Point-by-point response to reviewer comments
"Artificial intelligence in gastrointestinal oncology: current applications and future directions "

We thank the editors and reviewers for their time and for the thoughtful commentary regarding our manuscript. Please see our point-by-point response to the reviewers below, with comments in italics. The changes made in the revised manuscript are bolded.

Reviewer #1:

The topic describe in the article is very interesting and very well documented. For sure AI represents a good help in our practice especially in early detection of gastrointestinal neoplasia and I sure hope it will represent the near future.

We thank the reviewer for a thorough review of our manuscript and for their feedback.

Reviewer # 2

Comment 1: It is better that the authors include and compare CNN architectures of recent studies related to gastrointestinal malignancies.

We thank the reviewer for a comprehensive review of our manuscript and for their feedback. We have created a new table (Table 1, produced at the bottom of this letter), which includes the CNN architectures used in all randomized trials applying CADe to colonoscopy.

Comment 2: The authors must include the detail of datasets which commonly used to train the deep learning-based models in recent articles (present in Table).

We agree with the reviewer that this is valuable information to include in our study. We have created a new table (see above), which includes the details of datasets used to train the models used in all randomized trials applying CADe to colonoscopy.

Comment 3: The authors can write one paragraph about the standard endoscopy vs capsule endoscopy.

As per the suggestion of the reviewer, we have included the following paragraph comparing standard endoscopy to capsule endoscopy in the capsule endoscopy section:

Traditional endoscopic techniques allow for the visualization of the esophagus, stomach, duodenum, terminal ileum, and colon. With the advent of push enteroscopy, we have the ability to reach the proximal jejunum, but are still unable to explore most of the small intestine. Capsule endoscopy (CE) uses a 26 x 11 mm pill sized video camera that is swallowed and allows for the wireless transmission of video from the whole GI tract. CE allows for visualization of portions of the jejunum and ileum previously unreachable. Unlike traditional endoscopy, CE is unable to be controlled by an operator so important pathology can be missed,

and there is no way to intervene immediately if an abnormality is identified. CE is also limited by an eight-hour battery life and the risk of obstruction in patients with strictures. Even with its limitations, CE has become an important tool for the diagnosis of GI pathology.

Comment 4: Which performance evaluation metrics are usually utilizing in gastrointestinal abnormalities studies?

As per the suggestion of the reviewer, we have included the following passage discussing performance evaluation metrics in GI studies:

Trials applying AI in GI oncology typically report the following metrics: sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, precision and area under the receiver operating characteristic curve (AuROC). In order to measure the performance of a detection method or segmentation task, the intersection over union (IoU) can be calculated by dividing the area of overlap (overlap of prediction label and ground-truth labels) by the area of union (area of both the predicted and ground-truth labels). The IoU varies from study to study, and a predetermined threshold is typically set to determine true positive (TP) and false positive (FP). Often an IoU ≥ 0.25-0.5 defines a true positive (TP) and an IoU < 0.25-0.5 is considered a false positive (FP). Many prospective studies use a clinical definition of true positive as the number of correctly identified lesions by either AI or endoscopists. Using the discussed parameters, various AI based approaches for the detection of GI cancers can be compared.

Comment 5: In this article, the focus of the authors on traditional machine learning algorithm such as SVM, it is suggested that the authors also include the recent articles related to human gastrointestinal tract abnormalities based on DCNN such as Imran Iqbal et al. and Timothy Cogan et al.

We would be happy to include the Imran Iqbal et al. article about DCNN applied to GI pathology but were unable to locate this article. We were able to find articles relating to skin lesions and morphological classification of human sperm heads by Dr. Aqbal. We have included the article by Cogan et al. entitled MAPGI: Accurate identification of anatomical landmarks and diseased tissue in gastrointestinal tract using deep learning. We have also cited papers implementing deep (> 10 layers) CNNs such as: Hirasawa et al., 2018; Li et al., 2020; Zhu et al., 2019; Wu et al., 2019; Wu et al., 2021; De Groof AJ et al., 2019; Hashimoto R et al., 2020; Horie et al., 2019; Cai et al., 2019; Guo et al., 2019; Ohmori et al., 2019; Liu et al., 2020; Fukuda et al., 2020.

Comment 6: The authors should mention which pre-processing and data augmentation operations are commonly applied in recent studies?

As per the suggestion of the reviewer, we have included the following passage describing pre-processing and data augmentation techniques used by recent studies in the Definitions section of our article:

Preprocessing refers to the methods applied to images prior to analysis by the machine learning model. Techniques include histogram equalization to adjust contrast and gaussian filtering to remove noise. Transformation of the images can be achieved via resizing and processing through multiple layers, where deeper layers typically contain an increasing number of dimensions.

Data augmentation is a process to artificially enlarge a dataset when developing an AI algorithm. It is typically performed via rotation, flipping, shear, and zoom of the original data, thus expanding the amount of data in the training dataset.

Comment 7: Which criteria in recent researches used to consider their result a TP (true positive) or FP (false positive)? Such as more than 50% IoU (Intersection over Union) between the GT (ground truth) and prediction is considering a TP.

As per the suggestion of the reviewer, we included a discussion of the criteria used when considering TP and FP as well as a discussion regarding IoU and a range of cutoffs typically used in the field.

In order to measure the performance of a detection method or segmentation task, the intersection over union (IoU) can be calculated by dividing the area of overlap (overlap of prediction label and ground-truth labels) by the area of union (area of both the predicted and ground-truth labels). The IoU varies from study to study, and a predetermined threshold is typically set to determine true positive (TP) and false positive (FP). Often an IoU ≥ 0.25-0.5 defines a true positive (TP) and an IoU < 0.25-0.5 is considered a false positive (FP). Many prospective studies use a clinical definition of true positive as the number of correctly identified lesions by either AI or endoscopists. Using the discussed parameters, various AI based approaches for the detection of GI cancers can be compared.

Comment 8: Traditionally, "random" biopsies were obtained with a relatively low diagnostic yield as lesions concerning for neoplasia in patients with Barrett's esophagus (BE) are often challenging to identify (figure 3). There are only two figures in the manuscript. There is no figure 3.

We were searching for a representative figure for Barrett's esophagus, but unfortunately were unable to obtain one. Our hope was that the journal could provide us with an image Barrett's esophagus that we can include as discussed in the cover letter. However, we have removed figure 3 and do not think that it detracts from the work in its current form.

Comment 9: Some of the sentences are too long which need to short enough to convey proper meaning.

As per the suggestion of the reviewer, we have shortened several sentences throughout the manuscript.

Comment 10: I would suggest that the authors must add the comparison and detail of "number of filters", "number of parameters" etc. of recent deep learning-based methods such as. Taruna

Agrawal et al., Konstantin Pogorelov et al., Timothy Cogan et al. and Imran Iqbal et al. methods for human gastrointestinal tract abnormalities.

We once again thank the reviewer for their time. We have included the above references and hope that we have included enough detail regarding CNN architecture and parameters discussed in the response to comments 1, 2 and 5.

Reviewer #3

Comment 1: There is no mention of the limitation of the subject.

We thank the reviewer for a comprehensive review of our manuscript and for their feedback. We agree with the reviewer that it is important to include the limitations of this subject. We have added to the following paragraph in the "Conclusion and future directions section":

*This field is growing rapidly, but it is still in its infancy ...**Additionally, it will be important to monitor the efficacy of these tools in the real-world setting. Finally, clinicians will need to collaborate with lawmakers and other stakeholders to determine how best to regulate these technologies and establish clear policies on accountability. In clinical practice today, AI serves as a "safety net" for physicians. It is there to serve as a second set of eyes to support a diagnosis only. We believe it will be many years before AI is used to make definitive diagnostic or drive management decisions.***

In addition, we hope we have addressed limitations of individual studies in the following passages of the manuscript:

"Although studies in this field have demonstrated excellent diagnostic characteristics, many have limited external validity."

"Although these findings are promising, these trials have several limitations. First, the augmented ADR seen in these trials was largely driven by improved detection of diminutive adenomas (size < 5 mm), the clinical benefit of which remains an area of active debate[29]. Secondly, only one trial was double-blinded[24]. In the single-blind trials, being observed may have facilitated a "competitive spirit" or Hawthorne effect in provider participants, leading to improved inspection techniques[10]. Third, all but one of these trials were performed at a single center[20]. Thus, the results of these studies may not be broadly generalizable."

"Kanesaka et al. demonstrated the power of SVM relating to detection of gastric cancer but their study was limited by its sample size (81 test images), lesion type (focused only on depressed-type lesions), and selection bias[41]"

"Major limitations include a small sample size, lack of validation and testing on video or live endoscopy, and the fact that the data was collected from a single center using a single type of endoscope."

"most studies in this field are still retrospective. Furthermore, the majority of datasets used to train the algorithms used in these studies were collected from single-center databases in heterogenous patient populations. Consequently, these studies are at high risk of selection bias and with models at risk for overfitting."

Comment 2: The scope of the title is too broad. In order to promote readers' understanding, the title should be changed to "Use of AI in diagnosis using endoscopes" since the target devices are limited.

We agree with your review and have changed the title of the paper to: Scoping out the future: The application of artificial intelligence to gastrointestinal endoscopy

Comment 3: In the text, there is little discussion on whether AI should be used to "make a diagnosis" or "support diagnosis". Although the policy may differ from country to country, it is necessary to discuss the issue including the opinions of clinical practice and ethical aspects. There are many objective evaluations, but I would like to see a description of the advantages and disadvantages of AI when compared to skilled doctors.

We thank the reviewer for bringing up this important point. We have made the following modification to our conclusion to address this point as also described above:

"In clinical practice today, AI serves as a "safety net" for physicians. It is there to serve as a second set of eyes to support a diagnosis only. We believe it will be many years before AI is used to make definitive diagnostic or drive management decisions."

Comment 4: Figure 3 is shown in the text, but the figure is not attached.

We were searching for a representative figure for Barrett's esophagus but unfortunately were unable to obtain one. Our hope was that the journal could provide us with an image Barrett's esophagus that we can include as discussed in the cover letter. However, we have removed figure 3 and do not think that it detracts from the work in its current form.

Comment 5: In the sentence referring to reference 39, there is a reference to Rie et.al. Is this the first name of the first author?

Rie was the first name of the author, we have corrected the reference and apologize for the oversight.

Comment 6: In the paragraph before Endoscopeic Ultrasound, the last sentence ends with "," instead of ".".

We have corrected this error and apologize for the overs