Reviewer #1: 1. Please include machine learning as a key word 2. What criteria were used to select literature to review?

Machine learning and deep learning are now added as a keyword. No specific criteria were used during the literature search as this is a opinion review. However, the study group did do a thorough research before the manuscript was written.

Reviewer #2: This opinion review presents an important and upcoming concept of utilizing artificial intelligence for the assessment of inflammatory bowel disease. This disorder has unclear etiology and heterogenous diagnostic tools with various confounding factors including skill of the endoscopist. Therefore diagnostic accuracy is required in order to stage and subsequently manage the condition to improve patients' quality of life as well as limit disease complications. However, to provide useful information, artificial intelligence models need to be user friendly and easy to apply, and hence this is an area where much more prospective research with larger data set is needed. The table proposed by the authors can also serve as a useful guide and standard for futures articles on the subject.

We agree upon the reviewer's assessment and we are glad that the reviewer agree on our findings.

Reviewer #3: The Abstract is very defective in describing the aim and the main contents of the article. What do you mean by the abbreviation CTI?

Thank you very much for your comments. We have now corrected the aim in the abstract to be more specific in regard to the main content. Furthermore, we have now spelled out CTI and MRI as these abbreviation were not clear.

Reviewer #4: Please find below my comments regarding this manuscript:

Line 115: "In addition, the group used only a training and a validation set but not a test set to assess whether the algorithm was overfitted". I do not completely agree with this statement. From what I understand from this manuscript, Maeda et al. have trained a model on a dataset denoted as the "training set", and have validated its results on a dataset, denoted as the "validation set", consisting of data unseen by the model during training. I agree that the term "validation set" commonly denotes the dataset used for optimizing a model (e.g. hyperparameters grid search), while the "test set" commonly denotes unseen data used only for assessing the performance of the final model. However, the way they describe their study makes me think that the dataset they denote as a "validation set" is actually a "test set", since they do not mention any data used for comparing architectures or hyperparameters configurations. However, it may be interesting and pertinent to highlight the fact that this study, as many others, did not validate its results on an independent cohort analyzed by independent experts, in order to test the performance of their model when compared to another population or to the point of view of different experts.

Thank you very much for the comment. We do agree with your perspective on the study after further dissection of the study. We do also agree with your point of view and therefore added a comment on this.

Line 205: "Often, only AUC is reported, which can be misleading as sensitivity, specificity and accuracy may be only modest." I believe this sentence does not exactly point out why the presentation of such metrics (sensitivity, specificity, accuracy…) is important, rather than presenting only the ROC-AUC. Usually, this is important because the ROC-AUC evaluates the performance of a model's output regardless of any threshold, and thus does not allow assessing the consistency of this output on different datasets. For instance, if a case-control prediction model consistently outputs 0% for control samples, but outputs 100% for case samples from the training set, and 50% for case samples from the test set: this model will have a ROC-AUC of 1 for both datasets, but its sensitivity will decrease from 100% on the training set to 0% on the test set. Maybe the authors could better highlight the fact that the ROC-AUC only allows assessing the performance of a model on a single cohort, and does not allow assessing the consistency of its results throughout multiple datasets.

Thank you for this comment. This is a statistical aspect of a model development, and we understand the importance of specifying the differences between some of the measures. We have now added a sentence on the subject and hopefully elaborating why all these metrics are important.

Minor comments: The exact formulation may be improved for points listed below. Please note that I am not a native English speaker. Therefore, please ignore my suggestions concerning English issues if those are not appropriate.

Line 31: maybe prefer "data analysis methods" to "data analyzing methods"

Done

Line 33: "[…] its ability to learn and optimize its from new inputs." sounds a little strange. Maybe replace by "[…] its ability to learn and optimize its predictions from new inputs."

Done

Line 100: "[…] methods such as the convolutional network […]" sounds really strange to me, since convolutional neural networks are rather a type of architecture which belongs to the field of machine learning (or more frequently deep learning, even if these arbitrary definitions are a bit vague) than a unique and homogeneous method. Prefer perhaps: "[**...**] methods such as convolutional neural networks" or even "[**...**] methods such as deep convolutional neural networks [**...**]".

We agree and it is now added

Line 111: "support vector machine" instead of "support vector machine learning" Table 1: maybe replace "Generalisability" by "Generalizability"

This is now corrected

2 Editorial Office's comments

1) Science Editor: 1 Scientific quality: The manuscript describes an Opinion Review of the artificial intelligence in inflammatory bowel disease. The topic is within the scope of the WJG. (1) Classification: Grade C, C, C and C; (2) Summary of the Peer-Review Report: This opinion review presents an important and upcoming concept of utilizing artificial intelligence. More prospective research with larger data set is needed. Some sentences need to be rephrased. The questions raised by the reviewers should be answered; (3) Format: There is 1 table; (4) References: A total of 50 references are cited, including 24 references published in the last 3 years; (5) Self-cited references: There is 1 self-cited references. 2 Language evaluation: Classification: Grade A, B, B and B. The manuscript is reviewed by a English editor. 3 Academic norms and rules: No academic misconduct was found in the Bing search. 4 Supplementary comments: This is an invited manuscript. No financial support was obtained for the study. The topic has not previously been published in the WJG. (13)5 Issues raised: PMID numbers are missing in the reference list. Please provide the PubMed numbers and DOI citation numbers to the reference list and list all authors of the references. Please revise throughout. 6 Recommendation: Conditional acceptance.