Development of artificial intelligence algorithm for screening colorectal cancer lesions in routine abdominopelvic CT without bowel preparation

Kim, Seung-seob

Department of Medicine
Graduate School
Yonsei University

Development of artificial intelligence algorithm for screening colorectal cancer lesions in routine abdominopelvic CT without bowel preparation

Advisor Lim, Joon Seok

A Dissertation Submitted
to the Department of Medicine
and the Committee on Graduate School
of Yonsei University in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy in Medical Science

Kim, Seung-seob

June 2025

Development of artificial intelligence algorithm for screening colorectal cancer lesions in routine abdominopelvic CT without bowel preparation

This Certifies that the Dissertation of
Kim, Seung-seob is approved

Committee Chair Shin, Sang Joon

Committee Member Lim, Joon Seok

Committee Member Choi, Kihwan

Committee Member Min, Byung Soh

Committee Member Kim, Sungjun

Department of Medicine
Graduate School
Yonsei University
June 2025

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Professor Joon Seok Lim, for his invaluable guidance and unwavering support throughout the course of this research. I am also sincerely grateful to the members of the thesis committee for generously taking time from their busy schedules to review my dissertation.

My heartfelt thanks also go to Professor Sungwon Kim, whose insights and direction helped lay a strong foundation in the early stages of this study. I am especially grateful to Dr. Hyunseok Seo and Professor Kihwan Choi for their pivotal roles in developing the core AI model that forms the basis of this research.

Lastly, my deepest thanks go to my beloved wife, Soyeon Yong, and our son, Jaewoo Kim, whose love and support have been a constant source of strength throughout this journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Development of artificial intelligence algorithm for screening colorectal cancer lesions in routine abdominopelvic CT without bowel preparation

Background: Unlike CT colonography, routine abdominopelvic CT (APCT) is performed without bowel preparation, which can lead to the occasional oversight of unsuspected colorectal cancer (CRC).

Objective: To develop an AI-based algorithm to detect CRC in contrast-enhanced APCT acquired without bowel preparation.

Methods: 2,662 patients with CRC who underwent APCT before treatment between January 2010 and December 2014 were enrolled to train the AI model. The model was retrospectively tested with internal and external datasets. Both testing datasets comprised APCTs from consecutive patients with or without CRC who underwent CT and colonoscopy within two months at two independent tertiary hospitals between January and June 2018. For reference standard annotation, an expert radiologist labeled bounding boxes enclosing colorectal cancer in each CT axial slice, referencing colonoscopic reports. For CRC detection, a contemporary transformer-based object detection network, i.e., DEtection with TRansformer (DETR), was adapted and trained. The alternative free-response receiver operating characteristic (AFROC) was used to evaluate the performance of the AI algorithm, which was then compared to that of two expert radiologists.

Results: In the internal 841-patient (mean age, 58 years; 92 patients with 93 CT-detectable CRCs) testing dataset, the area under the AFROC curve (AUAFROC) was 0.867. Sensitivity and specificity were 79.6% (74/93; per-lesion) and 91.2% (683/749; per-patient), respectively, at the point of maximal Youden index. In the external 442-patient (57 years; 26 patients with 26 CT-detectable CRCs) testing dataset, AUAFROC was 0.808. Sensitivity and specificity were 80.8% (21/26; per-lesion) and 90.9% (378/416; per-patient), respectively. Two expert radiologists showed sensitivities (73.1% [19/26] vs. 80.8% [21/26]) and specificities (98.3% [409/416] vs. 98.6% [410/416]) similar to each other. When compared to the AI, the sensitivities were similar ($p =$ 0.743 and 1.0, respectively), but the specificities were higher for the human readers ($p < 0.001$, both).

Conclusion: This study demonstrated the potential feasibility of an AI-based algorithm for detecting CRC in unprepared APCT.

Clinical Impact: By assisting radiologists in detecting cancer in patients not clinically suspected of having CRC, the model can improve outcomes, especially in settings with a shortage of expert radiologists.

Key Words: Colorectal Neoplasms; Artificial Intelligence; Deep Learning; DEtection with TRansformer (DETR); Computed Tomography; Automatic Detection.

1. INTRODUCTION

1.1. Colorectal Cancer and Routine Abdominopelvic CT

Colorectal cancer (CRC) is the third most common malignancy and the second most deadly cancer, with an estimated 1.9 million cases and 0.9 million deaths worldwide in 2020[1]. The U.S. Preventive Services Task Force recommends that adults aged 45–75 be screened for CRC by either optical colonoscopy or computed tomography (CT) colonography[2]. The sensitivity and specificity of CT colonography for detecting CRC larger than 1 cm were reported as 82–92% and 83–86%, respectively[3].

The major difference in scanning protocols between CT colonography and routine abdominopelvic CT (APCT) is whether the bowel is prepared with a cathartic agent and then insufflated. Some authors argued that routine unprepared APCT was also reasonably accurate in detecting CRC with the pooled sensitivity, specificity, and accuracy of 72.4%, 83.6%, and 80.3%, respectively[4]. However, their results were not indicative of real-world performance in that the readers were instructed to rate all colonic segments, which is often omitted during the routine CT interpretation process. Another study reported an overall sensitivity of 74.5% for detecting CRC on routine APCT. However, the sensitivity decreased to 65% for tumors measuring 2–3 cm and further dropped to 50% for tumors smaller than 2 cm[5]. To summarize, although routine APCT can detect and diagnose CRC to some extent, its diagnostic accuracy inevitably falls short of that of CT colonography, particularly when lesions are small.

Meanwhile, the reason radiologists miss CRCs on routine APCT is not always attributable to inherent limitations of the scanning protocol. According to one study, the CRC detection rate in routine APCT decreased further in a community hospital setting with general radiologists, with a reported sensitivity of approximately 66%[6]. Upon re-examination of these initially missed cases, 59% were detected in retrospect, increasing the sensitivity to 86%[6]. This result implies that a number of CRCs are likely being missed even when cancers are actually detectable in routine APCT. Two major factors may explain this phenomenon. First, the participating radiologists were general radiologists rather than gastrointestinal imaging specialists. Second, the radiologists may not have thoroughly examined the large bowel, as the clinical indications for the CT examinations in that study were not specifically related to CRC screening or detection. In this regard, some authors have insisted that searching for unsuspected CRC should be included in the routine APCT interpretation process regardless of the original purpose of the CT scan[4]. However, the number of expert radiologists is limited, and the continuous increase in workload, along with eventual burnout, further exacerbates the two aforementioned factors[7-9].

Artificial intelligence (AI) could potentially be used to complement human readers in automating the detection of CRC on routine APCT, reducing the frequency of missed cancers. Indeed, routine abdominopelvic CT has become established as one of the most commonly used imaging tests for a wide spectrum of clinical settings, resulting in accumulation of a massive amount of data for possible model creation[10,11]. The widespread clinical use of routine APCT highlights the potential large impact of an AI tool for CRC detection on these examinations in contrast with tools tuned specifically for evaluation of dedicated CT colonography examinations[12].

The purpose of our study was to develop an AI-based algorithm to automatically detect CRC in routine APCT scanned without bowel preparation, regardless of the reason that APCT was originally performed.

1.2. Object Detection Models: Historical Perspectives and Current Trends

Object detection is a fundamental computer vision task that involves identifying and localizing objects within images. Over the past two decades, object detection models have undergone a remarkable evolution, transitioning from early hand-crafted feature detectors to modern deep learning-based approaches. This evolution has been driven by key technological breakthroughs that improved detection accuracy and speed, enabling wide-ranging applications—from autonomous driving to medical imaging—where reliable object detection is critical. In the medical AI domain, these advances empower systems to detect anatomical structures or lesions in complex images with growing precision.

1.2.1. Early Era: Hand-Crafted Features and Limitations

The earliest object detectors relied on hand-crafted features and simple classifiers. A notable example is the Histogram of Oriented Gradients (HOG) descriptor introduced by Dalal and Triggs in 2005[13]. HOG features encodes local shape information (edge orientations) on a dense grid, which improved invariance to illumination and slight deformations, leading to substantial gains in tasks like pedestrian detection. Building on such features, Felzenszwalb et al. developed the Deformable Part-Based Model (DPM) around 2008[14]. DPM represented objects as a collection of parts (e.g., a car modeled by its wheels, windows, etc.), allowing some deformation, and used an ensemble of part detectors (a “mixture of star models”) for robust detection. DPM achieved state-of-the-art (SOTA) results and won multiple PASCAL Visual Object Classes (VOC) detection challenges (2007–2009), epitomizing the power of the pre-deep learning paradigm. However, by 2010 these methods began to plateau in performance. Despite incremental improvements (e.g., better hard-negative mining and bounding-box refinement in DPM variants), detection accuracy on challenging benchmarks stagnated in the 30–50% mean Average Precision (mAP) range. The limitations stemmed from the reliance on fixed features and exhaustive sliding-window search, which struggled with object variations and were computationally intensive. A new approach was needed to break this ceiling.

1.2.2. The Deep Learning Revolution: R-CNN and Two-Stage Detectors

The resurgence of neural networks in 2012 (exemplified by the success of AlexNet on ImageNet benchmark) hinted that Convolutional Neural Networks (CNNs) could learn richer features for detection[15,16]. Indeed, 2014 marked a turning point with the introduction of Region-CNN (R-CNN) by Girshick et al.[17]. R-CNN was a breakthrough model that brought deep learning to object detection: it generated region proposals using selective search, then extracted a CNN feature vector for each proposed region, and finally classified each region with a linear Suppor Vector Machine (SVM). This two-step approach (proposal then classification) yielded a massive jump in accuracy – for example, raising mAP on PASCAL VOC from ~33.7% (DPM) to 58.5%. Despite its accuracy, R-CNN had clear drawbacks: the need to run a deep CNN on ~2000 proposals per image made it extremely slow (approximately 14 seconds per image even with GPU acceleration). Researchers quickly sought improvements to streamline this process. One improvement was Spatial Pyramid Pooling Network (SPP-Net) by He et al. in 2014, which introduced a spatial pyramid pooling layer to the CNN[18]. SPP-Net allowed feature extraction from arbitrarily sized regions in a single pass, avoiding repeated CNN computations for each proposal. This sped up detection considerably by computing convolutional features once per image and pooling them for each region. Building on this idea, Girshick proposed Fast R-CNN in 2015, which further unified and accelerated the pipeline[19]. Fast R-CNN enabled end-to-end training of the detector by incorporating a Region of Interest (RoI) pooling layer on shared convolutional feature maps. This integration boosted accuracy (mAP ~70% on VOC2007, up from 58.5% with R-CNN) while running orders of magnitude faster. The next milestone was Faster R-CNN developed by Ren et al. in late 2015[20]. Faster R-CNN solved the last major bottleneck by introducing the Region Proposal Network (RPN), a small CNN that generates object proposals inside the network, replacing external proposal methods. By sharing convolutional features between the RPN and the detector head, Faster R-CNN achieved near real-time performance (e.g., 5–17 frames per second depending on the backbone) without sacrificing accuracy. This two-stage "proposal + refinement" framework became the de facto standard for high-accuracy detection, as it efficiently balances precision and speed. Faster R-CNN and its variants (e.g., R-FCN [Region-based Fully Convolutional Networks], Mask R-CNN) dominated benchmarks by achieving high mAP while being faster and more trainable than earlier methods[21,22].

1.2.3. One-Stage Detectors: YOLO and SSD – Emphasis on Speed

While two-stage detectors optimized accuracy, an alternative family of one-stage detectors emerged to maximize speed. The pioneer in this category was You Only Look Once (YOLO), introduced by Redmon et al. in 2015–2016[23]. YOLO formulates object detection as a single regression problem, feeding the entire image through a CNN that directly predicts bounding box coordinates and class probabilities in one evaluation. By eliminating the region proposal step, YOLO achieved unprecedented speed – the original YOLO could run at 45 frames per second (FPS), and a simplified version reached up to 155 FPS. This real-time performance came with a trade-off in localization accuracy and difficulty detecting small objects, as early YOLO versions were less precise than contemporary two-stage methods. Nonetheless, the paradigm shift was significant: object detection became feasible in time-critical applications. In the medical context, such real-time detection can be valuable (for instance, during surgery or live analysis of ultrasound/video endoscopy), provided accuracy meets acceptable levels. Following YOLO, the Single Shot MultiBox Detector (SSD) by Liu et al. in 2016 extended the one-stage idea with improved accuracy[24]. SSD introduced multi-scale feature maps and anchor boxes of various sizes ratios in a single network pass, allowing it to detect objects of different scales more effectively. By making predictions on multiple convolutional layers, each responsible for detecting objects within a certain size range, SSD significantly improved small object detection compared to YOLO, while still operating quickly (e.g., 59 FPS with mAP around 46–48% on Common Objects in COntext [COCO] dataset)[25]. These one-stage detectors democratized object detection, making it more accessible for widespread use. Subsequent versions of YOLO (v2, v3, and beyond) steadily closed the accuracy gap while retaining high speed, incorporating ideas like multi-scale predictions and better backbone networks. By the late 2010s, one-stage and two-stage detectors each offered compelling trade-offs, and the field began focusing on combining their strengths.

1.2.4. Further Advancements: Multiscale Detection and Anchor-Free Models

To further enhance detection performance, researchers addressed remaining challenges such as multi-scale detection and class imbalance. One influential development was the Feature Pyramid Network (FPN) by Lin et al. (2017), which created a top-down architecture to merge high-level semantic information with low-level spatial detail across multiple scales[26]. FPN became a common backbone component for both two-stage and one-stage detectors, bolstering their ability to detect small, subtle objects – a capability highly relevant for medical images (e.g., detecting tiny lesions). Around the same time, Lin et al. also introduced RetinaNet (2017), an one-stage detector that bridged the accuracy gap with two-stage models by addressing class imbalance in training[27]. RetinaNet's key contribution was the focal loss, a modified loss function that down-weights easy negatives and focuses training on hard examples. This innovation allowed one-stage detectors to achieve comparable accuracy to two-stage detectors on challenging datasets (RetinaNet reached ~59% mAP on COCO dataset), without sacrificing much speed. The idea of focusing on rare positive examples and difficult cases is particularly pertinent to medical AI, where positive findings (e.g., tumors) may be sparse in a sea of normal images.

Another trend was the move towards anchor-free detectors to simplify the detection pipeline. Traditional detectors (both two-stage and one-stage detectors) rely on predefined anchor boxes –

a set of default rectangles of various sizes/aspects – as reference points for predictions. Tuning these anchor settings can be tedious and may not generalize well to unusual object shapes. Anchor-free approaches sidestep this by detecting objects via keypoints. CornerNet (Law and Deng, 2018) was an early example that predicted the top-left and bottom-right corner points of bounding boxes and paired them to form detections[28]. It demonstrated that anchor boxes were not the only way, achieving competitive results (~57.8% mAP on COCO) without anchors. CenterNet (Zhou et al., 2019) further simplified this concept by predicting object centers on a heatmap and regressing to object size, essentially treating objects as single points[29]. By eliminating the anchor generation and Non-Maximum Suppression (NMS) steps, CenterNet provided a fully end-to-end pipeline that was both elegant and effective (reaching ~61% mAP on COCO). The success of anchor-free detectors suggested that with strong feature representations, explicit anchoring of boxes was optional. This is encouraging for medical imaging, where defining appropriate anchors for irregular anatomy or lesions can be challenging – letting the network learn to pinpoint objects directly could be advantageous.

1.2.5. Advent of Transformer Architecture

Most recently, transformer-based models have pushed object detection into a new era. Transformers, which excel at modeling long-range dependencies via self-attention, were introduced to vision tasks after their triumph in natural language processing. In 2020, Carion et al. proposed DEtection with TRansformer (DETR), the first fully end-to-end transformer-based object detector[30]. DETR treats object detection as a direct set prediction problem: it uses a transformer encoder-decoder architecture to globally reason over image features and outputs a set of object bounding boxes without needing hand-crafted components like anchor boxes or post-processing with NMS. This novel design proved that competitive detection performance can be achieved with a much simpler training pipeline, albeit with longer training times required for convergence. Follow-up work such as Deformable DETR introduced multi-scale attention mechanisms to improve convergence and performance[31]. Additionally, modern CNN/Transformer hybrid backbones (e.g., Swin Transformer by Liu et al. 2021) have further improved detection accuracy on benchmarks, indicating the continuing evolution of the field[32].

2. MATERIALS AND METHODS (EXPERIMENT #1)

2.1. Construction of the Training Dataset

Among patients histologically diagnosed with CRC, those who had APCT performed before treatment at a tertiary hospital (Severance Hospital) between January 2010 and December 2014 were retrospectively identified. The exclusion criteria are as follows: 1) APCT was performed without intravenous contrast injection, 2) surgical history of colonic resection, and 3) history of endoscopic mucosal resection (or submucosal dissection) of the colon or rectum. The original purpose of the APCT scans—whether they were performed for CRC diagnosis and staging or for reasons unrelated to CRC—was not considered. A total of 2,662 patients (1566 male, 1096 female; mean age, 63±12 years) with 419,059 axial CT slices of portal venous phase were identified. Among them, CRCs were shown in 31,364 axial slices.

2.2. Construction of the Internal Testing Dataset

We identified consecutive patients at the same tertiary hospital (Severance Hospital) who underwent both APCT and colonoscopy within an interval of less than 2 months between January and June in 2018. The exclusion criteria were as follows: 1) APCT was performed without intravenous contrast injection, 2) the colonoscopic result was incomplete for reasons including poor bowel preparation or failed scope passage until terminal ileum, 3) presence of malignant lesion on colonoscopy that was not confirmed as primary colorectal adenocarcinoma, 4) surgical history of colonic resection, and 5) history of endoscopic mucosal resection (or submucosal dissection) of the colon or rectum.

The diagnosis of CRC was determined based on the colonoscopy and pathology reports. Patients were then divided into two groups to test the model’s performance: those with CT-detectable CRC and those without, including cases where CRC was either not diagnosed or diagnosed but not detectable on CT. The location of the cancer was determined based on its most distal end, and its largest axial diameter was measured on CT. The morphology of the cancer was determined based on colonoscopic findings. It was considered polypoid when the height of the mass was over 50% of its lateral diameter. Otherwise, the mass was regarded as annular, and it was further divided based on whether the circumferential extent exceeded 50% of the bowel lumen.

2.3. Determination of the Reference Standard for CRC

A gastrointestinal expert radiologist with 8 years of experience labeled reference standard bounding boxes, enclosing and fitting the CRC as closely as possible in each axial CT slice where CRC was shown, referencing colonoscopic and/or surgical reports. When the tumor and metastatic lymph nodes were conglomerated and thus inseparable, they were labeled together. MIPAV (Medical Image Processing, Analysis, and Visualization, NIH, Bethesda, MD, USA) was used for bounding box labeling.

2.4. Development of the Initial Prototype Model Using the Hourglass Network

The hourglass network is a deep convolutional architecture featuring a symmetric encoder-decoder design that excels in capturing multi-scale features, making it widely used in medical AI for tasks such as image segmentation and anatomical landmark detection. The CT axial images were fed into the hourglass network as input without any preprocessing. The model was trained to place bounding boxes in areas suspected of CRC. Not all images from the training dataset were used; to address class imbalance, only a randomly selected subset of negative slices (without tumors) was included, ensuring approximate 1:1 ratio of positive to negative slices. The loss function was experimentally set to utilize mean squared error. To enhance model sensitivity, the loss function for tumor cases was scaled by a factor of two.

The performance of the trained model was evaluated using the internal testing dataset. The overlap between the model-generated bounding boxes and the reference standard boxes was evaluated on a per-slice basis. The Dice Similarity Coefficient (DSC), precision, sensitivity, and specificity were calculated.

3. RESULTS (EXPERIMENT #1)

3.1. Internal Testing Dataset

A total of 841 patients were enrolled in the internal testing dataset. Among them, 99 patients were histologically diagnosed with primary CRC, three of whom had two synchronous CRCs. Among the total of 102 CRCs, the expert radiologist failed to detect nine cancers from eight patients on APCT even after referencing the colonoscopic reports. The clinical and imaging characteristics of the internal testing dataset are summarized in Table 1.

Table 1. Clinical and imaging characteristics of the internal testing dataset

	Internal testing dataset
Age (y)[*]	58 ± 15
Male : Female	458 : 383
Patients with primary colorectal cancer	99/841 (12)
Number of colorectal cancers	102
Detectable on APCT	93/102 (91)
Undetectable on APCT	9/102 (9)
Patients with CT-detectable colorectal cancer	92/841 (11)
Cancer location, based on the most distal end	
Ascending colon	22/102 (21)
Hepatic flexure colon	13/102 (13)
Transverse colon	8/102 (8)
Splenic flexure colon	1/102 (1)
Descending colon	5/102 (5)
Sigmoid colon	18/102 (18)
Rectum	28/102 (27)
Anus	7/102 (7)
Cancer size at CT (cm)[†]	3.6 (2.7–4.5)
Same or smaller than 2 cm	8/93 (9)
Larger than 2 cm	85/93 (91)
Cancer morphology at CT	
Polypoid	3/93 (3)
Annular, < 50% of bowel lumen	21/93 (23)
Annular, > 50% of bowel lumen	69/93 (74)
Clinical T staging on CT	
cT1	0/93 (0)
cT2	17/93 (18)
cT3	68/93 (73)
cT4a	5/93 (5)
cT4b	3/93 (3)
Suspicious regional lymphnode metastasis on CT	71/99 (72)
Presence of distant metastasis	9/99 (9)

Unless otherwise noted, data are numbers of patients or lesions, with percentages in parentheses.

[*] Data are mean ± standard deviation.

[†] Data are medians, with the interquartile range in parentheses.

Detailed information on the nine undetectable cancers is summarized in Table 2. A majority of these undetectable cases were either polypoid or small in size with a circumferential tumor extent of less than 50% of bowel lumen.

Table 2. Summary of CT-undetectable cancers in the internal testing dataset

Dataset	Sex/Age	Location	Cancer morphology*	Size (cm)
Internal	M/58	Transverse colon	Polypoid	5.1
Internal	M/67	Anus	Annular, < 50%	2.0
Internal	M/65	Rectum	Annular, < 50%	2.0
Internal	F/58	Sigmoid colon	Polypoid	3.7
Internal	M/65	Descending colon	Annular, < 50%	2.0
Internal	M/74	Sigmoid colon	Annular, > 50%	3.0
Internal	F/73	Rectum	Annular, < 50%	1.7
Internal	F/73	Sigmoid colon	Annular, < 50%	1.5
Internal	M/52	Rectum	Annular, < 50%	2.7

*The morphology of the cancer was determined based on colonoscopic findings. It was determined as polypoid when the height of the mass was over 50% of its lateral diameter. Otherwise, the mass was regarded as annular, and it was further divided based on whether the circumferential extent exceeded 50% of the bowel lumen.

3.2. Performance of the Initial Prototype Model Using the Hourglass Network

The performance of the AI model is summarized in Table 3. Althrough specificity was very high (over 0.9), DSC, precision, and sensitivity all fell below 0.7, with particularly lower performance on tumor slices compared to non-tumor slices. This consistent pattern in the metrics suggests that the model has not yet achieved sufficient sensitivity in detecting CRC.

Table 3. Slice-based performance of the hourglass network

	All slices	Tumor slices	Non-tumor slices
DSC	0.6437	0.5527	0.7346
Precision	0.6565	0.5783	0.7346
Sensitivity	0.6926	0.6120	0.7731
Specificity	0.9883	0.9905	0.9860

DSC, dice similarity coefficient.

4. MATERIALS AND METHODS (EXPERIMENT #2)

4.1. Combination of Two Contrary Networks: DETR and Hourglass

To overcome the low sensitivity of the hourglass network, we devised a two-step strategy: first, a high-sensitivity model selects regions with even a slight possibility of CRC, and then only these selected regions are fed into the high-precision model. DETR is a cutting-edge deep learning method for object detection tasks that combines the power of transformers with object detection algorithms, achieving SOTA performance in the ImageNet benchmark[16,30]. DETR combines the power of transformers and object detection algorithms to perform object detection tasks. Traditionally, object detection systems relied on CNNs as the primary architecture. However, transformers, which have been highly successful in natural language processing tasks, have shown promise in computer vision tasks as well. The basic idea behind DETR is to leverage the attention mechanisms of transformers to capture global contextual information and model relationships between different objects in an image. This is in contrast to CNNs, which primarily focus on local features within the image. Transformers excel in modeling long-range dependencies and capturing global relationships, making them suitable for object detection tasks where understanding the context is crucial. Additionally, transformers enable end-to-end training, eliminating the need for intermediate steps like region proposal networks or anchor generation.

We tuned the DETR model to prioritize high sensitivity, while the hourglass model was tuned to favor high precision. The DETR model generated up to four bounding boxes per CT axial image, each with a probability score, to indicate regions suspected of CRC. Overlapping boxes had their scores summed, and only regions with a final score of 0.8 or higher were selected. A new larger box was then created to encompass the selected boxes, cropped, and used as input for the hourglass network.

The performance of the DETR–hourglass model was evaluated using the internal testing dataset on a per-slice basis. The performance of each model was also evaluated separately without combining them. Since the comparison was based on bounding box areas rather than pixel-level segmentation, Intersection over Union (IoU) was calculated instead of DSC.

5. RESULTS (EXPERIMENT #2)

5.1. Performance of the DETR−Hourglass Model

Table 4 summarizes the performance of the tested models. As intended, the DETR–hourglass model achieved the best performance across all metrics, including IoU, sensitivity, and precision. However, we noticed that the performance difference between our combined model and the DETR-only model was minimal. This suggests that attaching the hourglass model to DETR provided little to no performance gain.

Table 4. Slice-based performance of the DETR–hourglass model

	IoU	**Sensitivity**	**Precision**
Hourglass	0.43	0.54	0.56
DETR	0.55	0.66	0.67
DETR–Hourglass	0.56	0.67	0.68

Iou, intersection over union.

6. MATERIALS AND METHODS (EXPERIMENT #3)

6.1. Optimization of the DETR-only Model

We decided to build a new model using only DETR, without the hourglass model, and proceed with its optimization. We pretrained the model weights on the COCO dataset[25]. The number of predicted boxes was experimentally set to five per axial slice (q = 5), and only the box with the highest probability score of cancer presence was chosen in each axial slice. ResNet101 was used as a backbone network, and the dilated convolution method was used. The number of epochs was 20.

To optimize model performance and address class imbalance simultaneously, we adjusted the ratio of positive to negative slices from 1:1 to 2:1. To achieve this, we randomly selected 17,576 axial slices without a labeled reference standard box. We ensured that the number of slices extracted from each patient was as equal as possible to minimize intra-patient dependence. In total, 48,940 axial slices were used as the final training dataset.

6.2. Construction of the External Testing Dataset

We identified consecutive patients at another external tertiary hospital (Severance Hospital) who underwent both APCT and colonoscopy within an interval of less than 2 months between January and June in 2018. The same exclusion criteria used for constructing the internal testing dataset were applied. Clinical and imaging information were analyzed and recorded in the same manner as when constructing the internal testing dataset.

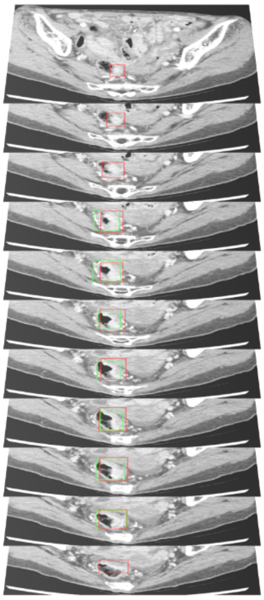
6.3. Improvement of the Model Performance Evaluation Method

6.3.1. Use of DSC Instead of IoU

We decided to use DSC instead of IoU for performance evaluation due to the following reasons. First, IoU is relatively sensitive to the errors in small boxes. Since a few pixel errors in predicting small boxes can result in low scores of IoU[33], IoU might be no longer a suitable metric to detect small lesions in our study. In order to mitigate this issue, DSC can be a good alternative with a large weight for overlapped area between predicted box and reference standard box[34]. Second, we are focused on assessing per-lesion performance of our model. While IoU tends to emphasize the prediction accuracy for large lesions, DSC is adequate for detecting lesions with different scales. Third, our task is closer to z-directional segmentation rather than 3D bounding box prediction. Since we annotated bounding box for each axial slice, the whole volumetric labels can be seen as coarse segmentation masks in coronal or sagittal views. In medical image analysis as well as classical computer vision, DSC is a widely-used metric to measure the segmentation performance[35].

6.3.2. Grouping of AI-predicted bounding boxes

To more accurately evaluate the AI model's performance, we improved the assessment method by shifting from slice-level evaluation to lesion-level evaluation. When the AI-predicted bounding boxes were contiguously present through multiple axial slices and the inner areas of those boxes were overlapped at least partially, those boxes were considered to belong to the same lesion (Figure 1). The arithmetic sum of predicted probability scores of bounding boxes belonging to the same lesion was regarded as the overall probability score corresponding to the lesion. We defined the average DSC as the sum of DSCs of AI-predicted bounding boxes divided by the number of total CT slices with respect to each lesion. An average DSC greater than 0.3 was regarded as true positive.



[Coordinates of Ground truth]		[Coordinates of AI-predicted box]		[AI-predicted Probability]	[DSC]
[Left upper]	[Right lower]	[Left upper]	[Right lower]		
(224 336)	(249 367)	none	none	none	0
(219 334)	(249 368)	none	none	none	0
(211 334)	(248 374)	none	none	none	0
(209 333)	(246 390)	(197 333)	(240 384)	0.66439	0.7409
(208 340)	(246 390)	(196 334)	(243 388)	0.66458	0.7622
(207 333)	(250 390)	(204 339)	(242 383)	0.75030	0.7523
(209 333)	(251 388)	(203 341)	(243 388)	0.67209	0.7678
(206 333)	(251 394)	(208 343)	(246 391)	0.66482	0.8024
(207 335)	(254 395)	(201 340)	(250 392)	0.66558	0.8361
(206 343)	(253 392)	(201 346)	(251 392)	0.66462	0.9014
(206 346)	(255 380)	none	none	none	0

Regarded as one single lesion
- Total number of slices: 7
- Sum of probabilities: 4.7464
- Average DSC: 0.7947

Figure 1. Example of AI grouping multiple bounding boxes into a single lesion. The AI-predicted bounding boxes are shown in green, while the red boxes indicate the reference standard boxes.
DSC, dice similarity coefficient.

6.3.3. Human Reader Study

Human reader study was done using the external testing dataset. Two gastrointestinal expert radiologists with 5 and 10 years of experience, respectively, were requested to independently detect CRC on the external testing dataset. They were not provided any further information regarding the dataset, including the prevalence of CRC patients. No additional CT images, such as other dynamic phases or coronal/sagittal planes, were provided. The expert radiologist who had labeled the reference standard bounding boxes determined whether the reviewers correctly localized CRC by referencing ground-truth images.

6.3.4. Statistical Analysis

Statistical analyses were performed using R, version 4.2.2 (R Foundation for Statistical Computing). The Student's t test was used for age, and a Mann-Whitney U test was used for cancer size at CT. Fisher's exact test was used for categorical variables. For the per-patient analysis, receiver operating characteristic (ROC) analysis was performed. To integrate the per-patient and per-lesion analyses, alternative free-response ROC (AFROC) was performed. To evaluate the localization performance of the model, localization ROC (LROC) analysis was additionally performed[36]. The cutoff for AI-predicted probability score was determined based on maximal Youden index calculated from the AFROC curve on the internal testing dataset. The determined cutoff was used for the external testing as well. $P < .05$ was considered to indicate statistical significance.

Both false-negative lesions (reference standard lesions without any AI-predicted bounding box) and true-negative patients (patients with neither reference standard nor AI-predicted bounding box) were regarded with an overall probability score of zero in the alternative free-response receiver operating characteristic (AFROC) analysis. In localization receiver operating characteristic (LROC) analysis, false-negative lesions were ignored from the analysis. Delong's method was used to estimate the 95% confidence internal (CI) for area under the receiver operating characteristic (AUROC) and area under the AFROC (AUAFROC). Bootstrapping was performed with 1,000 resampling iterations to estimate the 95% CI for area under the LROC (AULROC).

7. RESULTS (EXPERIMENT #3)

7.1. Summary of the All Three Datasets

An overall overview of the training, internal and external testing datasets is summarized in Figure 2. The internal testing dataset was used unchanged as constructed in Experiment #1.

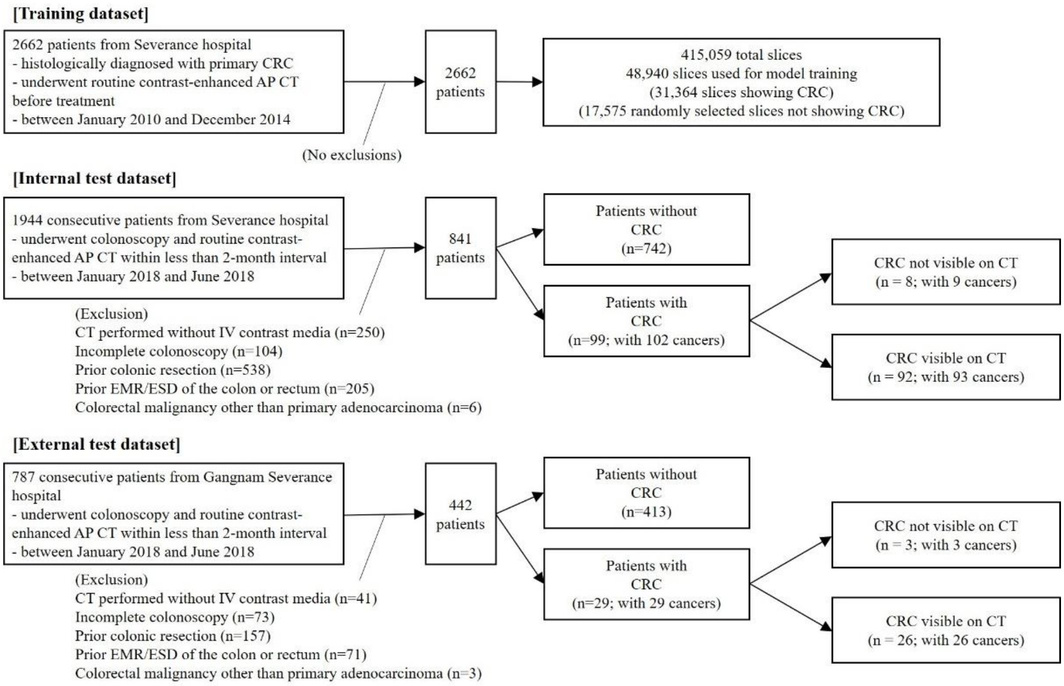

[Training dataset]
2662 patients from Severance hospital
- histologically diagnosed with primary CRC
- underwent routine contrast-enhanced AP CT before treatment
- between January 2010 and December 2014
(No exclusions)
2662 patients
415,059 total slices
48,940 slices used for model training
(31,364 slices showing CRC)
(17,575 randomly selected slices not showing CRC)
[Internal test dataset]
1944 consecutive patients from Severance hospital
- underwent colonoscopy and routine contrast-enhanced AP CT within less than 2-month interval
- between January 2018 and June 2018
(Exclusion)
CT performed without IV contrast media (n=250)
Incomplete colonoscopy (n=104)
Prior colonic resection (n=538)
Prior EMR/ESD of the colon or rectum (n=205)
Colorectal malignancy other than primary adenocarcinoma (n=6)
841 patients
Patients without CRC (n=742)
Patients with CRC (n=99; with 102 cancers)
CRC not visible on CT (n = 8; with 9 cancers)
CRC visible on CT (n = 92; with 93 cancers)
[External test dataset]
787 consecutive patients from Gangnam Severance hospital
- underwent colonoscopy and routine contrast-enhanced AP CT within less than 2-month interval
- between January 2018 and June 2018
(Exclusion)
CT performed without IV contrast media (n=41)
Incomplete colonoscopy (n=73)
Prior colonic resection (n=157)
Prior EMR/ESD of the colon or rectum (n=71)
Colorectal malignancy other than primary adenocarcinoma (n=3)
442 patients
Patients without CRC (n=413)
Patients with CRC (n=29; with 29 cancers)
CRC not visible on CT (n = 3; with 3 cancers)
CRC visible on CT (n = 26; with 26 cancers)

Figure 2. An overall overview of the training, internal and external testing datasets
CRC, colorectal cancer.

A total of 442 patients were enrolled in the external testing dataset. Among them, 29 patients were histologically diagnosed with primary CRC. Each of the 29 cancer patients was found to have a single lesion upon undergoing colonoscopy. The clinical and imaging characteristics of the internal and external testing datasets are summarized and compared in Table 5.

Table 5. Clinical and imaging characteristics of the internal and external testing sets

	Internal testing set	**External testing set**	***P***
Age (y)*	58 ± 15	57 ± 15	.33
Sex			.44
M	458	251	
F	383	191	
Patients with primary colorectal cancer	99/841 (12)	29/442 (7)	.003
Number of colorectal cancers	102	29	
Detectable on APCT	93/102 (91)	26/29 (90)	
Undetectable on APCT	9/102 (9)	3/29 (10)	
Patients with CT-detectable colorectal cancer	92/841 (11)	26/442 (6)	.003
Cancer location, based on the most distal end			.46
Ascending colon	22/102 (21)	5/29 (17)	
Hepatic flexure colon	13/102 (13)	2/29 (7)	
Transverse colon	8/102 (8)	2/29 (7)	
Splenic flexure colon	1/102 (1)	1/29 (3)	
Descending colon	5/102 (5)	0/29 (0)	
Sigmoid colon	18/102 (18)	9/29 (31)	
Rectum	28/102 (27)	10/29 (35)	
Anus	7/102 (7)	0/29 (0)	
Cancer size at CT (cm)†	3.6 (2.7–4.5)	3.2 (2.7–3.9)	.42
Same or smaller than 2 cm	8/93 (9)	2/26 (8)	
Larger than 2 cm	85/93 (91)	24/26 (92)	
Cancer morphology at CT			.30
Polypoid	3/93 (3)	2/26 (8)	
Annular, < 50% of bowel lumen	21/93 (23)	8/26 (31)	
Annular, > 50% of bowel lumen	69/93 (74)	16/26 (61)	
Clinical T staging on CT			
cT1	0/93 (0)	5/26 (19)	
cT2	17/93 (18)	7/26 (27)	
cT3	68/93 (73)	11/26 (42)	
cT4a	5/93 (5)	0/26 (0)	
cT4b	3/93 (3)	3/26 (12)	
Suspicious regional LN metastasis on CT	71/99 (72)	13/29 (45)	
Presence of distant metastasis	9/99 (9)	8/29 (28)	

Unless otherwise noted, data are numbers of patients or lesions, with percentages in parentheses.
*Data are mean ± standard deviation.
†Data are medians, with the interquartile range in parentheses.

The expert radiologist failed to detect three cancers on APCT in the external testing dataset. Detailed information on those three patients is summarized in Table 6. Two of them had polypoid cancers and the remaining patient had small-sized cancer with circumferential tumor extent less than 50% of bowel lumen.

Table 6. Summary of CT-undetectable cancers in the external testing dataset

Dataset	Sex/Age	Location	Cancer morphology*	Size (cm)
External	M/59	Sigmoid colon	Polypoid	2.2
External	F/59	Ascending colon	Polypoid	2.3
External	M/72	Rectum	Annular, < 50%	1.5

*The morphology of the cancer was determined based on colonoscopic findings. It was determined as polypoid when the height of the mass was over 50% of its lateral diameter. Otherwise, the mass was regarded as annular, and it was further divided based on whether the circumferential extent exceeded 50% of the bowel lumen.

7.2. Internal and External Testing

The ROC curves (Figure 3a) and AFROC curves (Figure 3b) were drawn by varying the cutoff for AI-predicted probability score. Regarding the internal testing dataset, when the cutoff for AI-predicted probability score was set to 3.321, Youden index reached its maximum value on the AFROC curve with a per-lesion sensitivity and per-patient specificity of 75.3% (70/93) and 95.1% (712/749), respectively. When the same cutoff was applied to the external testing dataset, sensitivity and specificity were 76.9% (20/26) and 71.2% (296/416), respectively. On LROC analysis, AULROC was calculated as 0.886 (95% CI: 0.829, 0.926) and 0.801 (95% CI: 0.681, 0.876) for the internal and external testing datasets, respectively.

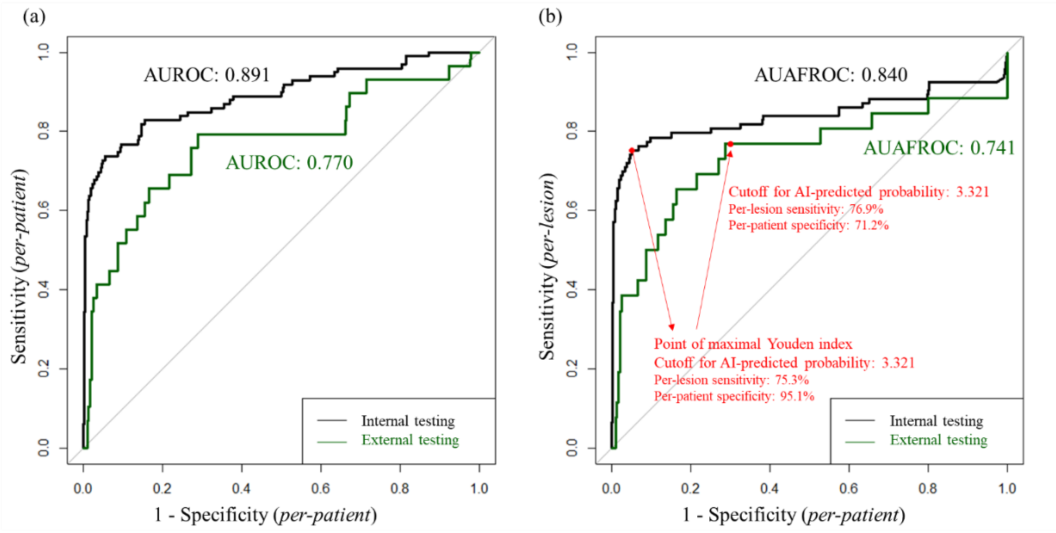


Figure 3. ROC and AFROC analyses (Experiment #3). (a) ROC curves based on the colonoscopic results of internal (black) and external (dark green) testing datasets are shown. AUROC were 0.891 and 0.770 for internal and external testing datasets, respectively. (b) AFROC curves based on the CT-based reference standard box of internal (black) and external (dark green) testing datasets are shown. AUAFROC were 0.840 and 0.741 for internal and external testing datasets, respectively. When the cutoff for AI-predicted probability score was set to 3.321, Youden index reached its maximum value on the AFROC curve of internal testing dataset with a per-lesion sensitivity and per-patient specificity of 75.3% and 95.1%, respectively. When the same cutoff value was applied to the external testing dataset, per-lesion sensitivity was 76.9% and per-patient specificity was 71.2%.
ROC, receiver operating characteristics; AUROC, area under the ROC; AFROC, alternative free-response ROC; AUAFROC, area under the AFROC.

The results of the per-lesion analysis using the cutoff of 3.321 are summarized in Table 7. The AI model showed a sensitivity of 75.3% (70/93) and 76.9% (20/26) for the internal and external testing datasets, respectively.

Table 7. Per-lesion analyses (Experiment #3)

	Internal testing	**External testing**	***P* Value**
True positive	70	20	
False negative	23	6	
Sensitivity*	75.3%	76.9%	.87
False positive	44	179	
Number of false positive lesions per patient			
0	801/841 (95.3)	312/442 (70.6)	
1	37/841 (4.4)	91/442 (20.6)	
2	2/841 (0.2)	30/442 (6.8)	
3	1/841 (0.1)	8/442 (1.8)	
4	0/841 (0.0)	1/442 (0.2)	

Unless otherwise noted, data are numbers of lesions, with percentages in parentheses.
*Sensitivity = True positive / (True positive + False negative)

The model falsely detected 44 lesions in 40 patients and 179 lesions in 130 patients in the internal and external testing datasets, respectively. The detailed locations of false positive lesions are summarized in Table 8.

Table 8. Detailed locations of false positive lesions (Experiment #3)

	Internal testing dataset	**External testing dataset**
Large bowel	25/44 (56.8)	111/179 (62.0)
Ascending colon	14/25 (56.0)	41/111 (36.9)
Hepatic flexure colon	2/25 (8.0)	4/111 (3.6)
Transverse colon	0/25 (0.0)	1/111 (0.9)
Splenic flexure colon	0/25 (0.0)	2/111 (1.8)
Descending colon	0/25 (0.0)	0/111 (0.0)
Sigmoid colon	2/25 (8.0)	23/111 (20.7)
Rectum	7/25 (28.0)	39/111 (35.1)
Anus	0/25 (0.0)	1/111 (0.9)
Stomach	1/44 (2.3)	20/179 (11.2)
Small bowel	10/44 (22.7)	30/179 (16.8)
Uterus	4/44 (9.1)	11/179 (6.1)
Ovary	0/44 (0.0)	1/179 (0.6)
Omentum	0/44 (0.0)	2/179 (1.1)
Kidney	0/44 (0.0)	2/179 (1.1)
Liver	1/44 (2.3)	1/179 (0.6)
Gallbladder	0/44 (0.0)	1/179 (0.6)
Unspecified location	3/44 (6.8)	0/179 (0.0)

Data are numbers of lesions, with percentages in parentheses.

7.3. Human Reader Study

The results of two expert radiologists are summarized in Table 9. Both readers showed similar sensitivities and specificities to each other (sensitivity: 73.1% [19/26] vs. 80.8% [21/26], p = .51; specificity: 98.3% [409/416] vs. 98.6% [410/416], p = .73). When compared to the performance of the AI model, both radiologists showed comparable sensitivity (p = .75 and .73, respectively) but significantly higher specificity (p < .001, both).

Table 9. Performance comparison between radiologists and AI (Experiment #3)

	Reader #1	**Reader #2**	**AI**			
	(5 years of	**(10 years of**	**(cutoff value:**	***P*‡**	***P*§**	***P*‖**
	experience)	**experience)**	**>3.321)**			
Per-patient analysis						
True positive	19	21	20			
False negative	7	5	6			
Sensitivity*	73.1%	80.8%	76.9%	.51	.75	.73
False positive	7	6	120			
True negative	409	410	296			
Specificity†	98.3%	98.6%	71.2%	.73	< .001	< .001
Per-lesion analysis						
True positive	19	21	20			
False negative	7	5	6			
Sensitivity*	73.1%	80.8%	76.9%	.51	.75	.73
False positive	7	6	179			

Unless otherwise noted, data are numbers of lesions.
*Sensitivity = True positive / (True positive + False negative)
†Specificity = True negative / (True negative + False positive)
‡Compared between reader #1 and #2.
§Compared between reader #1 and AI.
‖Compared between reader #2 and AI.

It was highly encouraging that sensitivity improved to a level where it showed no significant difference compared to human expert readers. However, as a trade-off, the increased number of false positives led to lower specificity, which remains a challenge to be addressed.

8. MATERIALS AND METHODS (EXPERIMENT #4)

8.1. Advancement of Model Architecture: Combination of Two DETR Networks and the TotalSegmentator

We explored ways to reduce the number of false positives and noted that, in Experiment #3, approximately 40% of false-positive lesions were located in organs other than the large bowel. Therefore, we refined the model by incorporating a colorectal mask and discarding predictions where the DETR-generated bounding box had no overlap with it.

We cascaded two DETR models for end-to-end training and bounding box prediction, as shown in Figure 4. The first DETR model estimated the coarse bounding box on the original APCT images. We also used prior delineation information to focus on the colorectal area, employing 3D Slicer (version 5.6.0)[37] with the TotalSegmentator tool (version dcfa716b), an AI model that automatically segments 104 anatomical structures from CT images[38]. If there was an overlapping region between the first DETR-estimated box and the colorectal mask, a new rectangle box was generated to encompass both boxes. The longer side of this new box was increased by 60 pixels, and then the shorter side was extended to the same length, converting the box into a larger square. The image inside this larger square box was cropped and interpolated to a resolution of 512 x 512 pixels (the original CT resolution). Then, the second DETR model predicted the fine bounding box on this new input. If the first DETR model did not predict any box, or if there was no overlap between the predicted box and the colorectal mask, a new rectangle box was created to encompass only the colorectal mask. The subsequent steps were carried out in the same manner as previously described: the box size was increased to form a larger square, the image within was cropped and interpolated, and then it was used as input for the second DETR model. After recalculating the coordinates of the second DETR-predicted box from the cropped image to the original CT slice, the overlap with the reference standard box was evaluated. Hyperparameters were unchanged from the previous DETR-only model used in Experiment #3, except that the number of epochs was increased from 20 to 30.

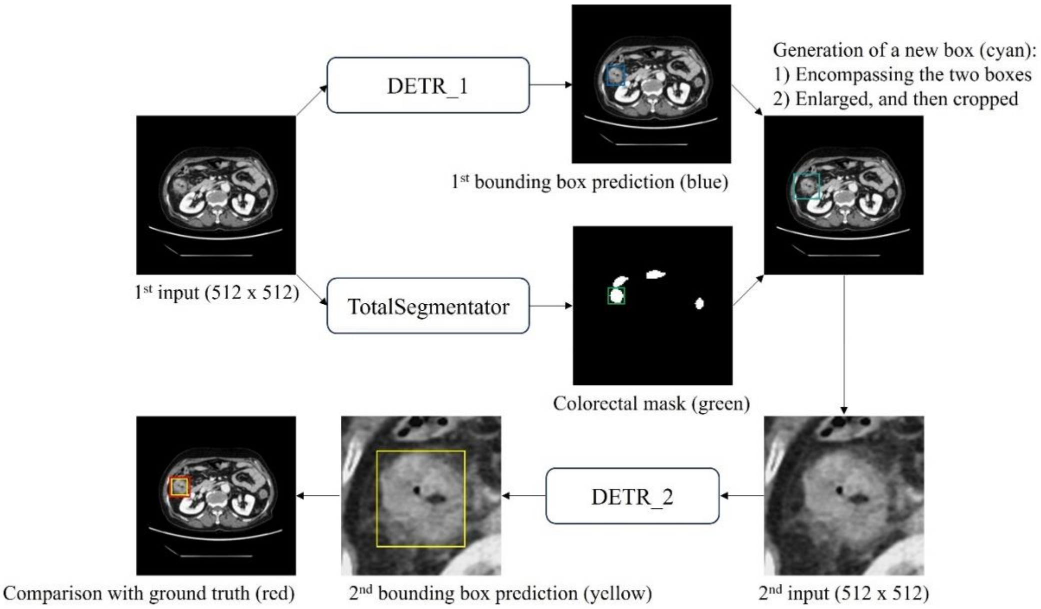

DETR_1
Generation of a new box (cyan):
1) Encompassing the two boxes
2) Enlarged, and then cropped
1st bounding box prediction (blue)
1st input (512 x 512)
TotalSegmentator
Colorectal mask (green)
DETR_2
Comparison with ground truth (red)
2nd bounding box prediction (yellow)
2nd input (512 x 512)

Figure 4. Final Architecture of Our AI model
In our framework, we cascaded two DETR models to learn and predict the bounding box enclosing CRC in axial CT images in an end-to-end fashion. The first DETR model estimated the coarse bounding box (blue) on the original APCT images. We also used prior delineation information to focus on the colorectal area, employing 3D Slicer (version 5.6.0) with the TotalSegmentator tool (version dcfa716b). If there was an overlapping region between the DETR-estimated box (blue) and the colorectal mask (green), a new rectangle box was generated to encompass both boxes. The longer side of this new box was increased by 60 pixels, and then the shorter side was extended to the same length, converting the box into a larger square (cyan). The image inside this larger square box was cropped and interpolated to a resolution of 512 x 512 pixels (the original CT resolution). Finally, the second DETR model predicted the fine bounding box (yellow) on this new input. After recalculating the coordinates of the second DETR-predicted box (yellow) from the cropped image to the original CT slice, the overlap with the ground truth box (red) was evaluated.
CRC, colorectal cancer; DETR, DEtection with TRansformer

8.2. Update of the Model Performance Evaluation Method

When the AI-predicted bounding boxes were contiguously present across multiple contiguous axial slices, the previous approach of considering them as a single lesion was maintained. However, we decided to no longer use quantitative metrics such as DSC or IoU to assess the degree of overlap with the reference standard. Instead, we adopted a binary evaluation method that simply determines whether overlap is present or not. This decision was based on previous experiments, which revealed that even when the degree of overlap with the reference standard box was not high, the model still accurately identified and marked the lesions in most cases. Therefore, a lesion containing any number of AI-predicted boxes that overlap with the reference standard boxes was considered a true positive detection. The previous approach of considering the arithmetic sum of predicted scores of bounding boxes belonging to the same lesion as the overall score of that lesion was maintained. Representative CT slices with the labeled reference standard and the final predicted bounding boxes are shown in Figure 5.

Regarding the statistical analysis method, we decided not to perform the ROC and LROC analyses previously conducted in Experiment #3; instead, we chose to perform only the AFROC analysis. This decision was made because the AFROC analysis alone was considered sufficient to evaluate per-patient and per-lesion aspects, as well as localization performance.

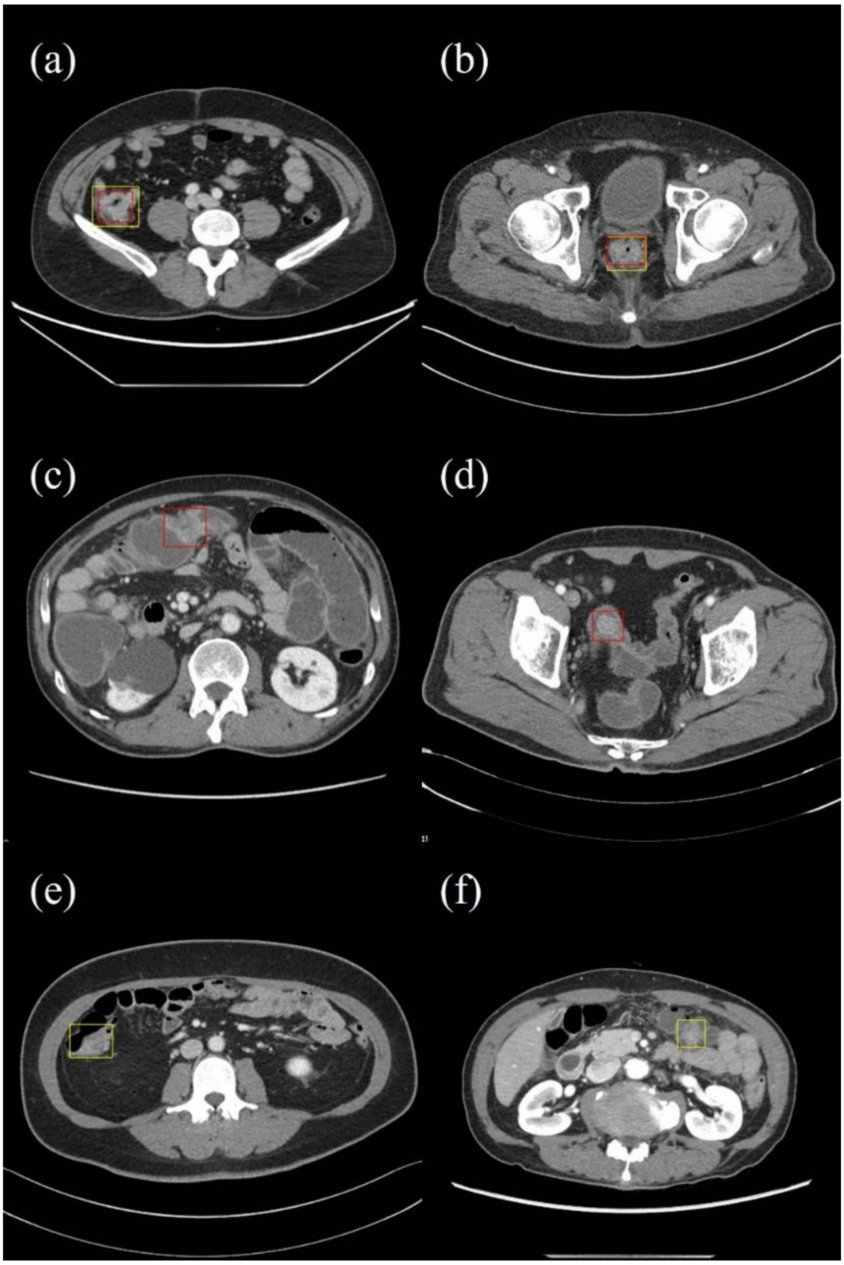


Figure 5. Representative images of reference standard (red) and AI-predicted (yellow) boxes. (a, b) Examples of true positive detections are shown. (c, d) Examples of false negative detections are shown. AI failed to propose a bounding box over the reference standard lesion. (e, f) Examples of false positive detections are shown. AI incorrectly proposed bounding boxes at the area where reference standard box was absent.

9. RESULTS (EXPERIMENT #4)

9.1. Internal and External Testing

The AFROC curves (Figure 6) were drawn by varying the cutoff for the AI-predicted score. The AUAFROC was 0.867 (95% CI: 0.809, 0.924) and was 0.808 (95% CI: 0.661, 0.955) in the internal and external testing datasets, respectively. Regarding the internal testing dataset, when the cutoff for AI-predicted score was set to 3.9996, Youden index reached its maximum value on the AFROC curve with a per-lesion sensitivity and per-patient specificity of 79.6% (74/93) and 91.2% (683/749), respectively. When the same cutoff was applied to the external testing dataset, sensitivity and specificity were 80.8% (21/26) and 90.9% (378/416), respectively.

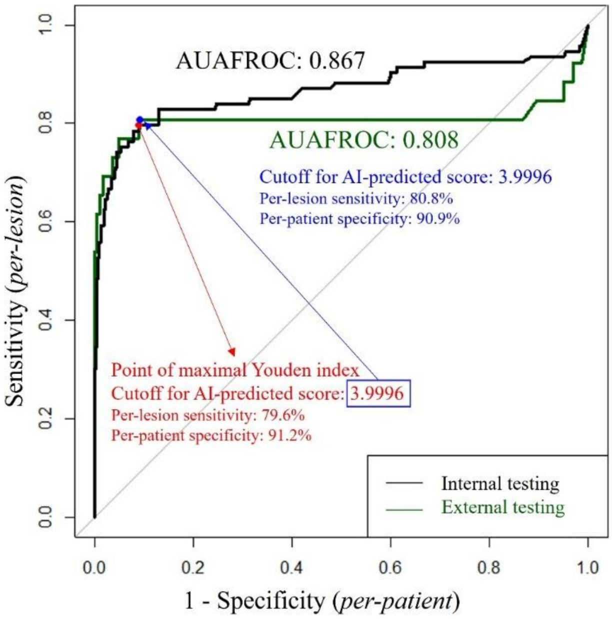


Figure 6. AFROC analysis of the final AI model (Experiment #4).

The results of the per-lesion analysis using the cutoff of 3.9996 are summarized in Table 10. The AI model showed a sensitivity of 79.6% (74/93) and 80.8% (21/26) for the internal and external testing datasets, respectively. Regarding the false positives, the model falsely detected 84 lesions in 70 patients and 52 lesions in 40 patients in the internal and external testing datasets, respectively.

Table 10. Per-lesion analyses (Experiment #4)

	Internal testing	**External testing**
True positive	74	21
False negative	19	5
Sensitivity*	79.6%	80.8%
False positive	84	52
Number of false positive lesions per patient		
0	771/841 (91.7)	402/442 (91.0)
1	60/841 (7.1)	32/442 (7.2)
2	7/841 (0.8)	6/442 (1.4)
3	2/841 (0.2)	0/442 (0.0)
4	1/841 (0.1)	2/442 (0.5)

Unless otherwise noted, data are numbers of lesions, with percentages in parentheses.
*Sensitivity = True positive / (True positive + False negative)

9.2. Performance Comparison between Two Human Readers and the AI Model

Although the specificity of our AI model improved up to 90.9%, yet it still did not reach the performance level of human readers (Table 11). Regarding sensitivity, the AI model continued to show performance comparable to or better than that of human readers.

Table 11. Performance comparison between radiologists and AI (Experiment #4)

	Reader #1 (5 years of experience)	**Reader #2 (10 years of experience)**	**AI (cutoff value: >3.9996)**	***P*[‡]**	***P*[§]**	***P*[‖]**
Per-lesion analysis						
True positive	19	21	21			
False negative	7	5	5			
Sensitivity[*]	73.1%	80.8%	80.8%	0.743	0.743	1.0
False positive	7	6	52			
Per-patient analysis						
False positive	7	6	38			
True negative	409	410	378			
Specificity[†]	98.3%	98.6%	90.9%	1.0	<0.001	<0.001

Unless otherwise noted, data are numbers of lesions.

[*] Sensitivity = True positive / (True positive + False negative)

[†] Specificity = True negative / (True negative + False positive)

[‡] Compared between reader #1 and #2.

[§] Compared between reader #1 and AI.

[‖] Compared between reader #2 and AI.

Table 12 summarizes information on the nine CRCs missed by at least one of the two readers, including three CRCs missed by both readers. AI correctly detected five of these nine CRCs, including one of the three CRCs missed by both readers. Figures 7 and 8 show examples of CRCs missed by both readers and by the AI model; Figures 9 and 10 show examples of CRC missed by at least one reader but detected by the AI model.

Table 12. Cancers missed by at least one of the two expert radiologists

Reader #1 (5 years of experience)	**Reader #2 (10 years of experience)**	**AI (cutoff value: >3.9996)**	**Location**	**Cancer morphology**	**Size (cm)**
X	X	X	Transverse colon	Annular, < 50%	2.8
X	X	X	Rectum	Annular, < 50%	1.8
X	O	X	Sigmoid colon	Annular, > 50%	4.0
X	O	X	Sigmoid colon	Annular, < 50%	2.9
X	X	O	Sigmoid colon	Annular, > 50%	2.7
X	O	O	Hepatic flexure colon	Annular, > 50%	2.8
X	O	O	Rectum	Annular, > 50%	3.2
O	X	O	Rectum	Annular, < 50%	2.8
O	X	O	Rectum	Annular, < 50%	3.2

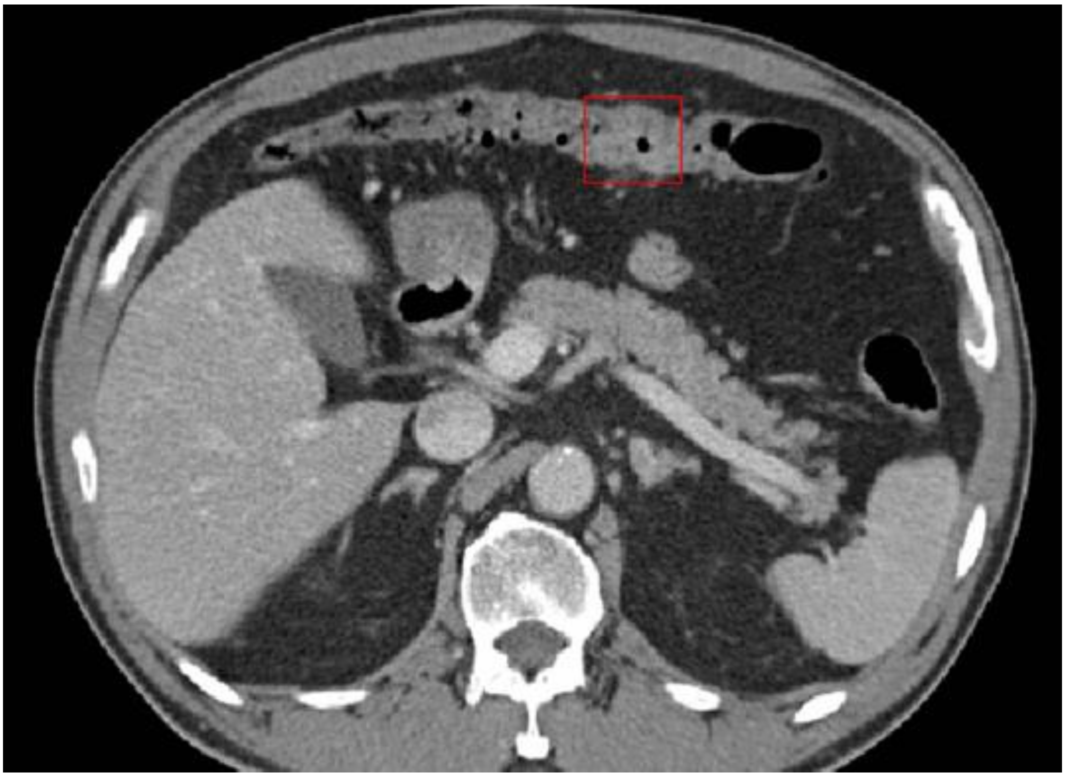


Figure 7. Example of a CRC case missed by both readers and the AI model. This is an axial image from routine abdominopelvic CT examination in 71-year-old patient from external test set with histologically confirmed CRC involving transverse colon (box). Lesion measured 2.8 cm and had annular (not exceeding 50% of bowel lumen) morphology. Lesion was missed by both readers and by AI model.
CRC, colorectal cancer.

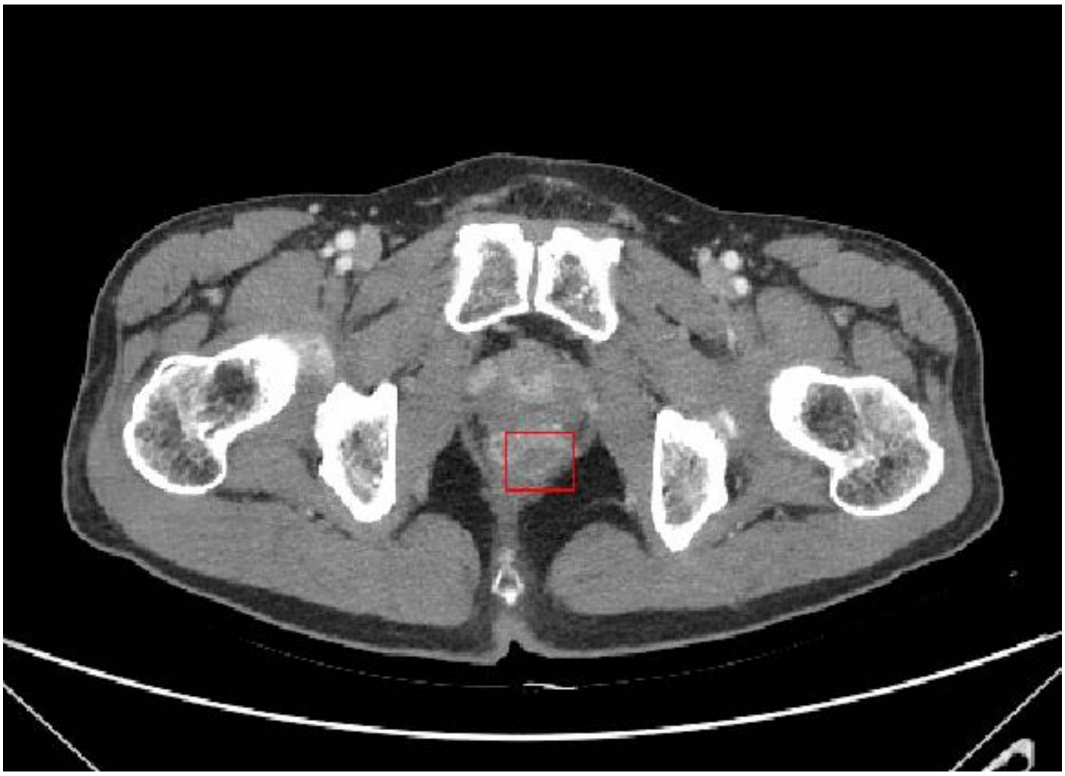


Figure 8. Example of a CRC case missed by both readers and the AI model. This is an axial image from routine abdominopelvic CT examination in 61-year-old patient from external test set with histologically confirmed CRC involving rectum (box). Lesion measured 1.8 cm and had annular (not exceeding 50% of bowel lumen) morphology. Lesion was missed by both readers and by AI model.
CRC, colorectal cancer.

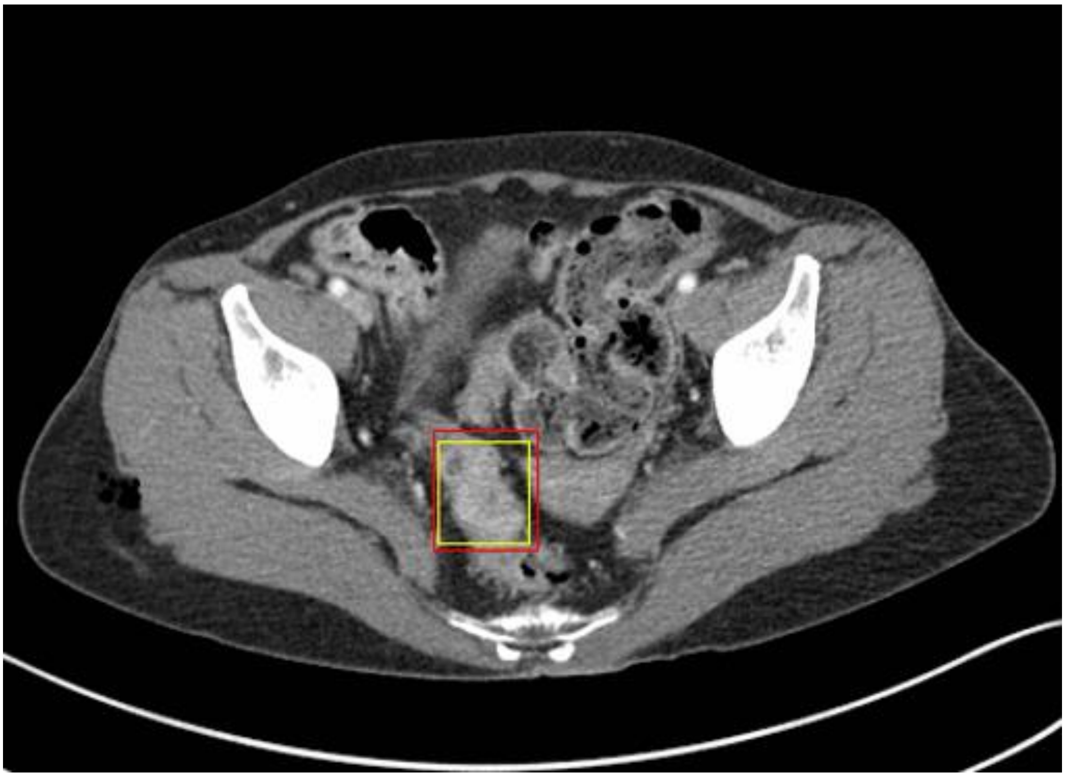


Figure 9. Example of a CRC case missed by at least one reader but detected by the AI model. This is an axial image from routine abdominopelvic CT examination in 57-year-old patient from external test set with histologically confirmed CRC involving sigmoid colon (red box). Lesion measured 2.7 cm and had annular (exceeding 50% of bowel lumen) morphology. Lesion was missed by both readers but detected by AI model (yellow box).

CRC, colorectal cancer.

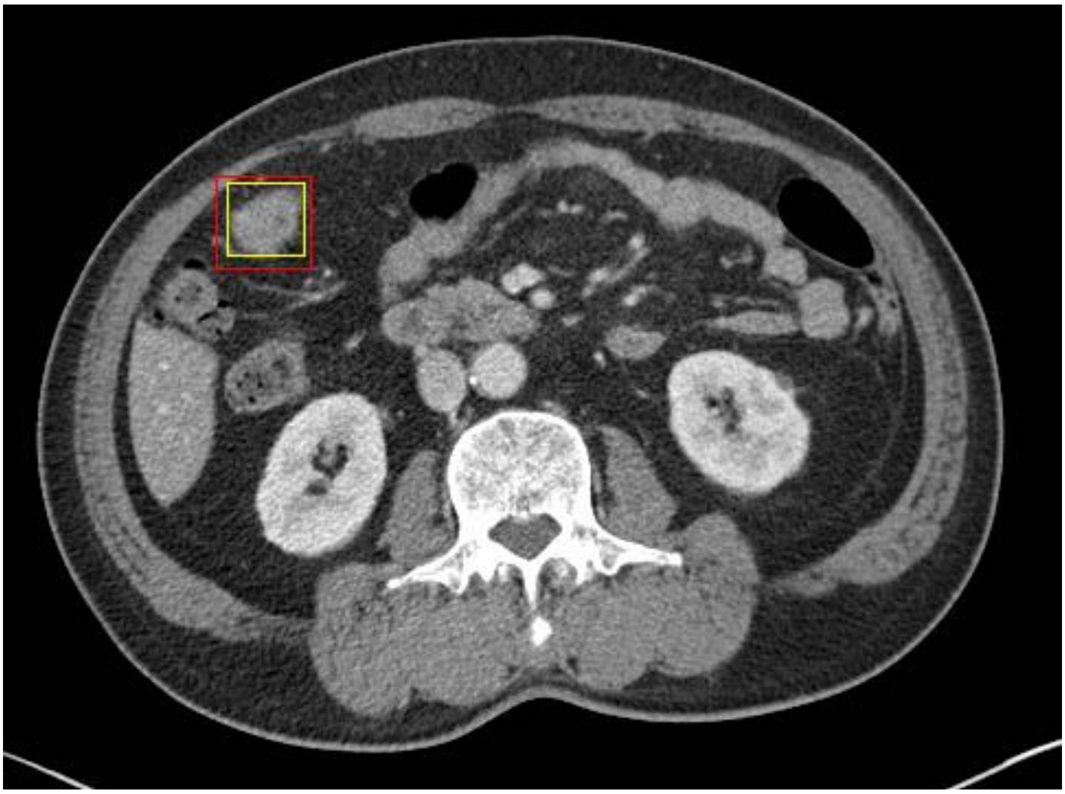


Figure 10. Example of a CRC case missed by at least one reader but detected by the AI model. This is an axial image from routine abdominopelvic CT examination in 76-year-old patient from external test set with histologically confirmed CRC involving hepatic flexure (box). Lesion measured 2.8 cm and had annular (exceeding 50% of bowel lumen) morphology. Lesion was detected by one of two readers and by AI model.
CRC, colorectal cancer.

9.3. Subgroup Analysis

Subgroup analyses of internal and external testing datasets were performed according to the size, morphology, and location of the cancer based on CT (Table 13). Per-lesion sensitivity was higher when diagnosing the annular cancers involving more than 50% of the bowel lumen than those involving less than 50%. Regarding the location of the cancer, the model showed the lowest sensitivity to transverse colon cancer.

Table 13. Subgroup analyses of the internal testing and external testing datasets

	Internal testing dataset	**External testing dataset**
	Sensitivity[*]	**Sensitivity**[*]
All CT-detectable cancers	79.6% (74/93)	80.8% (21/26)
Cancer size at CT (cm)		
Same or smaller than 2 cm	50.0% (4/8)	0.0% (0/2)
Larger than 2 cm	82.4% (70/85)	87.5% (21/24)
Cancer morphology at CT		
Polypoid, or annular (<50%)	58.3% (14/24)	60.0% (6/10)
Annular (>50%)	87.0% (60/69)	93.8% (15/16)
Cancer location, based on the most distal end		
Ascending colon	86.4% (19/22)	100.0% (4/4)
Hepatic flexure colon	84.6% (11/13)	100.0% (2/2)
Transverse colon	28.6% (2/7)	50.0% (1/2)
Splenic flexure colon	100.0% (1/1)	100.0% (1/1)
Descending colon	80.0% (4/5)	*n/a* (0/0)
Sigmoid colon	71.4% (10/14)	62.5% (5/8)
Rectum	92.0% (23/25)	88.9% (8/9)
Anus	66.7% (4/6)	*n/a* (0/0)

[*] Sensitivity = True positive / (True positive + False negative)

10. DISCUSSION

Unanticipated CRCs can easily be missed on routine APCT scans performed without bowel preparation. This issue can be particularly problematic when the primary purpose of the scans is unrelated to CRC screening or when interpretations are performed by general radiologists rather than gastrointestinal imaging specialists. Our final AI model demonstrated the potential for the automatic detection of CRC, achieving AUAFROC over 0.8 in both internal and external testing datasets. The sensitivity of the AI model, at approximately 80%, was comparable to that of human expert radiologists; however, its specificity was still lower, at around 90% compared to 98%.

Regarding the computer-aided detection of CRC, the majority of previous works focused on either optical colonoscopy videos or CT colonography images[39,40]. The objectives of those previous studies using CT colonography were mostly to detect or characterize the polyps, not the cancer itself[41]. Because the detection rate of CRC may be lower in routine APCT than that of CT colonography, there have been unmet needs for the development of computer-aided cancer-detection algorithms based on routine APCT[4,6,42]. Furthermore, routine APCTs generally outnumber CT colonography exams because 1) they do not require any special scanning protocol such as bowel preparation or gaseous distension, and 2) they are more frequently performed for various clinical purposes, not limited to CRC screening. In this context, routine APCT has the advantage of providing large volumes of data required to train deep-learning-based AI models. Furthermore, AI models trained on routine APCT have considerably broader applicability compared to those based on CT colonography.

Recently, a few studies have started to explore AI models for detecting CRC using routine APCT scans performed without bowel preparation. Among these, two studies assessed the model’s performance exclusively in patients with confirmed CRC, without evaluating its efficacy in patients without CRC[43,44]. As a result, the model's real-world performance remains unclear. The other two papers by Yao et al. validated their model performance using a real-world dataset but classified all patients without detectable CRC on CT as normal, introducing a significant limitation[45,46]. A key advantage of AI is its potential to identify CRC cases that radiologists might overlook. Consequently, the patients classified as normal in these studies may have included those with undiagnosed CRC that were missed by radiologists. In our study, we addressed this limitation by including only patients who underwent colonoscopy within two months of their CT scan and classified as normal only those whose colonoscopy results showed no evidence of CRC. This approach allowed for more precise patient categorization and a more accurate performance evaluation.

Our AI model demonstrated good performance in detecting CRC, as the AUAFROC was 0.867 and 0.808 on internal and external testing sets, respectively. Furthermore, the sensitivity of our model, at approximately 80%, was comparable to that of the human expert radiologists. The human readers involved in this study were not simply instructed to interpret the CT scans; they were specifically directed to look for CRC. Consequently, the readers likely concentrated more on the large bowel while reviewing the images than they would in routine real-world practice, which might have resulted in their improved sensitivity. Considering this, the sensitivity of our AI model appears to be quite sufficient for real-world CT interpretation sessions.

In approximately 91% of patients in both testing datasets, there were no false positive lesions. This result is better than that of previous studies regarding CT colonography, which reported two or more false positive lesions per patient[47,48]. One possible explanation for the lower false positivity of our model may lie in our lesion-based approach, which aggregates multiple bounding boxes as a single lesion unit. As we arithmetically added the scores of each slice to calculate the final score for that lesion, the lesions that consisted of a high number of slices were weighted more than the lesions with a small number of slices. We speculate that the underlying mechanism for this approach reflects the way human radiologists determine between pseudo-lesions and true lesions — true lesions usually show their presence consistently across multiple slices. The two-step approach using the colorectal mask obtained through the TotalSegmentator, also possibly contributed to reducing false positives. If the first DETR model suggested a bounding box on organs other than the large bowel, it was discarded before the second DETR model because it failed to overlap with the colorectal mask.

The performance of our AI model was better for annular cancers with circumferential tumor extent exceeding 50% of the bowel lumen, achieving per-lesion sensitivities of 87.0% and 93.8% in the internal and external testing datasets, respectively. Our results are consistent with previous studies that found that longer circumferential tumor extent was associated with more advanced TNM stages and was thus easier to detect[49,50]. Conversely, for small CRCs measuring 2 cm or less, both our AI model and expert readers demonstrated low sensitivities of 50% or less. This is consistent with previous reports, indicating that there are inherent limitations to improving detection sensitivity for such small CRCs using routine CT alone[5]. Regarding the location of the cancer, our model showed the worst performance when the cancer was located in the transverse colon. Although the exact reason that our AI model showed the worst performance in such a location is not clear, one possible explanation is that axial slices of the transverse colon appear rather cylindrical, whereas those of the other colonic parts look circular. In this regard, we speculate that the performance of our model could be improved if sagittal images, where the transverse colon appears circular, were also considered. Investigation of such different views remains as a topic of our future research.

Our research has several limitations. First, the AI model was trained and tested retrospectively. The model's ability to reduce missed cancers in a prospective real-world setting was not evaluated. Second, patients who did not undergo both CT and colonoscopy within a 2-month interval were excluded from the testing dataset, which constitutes a selection bias. However, we believe it is the only retrospective design that allows for constructing the accurate reference standard for non-cancer patients. Third, the impact of false-positive AI results, even if uncommon, was not evaluated. Prospective cost-effectiveness analysis is needed to determine whether the benefits outweigh the costs associated with false positive suggestions. Fourth, coronal/sagittal CT images were not used in this study, which might have further improved the performance of the human readers and the AI. Fifth, this study evaluated performance for detection of only CT-visible CRCs, classifying patients with nonvisible CRCs as negative by the reference standard. This approach is expected to have reduced the prevalence in the study sample of lesions that may be particularly difficult to detect. Finally, the analysis focused only on CRCs and did not consider detection of adenomas or other precancerous lesions. Thus, while the explored approach may help reduce the frequency with which CT-visible cancers are missed, it does not represent an opportunistic screening method for CRC or a substitute for CRC screening tests.

In conclusion, our study demonstrated the potential utility of an AI model for detecting CRC on routine APCT examinations, performed without bowel preparation for reasons unrelated to CRC detection and staging.

References

1. Xi Y, Xu P. Global colorectal cancer burden in 2020 and projections to 2040. *Transl Oncol*. Oct 2021;14(10):101174. doi:10.1016/j.tranon.2021.101174

2. U.S. Preventive Services Task Force - Colorectal Cancer: Screening. Accessed June 22, 2022. https://www.uspreventiveservicestaskforce.org/uspstf/recommendation/colorectal-cancer-screening

3. Johnson CD, Herman BA, Chen MH, et al. The National CT Colonography Trial: assessment of accuracy in participants 65 years of age and older. *Radiology*. May 2012;263(2):401-8. doi:10.1148/radiol.12102177

4. Ozel B, Pickhardt PJ, Kim DH, Schumacher C, Bhargava N, Winter TC. Accuracy of routine nontargeted CT without colonography technique for the detection of large colorectal polyps and cancer. *Dis Colon Rectum*. Jun 2010;53(6):911-8. doi:10.1007/DCR.0b013e3181d5de13

5. Myo K, Manda V, Qi LJ, Rawlings D, Leung E. Sensitivity of routine CT abdomen and pelvis for detecting colorectal concer. *International Journal of Surgery*. 2016/11/01/ 2016;36:S60. doi:https://doi.org/10.1016/j.ijsu.2016.08.149

6. Mangat S, Kozoriz MG, Bicknell S, Spielmann A. The Accuracy of Colorectal Cancer Detection by Computed Tomography in the Unprepared Large Bowel in a Community-Based Hospital. *Can Assoc Radiol J*. Feb 2018;69(1):92-96. doi:10.1016/j.carj.2017.12.005

7. Do KH, Beck KS, Lee JM. The Growing Problem of Radiologist Shortages: Korean Perspective. *Korean J Radiol*. Dec 2023;24(12):1173-1175. doi:10.3348/kjr.2023.1010

8. Fawzy NA, Tahir MJ, Saeed A, et al. Incidence and factors associated with burnout in radiologists: A systematic review. *Eur J Radiol Open*. Dec 2023;11:100530. doi:10.1016/j.ejro.2023.100530

9. Bailey CR, Bailey AM, McKenney AS, Weiss CR. Understanding and Appreciating Burnout in Radiologists. *Radiographics*. Sep-Oct 2022;42(5):E137-E139. doi:10.1148/rg.220037

10. Federle MP. CT of the acute (emergency) abdomen. *Eur Radiol*. Nov 2005;15 Suppl 4:D100-4. doi:10.1007/s10406-005-0123-8

11. Dan Lantsman C, Barash Y, Klang E, Guranda L, Konen E, Tau N. Trend in radiologist workload compared to number of admissions in the emergency department. *Eur J Radiol*. Apr 2022;149:110195. doi:10.1016/j.ejrad.2022.110195

12. Pickhardt PJ, Summers RM, Garrett JW, et al. Opportunistic Screening: Radiology Scientific Expert Panel. *Radiology*. May 23 2023:222044. doi:10.1148/radiol.222044

13. Dalal N, Triggs B. Histograms of oriented gradients for human detection. Ieee; 2005:886-893.

14. Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model. Ieee; 2008:1-8.

15. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*. 2012;25

16. Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*. 2015/12/01 2015;115(3):211-252. doi:10.1007/s11263-015-0816-y

17. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. 2014:580-587.

18. He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*. 2015;37(9):1904-1916.

19. Girshick R. Fast r-cnn. 2015:1440-1448.

20. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*. 2016;39(6):1137-1149.

21. Dai J, Li Y, He K, Sun J. R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems*. 2016;29

22. He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. 2017:2961-2969.

23. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. 2016:779-788.

24. Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector. Springer; 2016:21-37.

25. Lin T-Y, Maire M, Belongie S, et al. Microsoft COCO: Common Objects in Context. Springer International Publishing; 2014:740-755.

26. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. 2017:2117-2125.

27. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. 2017:2980-2988.

28. Law H, Deng J. Cornernet: Detecting objects as paired keypoints. 2018:734-750.

29. Zhou X, Wang D, Krähenbühl P. Objects as points. *arXiv preprint arXiv:190407850*. 2019;

30. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-End Object Detection with Transformers. Springer International Publishing; 2020:213-229.

31. Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:201004159*. 2020;

32. Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows. 2021:10012-10022.

33. Cao C, Wang B, Zhang W, et al. An Improved Faster R-CNN for Small Object Detection. *IEEE Access*. 2019;7:106838-106846. doi:10.1109/ACCESS.2019.2932731

34. Seo H, Bassenne M, Xing L. Closing the Gap Between Deep Neural Network Modeling and Biomedical Decision-Making Metrics in Segmentation via Adaptive Loss Functions. *IEEE Trans Med Imaging*. Feb 2021;40(2):585-593. doi:10.1109/tmi.2020.3031913

35. Eelbode T, Bertels J, Berman M, et al. Optimization for Medical Image Segmentation: Theory and Practice When Evaluating With Dice Score or Jaccard Index. *IEEE Transactions on Medical Imaging*. 2020;39(11):3679-3690. doi:10.1109/TMI.2020.3002417

36. Wunderlich A, Noo F. A nonparametric procedure for comparing the areas under correlated LROC curves. *IEEE Trans Med Imaging*. Nov 2012;31(11):2050-61. doi:10.1109/TMI.2012.2205015

37. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn Reson Imaging*. Nov 2012;30(9):1323-41. doi:10.1016/j.mri.2012.05.001

38. Wasserthal J, Breit HC, Meyer MT, et al. TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. *Radiol Artif Intell*. Sep 2023;5(5):e230024. doi:10.1148/ryai.230024

39. Goyal H, Mann R, Gandhi Z, et al. Scope of Artificial Intelligence in Screening and Diagnosis of Colorectal Cancer. *J Clin Med*. Oct 15 2020;9(10)doi:10.3390/jcm9103313

40. Pacal I, Karaboga D, Basturk A, Akay B, Nalbantoglu U. A comprehensive review of deep learning in colon cancer. *Comput Biol Med*. Nov 2020;126:104003. doi:10.1016/j.compbiomed.2020.104003

41. Hegde N, Shishir M, Shashank S, Dayananda P, Latte MV. A Survey on Machine Learning and Deep Learning-based Computer-Aided Methods for Detection of Polyps in CT Colonography. *Curr Med Imaging*. 2021;17(1):3-15. doi:10.2174/2213335607999200415141427

42. Than M, Witherspoon J, Shami J, Patil P, Saklani A. Diagnostic miss rate for colorectal cancer: an audit. *Ann Gastroenterol*. Jan-Mar 2015;28(1):94-98.

43. Han YE, Cho Y, Park BJ, et al. Development and multicenter validation of deep convolutional neural network-based detection of colorectal cancer on abdominal CT. *Eur Radiol*. Sep 2024;34(9):6182-6192. doi:10.1007/s00330-023-10452-2

44. Sahoo PK, Gupta P, Lai YC, et al. Localization of Colorectal Cancer Lesions in Contrast-Computed Tomography Images via a Deep Learning Approach. *Bioengineering (Basel)*. Aug 17 2023;10(8)doi:10.3390/bioengineering10080972

45. Yao L, Li S, Tao Q, et al. Deep learning for colorectal cancer detection in contrast-enhanced CT without bowel preparation: a retrospective, multicentre study. *EBioMedicine*. Jun 2024;104:105183. doi:10.1016/j.ebiom.2024.105183

46. Yao L, Xia Y, Chen Z, et al. A Colorectal Coordinate-Driven Method for Colorectum and Colorectal Cancer Segmentation in Conventional CT Scans. *IEEE Trans Neural Netw Learn Syst*. Apr 30 2024;PPdoi:10.1109/TNNLS.2024.3386610

47. Ren Y, Ma J, Xiong J, Chen Y, Lu L, Zhao J. Improved False Positive Reduction by Novel Morphological Features for Computer-Aided Polyp Detection in CT Colonography. *IEEE J Biomed Health Inform*. Jan 2019;23(1):324-333. doi:10.1109/JBHI.2018.2808199

48. Ziemlewicz TJ, Kim DH, Hinshaw JL, Lubner MG, Robbins JB, Pickhardt PJ. Computer-Aided Detection of Colorectal Polyps at CT Colonography: Prospective Clinical Performance and Third-Party Reimbursement. *AJR Am J Roentgenol*. Jun 2017;208(6):1244-1248. doi:10.2214/AJR.16.17499

49. Tsurumaru D, Takatsu N, Kai S, Oki E, Ishigami K. Measurement of circumferential tumor extent of colorectal cancer on CT colonography: relation to clinicopathological features and patient prognosis after surgery. *Jpn J Radiol*. Oct 2021;39(10):966-972. doi:10.1007/s11604-021-01141-5

50. Horie H, Togashi K, Utano K, Miyakura Y, Lefor AT, Yasuda Y. Predicting rectal cancer T stage using circumferential tumor extent determined by computed tomography colonography. *Asian J Surg*. Jan 2016;39(1):29-33. doi:10.1016/j.asjsur.2015.03.002

Abstract in Korean

장처치 없는 일반 복부골반CT에서 직대장암 발견을 위한 인공지능 알고리듬의 개발

연구배경: CT colonography와 달리 일반 복부골반CT는 장 정결 없이 촬영되기 때문에, 임상적으로 의심하지 않았던 직대장암은 종종 놓쳐지곤 한다.

연구목적: 장 정결 없이 시행된 조영증강 복부골반CT 영상에서 직대장암을 자동 탐지할 수 있는 인공지능 기반 알고리즘을 개발하고자 한다.

연구방법: 2010년 1월부터 2014년 12월까지 치료 시작 전에 복부골반CT를 시행받았던 직대장암 환자 2,662명을 대상으로 인공지능 모델을 학습시켰다. 개발된 모델의 성능은 내부 및 외부 데이터셋을 이용하여 후향적으로 검증하였다. 두 검증 데이터셋 모두 2018년 1월부터 6월 사이, 각각의 3차 병원에서 2개월 이내에 CT와 대장내시경 둘 다를 시행받았던 모든 환자들의 CT 영상으로 구성되었고, 따라서 직대장암 환자와 정상 환자가 모두 포함되었다. 표준 참조(reference standard)를 위한 병변 표지는, 소화기영상의학 세부전공 전문의가 대장내시경 결과지를 참고해가며 병변이 포함되어 있는 모든 CT 축상 단면 각각에서 직대장암을 최대한 정확히 감싸도록 네모 표시를 하는 식으로 구성하였다. 직대장암 탐지를 위한 인공지능 모델로는 transformer 기반 최신 객체 탐지 네트워크인 DETR(DEtection with TRansformer)을 이용하였다. 모델 성능은 Alternative free-response receiver operating characteristic(AFROC) 분석을 이용하여 평가하였고, 두 명의 소화기영상의학 세부전공 전문의의 성능과도 비교하였다.

결과: 내부 검증 데이터셋은 총 841명(평균 연령 58세)의 환자로 구성되었으며, 이 중 92명의 환자에서 93개의 CT에서 발견 가능한 직대장암 병변이 존재하였다. AFROC 곡선하면적(AUAFROC)은 0.867이었다. 최대 Youden index 지점에서 민감도(병변별)는 79.6%(74/93), 특이도(환자별)는 91.2%(683/749)였다. 외부 검증 데이터셋은 총 442명의 환자(평균 연령 57세)로 구성되었으며, 이 중 26명의 환자에서 26개의 CT로 탐지 가능한 직대장암 병변이 있었다. 이 데이터셋에서 AUAFROC는 0.808이었으며, 민감도(병변별)는 80.8%(21/26), 특이도(환자별)는 90.9%(378/416)였다. 두 명의 영상의학 전문의의 민감도는 각각 73.1%(19/26) 및 80.8%(21/26)였으며, 특이도는 각각 98.3%(409/416) 및 98.6%(410/416)로 서로 유사하였다. 인공지능 모델의 성능과 전문의들의 성능을 비교하였을 때, 민감도는 유의한 차이를 보이지 않았으나(각각 $p = 0.743$ 및 1.0), 특이도는 전문의들의

성능이 인공지능 보다 유의하게 높았다(두 전문의 모두 $p < 0.001$).

결론: 본 연구는 장 정결 없이 시행된 일반 복부골반CT에서 직대장암을 자동 탐지할 수 있는 인공지능 모델의 임상 적용 가능성을 보여주었다.

임상적 의의: 본 인공지능 모델은 임상적으로 직대장암이 의심되지 않았던 환자들에 대한 영상의학과 의사의 판독 업무를 보조하는 식으로 활용될 수 있다. 이는 특히 소화기영상의학 세부전공 전문의 인력이 부족한 의료 기관에서 더욱 유용할 것으로 기대된다.

핵심되는 말: 직대장암; 인공지능; 심층학습; DETR; CT; 자동 탐지.