Deep learning for colorectal cancer detection in contrast-enhanced CT without bowel preparation: a retrospective, multicentre study

27 authors, including:

Quan Tao
Southern Medical University
12 PUBLICATIONS 98 CITATIONS

Chu Han
Guangdong provincial people's hospital
109 PUBLICATIONS 1,976 CITATIONS

Bingjiang Qiu
Guangdong Academy of Medical Sciences and General Hospital
35 PUBLICATIONS 582 CITATIONS

Yanqi Huang
Guangdong Provincial People's Hospital
59 PUBLICATIONS 4,541 CITATIONS

Deep learning for colorectal cancer detection in contrast-enhanced CT without bowel preparation: a retrospective, multicentre study

Lisha Yao,[a,b,c,t] Suyun Li,[a,c,d,t] Quan Tao,[e,t] Yun Mao,[f,t] Jie Dong,[g] Cheng Lu,[a,c,h] Chu Han,[a,c,h] Bingjiang Qiu,[a,c,i] Yanqi Huang,[a,c] Xin Huang,[a,c,j] Yanting Liang,[a,c,d] Huan Lin,[a,b,c] Yongmei Guo,[k] Yingying Liang,[k] Yizhou Chen,[l] Jie Lin,[l] Enyan Chen,[l] Yanlian Jia,[m] Zhihong Chen,[n] Bochi Zheng,[o] Tong Ling,[a,c] Shunli Liu,[p] Tong Tong,[q,r] Wuteng Cao,[s] Ruiping Zhang,[g,*] Xin Chen,[k,**] and Zaiyi Liu[a,b,c,***]

[a]Department of Radiology, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, China
[b]School of Medicine, South China University of Technology, Guangzhou, China
[c]Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application, Guangzhou, China
[d]School of Medicine, South Medical University, Guangzhou, China
[e]Department of Rehabilitation Medicine, Zhujiang Hospital, Southern Medical University, Guangzhou, China
[f]Department of Radiology, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China
[g]Department of Radiology, Shanxi Bethune Hospital (Shanxi Academy of Medical Sciences), The Third Affiliated Hospital of Shanxi Medical University, Taiyuan, China
[h]Medical Research Institute, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, China
[i]Guangdong Cardiovascular Institute, Guangdong Provincial People's Hospital (Guangdong Academy of Sciences), Guangzhou, China
[j]School of Medicine, Shantou University Medical College, Shantou, China
[k]Department of Radiology, Guangzhou First People's Hospital, South China University of Technology, Guangzhou, China
[l]Department of Radiology, Puning People's Hospital, Southern Medical University, Jieyang, China
[m]Department of Radiology, Liaobu Hospital of Guangdong, Dongguan, China
[n]Institute of Computing Science and Technology, Guangzhou University, Guangzhou, China
[o]Department of Biomedical Engineering, Southern University of Science and Technology, Shenzhen, China
[p]Department of Radiology, The Affiliated Hospital of Qingdao University, Qingdao, China
[q]Department of Radiology, Fudan University Shanghai Cancer Center, Shanghai, China
[r]Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, China
[s]Department of Radiology, The Sixth Affiliated Hospital of Sun Yat-sen University, Guangzhou, China

Summary

Background Contrast-enhanced CT scans provide a means to detect unsuspected colorectal cancer. However, colorectal cancers in contrast-enhanced CT without bowel preparation may elude detection by radiologists. We aimed to develop a deep learning (DL) model for accurate detection of colorectal cancer, and evaluate whether it could improve the detection performance of radiologists.

Methods We developed a DL model using a manually annotated dataset (1196 cancer *vs* 1034 normal). The DL model was tested using an internal test set (98 *vs* 115), two external test sets (202 *vs* 265 in 1, and 252 *vs* 481 in 2), and a real-world test set (53 *vs* 1524). We compared the detection performance of the DL model with radiologists, and evaluated its capacity to enhance radiologists' detection performance.

Findings In the four test sets, the DL model had the area under the receiver operating characteristic curves (AUCs) ranging between 0.957 and 0.994. In both the internal test set and external test set 1, the DL model yielded higher accuracy than that of radiologists (97.2% *vs* 86.0%, p < 0.0001; 94.9% *vs* 85.3%, p < 0.0001), and significantly improved the accuracy of radiologists (93.4% *vs* 86.0%, p < 0.0001; 93.6% *vs* 85.3%, p < 0.0001). In the real-world test set, the DL

---

*Corresponding author. Department of Radiology, Shanxi Bethune Hospital (Shanxi Academy of Medical Sciences), Shanxi Medical University, Taiyuan, 030032, China.
**Corresponding author. Department of Radiology, Guangzhou First People's Hospital, School of Medicine, South China University of Technology, Guangzhou, 510180, China.
***Corresponding author. Department of Radiology, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, 510080, China.
E-mail addresses: zrp_7142@sxmu.edu.cn (R. Zhang), wolfchenxin@163.com (X. Chen), liuzaiyi@gdph.org.cn (Z. Liu).
[t]Contributed equally.

model delivered sensitivity comparable to that of radiologists who had been informed about clinical indications for most cancer cases (94.3% *vs* 96.2%, p > 0.99), and it detected 2 cases that had been missed by radiologists.

Interpretation The developed DL model can accurately detect colorectal cancer and improve radiologists' detection performance, showing its potential as an effective computer-aided detection tool.

Research in context

Evidence before this study
We searched PubMed for relevant articles published in any language from database inception to October 15, 2023, using the search terms "deep learning" OR "artificial intelligence" OR "machine learning" AND "colorectal cancer" AND "detection" OR "diagnosis". Deep learning has been reported for the diagnosis or detection of colorectal cancer in histopathology and colonoscopy. No study reported deep learning-based model for the detection of colorectal cancer in contrast-enhanced CT without bowel preparation.

Added value of this study
To the best of knowledge, this study is the first to develop and validate a deep learning-based model for the detection of colorectal cancer in contrast-enhanced CT without bowel preparation. The DL model achieved satisfactory detection performance in three multicentre datasets and a real-world dataset. It demonstrated superior performance to that of radiologists in two independent datasets, and even improved the detection accuracy of radiologists. In the real-world dataset, the DL model achieved comparable detection

sensitivity to that of radiologists possessing knowledge of clinical indications for the majority of colorectal cancer cases, and detected 2 colorectal cancer cases that were missed by radiologists.

Implications of all the available evidence
Contrast-enhanced CT is a routine imaging modality for various clinical indications, providing a method to detect incident colorectal cancer and improve patient outcomes. However, it is usually conducted without bowel preparation, thus colorectal cancers with subtle features could be missed by radiologists. In addition, the detection performance of radiologists is susceptible to their workload and experience. The developed DL model not only distinguished patients with colorectal cancer from individuals without colorectal abnormality, but also identified the colorectal cancer regions to aid accurate diagnosis and treatment. Multicentre validation, comparison with radiologists, and real-world validation indicate that the DL model has the potential to be a clinical diagnosis support tool to assist radiologists for more accurate detection of colorectal cancer.

Introduction

Colorectal cancer (CRC) is the third most common cancer and the second leading cause of cancer-related death.[1,2] Early-stage CRC has a favorable prognosis with a five-year survival rate of 91%; however, as CRC progresses to advanced stages and metastasizes to distant organs, the prognosis significantly worsens with a five-year survival rate of only 14%.[2] Various methods such as colonoscopy, fecal immunochemical tests (FITs), and CT colonography, have been proposed for the detection of CRC.[3,4] Contrast-enhanced CT scans are routinely performed for a broad range of clinical indications, providing a means to detect unsuspected CRC.[5,6] Annually, the number of contrast-enhanced CT examinations approximates 120 million with a growth

rate of 4% worldwide.[7] Additionally, contrast-enhanced CT is a robust method for staging diagnosed CRC to aid appropriate treatment planning.[8] Therefore, accurate detection of CRC in contrast-enhanced CT could improve patient outcomes and decline associated mortality.[6]

However, routine contrast-enhanced CT examinations are usually conducted without bowel preparation, potentially resulting in the obscuration of small tumors by colorectal contents and collapsed segments.[9] A previous study has shown that tumors with an average size of 3.3 cm could potentially evade detection by radiologists.[10] In addition, incidental tumors with subtle features may exhibit attenuations similar to the surrounding colorectal tissues. In the absence of additional

specific indications, such as mesenteric infiltration and lymph node enlargement, the detection of these tumors presents more significant challenges for radiologists.[9,10] Since the colorectum occupies a large portion in the abdomen and is surrounded by similar soft-tissue organs, searching for abnormality within the colorectum slice by slice is a laborious and time-consuming task. Meanwhile, the increasing clinical workload often prompts radiologists to prioritize the identification of the listed indications for CT examinations, leading to a lack of comprehensive exploration of the colorectum.[9] Furthermore, significant variations in the diagnosis of CRC exist among radiologists with different levels of experience.[11]

In the past decade, deep learning (DL) has demonstrated considerable promise in medical image analysis.[12,13] It has delivered impressive outcomes in the automated detection and diagnosis of pancreatic cancer,[14] ovarian cancer,[15] and breast cancer,[16] showing comparable or even superior diagnostic performance to that of radiologists. Furthermore, the enhanced diagnostic capabilities observed in DL-assisted radiologists suggest that deep learning has the potential to evolve into clinical decision support systems.[15,16] Nevertheless, to the best of knowledge, the potential of deep learning in the detection of CRC in contrast-enhanced CT without bowel preparation has not been investigated.

We hypothesized that leveraging advanced deep learning algorithms, high-quality training images and annotations could facilitate the development of a model for accurate detection of CRC. In this study, we aimed to develop and validate a deep learning model for CRC detection using multicentre datasets, and to compare its detection performance with that of 9 radiologists. We also aimed to assess the potential of the DL model to enhance the detection performance of radiologists in a reader study and in real clinical scenarios.

Methods

Ethics
This study was approved by the ethic committee of Guangdong Provincial People's Hospital (KY-Q-2022-418-01). Considering the retrospective nature of this study, the informed consent was waived.

Study design
In this study, we developed a deep learning-based model to detect CRC using the datasets from four institutions (Fig. 1A). The detection model was modified from a segmentation framework, providing predictions for CRC regions as detection results (Fig. 1B). We tested the DL model using an internal test set, two independent external test sets, and a real-world test set (Fig. 1C). Within the internal test set and external test set 1, we compared the DL model with 9 radiologists in terms of both overall detection performance and subgroup detection sensitivity. Furthermore, we explored whether the DL model could improve the detection accuracy of radiologists. In the real-world test set, the DL model was compared with the original radiology reports to investigate its ability and practicality for detecting CRC (Fig. 1D). This study was approved by the ethic committee of Guangdong Provincial People's Hospital. Considering the retrospective nature of this study, the informed consent was waived.

Participant datasets
From Guangdong Provincial People's Hospital, we included 984 patients with CRC between January 2008
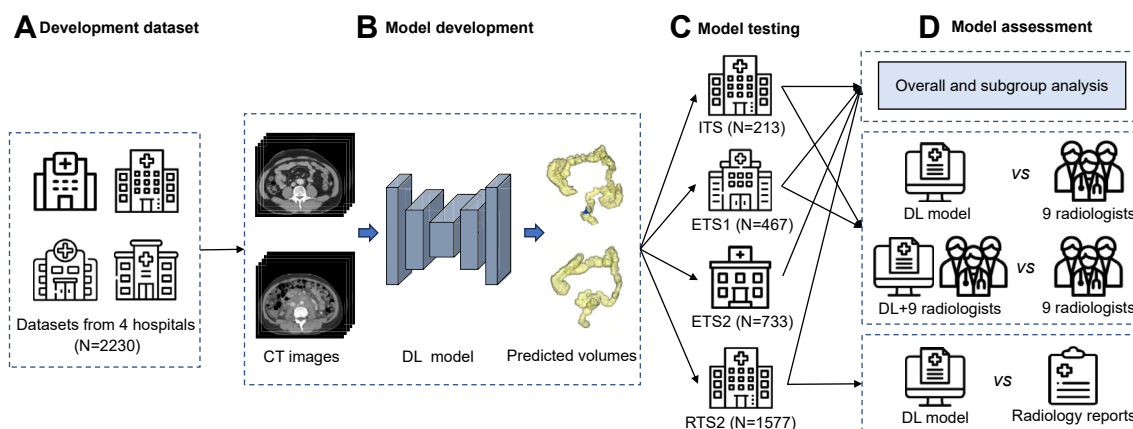
Fig. 1: Overview of this study. (A) A dataset collected from four hospitals was used to train a DL model to detect CRC. (B) The DL model was tested on an internal test set, two external test sets as well as a real-world test set. (C) Overall detection performance and subgroup sensitivity analysis were conducted to assess the performance of the DL model. (D) The DL model was compared with 9 radiologists in internal test set and external test set 1. DL-assisted CT interpretation was performed to evaluate whether the DL model could improve the detection performance of radiologists. The DL model was also compared with the original radiology reports in real-world clinical settings. ITS, Internal test set; ETS1, External test set 1; ETS2, External test set 2; RTS, Real-world test set.

and December 2018, and 1073 individuals without colorectal abnormalities between January 2018 and June 2018. Without overlap, these participants were randomly divided into a training set (788 patients with CRC and 919 normal individuals), a validation set (98 patients with CRC and 115 normal individuals), and an internal test set (98 patients with CRC and 115 normal individuals), according to a ratio of 8:1:1. Given the subtle features and limited sample size of T1-stage CRC data in comparison to other T-stage CRC data, we additionally enrolled 310 patients with T1-stage CRC from four hospitals (38 patients from Fudan University Shanghai Cancer Center between May 2016 and May 2021, 43 patients from the First Affiliated Hospital of Qingdao University between October 2016 and November 2022, 42 patients from Guangdong Provincial People's Hospital between March 2011 and March 2021, and 187 patients from the Sixth Affiliated Hospital of Sun Yat-sen University between December 2012 and September 2022) to enrich the training set. Consequently, the sample sizes of the training, validation, and internal testing sets are 2017, 213, and 213 respectively, resulting in a ratio of 8.26:0.87:0.87. Two independent datasets were used for external testing. The external test set 1 comprised 202 patients with CRC and 265 normal individuals from the First Affiliated Hospital of Chongqing Medical University between January 2018 and December 2018. The external test set 2 included 252 patients with CRC and 481 normal individuals from Shanxi Bethune Hospital between January 2016 and September 2022 (Supplementary Figure S1).

All patients with CRC underwent contrast-enhanced abdominal CT scans within one month before surgery, with confirmation provided by postoperative pathology reports. Contrast-enhanced abdominal CT scans of normal individuals were selected according to the statement of no obvious colorectal abnormalities in the original radiology reports, further validated through a 12-month follow-up. Exclusion criteria included: (1) duplicated CT examinations of the same patients, (2) CT examinations not covering the entire abdomen and pelvis, (3) CT examinations containing severe artifacts, (4) patients presenting severe ascites, (5) patients with residual positive contrast agent within the colorectum.

To explore the detection performance of the DL model in real-world settings, we collected a consecutive cohort of patients who underwent contrast-enhanced CT scans from Guangdong Provincial People's Hospital between February 22 and March 28, 2023. After inclusion and exclusion, the real-world cohort included 53 patients with CRC confirmed by postoperative or biopsy pathology reports, and 1524 normal individuals without radiologically suspected colorectal abnormalities (Supplementary Figure S2).

CT examinations were conducted using one of the 22 scanner devices (Supplementary Figure S3). Automatic tube current modulation was used during CT scanning.

The imaging matrix was 512 × 512. Detailed imaging parameters were provided in Supplementary Table S1. Following the non-contrast CT scan, arterial and venous phase CT images were acquired at 20–30 s and 70–80 s after intravenous administration of contrast medium (3–3.5 ml/s, total amount of 90–150 ml).

Model development

In this study, the colorectum in all cases and tumors in all cancer cases were annotated on venous phase CT images to train and validate the detection model. To facilitate efficient annotation, we employed an automated segmentation model for colorectum segmentation.[17] The segmentation results subsequently underwent simply correction to provide rough colorectum annotations. Tumors were manually annotated by one of three radiologists from two institutions (Guangdong Provincial People's Hospital, Shanxi Bethune Hospital) and reviewed by a colorectal specialist with 10 years of experience from Guangzhou First People's Hospital. All radiologists annotated the tumors by referring to radiological and pathological diagnostic information, such as the tumor location and size. Both the colorectum correction and tumor annotation processes were conducted using the ITK-SNAP software.[18]

We developed the detection model based on a segmentation framework for both colorectum and CRC segmentation. Detection outcomes were obtained based on segmented tumors. The detection model was modified from nnUNet,[19] an advanced deep learning-based model for medical image segmentation. The nnUNet leveraged a traditional U-Net architecture for feature extraction and context integration. It provided automated configuration for the entire segmentation pipeline to adapt to diverse medical image datasets. Moreover, we incorporated an auxiliary classification branch that integrated multi-scale features from the nnUNet to achieve patch-level classification (i.e., CRC or non-CRC) for more accurate identification of CRC.[20] Details of the model architecture and training process are provided in Supplementary Figure S4.

Reader study

Nine radiologists from four institutions participated in a reader study using all data from both the internal test set and external test set 1. The reader group consisted of six general radiologists and three colorectal specialists (Appendix SupplementaryTable S2). Each radiologist employed a second-reader paradigm, initially interpreting all 680 CT scans in the reader study without DL assistance, followed by interpretation of the same 680 CT scans with DL assistance. None of the radiologists were involved in the tumor annotation process, and they were unaware of the distribution of cancer and normal cases. All radiologists were trained to use the ITK-SNAP software to visualize the CT images and annotate suspected CRC. During the standalone interpretation

process, radiologists were provided only with the CT images and were blinded to any clinical, radiological and pathological diagnostic information. They confirmed the presence or absence of CRC in each CT scan and provided rough annotations for the suspected tumor region within the axial slice encompassing the largest tumor area. In the DL-assisted CT interpretation, both the CT images and the segmentation outputs of the DL model were displayed on the ITK-SNAP software for interpretation by radiologists. The detection performance of both the DL model and standalone interpretation was not revealed to the radiologists. The interval between the CT interpretation with and without DL assistance exceeded 30 days.

Statistical analyses

Receiver operating characteristic (ROC) curves, the area under the receiver operating characteristic (AUC) curves, detection sensitivity, specificity, accuracy, positive, and negative predictive values were conducted to evaluate the detection performance of the DL model and radiologists. For the DL model, detection was deemed correct if the predicted tumor volume exceeded a cutoff value, and the Dice index between the predicted tumor and annotated tumor exceeded 0.[21] The cutoff value was determined as the tumor volume yielding the highest Youden index from the ROC curve built using the training and validation sets. In the CT interpretation, a sample was considered correctly detected if the Dice index between the roughly interpreted tumor and the annotated tumor was greater than 0. The Clopper-Pearson method was used to estimate 95% confidence intervals (CIs) of AUC and the detection metrics.[22] Detection agreement was evaluated using Kappa statistics. The Obuchowski-Rockette and Dorfman-Berbaum-Metz method for multireader multicase analysis was used to evaluate the overall detection performance of all radiologists.[23,24] The McNemar test was employed to compare the detection performance of the DL model and radiologists.[25] A statistically significant difference was set as p < 0.05. All statistical analysis were performed using Python version 3.7.11 with Sklearn package and Statsmodels package.

Role of funders

The funders played no role in the study design, data collection, data analysis, data interpretation, and writing of the report.

Results
Baseline characteristics

Baseline characteristics of the participants were shown in Table 1. A total of 5220 patients were included in this retrospective study. The training set included 2017 patients (median age of 50 years, 56.7% men), and the validation set included 213 patients (median age of 62

years, 64.8% men). The test sets comprised 213 patients (median age of 57 years, 52.1% men) in the internal test set, 467 patients (median age of 64 years, 63.2% men) in the external test set 1, 733 patients (median age of 59 years, 44.5% men) in the external test set 2, and 1577 patients (median age of 58 years, 52.6% men) in the real-world test set. Because of the inclusion of an additional cohort consisting of 310 patients with T1-stage CRC, the training set exhibited a higher proportion of cancer cases in the T1 stage (30.3%). The real-world test set had a relatively low proportion of patients with CRC (3.4%).

Cancer detection performance of the DL model

The DL model had an AUC of 0.988 (95% CI 0.963–0.998) in the validation set, 0.994 (95% CI 0.972–1.000) in the internal test set, 0.965 (95% CI 0.944–0.980) in external test set 1, 0.957 (95% CI 0.940–0.971) in external test set 2, and 0.968 (95% CI 0.958–0.976) in the real-world test set (Fig. 2). The cutoff yielding the highest Youden index was 250 mm^3. In validation set, the DL model achieved a sensitivity of 95.9% (95% CI: 89.9–98.9), specificity of 97.4% (95% CI: 92.6–99.5), and accuracy of 96.7% (95% CI: 93.4–98.7) to detect the CRC in contrast-enhanced CT images without bowel preparation. The DL model obtained a sensitivity of 96.9% (95% CI: 91.3–99.4), specificity of 97.4% (95% CI: 92.6–99.5), and accuracy of 97.2% (95% CI: 94.0–99.0) in internal test set. In external test set 1, the DL model had a sensitivity of 93.6% (95% CI: 89.3–96.5), specificity of 95.9% (95% CI: 92.7–97.9), and accuracy of 94.9% (95% CI: 92.5–96.7). In external test set 2, the DL model achieved a sensitivity of 91.7% (95% CI: 87.5–94.8), specificity of 95.6% (95% CI: 93.4–97.3), and accuracy of 94.3% (95% CI: 92.3–95.8). In the real-world test set, the DL model achieved a sensitivity of 94.3% (95% CI: 84.3–98.8), specificity of 97.4% (95% CI: 96.5–98.2), and accuracy of 97.3% (95% CI: 96.4–98.1) to distinguish the patients with CRC from normal individuals (Table 2).

Comparison in cancer detection between the DL model and radiologists

The detection performance of 9 individual radiologists in internal test set and external test set 1 was shown in Fig. 2 and Supplementary Tables S3 and S4. The operating points of each radiologist and the operating points of overall radiologists were located under the ROC curves of the DL model in internal test set and external test set 1, indicating that the detection performance of the DL model is superior to all radiologists (Fig. 2). Among 9 radiologists, the accuracy in the internal test set varied from 79.3% (95% CI 73.3–84.6) to 90.6% (95% CI 85.9–94.2), with inter-observer agreements ranging between 0.49 and 0.87. In external test set 1, the accuracy varied from 75.6% (95% CI 71.4–79.4) to 91.4% (95% CI 88.5–93.8), with inter-observer

	Training set (n = 2017)	Validation set (n = 213)	ITS (n = 213)	ETS1 (n = 467)	ETS2 (n = 733)	RTS (n = 1577)
Sample characteristics, n (%)						
Normal	919 (45.6%)	115 (54.0%)	115 (54.0%)	265 (56.7%)	481 (65.6%)	1524 (96.6%)
CRC	1098 (54.4%)	98 (46.0%)	98 (46.0%)	202 (43.3%)	252 (34.4%)	53 (3.4%)
Sex						
Male	1144 (56.7%)	138 (64.8%)	111 (52.1%)	295 (63.2%)	326 (44.5%)	830 (52.6%)
Female	873 (43.3%)	75 (35.2%)	102 (47.9%)	172 (36.8%)	407 (55.5%)	747 (47.4%)
Age (years)						
	50 (59–68)	62 (51–70)	57 (48–66)	64 (55–69)	59 (46–67)	58 (50–66)
Tumor T stage						
T1	333 (30.3%)	5 (5.1%)	4 (4.1%)	6 (3.0%)	12 (4.8%)	4 (7.5%)
T2	105 (9.6%)	13 (13.3%)	14 (14.3%)	23 (11.4%)	37 (14.7%)	7 (13.2%)
T3	581 (52.9%)	71 (72.4%)	68 (69.4%)	146 (72.3%)	113 (44.8%)	24 (45.3%)
T4	79 (7.2%)	9 (9.2%)	12 (12.2%)	27 (13.3%)	90 (35.7%)	1 (1.9%)
Unknown	–	–	–	–	–	17 (32.1%)
Tumor location						
Ascending colon	240 (21.9%)	23 (23.5%)	23 (23.5%)	31 (15.4%)	39 (15.5%)	8 (15.1%)
Transverse colon	68 (6.2%)	6 (6.1%)	9 (9.2%)	21 (10.4%)	20 (7.9%)	5 (9.5%)
Descending colon	79 (7.2%)	7 (7.1%)	8 (8.2%)	15 (7.4%)	15 (6.0%)	4 (7.5%)
Sigmoid colon	298 (27.1%)	26 (26.5%)	22 (22.4%)	36 (17.8%)	49 (19.4%)	8 (15.1%)
Sigmoid colon-Rectum	170 (15.5%)	23 (23.5%)	17 (17.3%)	44 (21.8%)	61 (24.2%)	4 (7.5%)
Rectum	243 (22.1%)	13 (13.3%)	19 (19.4%)	55 (27.2%)	68 (27.0%)	24 (45.3%)
Tumor size (cm)						
	4.1 (2.8–5.2)	4.5 (3.4–5.2)	4.7 (3.5–5.6)	4.5 (3.3–5.2)	4.8 (3.5–6.0)	4.0 (2.8–4.5)

Data are n (%) or median (IQR). ITS, Internal test set; ETS1, External test set 1; ETS2, External test set 2; RTS, Real-world test set; CRC, Colorectal cancer; IQR, inter quartile range. Ascending colon: including ascending colon and cecum. Transverse colon: including transverse colon and hepatic flexure. Descending colon: including descending colon and splenic flexure. Sigmoid colon-Rectum: the junction of the sigmoid colon and rectum. Unknown: The CRC cases were confirmed by biopsy pathology, and the information of T stage and tumor size was not reported. Tumor size was the maximum axial diameter of the tumor reported in the pathology reports or measured on the venous phase CT images.

Table 1: Patient characteristics.

agreements ranging between 0.30 and 0.85 (Supplementary Figure S5 and Supplementary Tables S3 and S4). Compared with the overall radiologists, the DL model had significantly higher sensitivity in both internal test (96.9% *vs* 74.7%, p < 0.0001) and external test set 1 (93.6% *vs* 74.8%, p < 0.0001). The DL model had higher specificity than that of overall radiologists in the internal test set (97.4% *vs* 95.6%, p = 0.032) and the external test set 1 (95.9% *vs* 93.3%, p < 0.0001). The accuracy of the DL model significantly exceeded that of radiologists in internal test (97.2% *vs* 86.0%, p < 0.0001) and external test set 1 (94.9% *vs* 85.3%, p < 0.0001). In the real-world test set, the sensitivity of the DL model had no significant difference with that of the original radiology reports (94.3% *vs* 96.2%, p > 0.99) (Table 2). Moreover, the DL model detected 2 CRC cases that were not reported in the original radiology reports (Supplementary Table S12).

Comparison in cancer detection between radiologists with and without DL assistance

With the DL model assistance, the accuracy of 9 radiologists in the internal test set varied from 83.6% (95% CI 77.9–88.3) to 96.7% (95% CI 93.3–98.7), with improved inter-observer agreements ranging between 0.64 and 0.94. In the external test set 1, the accuracy varied from 89.7% (95% CI 86.6–92.3) to 95.9% (95% CI 93.7–97.5), with inter-observer agreements ranging between 0.75 and 0.94 (Supplementary Figure S5 and Supplementary Table S3 and S4). In both the internal test set and external test set 1, the DL model improved the accuracy of both the general radiologists (93.5% *vs* 84.8%, p < 0.0001; 93.4% *vs* 83.4%, p < 0.0001) and the specialists (93.3% *vs* 83.3%, p < 0.0001; 93.9% *vs* 89.0%, p < 0.0001). Notably, larger improvements of the accuracy were observed in the general radiologists than those in the specialists. Consequently, the DL-assisted radiologists yielded substantially higher accuracy in internal test set (93.4% *vs* 86.0%, p < 0.0001) and external test set 1 (93.6% *vs* 85.3%, p < 0.0001).

Subgroup sensitivity analysis

Sensitivity analysis associated with different tumor sizes, locations, and T stages was conducted to further evaluate the difference between the DL model and radiologists (Table 3 and Supplementary Table S5–S11). For the internal test set and external test set 1 combined, the DL model yielded higher sensitivity than that of radiologists in detecting CRC with different sizes, locations, and T stages (p < 0.0001). Furthermore, the
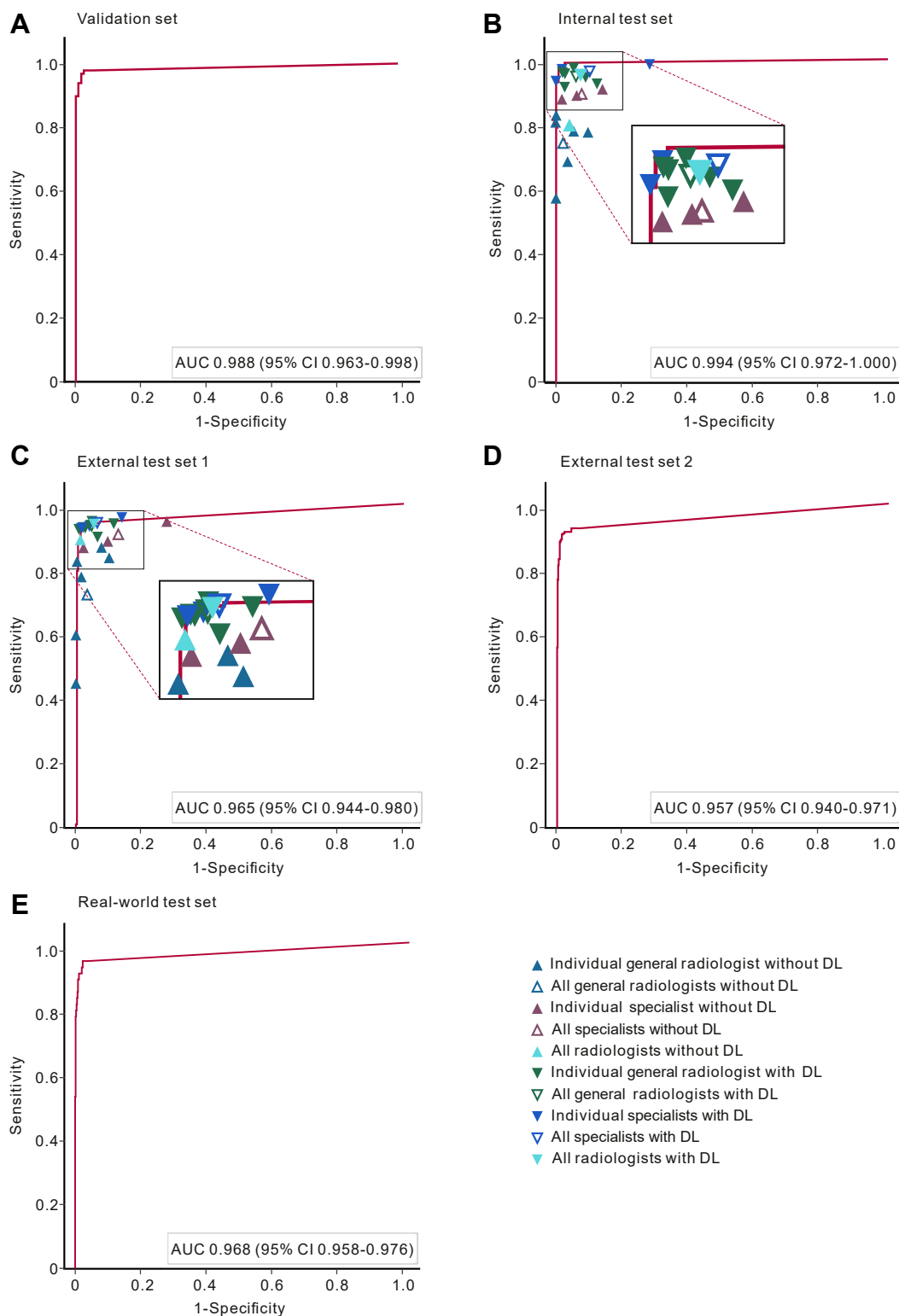
Fig. 2: Receiver operating characteristic curves. (A) ROC of the DL model in the validation set. (B) ROC of the DL model and the operating points of radiologists with and without DL assistance in the internal test set. (C) ROC of the DL model and the operating points of radiologists with and without DL assistance in the external test set 1. (D) ROC of the DL model in the external test set 2. (E) ROC of the DL model in the real-world test set. ROC, Receiver operating characteristic curve; AUC, area under the receiver operating characteristic curve; DL, Deep learning.

	Sensitivity	p value	Specificity	p value	Accuracy	p value	PPV	NPV
Validation set								
DL Model	95.9 (89.9–98.9)		97.4 (92.6–99.5)		96.7 (93.4–98.7)		96.9 (91.2–99.4)	96.6 (91.4–99.1)
Internal test set								
Radiologists without DL[a]								
General	73.8 (70.1–77.3)	<0.0001	94.2 (92.2–95.8)	0.005	84.8 (82.7–86.7)	<0.0001	91.6 (88.7–93.9)	80.8 (78.0–83.5)
Specialist	76.5 (71.3–81.3)	<0.0001	98.3 (96.3–99.4)	0.607	88.3 (85.5–90.7)	<0.0001	97.4 (94.4–99.0)	83.1 (79.1–86.6)
All	74.7 (71.7–77.6)	<0.0001	95.6 (94.1–96.7)	0.032	86.0 (84.3–87.5)	<0.0001	93.5 (91.4–95.2)	81.6 (79.3–83.7)
DL model	96.9 (91.3–99.4)		97.4 (92.6–99.5)		97.2 (94.0–99.0)		96.9 (91.3–99.4)	97.4 (92.6, 99.5)
Radiologists with DL[b]								
General	94.7 (92.6–96.4)	<0.0001	92.5 (90.2–94.3)	0.182	93.5 (92.0–94.8)	<0.0001	91.5 (89.0–93.6)	95.4 (93.5–96.8)
Specialist	92.2 (88.5–95.0)	<0.0001	94.2 (91.2–96.4)	0.007	93.3 (91.0–95.1)	<0.0001	93.1 (89.6–95.8)	93.4 (90.2–95.8)
All	93.9 (92.1–95.4)	<0.0001	93.0 (91.3–94.5)	0.009	93.4 (92.2–94.5)	<0.0001	92.0 (90.0–93.7)	94.7 (93.1–96.0)
External test set 1								
Radiologists without DL[a]								
General	72.5 (69.9–75.0)	<0.0001	91.7 (90.2–93.0)	<0.0001	83.4 (82.0–84.8)	<0.0001	86.9 (84.7–89.0)	81.4 (79.5–83.2)
Specialist	79.2 (75.8–82.4)	<0.0001	96.5 (94.9–97.6)	0.597	89.0 (87.3–90.6)	<0.0001	94.5 (92.1–96.3)	85.9 (83.4–88.1)
All	74.8 (72.7–76.7)	<0.0001	93.3 (92.2–94.3)	<0.0001	85.3 (84.2–86.3)	<0.0001	89.5 (87.8–91.0)	82.9 (81.4–84.3)
DL model	93.6 (91.3–99.4)		95.9 (92.7–97.9)		94.9 (92.5–96.7)		94.5 (90.4–97.2)	95.1 (91.8–97.4)
Radiologists with DL[b]								
General	92.2 (90.5–93.6)	<0.0001	94.3 (93.1–95.4)	0.001	93.4 (92.4–94.3)	<0.0001	92.5 (90.9–94.0)	94.0 (92.8–95.2)
Specialist	92.6 (90.2–94.5)	<0.0001	95.0 (93.2–96.4)	0.155	93.9 (92.6–95.1)	<0.0001	93.3 (91.0–95.2)	94.4 (92.5–95.9)
All	92.3 (91.0–93.5)	<0.0001	94.5 (93.6–95.4)	0.041	93.6 (92.8–94.3)	<0.0001	92.8 (91.5–94.0)	94.2 (93.1–95.1)
External test set 2								
DL model	91.7 (87.5–94.8)		95.6 (93.4–97.3)		94.3 (92.3–95.8)		91.7 (87.5–94.8)	95.6 (93.4–97.3)
Real-world test set								
DL model	94.3 (84.3–98.8)		97.4 (96.5–98.2)		97.3 (96.4–98.1)		56.2 (48.3–63.8)	99.8 (99.4–99.9)
Radiology reports	96.2 (87.0–99.5)	>0.99[c]	100.0 (99.8–100.0)	<0.0001[d]	99.9 (99.5–100.0)	<0.0001	100.0 (93.0–100.0)	99.9 (99.5–100.0)

Data are n (95% CI) and displayed as percentages (%). CI, confidence interval; DL model, Deep learning model; PPV, positive predictive value; NPV, negative predictive value. [a]Comparison between the DL model and radiologists without DL assistance. [b]Comparison between the radiologists with and without DL assistance. [c]Comparison the detection sensitivity for CRC cases between the DL model and the original radiology reports, where the pathology reports serve as the ground truth. [d]Comparison the detection specificity for normal cases between the DL model and the original radiology reports, where the original radiology reports serve as the ground truth.

Table 2: Detection performance of the DL model and radiologists with and without DL assistance.

improved sensitivity of the DL-assisted radiologists contributed to the above superiority of the DL model.

Discussion

In this study, we developed and validated a deep learning-based model that can accurately detect CRC in contrast-enhanced CT images without bowel preparation. The DL model achieved superior detection performance to that of 9 radiologists and even improved the radiologists' detection accuracy in two independent datasets. In real-world settings, the DL model showed comparable detection sensitivity to that of the original radiology reports, and it detected 2 CRC cases that were missed by the radiology reports.

The DL model developed in this study exhibits several advantages. First, it supports automated and stable clinical diagnosis workflow. In practice, the diagnosis performance of radiologists is susceptible to their workload and fatigue. Inconsistencies are observed among radiologists with varying levels of experience in

detecting CRC.[9] The DL model could be seamlessly integrated into various workstations, enabling automated and real-time detection of CRC. The DL model's detection performance remains stable and is unaffected by variations in workload and inter-observer variances. Second, the DL model demonstrated promising detection performance in multicentre datasets. To ensure the robustness of model training, all of the data were confirmed through pathology reports or follow-up. Pixel-level tumor annotations yielded by radiologists helped the DL model accurately identify CRC regions. An auxiliary patch-level classification branch was integrated into the DL model to further improve the accuracy of CRC identification. In addition, the proportion of tumor data to control data is a crucial factor for model performance in our experiments. In the two external test sets, the DL model produced consistent results with no significant deterioration in comparison to that obtained in the internal test set. However, since the real-world test set were enriched with a substantial number of normal cases, the DL model yielded a PPV of only

	Radiologists without DL	DL model	p value[a]	Radiologists with DL	p value[b]
T stage					
T1	10.0 (4.7–18.1)	70.0 (34.8–93.3)	<0.0001	61.1 (50.3–71.2)	<0.0001
T2	59.5 (54.0–64.8)	91.9 (78.1–98.3)	<0.0001	85.6 (81.3–89.2)	<0.0001
T3	78.5 (76.5–80.3)	95.3 (91.6–97.7)	<0.0001	94.6 (93.5–95.6)	<0.0001
T4	85.5 (81.3–89.0)	100.0 (91.0–100.0)	<0.0001	98.0 (95.9–99.2)	<0.0001
Tumor location					
Ascending colon	85.0 (81.5–88.0)	98.1 (90.1–100.0)	<0.0001	97.5 (95.7–98.7)	<0.0001
Transverse colon	81.9 (76.7–86.3)	90.0 (73.5–97.9)	0.0003	90.0 (85.8–93.3)	0.0003
Descending colon	74.4 (67.9–80.2)	82.6 (61.2–95.0)	0.0005	83.6 (77.8–88.3)	0.0001
Sigmoid colon	76.2 (72.4–79.8)	100.0 (93.8–100.0)	<0.0001	96.9 (95.1–98.2)	<0.0001
Sigmoid colon-Rectum	79.6 (76.0–82.9)	98.4 (91.2–100.0)	<0.0001	96.9 (95.1–98.2)	<0.0001
Rectum	59.3 (55.5–63.1)	90.5 (81.5–96.1)	<0.0001	86.8 (84.0–89.3)	<0.0001
Tumor size					
<3 cm	40.7 (35.9–45.7)	80.0 (65.4–90.4)	<0.0001	73.8 (69.3–78.0)	<0.0001
3–4 cm	70.6 (67.3–73.8)	96.6 (90.3–99.3)	<0.0001	94.5 (92.7–96.0)	<0.0001
>4 cm	86.0 (84.1–87.7)	97.6 (94.0–99.3)	<0.0001	97.0 (96.0–97.8)	<0.0001

Data are n (95% CI) and displayed as percentages (%). CI, confidence interval; DL model, Deep learning model. [a]Comparison between the DL model and radiologists without DL assistance. [b]Comparison between the radiologists with and without DL assistance.

Table 3: Subgroup sensitivity of the DL model and radiologists with and without DL assistance in the internal test set and external test set 1 combined.

56.2%, which is lower than the results on other datasets. Nevertheless, the DL model maintained similar sensitivity and specificity to those of other test sets in the real-world test set.

Another advantage of the DL model is its ability to improve the detection performance of radiologists. Detecting CRC in contrast-enhanced CT is a challenging task for radiologists, particularly in the absence of bowel preparation. The imaging features of CRC, such as large obstructing mass, mesenteric infiltration and lymph node enlargement, provide distinct diagnostic indications.[6,26] However, small masses and polyps are susceptible to be obscured by colorectal contents and collapsed colorectal segments.[9] Tumors exhibiting intestinal wall thickening are difficult to be distinguished from non-distended normal colorectal wall tissue. Deep learning leverages neural networks to learn latent features and structures that are imperceptible to human eyes, potentially enabling more precisely approximate the data distribution and generate more accurate outcomes.[27] For the internal test set and external test set 1 combined, radiologists failed to detect 54 CRC cases, whereas the DL model detected 39 out of the 54 cases. Only 1 case was missed by the DL model but detected by radiologists. Furthermore, the DL model exhibited superior performance compared to radiologists in detecting CRC with different sizes, locations, and T stages. With DL assistance, radiologists successfully detected 38 out of 54, showing significant improvements in the detection of CRC. In real clinical scenarios, radiologists had possessed knowledge of clinical history and indications for the majority of CRC cases (Supplementary Table S12).

Therefore, the detection sensitivity of the original radiology reports in the real-world test set is higher than that of the radiologists in the reader study. However, in the real-world test set, radiologists failed to detect 2 CRC cases that lacked clear clinical indications. The DL model showed comparable performance to radiologists and effectively detected the 2 cases that had escaped detection by radiologists. These findings indicate that the integration of deep learning and the expertise of radiologists has the capacity to enhance the accuracy of detecting CRC in contrast-enhanced CT scans. Additionally, the DL model detected 3 out of the 9 suspected CRC cases (Supplementary Table S13), and 1 out of 12 suspected colorectal abnormality cases by radiology reports (Supplementary Table S14), suggesting a high level of confidence in the radiologists' diagnosis. These patients, with double-suspected CRC or abnormality, may necessitate enhanced monitoring and follow-up.

To the best of our knowledge, most previous deep learning-based studies were restricted to the diagnosis or detection of colorectal lesions in histopathology,[28,29] and colonoscopy.[30,31] This study is the first to explore and validate the potential of deep learning in the detection of CRC in contrast-enhanced CT without bowel preparation. The subjective assessments of radiologists in this study were similar to those reported in previous studies.[6,9–11] Despite the proposal of several approaches aimed at enhancing the performance of morphologic identification in CRC,[6,9] subjective assessments remain susceptible to the influence of workload and varying levels of experience. In this study, we developed a deep learning model for automated and

9

accurate detection of CRC. This model not only distinguished patients with CRC from individuals without colorectal abnormality, but also identified the CRC regions to aid more accurate diagnosis and treatment.[32]

Our developed model aims to enhance radiologists' ability to detect CRC in CT interpretation, instead of replacing established first-line tools in CRC screening, such as colonoscopy. In clinical practice, bowel preparation is usually performed for patients with indications of gastrointestinal diseases. Our model is applicable to a larger volume of contrast-enhanced CT scans for various clinical indications, including both with and without bowel preparation, offering broader clinical value for the detection or opportunistic screening of CRC.[33] Although the detection performance of radiologists was significantly improved with DL assistance, their specificity was decreased, which may lead to overdiagnosis in clinical practice. Notably, the DL model correctly classified the majority of false-positive cases diagnosed by radiologists. Therefore, the decrease may be attributed to an overlap of imaging features visualized in venous phase CT images. One potential solution is to integrate clinical information and multiphase images for detection performance enhancements.

This study had some limitations. First, since the DL model did not take into account colorectal benign lesions, it demonstrated poor performance in the detection of benign lesions (Supplementary Table S15), hindering the applicability of the DL model in real clinical scenarios. Nevertheless, the DL model delivered satisfactory performance in the detection of CRC. Second, despite the favorable performance of the DL model on three independent datasets and a real-world dataset, further validation in prospective settings and diverse ethnicities should be conducted. Third, the reader study may not fully reflect the true clinical application potential of the DL model, given its difference from real-world reading scenarios, particularly in the integration of multiphase images and clinical information. Future work will focus on the detection of colorectal lesions, prospective validation, and the implementation of more realistic reading scenarios.

In conclusion, we developed a deep learning model that provided an automated and accurate workflow for the detection of CRC in contrast-enhanced CT images without bowel preparation. With further multicentre validation and comparison with radiologists, the DL model could potentially serve as an effective computer-aided tool for CRC detection.

References

1 Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2021;71:209–249.
2 Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. *CA Cancer J Clin.* 2023;73:17–48.
3 Helsingen LM, Kalager M. Colorectal cancer screening — approach, evidence, and future directions. *NEJM Evid.* 2022;1:1–13.
4 Lin JS, Perdue LA, Henrikson NB, Bean SI, Blasi PR. Screening for colorectal cancer: updated evidence report and systematic review for the US preventive services task force. *JAMA.* 2021;325:1978–1998.
5 Mettler FA, Bhargavan M, Faulkner K, et al. Radiologic and nuclear medicine studies in the United States and worldwide: frequency, radiation dose, and comparison with other radiation sources - 1950-2007. *Radiology.* 2009;253:520–531.
6 Rodriguez R, Perkins B, Y.S. Park P, Koo P J, Kundranda M, Chang J C. Detecting early colorectal cancer on routine CT scan of the abdomen and pelvis can improve patient's 5-year survival. *Arch Biomed Clin Res.* 2019;1:1–6.
7 Schöckel L, Jost G, Seidensticker P, Lengsfeld P, Palkowitsch P, Pietsch H. Developments in X-ray contrast media and the potential impact on computed tomography. *Invest Radiol.* 2020;55:592–597.
8 Nørgaard A, Dam C, Jakobsen A, Pløen J, Lindebjerg J, Rafaelsen SR. Selection of colon cancer patients for neoadjuvant chemotherapy by preoperative CT scan. *Scand J Gastroenterol.* 2014;49:202–208.
9 Johnson CD, Flicek KT, Mead-Harvey C, Quillen JK. Strategies for improving colorectal cancer detection with routine computed tomography. *Abdom Radiol.* 2023;48:1891–1899.
10 Klang E, Eifer M, Kopylov U, et al. Pitfalls in diagnosing colon cancer on abdominal CT. *Clin Radiol.* 2017;72:858–863.
11 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.* 2010;53:911–918.
12 Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. *Annu Rev Biomed Eng.* 2017;19:221–248.
13 Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med.* 2022;28:31–38.
14 Liu KL, Wu T, Chen PT, et al. Deep learning to distinguish pancreatic cancer tissue from non-cancerous pancreatic tissue: a

retrospective study with cross-racial external validation. *Lancet Digit Heal*. 2020;2:e303–e313.

15 Gao Y, Zeng S, Xu X, et al. Deep learning-enabled pelvic ultrasound images for accurate diagnosis of ovarian cancer in China: a retrospective, multicentre, diagnostic study. *Lancet Digit Heal*. 2022;4:e179–e187.

16 Kim HE, Kim HH, Han BK, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digit Heal*. 2020;2:e138–e148.

17 Yao L, Xia Y, Zhang H, et al. DeepCRC: colorectum and colorectal cancer segmentation in CT scans via deep colorectal coordinate transform. In: Wang L, Dou Q, Fletcher PT, Speidel S, Li S, eds. *Medical image computing and computer assisted intervention – MICCAI 2022*. Cham: Springer Nature Switzerland; 2022:564–573.

18 Yushkevich PA, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage*. 2006;31:1116–1128.

19 Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18:203–211.

20 Zhou Y, Chen H, Li Y, et al. Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images. *Med Image Anal*. 2021;70:101918.

21 Crum WR, Camara O, Hill DLG. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Trans Med Imaging*. 2006;25:1451–1461.

22 Zieliński W. The shortest clopper–pearson confidence interval for binomial probability. *Commun Stat Simul Comput*. 2009;39:188–193.

23 Obuchowski NA, Rockette HE. Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: an anova approach with dependent observations. *Commun Stat Simul Comput*. 1995;24:285–308.

24 Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method. *Invest Radiol*. 1992;27(9):723–731.

25 Zhou XH, Obuchowski NA, McClish DK. *Statistical methods in diagnostic medicine*. John Wiley & Sons; 2011;712:1–545.

26 Balthazar EJ, Megibow AJ, Hulnick D, Naidich DP. Carcinoma of the colon: detection and preoperative staging by CT. *AJR Am J Roentgenol*. 1988;150:301–306.

27 Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer*. 2018;18:500–510.

28 Skrede OJ, De Raedt S, Kleppe A, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *Lancet*. 2020;395:350–360.

29 Echle A, Grabsch HI, Quirke P, et al. Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning. *Gastroenterology*. 2020;159:1406–1416.e11.

30 Urban G, Tripathi P, Alkayali T, et al. Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. *Gastroenterology*. 2018;155:1069–1078.e8.

31 Zheng S, Lin X, Zhang W, et al. MDCC-Net: multiscale double-channel convolution U-Net framework for colorectal tumor segmentation. *Comput Biol Med*. 2021;130. https://doi.org/10.1016/j.compbiomed.2020.104183.

32 Argilés G, Tabernero J, Labianca R, et al. Localised colon cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol*. 2020;31:1291–1305.

33 Pickhardt PJ. Value-added opportunistic CT screening: state of the art. *Radiology*. 2022;303:241–254.