic score (MES) reflects the mucosal involvement directly by endoscopic examination and provides an accurate severity assessment. However, this evaluation depends on the endoscopist's own experience and education, making the assessment subjective. Previous studies have shown substantial intra- and interobserver differences in the grading of endoscopic severity with the level of experience.[2,3] Therefore, reproducibility and reliability remain as 2 main problems in grading the endoscopic severity of UC. Computer-assisted diagnosis using advanced artificial intelligence (AI) algorithms can be used to help endoscopists and alleviate these issues.

Deep learning (DL) has been proven successful in many tasks in recent years with the emergence of big datasets, availability of computation power, and advancements in artificial neural network algorithms and shows a great promise for health applications.[4] In recent studies, DL showed significant success in polyp and artifact detection, *Helicobacter pylori* infection diagnosis, gastrointestinal cancer analysis, and hemorrhage segmentation.[5-11] CNN is a commonly used class of DL methods in image-based tasks. In a supervised learning setting, a CNN model requires the annotations along with the input image and makes use of the relationships between input images and corresponding output labels to learn effective representations during model training.

Received for publication: June 6, 2022. Editorial Decision: October 2, 2022

Key Messages

- What is already known? The Mayo endoscopic score reflects the mucosal involvement directly by endoscopic examination and provides an accurate severity assessment. However, this evaluation depends on the endoscopist's own experience and education and it makes the assessment subjective.
- What is new here? We reviewed 19 537 endoscopic images from 1043 colonoscopy procedures of 572 ulcerative colitis patients and developed a new computer-aided estimation system for the assessment of endoscopic activity to classify ulcerative colitis severity with an quadratic weighted kappa score of 0.854 (0.842-0.867).
- How can this study help patient care? This approach has potential to decrease the interobserver variability and increase the accuracy of diagnosis, standardization of follow-up, and evaluation of the disease activity.

Evaluation of disease severity, especially the MES, plays a key role in UC follow-up; therefore, a reliable computer-assisted diagnosis can be very helpful by reducing subjectivity and improving the standardization of the process. In this study, our aim was to create a DL algorithm with a high classification performance and reliability to estimate the MES accurately from endoscopic images of UC patients. In line with this aim, there are 3 main contributions of this work. First, a new dataset has been collected and annotated to train AI algorithms and it has been made publicly available. The released dataset, Labeled Images for Ulcerative Colitis (LIMUC), is the largest publicly available labeled dataset for the UC. Moreover, we have made our code implementations publicly available for transparency and reproducibility and to facilitate future studies in the field. Second, we conducted a comprehensive baseline analysis by experimentally evaluating the existing approaches in the literature on the LIMUC dataset. We provide an objective comparison of their performances as a future reference. Third, we have proposed a regression-based approach that incorporates the domain knowledge. Experimental results show that the proposed regression-based framework compares favorably to the previously proposed classification-based approaches in the literature.

METHODS

Study cohort and labeling

A total of 19 537 endoscopic images from 1043 colonoscopy procedures of 564 UC patients who underwent colonoscopy at Marmara University Institute of Gastroenterology between December 2011 and July 2019 were collected. All images were acquired by a Pentax EPK-i video processor and Pentax EC-380LKp video colonoscope (Pentax, Tokyo, Japan) and downsized to a resolution of 352 × 288 when recorded to the database. The images were selected at different times by the operator during the colonoscopy; therefore, there is no spatial relationship among the images of the same patient in the cohort, which increases the heterogeneity, resulting in a more diverse dataset. The study design and all information obtained from the electronic health records were approved by the Marmara University School of Medicine ethical review board (Study Protocol No: 09.2020.627, Approval date: 12.06.2020). Characteristics of the study cohort are given in Table 1.

The images that were not suitable for evaluation due to debris, inadequate bowel preparation, artifacts, retroflexion, and poor image quality were excluded from the study. All patient information, software outputs, and date and time information on the images were masked to prevent any bias. The 2 experienced gastroenterologists blindly reviewed and classified all images according to the MES and consistently labeled images were included in the dataset. The interreader reliability for MES labeling was measured with the quadratic weighted kappa (QWK) and obtained as 0.781. A total of 7652 images were labeled differently by 2 reviewers. A third reviewer, who did not have any access to the previous labels, independently labeled these differently labeled images. Final scores of which were determined using majority voting. Images that were labeled differently by all 3 reviewers were excluded from the study (Figure 1). More detailed statistics and a breakdown of the annotation process are provided in the Supplementary Appendix. The reviewers evaluated the masked images without any knowledge of their clinical information.

A total of 8060 images were found to be unsuitable for the assessment for MES evaluation and 201 images were excluded, as they were labeled differently by each of the 3 reviewers. As a result, the final dataset contained a total of 11 276 images with the following distribution: MES 0, 6105 (54.14%); MES 1, 3052 (27.07%); MES 2, 1254 (11.12%); and MES 3, 865 (7.67%). The resulting dataset, LIMUC,[12] has been made publicly available, to the best of our knowledge, making it the largest publicly available labeled UC images dataset. The LIMUC dataset can be used to develop better algorithms and generate out-of-sample sets to evaluate their performances. We report the results of commonly used CNN architectures on the LIMUC to demonstrate baseline performances and the quality of the dataset.

Model development and evaluation

We randomly split the 15% of images (1686 images from 85 patients) as the test set by keeping the class (ie, MES) ratios the same as the study group. The splitting was done at the patient level (ie, assigning all images of a patient either to the test set or to the model development set). The rest, 85% of the images (9590 images from 479 patients), have been used to perform 10-fold cross-validation. For each fold, the training and validation split was performed at the patient level, randomly, and preserving the class ratios, as for the test set. CNN

Table 1. Characteristics of the cohort (n = 561)

Male/female	317 (56.5)/244 (43.5)
Age, y	43.3 ± 13.7
Male	44.3 ± 13.7
Female	42.1 ± 13.7
Colonoscopies per patient	2.0 ± 1.2
Male	2.0 ± 1.2
Female	1.9 ± 1.2

Values are n (%) or mean ± SD. Three patients had no information.

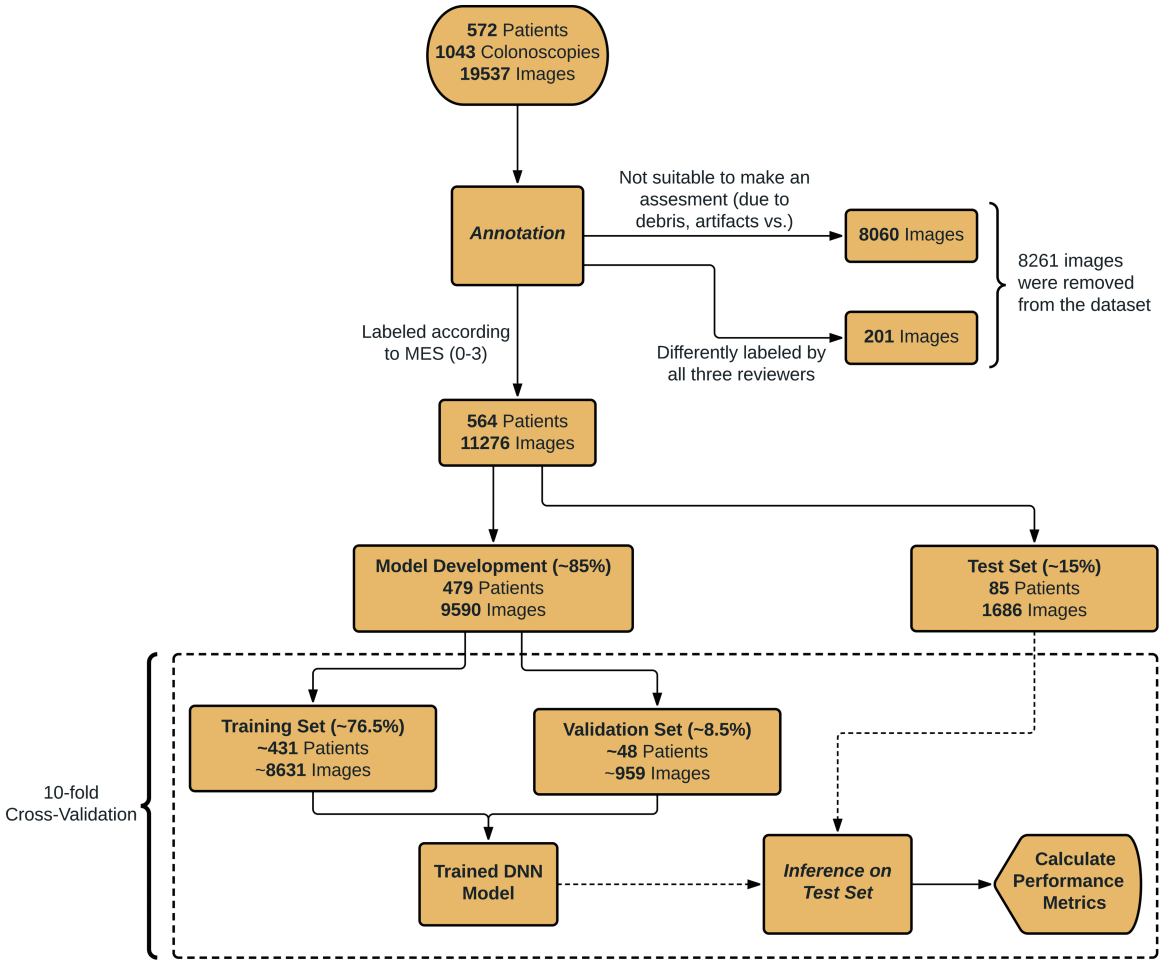

Figure 1. Overall design of the study. DNN, deep neural network; MES, Mayo endoscopic score.

architectures trained using different cross-validation folds were evaluated on a separately held-out test set (see Figure 1). Reported performance metrics are the mean values of 10-fold cross-validation results.

Model performances were evaluated on 2 baselines: full 4-level Mayo score classification and remission state classification, in which there are 2 levels: remission Mayo 0 or 1 and Mayo 2 or 3. CNNs were only trained for full Mayo scores, and model predictions were converted to remission states for remission classification. Model predictions were compared with the ground-truth labels provided by the human experts for the performance measurements. Because there are class imbalances and there is an ordinal relation between them, the QWK score was determined as the main performance metric along with the macro F1 score for the classification of all Mayo subscores. The QWK is one of the commonly used statistics for the assessment of agreement on an ordinal scale, and it is one of the most suitable singular performance metrics for this problem regarding the class imbalances. For the classification of remission, Cohen's kappa and F1 scores are reported. Because it is intended to compare different models, kappa and F1 scores are better singular metrics than the sensitivity, specificity, precision, and recall, which are highly affected by either false positives or false negatives. Nevertheless, accuracy, sensitivity, and specificity values are given for reference in the Supplementary Appendix.

Regression-based approach

In classification-based approaches, the CNN model assigns a probability for each MES. The model's prediction was determined by the score with the highest likelihood. Cross-entropy loss functions used to train CNN architectures only evaluate the probability value of the ground-truth MES (ie, a lower ground-truth probability leads to more punishment and vice versa). However, because there is a ranking (or order) relation among the Mayo scores, we claim that the probability values of the other classes should also be taken into account. For example, consider a hypothetical case in which the ground-truth score is Mayo 3, and 2 different model predictions give the same probability for Mayo 3. Assuming the first prediction gives a higher probability to Mayo 0, while the second prediction gives a higher probability to Mayo 2, it can be said that the first one is a worse prediction because Mayo 0 is more distant to Mayo 3, which is the ground truth. On the other hand, classification-based approaches do not take this into account and result in same punishment level for the system. Therefore, we transformed the classification problem into a regression problem in which the model only predicts a single continuous value. In the regression-based approach, a single node without any activation function at the output layer is employed to predict a continuous value. Then, the predicted value is compared with the ground truth value using the mean squared error. As the predicted value moves away from the

true value, the margin of error increases. Because the model provides a continuous numeric value, the predicted Mayo score is then determined using the thresholds calculated as the average of neighbor scores (eg, for Mayo 2 score, the lower and upper thresholds are 1.5 and 2.5, respectively). Further details are provided in the Supplementary Appendix.

The proposed change at the output layer of the network was applied to commonly used CNN architectures such as ResNet18,[13] ResNet50,[13] DenseNet121,[14] Inception-v3,[15] and MobileNet-v3-larg,e[16] and it improved the performance for all models.

Training setup for DL models

During training, image augmentation was applied using horizontal flipping and random rotation between 0° and 360° to increase variance in the dataset and prevent overfitting. Original image sizes of 352 × 288 were used for all the networks except Inception-v3, which accepts 299 × 299 resolution images. Each image channel was normalized with the mean and SD obtained from the training set. Due to the class imbalances, minority classes were oversampled so that all classes had an equal number of samples in each batch during the training. Weights of all models were initialized using the weights of pretrained networks on the ImageNet[17] dataset. The Adam optimizer[18] with a learning rate of 2×10^{-4} was used and learning rate scheduling was applied by scaling the learning rate with a factor of 0.2 if there was no improvement in the accuracy for the last 10 epochs. If there was no improvement in the accuracy of the validation set for the last 25 epochs, training was terminated. The weights of the checkpoint that has the best accuracy on the validation set was selected as the best model for that fold, and it was used for inference on the test set. CNN models implemented using the PyTorch[19] framework in the torchvision package were used for the training. All scripts for preprocessing and training models are publicly available online (https://github.com/GorkemP/labeled-images-for-ulcerative-colitis).

RESULTS

Evaluation of baseline models

Several state-of-the-art classification-based CNN models have been trained on the dataset to obtain baseline results. Each CNN architecture underwent 10-fold cross-validation, resulting in 10 different trained models, each evaluated on the initially allocated test set. Average values of the cross-validation results are reported in Table 2. The 95% confidence intervals (CIs), which were calculated according to the Student *t* distribution, are given in parentheses next to the average scores. Both QWK and kappa scores were >0.8 for all models, indicating almost a perfect agreement between the ground truth and the model predictions. DenseNet121 had the highest score for all performance metrics except macro F1 score, for which Inception-v3 model had the highest score. Nevertheless, there was no statistically significant difference between the model performances, except for the worst-performing (ResNet18) and best-performing (DenseNet121) models, which had a *P* value of .049. Baseline results show that the employed CNN models can differentiate different MES classes reliably, and because they always make the same inference from the same images, they can be used as an objective approach during the colonoscopy procedure.

Regression-based models

The same CNN models used in baseline evaluations were modified as proposed to make them regression-based, and the same training and testing procedure using 10-fold cross-validation was applied. Table 3 shows the mean performance metrics and 95% CIs. When the MES prediction was handled as a regression problem, we observed a statistically significant improvement in performances of most of the CNN models evaluated in this study. In addition, CIs narrowed that point to the robustness of models in terms of performance metrics. The Wilcoxon rank sum test was used to compare classification-based and regression-based approaches to measure whether there was a statistically significant difference (see Table 4). Except for DenseNet121 for the classification of all Mayo subscores, the performance increases of all architectures were statistically significant. When all classification-based results and regression-based results were merged within themselves and statistical significance was measured between 2 groups (50 classification-based results vs 50 regression-based results) *P* values <.001 for all Mayo subscores and remission were obtained. Meanwhile, there was no statistically significant difference between the performances of different regression-based models. The regression-based approach improved the performance and produced more stable results. In Figure 2, the vertical axis shows the average performance of each model (obtained through 10-fold cross-validation), while the horizontal axis shows the SD. Ideally, we would like to use a model with high performance and low SD, which refers to the upper left region of the graph. When the regression-based approach was used to train the CNN models, we observed a shift toward the upper left region for all CNN models for

Table 2. Baseline performance metrics for the 5 state-of-the-art convolutional neural network models

Model	All Mayo Subscores		Remission	
	QWK (95% CI)	Macro F1 (95% CI)	Kappa (95% CI)	Macro F1 (95% CI)
ResNet18	0.830 (0.797-0.861)	0.672 (0.614-0.730)	0.808 (0.756-0.859)	0.842 (0.801-0.883)
ResNet50	0.836 (0.801-0.870)	0.672 (0.622-0.723)	0.814 (0.765-0.864)	0.848 (0.806-0.889)
DenseNet121	0.844 (0.814-0.874)[a]	0.681 (0.635-0.726)	0.827 (0.781-0.873)[a]	0.858 (0.820-0.895)[a]
Inception-v3	0.836 (0.812-0.860)	0.683 (0.631-0.735)[a]	0.807 (0.760-0.855)	0.842 (0.803-0.881)
MobileNet-v3-large	0.830 (0.798-0.862)	0.667 (0.605-0.729)	0.812 (0.771-0.853)	0.845 (0.809-0.881)

Reported metrics refer to the mean results of the different cross-validation folds.
Abbreviation: CI, confidence interval.
[a]Best value.

Table 3. Regression-based results for the 5 state-of-the-art convolutional neural network models

Model	All Mayo Subscores		Remission	
	QWK (95% CI)	Macro F1 (95% CI)	Kappa (95% CI)	Macro F1 (95% CI)
ResNet18	0.854 (0.839-0.869)[a]	0.693 (0.660-0.725)	0.841 (0.812-0.870)	0.869 (0.846-0.893)
ResNet50	0.849 (0.833-0.866)	0.693 (0.651-0.735)	0.852 (0.825-0.879)[a]	0.878 (0.855-0.901)[a]
DenseNet121	0.854 (0.842-0.867)[a]	0.697 (0.664-0.730)[a]	0.850 (0.817-0.883)	0.876 (0.848-0.904)
Inception-v3	0.852 (0.836-0.867)	0.688 (0.659-0.717)	0.840 (0.802-0.879)	0.869 (0.838-0.900)
MobileNet-v3-large	0.847 (0.835-0.859)	0.695 (0.670-0.719)	0.834 (0.807-0.861)	0.863 (0.842-0.885)

Reported metrics refer to the mean results of the different cross-validation folds. All model results have improved compared with their baseline performances. Moreover, CIs are in a narrower range, indicating that the results are more stable.
Abbreviation: CI, confidence interval.
[a]Best value.

Table 4. *P* values for QWK and kappa scores calculated using the Wilcoxon rank sum test

Model	QWK	Kappa
ResNet18	.0005	.0025
ResNet50	.0494	.0012
DenseNet121	.0589	.0156
Inception-v3	.0012	.0019
MobileNet-v3-large	.0052	.0082
Mean	.0230	.0059
All models combined	4.27×10^{-9}	1.03×10^{-10}

Except for the DenseNet121 QWK scores, there is a statistically significant difference for all models between the baseline and proposed approach. Result of classification and regression-based models when all model results are merged (50 classifications vs 50 regressions) is given as all models combined.
Abbreviation: QWK, quadratic weighted kappa.

both MES classification and remission classification. As a result, the regression-based method improved the performance and made the results more stable, which instills more trust in experts.

Discussion

Recent studies employing DL methods to estimate the endoscopic activity of UC from recorded still images used general-purpose state-of-the-art CNN models for the classification.[20-25] These studies treated MES classes as nominal categories; however, UC endoscopic activity severity progresses from mild to severe, which makes the prediction classes ordinal. Exploiting this domain knowledge and incorporating it into the architecture can positively affect classification performance. There is only 1 study that utilized this information in the architectural design of the CNN model.[26] However, their approach was not flexible for different datasets and requires tuning of additional internal parameters. We framed the grading task as a regression problem and performed extensive experiments using both types of approaches. The experiments showed that there was no major performance difference between different regular classification-based (Table 2) and between different regression-based (Table 3) CNN models. On the other hand, there was statistically significant improvement when regression-based models were compared against their classification-based counterparts using the same CNN model (Table 4). Moreover, regression-based models provided more reliable results in terms of robustness to dataset distributions, an important indicator for clinical translation.

Several stochastic factors in the learning process, such as weight initialization, order of inputs when feeding to the network, and distribution of samples in training and validation sets caused by splitting of the dataset, lead to fluctuating performance of a DL model.[27] Therefore, the reported performance on a test set is highly dependent on training settings. However, such ambiguity is unfavorable for the domain experts who would want to assess this technology for adoption in clinical environments.[28] Although a model's average performance may be high, clinicians may prefer a more reliable model that performs similarly when trained under different conditions. The cross-validation technique can be used to reduce the ambiguity and to measure the model performance more reliably. Models trained on each fold differ from one another, and the greater the deviation between model performances is, the poorer the model's reliability is. The proposed regression-based approach for the UC severity estimation problem improved model robustness, bringing this technology closer to actual implementation in clinics (Figure 2).

The first major contribution is the release of the largest publicly available dataset, LIMUC, which was also used in this study. Annotation of a medical dataset is tedious work that involves many experts and should be handled rigorously. However, the common practice of using private datasets prevents reproducibility and fair comparison of different approaches. The performance results of the machine learning algorithms heavily depend on the dataset on which they were trained, validated, and tested; therefore, the availability of a publicly available dataset would make it possible to evaluate with different methods on common grounds, understanding the current state of the technology and its applicability in clinical practice. We have made the LIMUC, which is the largest publicly available labeled UC images dataset to the best of our knowledge, publicly available to alleviate these problems and provide a transparent and fair comparison. The LIMUC dataset can be used for both model development and external validation to test the generalization ability of the proposed approaches. Along with the dataset, several code scripts to help researchers to quickly get started using the LIMUC dataset or reproduce the experiments in this study are shared. We provided code scripts for forming the same 10 cross-validation folds for transparency and fair comparison of the algorithms.

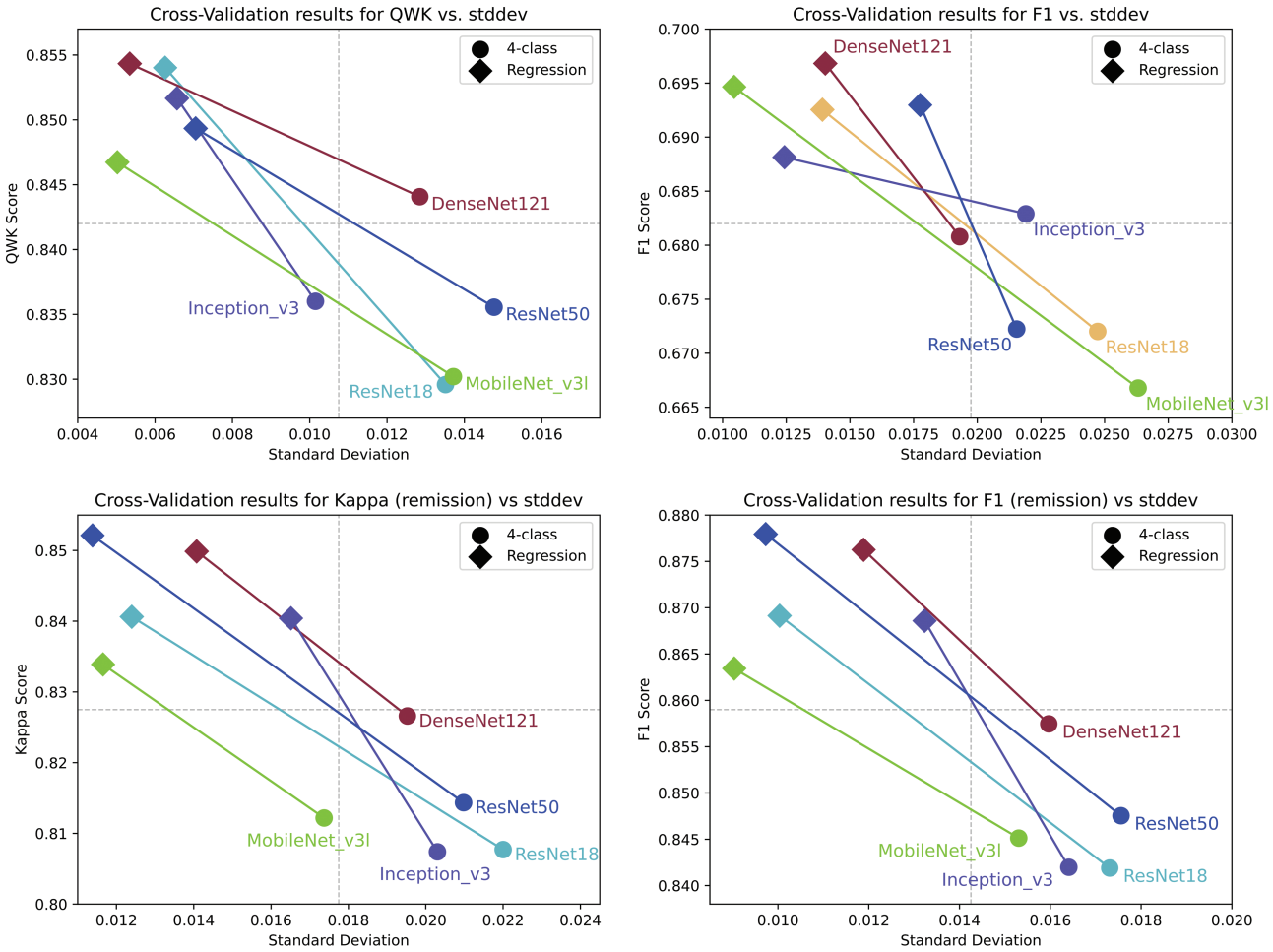

Figure 2. Performance metric vs standard deviation graphs. The upper left area is the desired region that indicates higher average performance and lower standard deviation among different cross-validation folds. All models, for both all Mayo subscore classifications and remission classifications, have an improvement with regression-based convolutional neural networks. stddev, standard deviation.

The second major contribution of this study is benchmarking of different CNN models for the UC severity estimation problem. All previous works have used a single CNN model for the UC severity estimation; for example, Inception-v3[15] was commonly employed by several studies,[21,23,24,29] Gutierrez Becker et al[25] used ResNet50 (13), Bhambhvani et al[22] used ResNext-101,[30] Ozawa et al[20] used GoogLeNet,[31] and Sutton et al[32] used DenseNet121.[14] This prevents comparative assessment of different models on a fair ground and leads to confusion among practitioners. We have evaluated the performance of various models by training them on a common dataset and using the same objective metrics. Experimental results showed that DenseNet121 gives the highest performance.

The third major contribution is adoption of a regression-based approach, which respects the ordinal structure of the MES, effectively making CNN models learn better to distinguish between different Mayo classes. When the proposed approach was employed, a statistically significant improvement was observed among all CNN models except DenseNet121 (which had a P value of .0589). DenseNet121 had the highest performance on the classification of all MESs, though ResNet50 had better performance on remission classification. Although it was not possible to directly compare the proposed approach with previous studies due to different datasets and experimental conditions, we present the reported performance results of automated MES estimation systems in Table 5.

Our study has several limitations. First, the retrospective design and labeling of the images is not ideal, and labeling during the colonoscopy procedure could be better for ground-truth label accuracy. Second, the relatively lower resolution of images may not be sufficient for observation of all details relevant to the diagnosis of disease. While the image resolutions are consistent with those in the literature using AI methods, capturing higher-resolution images (eg, 720p, 1080p) using more recent equipment has a potential to improve the performance of the algorithm due to better representation of the visual patterns.

The classification-based approach is able to provide a confidence value for each Mayo subscore. When the confidence values of all Mayo subscores are close, AI is not able to make a confident decision on the class, and in such a case, validity of choosing the Mayo subscore with the highest confidence may be challenged. Confidence values that could be provided by classification-based approaches, alongside the classification result, may prove to be of further utility in clinical use. On the other hand, the regression-based approach is unable to give confidence scores for each Mayo subscore directly. One of the topics for future research may be deriving a confidence metric from the regression results.

Table 5. Comparison of previous studies and the proposed approach with similar performance metrics

Study (Year)	Model Development Set	Test Set	Model	Outcome	MES Estimation	Remission Estimation (Mayo 0-1 vs Mayo 2-3)
Ozawa et al (2019)[20]	26 304 images from 444 patients	3981 images from 114 patients	GoogLeNet	MES (Mayo 0, Mayo 1, and Mayo 2-3)	Accuracy: 0.704[a]	AUROC: 0.980 Accuracy: 0.946[a]
Stidham et al (2019)[21]	14 862 images from 2778 patients	1652 images from 304 patients 11 432 images from 30 videos	Inception-v3	MES	Kappa: 0.840 Accuracy: 0.778[a]	AUROC: 0.970 Accuracy: 0.917[a] Sensitivity: 0.830 Specificity: 0.960
Takenaka et al (2020)[23]	40 758 images from 2012	4187 images from 875 patients	Inception-v3	Histologic remission Endoscopic remission[b] UCEIS	—	Kappa: 0.798 Accuracy: 0.901 Sensitivity: 0.933 Specificity: 0.878
Maeda et al (2019)[33,c]	12 900 images from 87 patients	9935 images from 100 patients	SVM	Histologic inflammation estimation (active vs healing) MES 0 vs MES 1[b]	Accuracy: 0.910 Sensitivity: 0.650 Specificity: 0.980	—
Bhambhvani et al (2020)[22]	90% of 777 images from 777 patients	10% of 777 images from 777 patients	ResNext-101	MES estimation (Mayo 1, Mayo 2, and Mayo 3)	Accuracy: 0.772 Sensitivity: 0.724 Specificity: 0.857	—
Gottlieb et al (2021)[34,d]	80% of 795 videos from 249 patients[e]	20% of 795 videos from 249 patients[e]	Proprietary algorithm, RNN	MES[b] UCEIS	QWK: 0.844 Accuracy: 0.702[a] Sensitivity: 0.716[a] Specificity: 0.901[a]	Accuracy: 0.866[a]
Yao et al (2021)[29]	16 000 images from 3000 patients	51 videos (A) 264 videos from 157 patients (B)[f]	Inception-v3	MES	QWK (A): 0.840 Accuracy (A): 0.780 QWK (B): 0.590 F1 (B): 0.571	Accuracy (B): 0.837
Huang et al (2021)[24]	70% of 856 images from 54 patients	30% of 856 images from 54 patients	Inception-v3, SVM, k-NN	MES 0-1 vs MES 2-3[b] MES 0 vs MES 1	—	Accuracy: 0.945 Sensitivity: 0.892 Specificity: 0.963
Gutierrez Becker et al (2021)[25]	80% of 1672 videos from 1105 patients[e,f]	20% of 1672 videos from 1105 patients[e,f]	ResNet50	MES 0 vs MES 1-2-3 MES 0-1 vs MES 2-3[b] MES 0-1-2 vs MES 3	—	AUROC: 0.850 Precision: 0.850 Recall: 0.810
Schwab et al (2022)[26]	80% 1881 videos from 726 patients[e,f]	20% 1881 videos from 726 patients[e,f]	ResNet34	MES	QWK: 0.680 (video level) QWK: 0.660 (frame level)	—
Luo et al (2022)[35]	80% of 9928 images (A) 80% of 4378 images (B) A + B: 1317 patients	20% of 9928 images (A) 20% of 4378 images (B) A + B: 1317 patients	UC_DenseNet	MES	Accuracy (A): 0.906 F1 (A): 0.868 Accuracy (B): 0.916 F1 (B): 0.858	Accuracy (A): 0.976 F1 (A): 0.976 Accuracy (B): 0.989 F1 (B): 0.989
Sutton et al (2022)[32]	80% of 2642 images (A)[e] 80% of 840 images (B)[e]	20% of 2642 images (A)[e] 20% of 840 images (B)[e]	DenseNet121	MES 0 vs MES 1-2-3 (A) MES 0-1 vs MES 2-3 (B)[b]	—	F1: 0.913 Accuracy: 0.875 Sensitivity: 0.790 Specificity: 0.910

Table 5. Continued

Study (Year)	Model Development Set	Test Set	Model	Outcome	MES Estimation	Remission Estimation (Mayo 0-1 vs Mayo 2-3)
Polat et al (present study)	9590 images from 462 patients	1686 images from 85 patients	DenseNet121	MES	QWK: 0.854 F1: 0.697 Accuracy: 0.772 Sensitivity: 0.693 Specificity: 0.911	Kappa: 0.827 F1: 0.858 Accuracy: 0.957 Sensitivity: 0.974 Specificity: 0.876

Only MES-related performance results are shared for the studies with several outcome measures.
Abbreviations: AUROC, area under the receiver-operating characteristic curve; MES, Mayo endoscopic score; QWK, quadratic weighted kappa; UCEIS, Ulcerative Colitis Endoscopic Index of Severity.
[a]Performance metrics calculated using the reported numerical values in the study (ie, they were not obtained directly in the article).
[b]Study with several outcome measures. Results are shared for the marked outcome measure.
[c]This work is based on endocytoscopy data.
[d]Deep learning model is trained with video-level labels.
[e]Cross-validation is applied for model performance assessment.
[f]MES estimations were performed for the whole video, not still frames.

Conclusions

We developed a computer-aided estimation system for the assessment of endoscopic activity to classify UC severity with an QWK score of 0.854 (95% CI, 0.842-0.867). The proposed approach has the potential to decrease the interobserver variability, increase the accuracy of diagnosis and standardization of follow-up, and evaluate the disease activity. Furthermore, we made the dataset publicly available for the usage of future studies, and all code has been made open source for both transparency and to facilitate future work.

Supplementary Data

Supplementary data is available at *Inflammatory Bowel Diseases* online.

Author Contribution

Conception and design of the study: G.P., H.T.K., I.E., Y.O.A., A.T., O.A. Acquisition of the data: G.P., H.T.K., I.E., Y.O.A., O.A. Analysis and interpretation of data: G.P., A.T. Critical revisions for important intellectual content: G.P., H.T.K., I.E., Y.O.A., A.T., O.A. Final approval of the version to be submitted: G.P., H.T.K., I.E., Y.O.A., A.T., O.A.

Funding

The work of Gorkem Polat was supported by the TÜBİTAK 2211-A National Scholarship Program and YÖK 100/2000 scholarship for PhD students.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1 Satsangi J, Silverberg M, Vermeire S, Colombel J. The montreal classification of inflammatory bowel disease: controversies, consensus, and implications. *Gut*. 2006;55(6):749-753.
2 Osada T, Ohkusa T, Yokoyama T, et al. Comparison of several activity indices for the evaluation of endoscopic activity in UC: inter- and intraobserver consistency. *Inflamm Bowel Dis*. 2010;16(2):192-197.
3 Travis SP, Schnell D, Krzeski P, et al. Developing an instrument to assess the endoscopic severity of ulcerative colitis: the Ulcerative Colitis Endoscopic Index of Severity (UCEIS). *Gut*. 2012;61(4):535-542.
4 Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60-88.
5 Polat G, Sen D, Inci A, Temizel A. Endoscopic artefact detection with ensemble of deep neural networks and false positive elimination. Presented at: International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV2020), IEEE Symposium on Biomedical Imaging (ISBI2020); April 3, 2020. Vol 2595. Iowa City, Iowa: CEUR-WS.org: pp. 8-12. CEUR Workshop Proceedings.
6 Polat G, Işık Polat E, Kayabay K, Temizel A. Polyp detection in colonoscopy images using deep learning and bootstrap aggregation. Presented at: 3rd International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV 2021) co-located with the IEEE Symposium on Biomedical Imaging (ISBI 2021); April 13, 2021; Nice, France.
7 Du W, Rao N, Liu D, et al. Review on the applications of deep learning in the analysis of gastrointestinal endoscopy images. *IEEE Access*. 2019;7:142053-142069.

8 Min JK, Kwak MS, Cha JM. Overview of deep learning in gastrointestinal endoscopy. *Gut Liver* 2019;13(4):388-393.
9 Choi J, Shin K, Jung J, et al. Convolutional neural network technology in endoscopic imaging: artificial intelligence for endoscopy. *Clin Endosc* 2020;53(2):117-126.
10 Ali S, Ghatwary N, Jha D, et al. Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge. *arXiv* doi:10.48850/arXiv:2020.12031
11 Ali S, Dmitrieva M, Ghatwary N, et al. Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy. *Med Image Anal.* 2021;70:102002.
12 Polat G, Kani HT, Ergenc I, Alahdab YO, Temizel A, Atug O. Labeled Images for Ulcerative Colitis (LIMUC) Dataset. Accessed March 14, 2022. https://zenodo.org/record/5827695#.Y1nsMOxKgaM
13 He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE; 2016:770-778.
14 Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE; 2017:4700-4708.
15 Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE; 2016:2818-2826.
16 Howard A, Sandler M, Chu G, et al. Searching for mobilenetv3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE; 2019:1314-1324.
17 Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. *Int J Comput Vis.* 2015;115(3):211-252.
18 Kingma D, Ba J. Adam: a method for stochastic optimization. *arXiv* doi:10.48550/arXiv.1412.6980
19 Paszke A, Gross S, Massa F, et al. PyTorch: An imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, eds. *NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*. NY, USA: Curran Associates, Inc.; 2019:8026-8037.
20 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;89(2):416-421.e1.
21 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):e193963.
22 Bhambhvani HP, Zamora A. Deep learning enabled classification of Mayo endoscopic subscore in patients with ulcerative colitis. *Eur J Gastroenterol Hepatol.* 2021;33(5):645-649.
23 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;158(8):2150-2157.
24 Huang T-Y, Zhan S-Q, Chen P-J, Yang C-W, Lu HH-S. Accurate diagnosis of endoscopic mucosal healing in ulcerative colitis using deep learning and machine learning. *J Chin Med Assoc.* 2021;84(7):678-681.
25 Gutierrez Becker B, Arcadu F, Thalhammer A, et al. Training and deploying a deep learning model for endoscopic severity grading in ulcerative colitis using multicenter clinical trial data. *Ther Adv Gastrointest Endosc.* 2021;14:2631774521990623.
26 Schwab E, Cula GO, Standish K, et al. Automatic estimation of ulcerative colitis severity from endoscopy videos using ordinal multi-instance learning. *Comput Methods Biomech Biomed Eng Imaging Vis*. 2022;10(4):425-433.
27 Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Cambridge, MA: MIT Press; 2016.
28 Omoumi P, Ducarouge A, Tournier A, et al. To buy or not to buy—evaluating commercial AI solutions in radiology (the ECLAIR guidelines). *Eur Radiol.* 2021;31(6):3786-3796.
29 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.
30 Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE; 2017:1492-1500.
31 Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE; 2015:1-9.
32 Sutton RT, Zaïane OR, Goebel R, Baumgart DC. Artificial intelligence enabled automated diagnosis and grading of ulcerative colitis endoscopy images. *Sci Rep.* 2022;12(1):2748.
33 Maeda Y, Kudo S-E, Mori Y, et al. Fully automated diagnostic system with artificial intelligence using endocytoscopy to identify the presence of histologic inflammation associated with ulcerative colitis (with video). *Gastrointest Endosc.* 2019;89(2):408-415.
34 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.
35 Luo X, Zhang J, Li Z, Yang R. Diagnosis of ulcerative colitis from endoscopic images based on deep learning. *Biomed Signal Proc Control.* 2022;73:103443.