Received: 5 August 2022 | Accepted: 21 January 2023

DOI: 10.1111/eci.13960

ORIGINAL ARTICLE

WILEY

Artificial-intelligence-based decision support tools for the differential diagnosis of colitis

Pedro Guimarães[1,2] | Helen Finkler[3] | Matthias Christian Reichert[3] | Vincent Zimmer[3,4] | Frank Grünhage[3] | Marcin Krawczyk[3] | Frank Lammert[3,5] | Andreas Keller[1,6] | Markus Casper[3]

[1]Chair for Clinical Bioinformatics, Saarland University, Saarbrücken, Germany

[2]Coimbra Institute for Biomedical Imaging and Translational Research (CIBIT), Institute for Nuclear Sciences Applied to Health (ICNAS), University of Coimbra, Coimbra, Portugal

[3]Department of Medicine II, Saarland University Medical Center, Saarland University, Homburg, Germany

[4]Department of Medicine, Knappschaft Hospital Saar, Püttlingen, Germany

[5]Chair for Health Sciences, Hannover Medical School (MHH), Hannover, Germany

[6]Department of Neurology and Neurological Sciences, Stanford University School of Medicine, Stanford University, Stanford, California, USA

Correspondence
Markus Casper, Department of Medicine II, Saarland University Medical Center, Saarland University, Homburg D 66421, Germany.
Email: maca005@uni-saarland.de

Abstract

Background: Whereas Artificial Intelligence (AI) based tools have recently been introduced in the field of gastroenterology, application in inflammatory bowel disease (IBD) is in its infancies. We established AI-based algorithms to distinguish IBD from infectious and ischemic colitis using endoscopic images and clinical data.

Methods: First, we trained and tested a Convolutional Neural Network (CNN) using 1796 real-world images from 494 patients, presenting with three diseases (IBD [$n = 212$], ischemic colitis [$n = 157$], and infectious colitis [$n = 125$]). Moreover, we evaluated a Gradient Boosted Decision Trees (GBDT) algorithm using five clinical parameters as well as a hybrid approach (CNN+GBDT). Patients and images were randomly split into two completely independent datasets. The proposed approaches were benchmarked against each other and three expert endoscopists on the test set.

Results: For the image-based CNN, the GBDT algorithm and the hybrid approach global accuracies were .709, .792, and .766, respectively. Positive predictive values were .602, .702, and .657. Global areas under the receiver operating characteristics (ROC) and precision recall (PR) curves were .727/.585, .888/.823, and .838/.733, respectively. Global accuracy did not differ between CNN and endoscopists (.721), but the clinical parameter-based GBDT algorithm outperformed CNN and expert image classification.

Conclusions: Decision support systems exclusively based on endoscopic image analysis for the differential diagnosis of colitis, representing a complex clinical challenge, seem not yet to be ready for primetime and more diverse image datasets may be necessary to improve performance in future development. The clinical value of the proposed clinical parameters algorithm should be evaluated in prospective cohorts.

KEYWORDS
computer-aided detection, computer-aided diagnosis, endoscopy, infectious colitis, inflammatory bowel disease, ischemic colitis, neuronal network

1 | INTRODUCTION

Over the past decade artificial intelligence (AI)-based tools have increasingly been applied to various tasks in the field of gastroenterology.[1] Deep learning (DL) and especially algorithms based on convolutional neuronal networks (CNN) are the most promising approaches in medical image analysis, including gastrointestinal endoscopy,[2,3] whereas conventional machine learning approaches can effectively be used for analysis of well-structured data (e.g., clinical information or and blood test results). So far, in the field of endoscopy, CNN have primarily been used to identify and classify malignant lesions und its precursors in the upper and lower gastrointestinal tract (e.g., colorectal polyps, oesophageal and gastric cancer).[4,5] In contrast, inflammatory disorders diffusely affecting the gastrointestinal mucosa have not been the primary focus.[6,7] Recently, AI-based image analysis has also been applied to endoscopy in inflammatory bowel disease (IBD). This includes the diagnosis of IBD,[8,9] the assessment of disease activity,[10,11] and screening for dysplasia.[12]

The diagnosis of Crohn's disease and ulcerative colitis is often not straightforward and depends on the clinician's judgement based on symptoms, medical history, endoscopic and histopathological appearance, blood and microbiological tests as well as additional biomarkers.[13] Furthermore, there are several conditions that mimic IBD (e.g., infectious colitis, ischemic colitis, drug-induced colitis) and have to be included in the differential diagnosis of colitis. From an endoscopic point of view, it is particularly challenging to distinguish these different forms of colitis.

To the best of our knowledge, we the first time evaluated AI-based approaches to differentiate IBD with colonic manifestation from infectious and ischemic colitis using real-world endoscopic still images as well as the patients' clinical records and present the data for performance evaluation in an independent dataset.

2 | PATIENTS AND METHODS

2.1 | Datasets

Endoscopy reports (Department of Medicine II, Saarland University Medical Center) from 01/2009 to 01/2020 were screened using the search-term 'colitis' to identify patients. Adequate endoscopic images were available for 743 patients. Considering medical history, disease course in the follow-up period, endoscopy reports and endoscopic images, histopathology and microbiological tests patients were re-classified (HF, MC) as: (i) IBD with colonic involvement, (ii) infectious colitis, and (iii) ischemic colitis. When the final diagnosis was ambiguous or in case of inconsistent results or combinations (e.g., IBD with superinfection), 249 patients were excluded at this step of development. For the remainder all available representative images showing signs of active inflammation were used. Finally, 1796 images obtained during 584 colonoscopies in 494 patients were included in the study. The number of images per patient ranged from 1 to 10. Median age at endoscopy was 55 years (4–94 years), 226 patients (45.7%) were women, and 268 (54.3%) were men. Subjects were randomly split into datasets DS1 and DS2. DS2 for validation was completely independent from DS1. Dataset 1 (DS1) contained 1635 images from 444 subjects (482 images from 190 patients with IBD; 577 images from 142 patients with ischemic colitis; 576 images from 112 patients with infectious colitis). Dataset 2 (DS2) included 161 images from 50 subjects (42 images from 22 patients with IBD; 71 images from 15 patients with ischemic colitis; 48 images from 13 patients with infectious colitis). In DS1, 92 patients had Crohn's disease (223 images), 87 ulcerative colitis (238 images) and 11 unclassified IBD (21 images). DS2 consisted of 10 patients with Crohn's disease (15 images) and 12 with ulcerative colitis (27 images).

Several generations of Olympus scopes (CF-Q140; CF-Q160; PCF-Q160; CF-Q180; CF-H180; CF-H190; PCF-H190) were used for colonoscopies. Unaltered white-light non-standardized images (various scope positions, distances, angles, illumination, contaminations), taken from all colonic segments, were cropped, resized, and normalized. Table 1 summarizes DS1 and DS2 characteristics. Additional details can be found in the Appendix S1.

2.2 | Artificial intelligence-based classification

Three different approaches were tested: image-based classification, clinical data-based classification, and a hybrid classifier, combining both clinical data and images. Figure 1 summarizes the study workflow, which included data splitting (A), three different approaches to classification (B) as well as training, tuning, and testing procedures

TABLE 1 Dataset characteristics.

	Dataset 1 (DS1)	Dataset 2 (DS2)
Patients with IBD	N = 190	N = 22
Age (years)	35 (12–83)	30 (14–80)
Gender (male/female, %)	99/91 (52%/48%)	8/14 (36%/64%)
Number of colonoscopies	248	25
Number of colonoscopies per patient	1 (1–4)	1 (1–3)
Number of images	482	42
Number of images per patient	2 (1–10)	2 (1–5)
Patients with infectious colitis	N = 112	N = 13
Age (years)	60 (4–90)	72 (44–83)
Gender (male/female, %)	60/52 (54%/46%)	8/5 (62%/38%)
Number of colonoscopies	116	13
Number of colonoscopies per patient	1 (1–2)	1 (1)
Number of images	576	48
Number of images per patient	4 (1–22)	3 (1–12)
Patients with ischemic colitis	N = 142	N = 15
Age (years)	73 (22–94)	77 (37–88)
Gender (male/female, %)	84/58 (59%/41%)	9/6 (60%/40%)
Number of colonoscopies	163	19
Number of colonoscopies per patient	1 (1–3)	1 (1–4)
Number of images	577	71
Number of images per patient	4 (1–16)	4 (3–9)

Note: Characteristics for the datasets 1 and 2 (DS1 and DS2) are given. Data are presented as absolute numbers or as median and range (in parenthesis). As expected, IBD patients were younger than patients with ischemic or infectious colitis.

Abbreviation: IBD, inflammatory bowel disease.

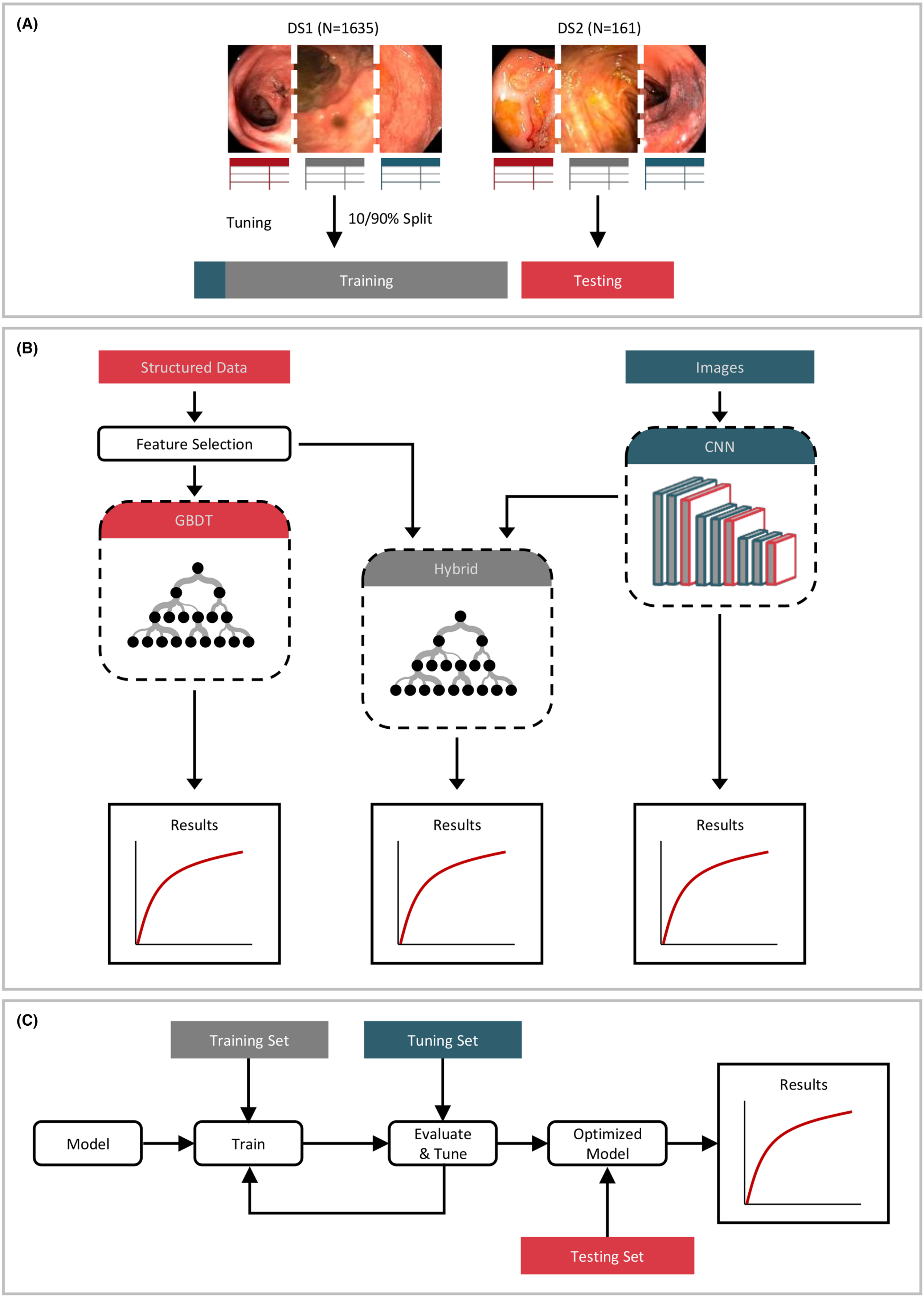
for each of these approaches (C). The models generated can be found in https://github.com/pedrogsc/colitis. The raw data supporting the conclusions of this article will be made available by the authors upon reasonable request. Datasets can be shared only after formal ethics approval.

2.3 | Deep-learning classification of endoscopic images

Deep learning is a subset of machine learning that differs in the way features are selected. Instead of hand-crafted feature selection, layer-to-layer more and more complex abstract representations are created. Like in this study, limited training data is a common challenge in the medical field, which can be addressed by transfer learning. The important issue in this context that must be controlled for is overfitting. Here, we used transfer-learning by fine-tuning a DenseNet[14] pre-trained on ImageNet.[15] A SoftMax output layer was used with categorical cross-entropy loss, and dropout regularization was added. Rotation, scaling as well as horizontal and vertical reflection were applied to training data only to artificially augment data availability. During testing variational dropout, an extension of regular dropout where optimal dropout rates are inferred from the data, with 30 prediction calls per classification was used. DS1 was used for training and tuning, and performance was evaluated in DS2. Final classification was obtained with the best performing hyperparameter combination (grid-search selecting learning-rate, momentum, dropout, and training steps), as assessed in the tuning set.

2.4 | Clinical parameter-based algorithm

For the whole dataset clinical records were analysed retrospectively and 16 clinical parameters were determined (demographic data: age, sex; blood tests: C-reactive protein [CRP], creatinine, red blood cell count, white blood cell count; short-term medical history: antibiotics use; major surgery, resuscitation; previous medical history: arterial hypertension, coronary heart disease, chronic kidney disease, diabetes, nicotine abuse, peripheral arterial occlusive disease, and stroke). Three classification methods were applied: Gradient Boosted Decision Trees (GBDT), Logistic Regression (LR), and Deep Feed



(A)
DS1 (N=1635)
DS2 (N=161)
Tuning
10/90% Split
Training
Testing
(B)
Structured Data
Images
Feature Selection
CNN
GBDT
Hybrid
Results
Results
Results
(C)
Training Set
Tuning Set
Model
Train
Evaluate & Tune
Optimized Model
Results
Testing Set

FIGURE 1 Workflow. For each approach images and/or clinical data of dataset 1 (DS1) were used for training and tuning (90%/10% split), whereas images and/or clinical data of dataset 2 (DS2) were used for hold-out testing. (A) The three different approaches tested: image-based classification, clinical data-based classification, and a hybrid classifier, combining both clinical data and images, are summarized in (B). For the three algorithms the workflow for each training, tuning, and testing cycle was performed as presented in (C).

Forward Neural Network (FFNN). Tuning was applied to each method and included forward stepwise feature selection to select the best subset of features and grid-search to select the best hyperparameter set. To evaluate the performance of the GBDT and LR approaches, we performed 30 prediction calls using random subsamples (80%) of the training set. For the FFNN approach, variational dropout was used to perform 30 prediction calls per classification. The rationale for both variational dropout and random subsampling was to create response groups of various slightly different models to assess performance.

2.5 | Hybrid approach

A hybrid approach combining clinical and image information was also developed. Here, we combined the clinical parameters that resulted from the feature selection and the output of the last fully connected layer of the same neural network as above (deep-learning classification of endoscopic images). All three classification methods for clinical parameters (GBDT, LR, and FFNN) were then tested. Tuning was specific for each classifier and was performed as described. Performance evaluation was done on DS2 using either variational dropout or random subsampling, depending on the classifier.

2.6 | Outcome definition and performance evaluation

Dataset 2 images were independently evaluated by three experienced expert endoscopists (FG > 10,000 colonoscopies, VZ > 10,000 colonoscopies, M > 2500 colonoscopies) blinded to the study aims and class distribution. The primary outcome was to evaluate the proposed approaches against endoscopists and against each other, for each class using a one versus rest (OvR) approach, and globally using the micro average OvR values. Accuracy, sensitivity and specificity were computed. Balanced accuracy, positive predictive value (PPV), negative predictive value (NPV) and F1-score were also calculated (for each class and global). Differences between the performance metrics obtained for the three classification approaches and endoscopists were assessed with Mann–Whitney *U* test (significance level .05) after excluding normality using Kolmogorov–Smirnov test. The receiving operating characteristics (ROC) and precision-recall (PR) curves were assessed for the proposed approaches.

2.7 | Interpretability

If machine learning is used in decision support systems, the interpretability of non-linear classification methods must be addressed:

Gradient Boosted Decision Trees and LR were applied in the clinical parameter-based and hybrid algorithms. For these two classifiers we computed the relative importance of each variable. For the former we used its coefficients, and for the latter, the total gain of splits that used the variable.

Deep learning-based approaches lack transparency in their decision-making process. We used deep Taylor decomposition to trace back each classification decision of the neural network model from output to input, to create relevance heatmaps that represent the relative importance of each pixel/input for the model's decision.[16] While still limited in scope, these maps over the entire testing dataset allow us to recognize patterns that the model uses to perform classifications.

2.8 | Ethical considerations

The study was approved by the ethics committee of Ärztekammer des Saarlandes (Saarbrücken, Germany; #36/19).

3 | RESULTS

Performance metrics for the different AI-based approaches and endoscopists are summarized in Figure 2. Additional metrics and discriminate *p* values can be found in the Appendix S1. Figure 3 illustrates ROC and PR curves for the different approaches and each class for DS2. Global ROC and PR curves are shown in Appendix S1.

3.1 | CNN-based approach

Classification took ~.5 s per prediction call on CPU (Intel® Core™ i7-8550U). A global accuracy of .709 [.682–.743] and areas under the ROC and PR curves of .727 [.687–.766] and .585 [.539–.635] were determined for the CNN approach

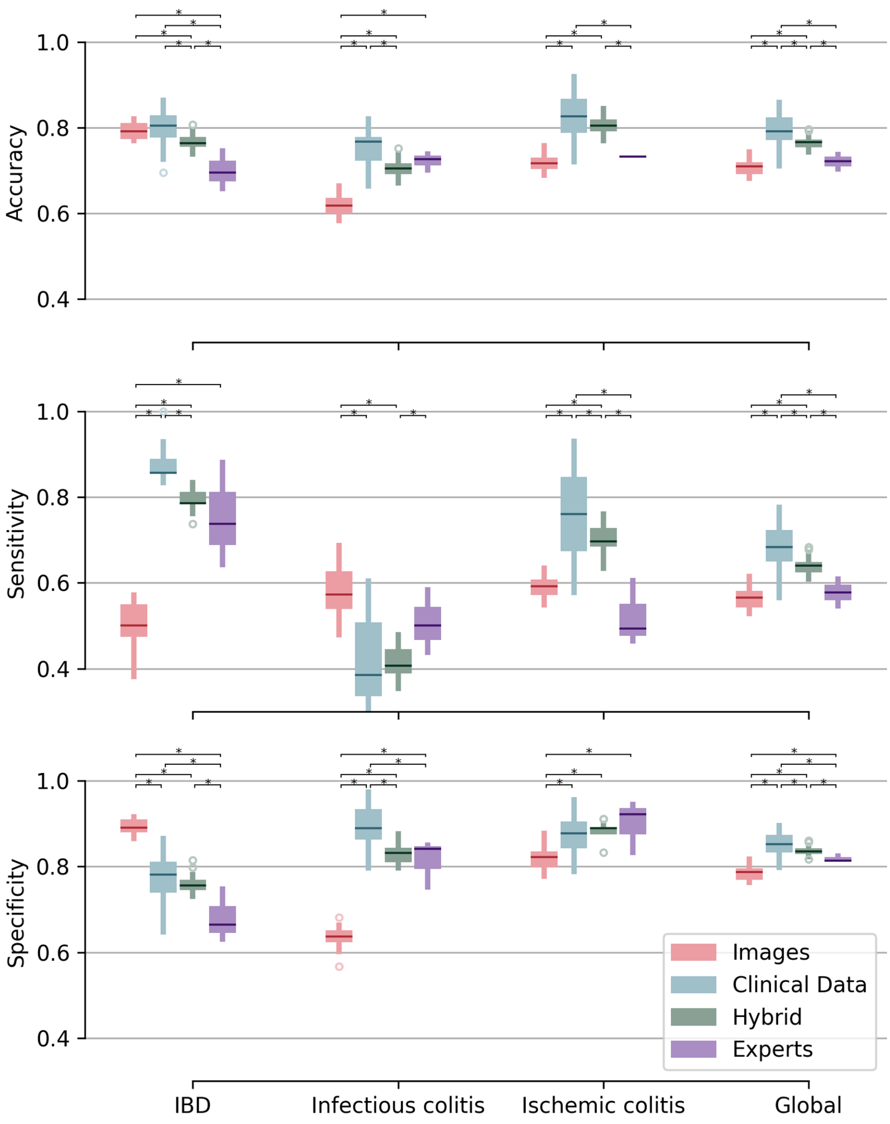


FIGURE 2 Performance metrics. Accuracy, sensitivity, and specificity boxplots as computed on dataset 2 (DS2) for the three different classes (inflammatory bowel disease, infectious colitis, ischemic colitis) were evaluated as one versus rest (median) for each algorithm. Global boxplots (micro average) are also shown. Differences between the performance metrics of the specific algorithms and in comparison, to endoscopists are marked by an asterisk. IBD, inflammatory bowel disease.

with DS2. Worst results were observed for the infectious colitis group with areas under the ROC and PR curves of .606 [.552–.676] and .344 [.305–.398] only, respectively. In the whole DS2 accuracy did not differ between CNN and endoscopists (p = .13). However, the CNN showed significantly better accuracy for IBD (.792 [.770–.820] vs. .696 [.658–.745]; p = .0034) but was significantly inferior for infectious colitis (.618 [.584–.665] vs. .727 [.702–.739] p = .0034). A low global sensitivity of .565 [.528–.615] with the worst performance for IBD (.500 [.381–.571]) and a global PPV of .602 [.560–.659] with the worst performance in the infectious colitis group (.399 [.365–.458]) are the most obvious limitations of this approach.

3.2 | Clinical parameter-based approach

The best performing approach was the algorithm based on GBDT. Figure 4 shows performance evolution as percentage of maximum performance when adding new features using the feed forward feature selection. As shown, with five features approximately 97.5% maximum performance was achieved, whereas with 10 features this value rose to almost 99%. In a clinical environment a low number of features means a lower probability of missing data. Thus, we selected only to use the five best features, theoretically achieving 97.5% maximum performance. For GBDT the best five feature set contained age, CRP, white blood cell count, previous major surgery, and antibiotics use. Areas under the ROC curves for each separate clinical feature on DS1 are shown in Appendix S1. Classification took <.001 s per prediction call on CPU (Intel® Core™ i7-8550U). Global accuracy was .792 [.711–.859] and global areas under the ROC and PR curves were .888 [.849–.913] and .823 [.762–.857], respectively. Global PPV was .702 [.577–.793]. This approach showed the weakest performance in the infectious colitis group again, with an accuracy of .767 [.665–.820], a sensitivity of .385 [.250–.604]

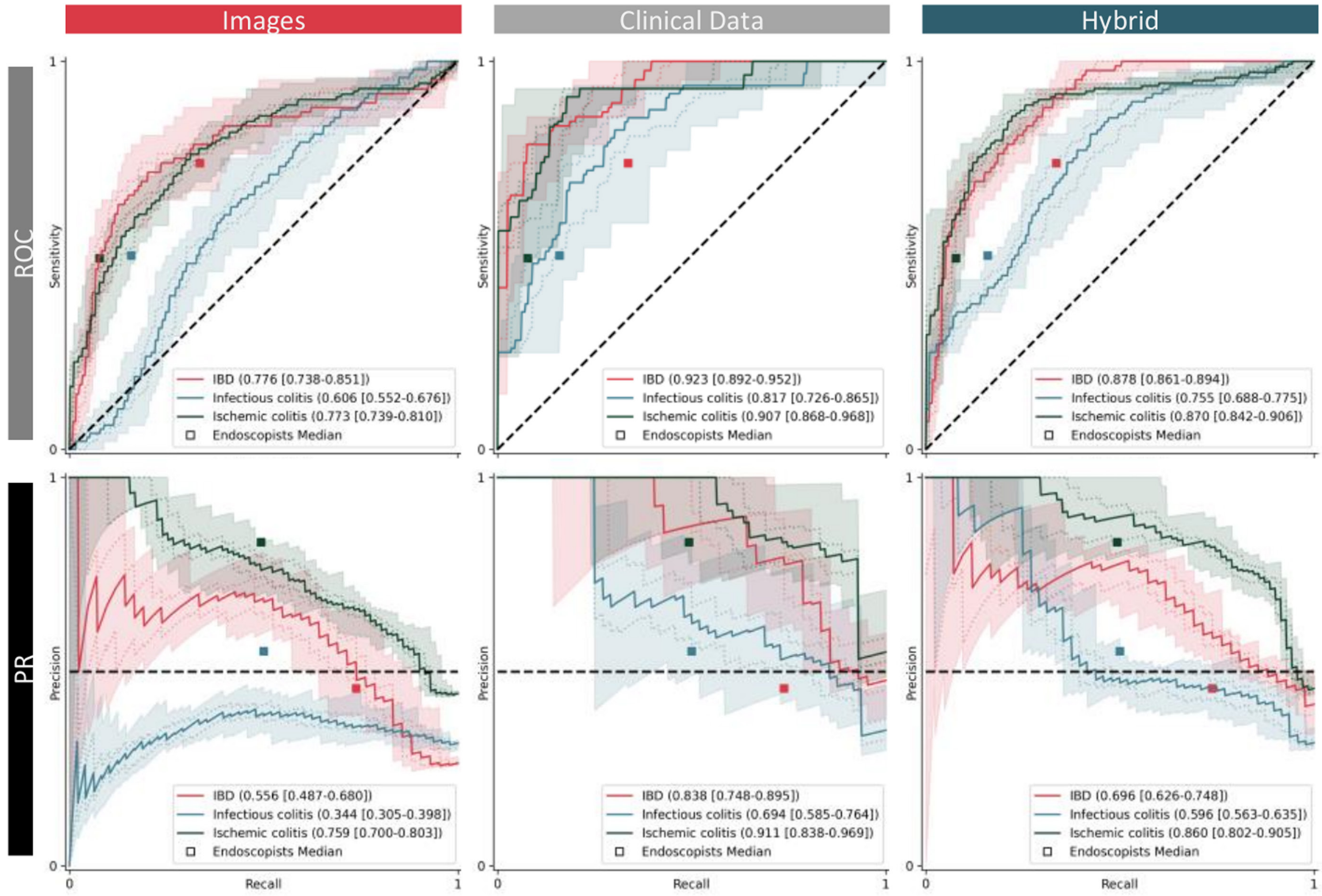


FIGURE 3 Performance curves. Receiving operating characteristics (ROC) and precision-recall (PR) curves for dataset 2 (DS2) are shown (top and bottom, respectively). Performance curves were evaluated as one versus rest. For reference, median results of the endoscopists are marked by squares. The area under each curve and range are provided in the legends. The range for each curve is shaded. Dotted lines represent the 25th and 75th quantiles.

and a specificity of .889 [.796–.973]. As illustrated in Figure 2, all global metrics were significantly higher as compared to CNN and endoscopists. The total gain of all the splits that use each clinical parameter was computed for each prediction call. Obtained results are summarized in the boxplot presented in Figure 5A. As shown, age is by far the most relevant clinical parameter, followed by CRP and white blood cell count.

3.3 | Hybrid approach

The approach considering image analysis and clinical parameters as a hybrid algorithm combining CNN and GBDT achieved a global accuracy of .766 [.743–.796]. Global areas under the ROC and PR curves were .838 [.811–.858] and .733 [.699–.778], respectively. Classification took ~.5 s per prediction call on CPU (Intel® Core™ i7-8550U). All global metrics were significantly higher as compared to CNN and endoscopists, but significantly lower as compared to the clinical data-based approach. Interestingly, the CNN's poor sensitivity for IBD and low specificity for infectious colitis significantly improved with the hybrid approach. As with the clinical parameters-based approach, the total gain of all the splits that use each clinical parameter was computed for each prediction call. Obtained results are summarized in the boxplot presented in Figure 5B. The gain for the output of the last fully connected layer of CNN was combined (sum) to ease visualization. As expected, age is still the most relevant parameter, followed by the combined gain of the CNN-derived features, CRP, and white blood cells count.

3.4 | Endoscopists

Evaluation of DS2 images by expert endoscopists resulted in an accuracy, sensitivity, and specificity of .721 [.704–.738], .578 [.547–.609] and .814 [.813–.824], respectively.

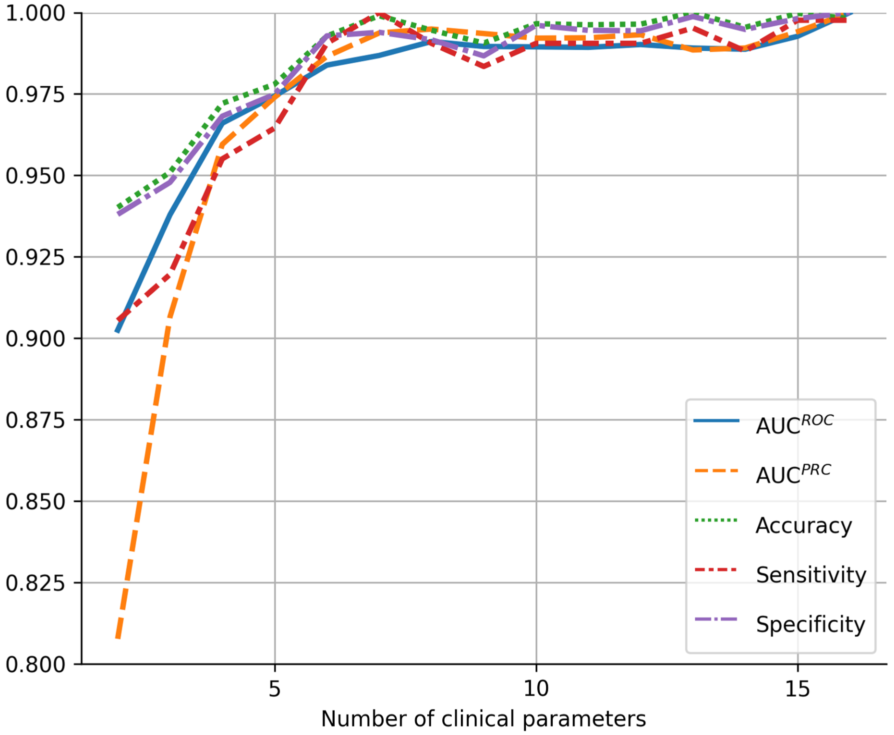


FIGURE 4 Evolution of the clinical data-based algorithm. The evolution of the Gradient Boosted Decision Trees (GBDT) algorithm's performance on dataset 1 (DS1) with the number of added features using the feed forward feature selection procedure is shown as percentage of peak performance.

4 | DISCUSSION

In patients with new-onset colitis who typically present with abdominal pain, cramping, (bloody) diarrhoea and occasionally fever it is clinically challenging to differentiate IBD from infectious or ischemic colitis as well as other colitis differential diagnoses.[17–19] Although there are several endoscopic findings (e.g., distribution and morphology of mucosal alterations) that may help to distinguish these entities,[19] endoscopy results are non-specific in most cases. Even when considering anamnestic information, histopathology, stool microbiology, blood tests, and imaging studies, it is often difficult to make the correct diagnosis.

Recently, AI has also been applied to various tasks in IBD.[20–22] However, IBD is still not in the center of AI-based research. AI should primarily be used for clinical constellations with suboptimal performance of conventional diagnostics. Thus, this is the first attempt to establish AI-based algorithms to distinguish IBD, infectious and ischemic colitis using endoscopic images and clinical data. Of note, our algorithms were not trained to distinguish colitis from healthy colonic mucosa or other gastrointestinal pathologies.[23,24]

With an overall accuracy of .709 [.682–.743] and a PPV of .602 [.560–.659] performance of our image-based CNN might appear low. Especially infectious colitis seems to be problematic. On the other hand, these results reflect the clinical reality for colitis differentiation as supported by the results of endoscopy experts that were comparable to CNN, and results will be worse among unexperienced endoscopists.

In comparison to many other tasks in the field of endoscopy that have already been successfully tackled by AI, colitis differentiation is one of the most complex medical challenges that have been tried to solve by CNN up to now. Moreover, multi-class classification approaches have rarely been reported. The most likely explanation for the results is the high variability of endoscopic presentation for each class. Thus, the performance of an algorithm highly depends on the quality and diversity of the training dataset (ideally the whole spectrum of endoscopic presentation must be covered under various endoscopic conditions). Image selection and correct classification of images used in the training set are crucial for classification performance. The simple number of images used to establish a CNN is not a quality criterion, since even very large datasets, with hundreds of images from each individual patient, may not be representative of the investigated disease and thus show poor performance in real-life scenarios. Ideally, a dataset is composed of as few images per patient of as many patients as possible. The strengths of our relatively small dataset are careful selection and image quality. The use of more than one image per patient might be a potential limitation, but all images were substantially different, and datasets completely independent. Since in case of colitis a considerable percentage of patients remain undefined in clinical routine it is difficult to establish large well characterized datasets. Particularly, in patients with

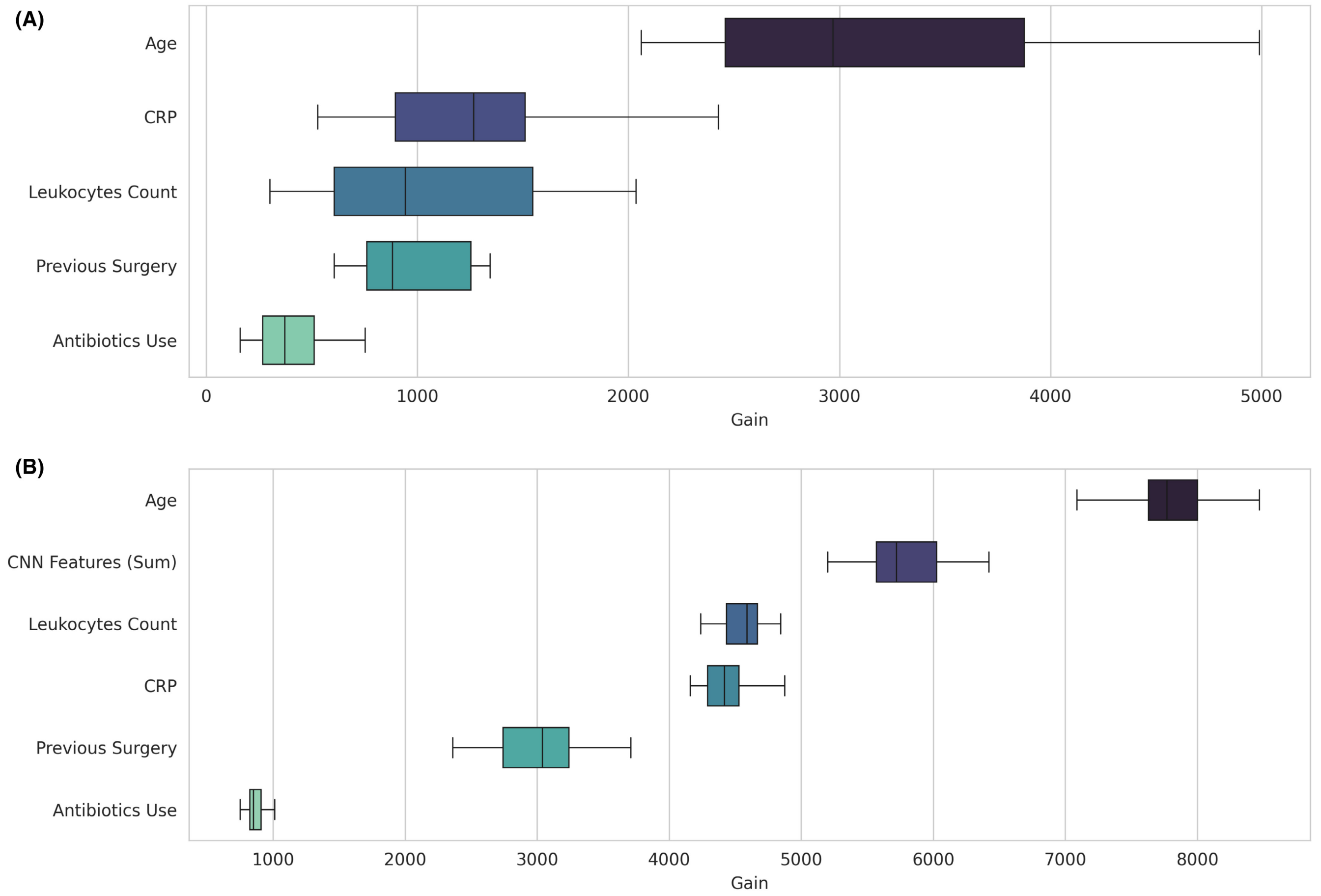


FIGURE 5 Feature importance for the clinical data-based and hybrid classification. Boxplots of the total gain of all the splits that use each clinical parameter. Results for the clinical data-based and hybrid classification are represented in A and B, respectively. Gains for each output of the last fully connected layer of convolution neural network (CNN) used to process the images were combined (sum) to ease visualization. CRP, C-reactive protein.

suspected infectious colitis endoscopy is not performed regularly. By using real-life images with obvious limitations (e.g., pronounced stool contamination because of emergency conditions in a high proportion of cases), we accept poorer performance on the test set, but ensure suitability in prospective real-world application. On the other hand, algorithms only using high-quality images usually show poor performance when confronted with suboptimal conditions in real-world.

Figure 6 shows representative examples of correctly classified and misclassified images overlayed with relevance heatmaps. Despite the lack of transparency of DL, these maps allow us to lift the veil and speculate on its decision process. In summary, the interpretability results mirror the difficulty of the classification task. Although the algorithm can evidently identify some signs of disease, a larger dataset, with not only more images, but also more diversity, would probably lead to improved performance. More images illustrating the difficulty associated with the classification tasks are given in the Appendix S1.

There are several limitations of the image-based algorithm to discuss: Essentially, the performance of the currently available algorithm cannot justify its use as a decision support tool in clinical practice right now but can be regarded as a first step in the process of development. Larger multicenter datasets covering the whole spectrum of endoscopic appearance may be an indispensable requirement to establish more robust algorithms. Here, we used still images taken by endoscopists in real-life to categorize colitis without consistent information on the localization. Moreover, the important information on distribution of findings throughout the colon is missing in still images. Alternatively, implementation of video-based algorithms could be considered but large datasets have yet to be curated and linked to clinical data. It must be kept in mind, that in real-life also other colitis etiologies (e.g., colitis associated with diverticulitis, drug-induced colitis) as well as combinations (e.g., IBD with superinfection) may occur in addition to the three major entities studied here, making the classification task much more complex. Furthermore, on-line evaluation by incorporation into

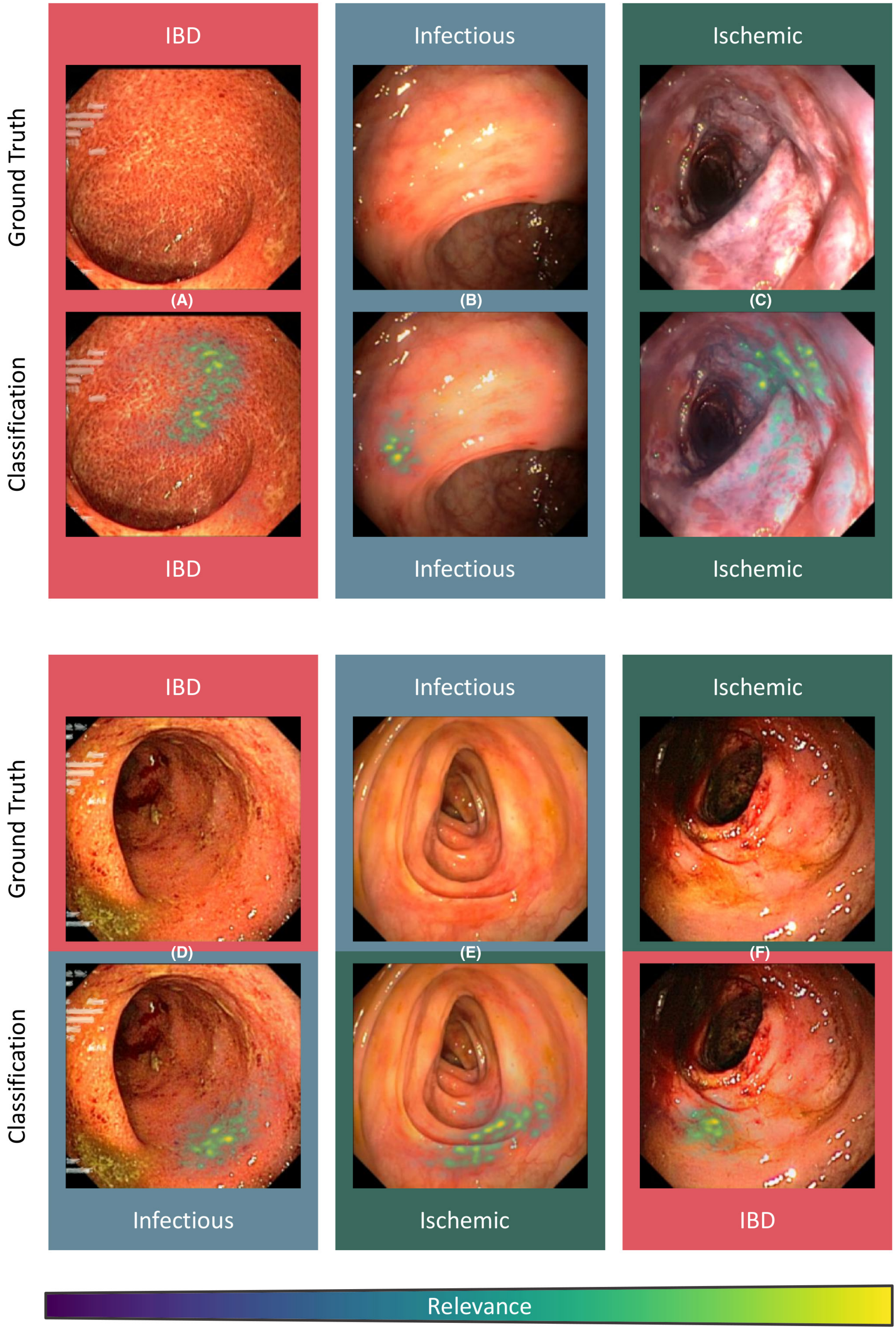

IBD
Infectious
Ischemic
Ground Truth
Classification
(A)
(B)
(C)
IBD
Infectious
Ischemic
IBD
Infectious
Ischemic
Ground Truth
Classification
(D)
(E)
(F)
Infectious
Ischemic
IBD
Relevance

FIGURE 6 Interpretability of convolution neural network (CNN)-based image analysis. Images from dataset 2 (DS2) for the three different classes are shown (six image pairs of original images on top, and on the bottom images overlayed with relevance heatmaps obtained by deep Taylor decomposition). For each pair, ground truth class is indicated on top, and classification by the proposed CNN approach is indicated on the bottom. In A, the relevance heatmaps show that the network correctly identified small erosions but did not focus on erythematous areas in ulcerative colitis with diffusely inflamed mucosa. An area with focal erythema and small erosions was identified in B and correctly classified as infectious colitis. (C) A livid background mucosa in severe ischemia. Interestingly, the network predominantly focused on whitish appearing ulcerative areas. In D, erosive mucosal lesions in ulcerative colitis (comparable with those in panel [A]) are highlighted but the model misclassified inflammatory bowel disease (IBD) as infectious colitis, while in E, the algorithm correctly identified inflamed regions but again mainly focused on whitish appearing areas beneath erythematous changes. The infectious colitis image was misclassified as ischemic colitis. In F, it appears that relevance was wrongly given to stool remnants to diagnose IBD in a patient with infectious colitis.

endoscopy processor software will be necessary in the future.

In a second approach we used the available clinical data from DS1 patients to establish a machine-learning algorithm. The GBDT algorithm using five clinical parameters outperformed the image-based CNN. Albeit the performance evolution observed during the feed-forward feature selection on DS1 suggested that the inclusion of more data might further improve performance, the algorithm using 10 parameters showed slightly worse results. That the algorithm was established on a tertiary care dataset only, is a potential limitation of this algorithm. In a recently published study, the authors established a machine learning prediction model based on 702 IBD patients and 315 healthy controls.[25] The best classifier based on 16 parameters (including age, haemoglobin, and faecal calprotectin) achieved a mean average precision of 91% for UC and 97% for MC but could not distinguish various types of colitis either.

In clinical reality gastroenterologists use several sources of information (e.g., medical history, endoscopy, stool and blood tests) to make the correct diagnosis. Therefore, we imitated this process using a hybrid algorithm including the image-based CNN and the clinical parameter-based GBDT algorithm with five features. Of note, this algorithm was inferior to the clinical information only approach due to the relatively poor performance of the image-based CNN. After a learning curve the human brain variably weighs the information coming from many sources and the physicians can overrule misleading inputs to increase the probability for a correct diagnosis. Theoretically, this could also be done by our hybrid approach, but current technical applications might not be as effective as needed.

5 | CONCLUSION

Although it is beyond question that AI has the capability of supporting endoscopists, we here show that difficult clinical decisions cannot straightforwardly be solved by AI either. An endoscopic image-based decision-support by the proposed state-of-the-art AI algorithms seems not to be ready for use in its current version. However, our results imply that more diverse data could lead to improvements. The clinical value of the clinical parameter-based algorithm should be evaluated in prospective cohorts.

AUTHOR CONTRIBUTIONS

Pedro Guimarães: programming of algorithms, data analysis, statistics, and manuscript preparation. Andreas Keller: revision of the manuscript and supervision of the artificial intelligence part. Frank Grünhage, Matthias Reichert, Marcin Krawczyk, and Vincent Zimmer: image evaluation, patient care, and revision of the manuscript. Helen Finkler: collection of endoscopic images and clinical characteristics, data analysis, and manuscript preparation. Frank Lammert: revision of the manuscript; supervision of patient care, and design of the study. Markus Casper: idea and design of the study, data analysis, manuscript preparation, collection of patients, images and clinical parameters, and patient care.

ACKNOWLEDGEMENT

Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflicts of interests.

ORCID

Matthias Christian Reichert https://orcid.org/0000-0002-8192-0575
Frank Grünhage https://orcid.org/0000-0002-0750-0329
Marcin Krawczyk https://orcid.org/0000-0002-0113-0777
Markus Casper https://orcid.org/0000-0002-1146-288X

REFERENCES

1. Le Berre C, Sandborn WJ, Aridhi S, et al. Application of artificial intelligence to gastroenterology and hepatology. *Gastroenterology*. 2020;158:76-94.e2.

2. Ebigbo A, Palm C, Probst A, et al. A technical review of artificial intelligence as applied to gastrointestinal endoscopy: clarifying the terminology. *Endosc Int Open*. 2019;7:E1616-E1623.
3. Okagawa Y, Abe S, Yamada M, Oda I, Saito Y. Artificial intelligence in endoscopy. *Dig Dis Sci*. 2021;67:1553-1572. doi:10.1007/s10620-021-07086-z
4. Byrne MF, Chapados N, Soudan F, et al. Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. *Gut*. 2019;68:94-100.
5. Horie Y, Yoshio T, Aoyama K, et al. Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks. *Gastrointest Endosc*. 2019;89:25-32.
6. Guimarães P, Keller A, Fehlmann T, Lammert F, Casper M. Deep-learning based detection of gastric precancerous conditions. *Gut*. 2020;69:4-6.
7. Guimarães P, Keller A, Fehlmann T, Lammert F, Casper M. Deep learning-based detection of eosinophilic esophagitis. *Endoscopy*. 2021;54:299-304. doi:10.1055/a-1520-8116
8. Mossotto E, Ashton JJ, Coelho T, Beattie RM, MacArthur BD, Ennis S. Classification of paediatric inflammatory bowel disease using machine learning. *Sci Rep*. 2017;7:2427.
9. Klang E, Barash Y, Margalit RY, et al. Deep learning algorithms for automated detection of Crohn's disease ulcers by video capsule endoscopy. *Gastrointest Endosc*. 2020;91:606-613.
10. 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;158:2150-2157.
11. 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;89:416-421.
12. Kandiah K, Subramaniam S, Thayalasekaran S, et al. Multicentre randomised controlled trial on virtual chromoendoscopy in the detection of neoplasia during colitis surveillance high-definition colonoscopy (the VIRTUOSO trial). *Gut*. 2021;70:1684-1690.
13. Feakins R, Torres J, Borralho-Nunes P, et al. ECCO topical review on clinicopathological spectrum and differential diagnosis of IBD. *J Crohns Colitis*. 2022;16:343-368. doi:10.1093/ecco-jcc/jjab141
14. Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Honolulu. 2017:2261–2269. doi:10.1109/CVPR.2017.243
15. Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. *Int J Comput Vis*. 2015;115:211-252.
16. Montavon G, Lapuschkin S, Binder A, Samek W, Müller KR. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognit*. 2017;65:211-222.
17. Schumacher G, Sandstedt B, Kollberg B. A prospective study of first attacks of inflammatory bowel disease and infectious colitis. Clinical findings and early diagnosis. *Scand J Gastroenterol*. 1994;29:265-274.
18. Chachu KA, Osterman MT. How to diagnose and treat IBD mimics in the refractory IBD patient who does not have IBD. *Inflamm Bowel Dis*. 2016;22:1262-1274.
19. Shivashankar R, Lichtenstein GR. Mimics of inflammatory bowel disease. *Inflamm Bowel Dis*. 2018;24:2315-2321.
20. Chen D, Fulmer C, Gordon IO, et al. Application of artificial intelligence to clinical practice in inflammatory bowel disease – what the clinician needs to know. *J Crohns Colitis*. 2021;16:460-471.
21. Javaid A, Shahab O, Adorno W, Fernandes P, May E, Syed S. Machine learning predictive outcomes modeling in inflammatory bowel diseases. *Inflamm Bowel Dis*. 2021;28:819-829.
22. Stafford IS, Gosink MM, Mossotto E, Ennis S, Hauben M. A systematic review of artificial intelligence and machine learning applications to inflammatory bowel disease, with practical guidelines for interpretation. *Inflamm Bowel Dis*. 2022;28:1573-1583.
23. Sutton RT, Zai Ane OR, Goebel R, Baumgart DC. Artificial intelligence enabled automated diagnosis and grading of ulcerative colitis endoscopy images. *Sci Rep*. 2022;12:2748.
24. Chierici M, Puica N, Pozzi M, et al. Automatically detecting Crohn's disease and ulcerative colitis from endoscopic imaging. *BMC Med Inform Decis Mak*. 2022;22:300.
25. Kraszewski S, Szczurek W, Szymczak J, Reguła M, Neubauer K. Machine learning prediction model for inflammatory bowel disease based on laboratory markers. Working model in a discovery cohort study. *J Clin Med*. 2021;10:4745.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Guimarães P, Finkler H, Reichert MC, et al. Artificial-intelligence-based decision support tools for the differential diagnosis of colitis. *Eur J Clin Invest*. 2023;53:e13960. doi:10.1111/eci.13960