MedAI: Transparency in Medical Image Segmentation

Steven A. Hicks¹,², Debes Jha¹,³, Vajira Thambawita¹,², Pål Halvorsen¹,², Bjørn-Jostein Singstad⁴, Sachin Gaur⁵,⁶, Klas H. Pettersen⁴, Morten Goodwin⁷, Sravanthi Parasa⁸, Thomas de Lange⁹,¹⁰,¹¹, and Michael A. Riegler¹,³


Abstract
MedAI: Transparency in Medical Image Segmentation is a challenge held for the first time at the Nordic AI Meet that focuses on medical image segmentation and transparency in machine learning (ML)-based systems. We propose three tasks to meet specific gastrointestinal image segmentation challenges collected from experts within the field, including two separate segmentation scenarios and one scenario on transparent ML systems. The latter emphasizes the need for explainable and interpretable ML algorithms. We provide a development dataset for the participants to train their ML models, tested on a concealed test dataset.

Keywords: artificial intelligence; machine learning; segmentation; transparency; medicine

Introduction
Medical image segmentation is a topic that has gained a lot of attention over the last few years. Compared to classification and object detection, segmentation gives a more precise region of interest for a given class. This method is immensely useful for medical doctors as it specifies that an image contains something interesting and where to look. Colonoscopies are a perfect use case for medical image segmentation as they contain a great variety of findings that may be easily overlooked during the procedure. Furthermore, transparent and interpretable machine learning (ML) systems are important to explain the why and the how of the predictions. Such systems are even more important in high-risk fields such as medicine, as conclusions based on wrong decisions resulting from biased or incorrect data, faulty evaluation, or simply a poor model could be fatal. Despite these pitfalls, medical artificial intelligence (AI) research is often published with closed-source software, private data, lackluster evaluation, and often poor or ambiguous method descriptions [1].

To promote more transparency in medical AI research, we present the MedAI: Transparency in Medical Image Segmentation task that aims to develop automatic segmentation systems for segmenting findings in the gastrointestinal (GI) tract. Submissions will be evaluated based on their predictive performance and the level of transparency of the work. The task is relevant to anyone working within the ML research community, and we especially welcome young researchers who want to participate and learn about developing transparent medical AI systems.

Dataset Details
We provide the participants with two open image segmentation datasets, one for polyp segmentation [2] and one for GI instrument segmentation [3]. The datasets are split between a development part and a testing part, where the testing part will be held secret until after the challenge has ended. Both datasets were collected from real colonoscopies performed at Bærum Hospital, Vestre...
Viken Hospital Trust in Norway, and the annotations were verified by expert gastroenterologists (clinicians). In the following, we give a brief description of each dataset and give some details on what the participants can expect from the testing parts.

**Polyp Segmentation Dataset**

The polyp segmentation dataset is based on the public dataset Kvasir-SEG\(^1\) [2], which contains 1,000 images of colon polyps with corresponding segmentation masks and bounding box coordinates. In the image masks, white pixels depict the parts of the image that contain a polyp, and the black pixels are background. The bounding box coordinates are defined as the outermost corners of the detected polyp. Examples from this dataset with corresponding image masks and bounding boxes can be seen in Figure 1. The testing part will consist of 300 images taken from a similar distribution to that of the development dataset.

**Instrument Segmentation Dataset**

The instrument segmentation dataset is based on the open dataset Kvasir-Instrument\(^2\) [3], which contains 590 images of endoscopic tools such as snares, balloons, and biopsy forceps. Each image is accompanied by a corresponding segmentation mask and bounding box coordinates. Examples from this dataset with corresponding image masks and bounding boxes can be seen in Figure 2. The testing part will contain 300 images containing the similar instruments as found in the development dataset.

---

\(^1\)https://datasets.simula.no/kvasir-seg
\(^2\)https://datasets.simula.no/kvasir-instrument

---

Figure 1: Examples taken from the development part of the polyp segmentation dataset Kvasir-SEG. Note that the images have been resized from their original image dimensions.

### Task Descriptions

We present three subtasks: the polyp segmentation task, the instrument segmentation task, and the transparency task. Each task targets a different requirement within automatic findings segmentation in GI image analysis. The participants are encouraged to submit to all subtasks, but it is not a requirement.

#### Polyp Segmentation Task

The *polyp segmentation task* aims to solve the requirement of providing precise segmentation masks of polyps found in the GI tract. This can be used by the doctors to both discover the polyp during the examination and in verifying that it has been fully removed once resected. Participants are asked to develop algorithms for generating precise segmentation masks for polyps in images taken from endoscopies. Submissions to this task should be a zip file containing the predicted masks in the same resolution as the input image for each image in the polyp part of the testing dataset. The filename of each mask should be the same as the input image and using the .png file format.

#### Instrument Segmentation Task

The *instrument segmentation task* targets the requirement of generating segmentation masks for GI instruments during live endoscopy procedures. These segmentation masks can be used to aid medical doctors in performing surgeries in the colon like polypectomies by giving the precise position of the instrument. Similar to the polyp segmentation task, the instrument segmentation task asks participants to develop algorithms for segmenting instruments present in colonoscopy videos, and submissions should be a zip file containing predicted masks...
Transparency Task
The transparency task addresses the requirement of transparent research in medical AI. The main focus for this task is to evaluate systems from a transparency point of view, meaning for example explanations of how the model was trained, what data was used, and interpretation of a model’s predictions. Submissions to this task require that the participants partake in either of the aforementioned segmentation tasks as they are used as the basis for measuring transparency. We leave it to the participants to decide what to deliver for this task. Some ideas include rigorous failure analysis of the model, detailed GitHub repository with clear steps for reproducibility, explaining model predictions, and thorough ablation studies.

Evaluation Methodology
Each task will be evaluated using appropriate metrics. Overall, there will be one first place and one second place winner of MedAI, where the choice will be determined based on a combination of the evaluations gathered from each task. In the following, we present how each task is evaluated. Note that the evaluation method for the polyp segmentation task and the instrument segmentation task have been merged under the same section as the evaluation method is the same.

Segmentation Masks
The submitted segmentation masks for the polyp segmentation task and the instrument segmentation task will be assessed using standard evaluation metrics commonly used for segmentation tasks. This includes pixel accuracy, precision, recall, the Dice coefficient, and intersection over union (IoU). Different metrics tell different stories about the performance of a model under specific conditions, which is why we use multiple metrics to evaluate the submissions. The metric which will be used to rank submissions will be the IoU as presented in Equation 5. Each team will receive a .csv file containing the aforementioned metrics, where the code used to evaluate the submissions for both tasks can be found on GitHub. In particular, we use the following metrics:

\[
\text{Pixel accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)
\]

The pixel accuracy is calculated by treating each pixel in the image as a binary classification mask and computing the accuracy from the results (as shown in Equation 1). In the equation above, TP is the true positives, TN is the true negatives, FP is the false positives, and FN is the false negatives for the pixel classification. The pixel accuracy ranges from 0 to 1, where 1 represents a perfect segmentation.

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (2)
\]

The precision (Equation 2) denotes the proportion of pixels that are correctly segmented positive (white) pixels against all positive pixels. The precision ranges from 0 to 1, where 1 means that the whole region containing the polyp was correctly segmented, and 0 denotes the opposite.

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (3)
\]
The recall (Equation 3) is similar to the precision, but is calculated based on the ratio of pixels that are correctly segmented positive (white) pixels against all correctly segmented pixels. The recall is bounded between 0 and 1, where 1 represents perfectly segmenting the polyp, and 0 would be missing the polyp completely.

\[
\text{Recall} = \frac{|A \cap B|}{|A|}
\]

The Sørensen–Dice coefficient (Equation 4), also called just the Dice coefficient or F1 Score, is a similarity metric used to gauge the similarity between two samples A and B. The metric ranges from 0 to 1, where 0 means the two samples are completely different, and 1 means that they are the same.

\[
\text{Dice} = \frac{2 \times |A \cap B|}{|A| + |B|}
\]

The IoU (Equation 5), also called the Jaccard index, is similar to the Dice score and measures the similarity between two samples. Like Dice, the IoU score ranges from 0 to 1.

\[
\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}
\]

Transparency

Submissions to the transparency task will be evaluated using a more qualitative approach compared to segmentation evaluations. A multi-disciplinary team will assess each submission that evaluates the transparency and understandability of the provided solutions. In this context, transparency will be measured by attributes like code availability, the thoroughness of the evaluation, and reproducibility. Each team will receive a report on the transparency of their solution, which will detail what parts were good and what parts may need more clarity.

Discussion and Outlook

This paper presented the first edition of MedAI: Transparency in Medical Image Segmentation task held at the 2021 Nordic AI Meet. The challenge aims to promote transparency in medical AI research by presenting three subtasks on developing segmentation models for detecting polyps and instruments in frames from GI colonoscopies. We hope that this challenge inspires established and young researchers in exploring medical AI research.

Conflict of interest

Pål Halvorsen is a board member of Augere Medical. Thomas de Lange is employed by Augere Medical and a shareholder. Sravanthi Parasa is a consultant at Covidien LP and on the Medical advisory board at Fujifilms.

References

