Digestive Endoscopy 2021; ••: ••–•• doi: 10.1111/den.14187

Original Article

Performance of artificial intelligence for detection of subtle and advanced colorectal neoplasia

Omer F. Ahmad,[1,2,3] Juana González-Bueno Puyal,[1,4] Patrick Brandao,[1,4] Rawen Kader,[1,2] Faisal Abbasi,[2] Mohamed Hussein,[1,2] Rehan J. Haidry,[2,3] Daniel Toth,[4] Peter Mountney,[4] Ed Seward,[3] Roser Vega,[3] Danail Stoyanov[1] and Laurence B. Lovat[1,2,3]

[1]*Wellcome/EPSRC Centre for Interventional and Surgical Sciences,* [2]*Division of Surgery and Interventional Sciences, University College London,* [3]*Gastrointestinal Services, University College London Hospital and* [4]*Odin Vision Ltd, London, UK*

Objectives: There is uncertainty regarding the efficacy of artificial intelligence (AI) software to detect advanced subtle neoplasia, particularly flat lesions and sessile serrated lesions (SSLs), due to low prevalence in testing datasets and prospective trials. This has been highlighted as a top research priority for the field.

Methods: An AI algorithm was evaluated on four video test datasets containing 173 polyps (35,114 polyp-positive frames and 634,988 polyp-negative frames) specifically enriched with flat lesions and SSLs, including a challenging dataset containing subtle advanced neoplasia. The challenging dataset was also evaluated by eight endoscopists (four independent, four trainees, according to the Joint Advisory Group on gastrointestinal endoscopy [JAG] standards in the UK).

Results: In the first two video datasets, the algorithm achieved per-polyp sensitivities of 100% and 98.9%. Per-frame sensitivities were 84.1% and 85.2%. In the subtle dataset, the algorithm detected a significantly higher number of polyps ($P < 0.0001$), compared to JAG-independent and trainee endoscopists, achieving per-polyp sensitivities of 79.5%, 37.2% and 11.5%, respectively. Furthermore, when considering subtle polyps detected by both the algorithm and at least one endoscopist, the AI detected polyps significantly faster on average.

Conclusions: The AI based algorithm achieved high per-polyp sensitivities for advanced colorectal neoplasia, including flat lesions and SSLs, outperforming both JAG independent and trainees on a very challenging dataset containing subtle lesions that could have been overlooked easily and contribute to interval colorectal cancer. Further prospective trials should evaluate AI to detect subtle advanced neoplasia in higher risk populations for colorectal cancer.

Key words: artificial intelligence, colonic polyps, colonoscopy, colorectal neoplasms, deep learning

INTRODUCTION

ARTIFICIAL INTELLIGENCE (AI) based systems for polyp detection have been shown to increase adenoma detection rate (ADR) in randomized controlled trials. To date, these have been limited to non-advanced adenomas.[1]

There remains significant uncertainty regarding the efficacy of AI software to detect advanced neoplasia, particularly flat lesions, due to low prevalence of these subtle abnormalities in both pre-clinical testing datasets and prospective trials.[2,3] A similar issue exists for sessile serrated lesions (SSLs).

Corresponding: Omer F. Ahmad, Wellcome/EPSRC Centre for Interventional and Surgical Sciences, Charles Bell House, 43-45 Foley Street, London W1W 7TS, UK. Email: ofahmad123@gmail.com
Received 10 July 2021; accepted 5 November 2021.

This issue is particularly important since there is debate about whether the increased detection of non-advanced adenomas alone translates to reductions in interval colorectal cancers (CRCs). Improving the performance of AI to detect more challenging and advanced lesions, was ranked as the second highest priority in a recent international research priority setting exercise for AI in colonoscopy.[4] In particular a recommendation was made to create enriched datasets with subtle lesions, especially in scenarios where perceptual errors can occur. This was further emphasized by a recent literature review.[3]

Although current research efforts predominantly focus on prospective evaluation of computer aided detection (CADe) software in clinical trials, there remains an important role for retrospective pre-clinical studies using video datasets. These allow for evaluation and improvement of standalone technical performance of the AI software, and comparison of performance against multiple endoscopists who view the

same videos. Current datasets are often limited by selection bias, largely containing lesions that are readily identified during routine clinic practice.[5]

In this study, we aimed to develop video datasets that were enriched with flat lesions, SSLs and advanced colorectal polyps, to evaluate AI technical performance, including a perceptually challenging video database to also allow for comparisons of AI performance against endoscopists.

METHODS

Datasets

Training and initial test set

TO DEVELOP THE deep-learning based algorithm, a training dataset was created which consisted of a combination of still colonoscopy images and videos (Dataset A and Dataset B). Dataset B is a public dataset containing 10,993 polyp-positive frames (CVC-ColonDB300, CVC-ClinicDB612, CVC-ClinicHDSegmentTrain and CVC-Video databases).[6–9] A video database was also created at our institution between August 2018 and March 2019 consisting of complete colonoscopy withdrawals from 50 patients (cecum to rectum) using Olympus (Tokyo, Japan) EVIS LUCERA CV290(SL) processors and colonoscopes, recorded at 25 frames per second. Patients with advanced CRC or inflammatory bowel disease were excluded. Procedures were performed by two expert national bowel cancer screening accredited colonoscopists (ADR >45%). All polyps were confirmed by histopathology. Polyp size, morphology and location were also recorded. Full-length videos (white light only) were divided into shorter polyp-positive and -negative sequences. Magnification or near-focus frames were excluded. Polyp-positive frames were annotated based on the methods described in Appendix S1. The 50 procedures containing 210 polyps were randomly split on a per-procedure basis to create training (Dataset A), tuning and initial test datasets (Dataset C) consisting of 33, two and 15 procedures, respectively. The datasets are described in further detail in Table 1. The tuning dataset was used for optimizing the model hyperparameters.

Prospective independent test datasets

Once the algorithm had been developed and initially evaluated in the first step above, we prospectively recorded a further 45 patient colonoscopy withdrawals using the same methods described previously at our institution between April 2019 and November 2019. Twenty of these procedures contained 88 polyps and 25 were negative. Based on the methods in Appendix S1, these generated 8950 polyp-positive (Dataset D) and 542,484 polyp-negative frames (Dataset E) which are described in further detail in Table 1.

'Subtle' and perceptually challenging dataset

To specifically evaluate the algorithm on perceptually challenging lesions, we prospectively collected colonoscopy polyp encounter videos, during routine clinical care, where two expert endoscopists identified subtle visual cues of polyps in 'near miss' scenarios. Short white light video sequences were generated. Initial early sequences of the polyp encounter, including the subtle visual cues of the polyp, were created. The median length of these videos was 9.5 s (interquartile range [IQR] 8.0–10.0). In these situations, the polyp was not immediately identified i.e. the operator continued to withdraw a few folds before noticing the subtle visual cue, or the lesion was in the periphery or distance of the visual field before being recognized. For the same polyp encounters, we also created paired late short sequences, where the same polyp had been brought close into view, and optimally positioned i.e. centered just prior to polypectomy. The median length of these videos was 4.0 s (IQR 3.0–6.0). All of these polyps were confirmed by histopathology. Frames were annotated according to the methods in Appendix S1. A total of 39 polyps were included from 30 patients resulting in a total of 7683 polyp-positive frames (Dataset F). We named this the University College London (UCL)-subtle polyp dataset.

External validation dataset

The ETIS-LARIB open database consists of 196 high-definition polyp-positive frames, from 44 different polyps involving 31 sequences, captured using Pentax 90i series, EPKi 7000 processors (Dataset G).[9,10]

All the datasets are summarized in Table 1; two were used for training (Datasets A and B) and five for testing purposes (Datasets C, D, E, F and G). The test datasets were independent of all training processes with no patient overlap.

Algorithm development

A fully convolutional network with a ResNet-101 backbone architecture was used. The model was trained with Pytorch on an NVIDIA GeForce RTX 2080 Ti GPU. Further algorithm development details are included in Appendix S1.

Evaluating the algorithm

The bounding box annotations were used as ground truth for polyp presence or absence, with all polyps included in the study confirmed by histopathology. Performance metrics for

Table 1 Description of all the datasets used to train and test the artificial intelligence algorithm

Training datasets		Initial test dataset (Dataset C)
Dataset A	Dataset B	
33 procedures 53,849 polyp-positive frames 5000 polyp-negative frames 158 polyps Mean size = 7.0 ± 4.9 mm **Paris classification** Protruded 46% (72) Flat/Flat elevated 54% (86) **Location** Right 65% (102) Left 32% (51) Rectum 3% (5) **Pathology** HGD adenoma 1% (1) LGD adenoma 78% (124) Sessile serrated lesion 17% (26) Hyperplastic 4% (7)	Public datasets (CVC-ColonDB300, CVC-ClinicDB612, CVC-ClinicHDSegmentTrain and CVC-Video databases) 10,993 polyp-positive video + static (still image) frames	15 procedures 18,481 polyp-positive frames 92,504 polyp-negative frames 46 polyps Mean size = 8.0 ± 5.0 mm **Paris classification** Protruded 57% (26) Flat/Flat elevated 43% (20) **Location** Right 74% (34) Left 13% (6) Rectum 13% (6) **Pathology** HGD adenoma 2% (1) LGD adenoma 52% (24) Sessile serrated lesion 39% (18) Hyperplastic 7% (3) Advanced colorectal polyp 35% (16)

Prospective independent validation test datasets		Perceptually challenging subtle test dataset (Dataset F)	External validation (Dataset G)
Dataset D	Dataset E		
20 procedures 8950 polyp-positive video frames 88 polyps Mean size = 8.8 ± 5.4 mm **Paris classification*** Protruded 45% (40) Flat/Flat elevated 55% (48) **Location** Right 70% (62) Left 26% (23) Rectum 3% (3) **Pathology** LGD adenoma 39% (34) Sessile serrated lesion 54% (48) Hyperplastic 7% (6) Advanced colorectal polyp 42% (37) *LST-G-H (IIa + Is) = 2, LST-G-M (IIa + Is) = 1, LST-NG-PD (IIa + IIc) = 1, LST-NG-F (IIa) = 1	25 negative procedures 542,484 non-polyp frames	30 procedures 7683 polyp-positive frames 39 polyps Mean size = 10.2 ± 7.3 mm **Paris classification*** Protruded 31% (12) Flat/Flat elevated 69% (27) **Location** Right 77% (30) Left 18% (7) Rectum 5% (2) **Pathology** LGD adenoma 46% (18) Sessile serrated lesion 54% (21) Advanced colorectal polyp 36% (14) *LST-NG-F (IIa) = 4, LST-G-H (IIa) = 1 LST-G-M (IIa + Is) = 1	ETIS-LARIB database 196 high definition frames (still images) 44 polyps from 31 sequences Polyp size, morphology and histopathology data not available

LST-G-H, laterally spreading tumor, granular homogeneous type; LST-G-M, laterally spreading tumor, granular mixed-nodular type; LST-NG-F, laterally spreading tumor, non-granular flat-elevated type; LST-NG-PD, laterally spreading tumor, non-granular pseudo-depressed type.

evaluating the algorithm performance included per-frame sensitivity, per-frame specificity, and per-frame positive predictive value. A true positive occurred when the algorithm bounding box overlapped with the ground truth bounding box. Per-polyp sensitivity was defined as the number of polyps correctly detected by the model in at least one frame

divided by the total number of polyps present in the test dataset. For the UCL-subtle dataset, time to detection was also calculated. Further detailed definitions are in Appendix S1.

We also utilized an existing published false positive CADe clinical classification system to categorize false positives.[11] For the purposes of this analysis, a total of 80 false positives were randomly extracted based on duration of appearance.

Endoscopist evaluation

To compare performance with the algorithm, and also evaluate the perceptual difficulty of the UCL-subtle polyp dataset, eight endoscopists from our institution reviewed the same 34 video clips containing 39 polyps. The endoscopists had never seen the lesions before. In these instances, only the challenging early video polyp sequences were included. Further methodological details are included in Appendix S1.

Two groups of endoscopists participated. The first consisted of four independent colonoscopists who had performed >1000 colonoscopies and were accredited according to the Joint Advisory Group on gastrointestinal endoscopy (JAG) national standards in the UK. The second group included four JAG non-independent (trainee) endoscopists who had performed <500 colonoscopies.

Statistical analysis

Parametric continuous variables are expressed as means with standard deviation and non-parametric variables as medians with IQR. Clopper–Pearson exact 95% confidence intervals (CIs) were calculated. Chi-squared, or Fisher's exact test where appropriate, was used to compare differences in categorical variables. The Mann–Whitney *U* test was used to compare differences in polyp detection reaction times between endoscopists and the convolutional neural network (CNN). $P < 0.05$ was considered to be statistically significant. All statistical analyses were performed using GraphPad Prism (version 8; San Diego, CA, USA).

Ethics

The study was approved by the Cambridge central research medical ethics committee (REC Reference No. 18/EE/0148).

RESULTS

Algorithm performance

THE ALGORITHM WAS first evaluated on the initial test set (Dataset C), which consisted of 15 colonoscopy procedures containing 46 polyps. The model achieved a per-polyp sensitivity of 100% (95% CI 92.3–100.0%). The per frame sensitivity was 84.1% (95% CI 83.6–84.6%) and per frame specificity was 79.6% (95% CI 79.3–79.8%).

Further evaluation was undertaken on the prospective independent test datasets (Datasets D and E). These included 20 procedures with 88 polyps and 25 completely negative withdrawals. The algorithm achieved a per-polyp sensitivity of 98.9% (95% CI 93.8–100.0%) and a per frame sensitivity of 85.2% (95% CI 84.5–85.9%). The per-frame specificity was 79.2% (95% CI 79.1–79.3%) for the negative withdrawals.

Evaluation on the perceptually difficult UCL-subtle dataset was split into performance on the early and late polyp sequences. For early sequences, the algorithm achieved a per-polyp sensitivity of 79.5% (95% CI 63.5–90.7%), a per-frame sensitivity of 18.8% (95% CI 17.6–20.1%). For late sequences, per-polyp sensitivity and per frame sensitivity were 100% (95% CI 91.0–100.0%) and 80.1% (95% CI 78.8–81.3%), respectively. There were statistically significant differences in all metrics when comparing early and late sequence performance.

The inference time for the algorithm was 53.5 ms for the testing dataset (using a NVIDIA Geforce RTX 2080 Ti GPU), meeting the requirements for real-time detection.

Figure 1 and Video S1 demonstrate detections of subtle polyps by the algorithm.

Performance in test datasets (Datasets C, D, E and G) and the subtle dataset (Dataset F) are summarized in Tables 2 and 3.

Further subgroup analyses based on polyp size, morphology and histopathology are included in Tables S2 and S3.

Endoscopist performance on perceptually difficult UCL-subtle dataset

The Joint Advisory Group independent endoscopists achieved a per-polyp sensitivity of 37.2% (95% CI 29.6–45.3%) and a per-polyp positive predictive value of 78.4% (95% CI 67.3–87.1%).

JAG non-independent (trainee) endoscopists achieved a per-polyp sensitivity of 11.5% (95% CI 7.0–17.6%) and a per-polyp positive predictive value of 45.0% (95% CI 29.3–61.5%).

The CNN detected significantly more polyps ($P < 0.0001$) than both independent and trainee endoscopists with a per-polyp sensitivity of 79.5% (95% CI 63.5–90.7%).

When considering only polyps that were detected by both the CNN and at least one endoscopist, after correcting for baseline endoscopist reaction times, the median detection times for JAG independent endoscopists and the CNN were

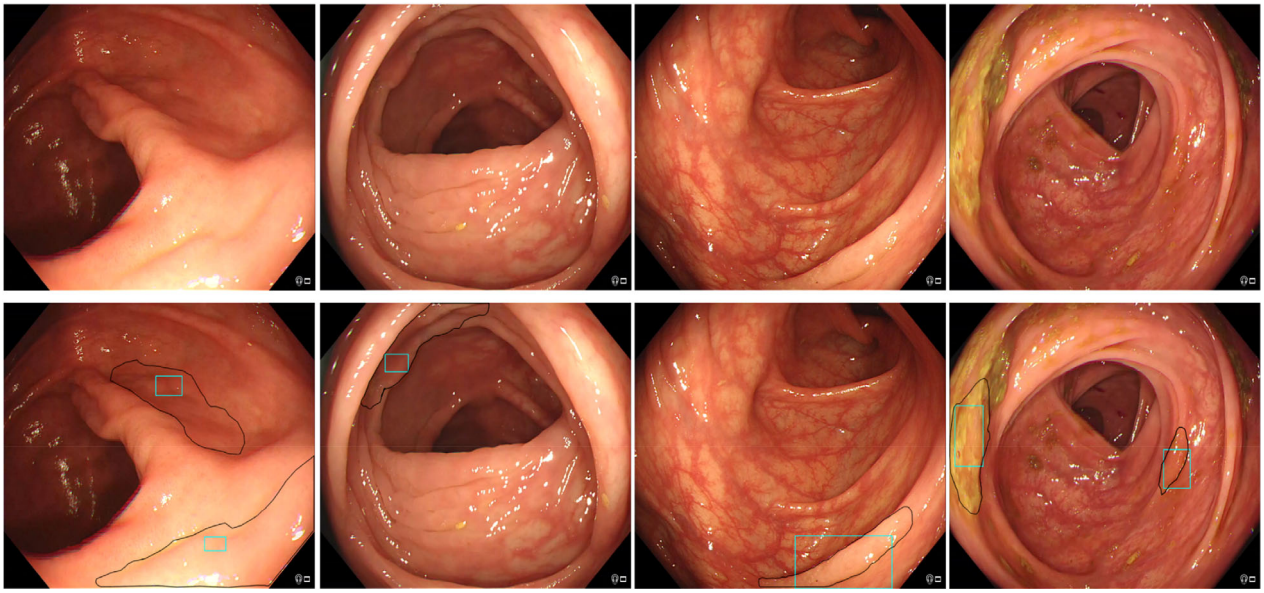

Figure 1 Examples of subtle polyp detections by algorithm. Top row contains raw images and bottom row contains corresponding images with the algorithm output (blue bounding box) and a black outline highlighting the polyp area. From left to right, the first image contains two LSTs (LST-G-H and LST-NG-F) subtypes in the cecum, the second image contains a LST-NG-F subtype in the transverse colon, the third contains a sessile serrated lesion (SSL) in the transverse colon and the final image contains two SSLs in the transverse colon. LST-G-H, laterally spreading tumor, granular homogeneous type; LST-NG-F, laterally spreading tumor, non-granular flat-elevated type.

Table 2 Algorithm performance in test datasets C, D, E and G

	Dataset C	Dataset D	Dataset E	Dataset G
Per-polyp sensitivity [95% CIs] (*n*)	100% [92.3–100.0] (46/46)	98.9% [93.8–100.0] (87/88)	N/A	N/A
Per frame sensitivity [95% CIs] (*n*)	84.1% [83.6–84.6] (16,450/19,560)	85.2% [84.5–85.9] 7847/9210	N/A	82.6% [76.9–87.5] (176/213)
Per frame specificity [95% CIs] (*n*)	79.6% [79.3–79.8] (76,611/96,301)	N/A	79.2% [79.1–79.3] (441,391/557,161)	N/A
Per frame positive predictive value [95% CIs] (*n*)	45.5% [45.0–46.0] (16,450/36,140)	90.8% [90.2–91.4] (7847/8643)	N/A	93.6% [89.1–96.7] (176/188)
F1-score [95% CIs] (*n*)	59.1% [58.5–59.6] (16,450/27,850)	87.9% [87.2–88.6] (7847/8926.5)	N/A	87.8% [82.2–91.8] (176/200.5)

CI, confidence interval; N/A, not available.

Table 3 Algorithm performance on subtle dataset split into early and late sequences

	Early sequences	Late sequences	*P*-value
Per-polyp sensitivity [95% CIs] (*n*)	79.5% [63.5–90.7] (31/39)	100% [91.0–100.0] (39/39)	0.0026
Per frame sensitivity [95% CIs] (*n*)	18.8% [17.6–20.1] (721/3828)	80.1% [78.8–81.3] (3087/3855)	<0.0001
Per frame positive predictive value [95% CIs] (*n*)	32.2% [30.3–34.2] (721/2237)	95.2% [94.5–96.0] (3087/3241)	<0.0001

CI, confidence interval.

1.78 s (IQR 0.85–3.20) and 0.28 s (IQR 0.04–0.83) respectively across 22 polyps. The CNN was significantly faster at detecting these polyps ($P < 0.0001$). The CNN was also significantly faster than JAG non-independent trainees with median detection times of 1.19 s (IQR 0.99–2.83) and 0.28 s (IQR 0.04–0.56) ($P < 0.0001$) across 11 polyps.

False positive analysis

Of the randomly selected 80 false positives that were reviewed, 86% ($n = 69$) were caused by artefacts from the bowel wall and 14% ($n = 11$) were caused by artefacts from bowel content. The subcategories leading to the highest proportion of false positives included folds 43.8% ($n = 35$), followed by normal mucosa 16.3% ($n = 13$) and ileocecal valve 15% ($n = 12$). The results are summarized in Table S1.

DISCUSSION

IN THIS STUDY we developed and validated an AI polyp detection system across multiple datasets, effectively creating high-risk populations enriched with flat lesions including a high proportion of SSLs and advanced colorectal polyps including laterally spreading tumors (LSTs). Our model was evaluated exclusively on video frames, demonstrating high per lesion sensitivities. Moreover, in our unique perceptually challenging video dataset, our CADe detected significantly more subtle polyps when compared to both JAG independent and trainee endoscopists.

To our knowledge, our study reports the largest video validation for CNN performance on SSLs to date. Most prior landmark studies that report polyp histopathology contain either none or fewer than five SSLs in their video test sets.[12–15] Hassan *et al.*[16] reported results using one of the largest video test datasets, containing 338 polyps, however absolute SSLs numbers were not described, performance reporting was grouped with adenomas and was on a per lesion basis. Zhou *et al.*[17] specifically addressed SSL validation on a dataset of 42 SSLs, with a per frame and per-polyp sensitivity of 84.1% and 100%, respectively. However, in the dataset described by Zhou *et al.*, 47% of SSLs were located in the rectum and sigmoid, and overall 69% were diminutive, which are less likely to contribute to interval CRC. Similarly, video performance evaluation on flat neoplasia, particularly advanced lesions including LSTs, is very limited in published studies.[3] Misawa *et al.* recently published an open-access video dataset, with a reported CNN flat per-lesion sensitivity of 98.3%, and a per-frame sensitivity of 86.7%. This dataset contained 100 lesions, including 34 flat lesions, one LST and four SSLs.[13] Yamada *et al.* produced a flat morphology enriched video dataset of 56 lesions, where 44 were slightly elevated or depressed, reporting an overall per frame and per-lesion sensitivity of 74% and 100%, respectively.[15] On the basis of per-frame sensitivities, our model performance was better than Yamada *et al.* and was comparable to Misawa *et al.* on all our datasets, excluding the subtle dataset. There is considerable variability in the definition for per-polyp sensitivity across studies, future consensus definitions could improve benchmarking.

Existing published retrospective CADe studies have rarely compared AI video performance with multiple endoscopists. Wang *et al.*[18] performed a post-hoc analysis, using 159 short video clips of missed polyps from a double-blind CADe RCT. Three experienced endoscopists retrospectively reviewed the video clips achieving an overall per-polyp sensitivity of 17%. Almost half of the missed polyps were hyperplastic, and 91% of the adenomas were diminutive and 99% were sessile. Furthermore, only five SSLs and one advanced adenoma (LST) were included. Livovsky *et al.*[19] evaluated CADe performance on a large testing video set, containing 1393 procedures, including a subgroup of 'subtle polyps' missed by endoscopists, although these were defined by re-analysis of false positives, without corresponding data for polyp size, morphology or histopathology. Our perceptually challenging dataset was enriched with lesions which are critical for CRC prevention. Our study is also the first to introduce the concept of separating analyses into early and late polyp encounter sequences. We demonstrate significantly lower sensitivities for early sequences. This emphasizes the importance of focusing video dataset design on the most challenging component of sequences. We also validated perceptual difficulty, by performing multi-reader studies on the polyp encounters, demonstrating a superior performance of our CNN against both JAG accredited independent endoscopists and trainees. The relatively low sensitivity of endoscopists in this study suggests that recognition errors for subtle advanced neoplasia could be an important factor in interval CRCs. In addition, our results suggest that a learning curve may exist.

When considering false positives, overall the per-frame specificity was approximately 80%, which is slightly lower than other video studies.[20] However, we did not exclude low-quality images in our non-polyp frames, unlike many other studies, which are a common source of false positives. However, per-frame metrics alone may not reflect the clinical relevance of false positives, therefore we classified a random selection of false positives using a published scheme.[11] The distribution of causes of false positives was similar to that published for another CADe system, mostly

represented by artefacts from the bowel wall, and a smaller proportion due to bowel content. Similarly, the main two subcategories were folds and normal mucosa. The previously published classification study suggested that most of these were readily discarded by endoscopists. False positives may, however, lead to poor adoption of CADe systems. Further research is warranted to identify methods to address this, such as the use of recurrent neural networks, whilst ensuring that sensitivity is maintained for the detection of subtle, advanced colorectal neoplasia.

Limitations of this study include its retrospective design, with results possibly being subject to selection bias, however this was minimized by using video data including low quality image frames from perceptually challenging lesions. Bowel preparation scores were not recorded for procedures, therefore its effect on CADe performance was not evaluated. Also, we did not evaluate AI-endoscopist interaction. The clinical impact of CADe systems will depend on the ability of operating endoscopists to recognize whether the AI output represents a true lesion, or it might be discarded incorrectly as a false positive. Furthermore, although we did perform an external validation using a still image dataset, it is very difficult to obtain video datasets enriched with subtle advanced lesions. Moreover, although we created a novel dataset, the absolute number of advanced subtle lesions was relatively small. Given the low prevalence of such lesions, large multi-center research collaborations will be required to overcome this limitation.

In conclusion, we evaluated the technical performance of a CADe algorithm to detect flat neoplasia, SSLs and advanced polyps demonstrating high sensitivity in a video dataset. Using a novel perceptually challenging dataset enriched with advanced lesions, the algorithm detected significantly more polyps than endoscopists. Prospective clinical trials should assess the ability of CADe systems to detect subtle advanced neoplasia in higher risk populations for CRC. However, ultimately population-based trials, targeting 'average-risk' individuals are required to establish the value of AI in CRC prevention.[21]

CONFLICT OF INTEREST

L.B.L. IS A minor shareholder in Odin Vision. He has received research grants from Medtronic, Pentax Medical and DynamX, and is on the Scientific Advisory Board of Dynamx, Odin Vision and Ninepoint Medical. D.S. is a shareholder in Odin Vision and Digital Surgery Ltd. P.M., J.G.-B.P., P.B. and D.T. are employees of Odin Vision Ltd. R.J.H. has received educational grants to support research infrastructure from Medtronic, Cook Medical (fellowship support), Pentax Europe, C2 Therapeutics, Beamline Diagnostics and Fractyl Ltd. The other authors declare no conflict of interest for this article.

FUNDING INFORMATION

L.B.L. IS SUPPORTED by the National Institute for Health Research University College London Hospitals Biomedical Research Centre. This work was supported by the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) at UCL [203145Z/16/Z].

REFERENCES

1 Barua I, Vinsard DG, Jodal HC *et al*. Artificial intelligence for polyp detection during colonoscopy: A systematic review and meta-analysis. *Endoscopy* 2021; **53**: 277–84.
2 Mori Y, Neumann H, Misawa M, Kudo S, Bretthauer M. Artificial intelligence in colonoscopy - now on the market. What's next? *J Gastroenterol Hepatol* 2021; **36**: 7–11.
3 Hassan C, Bhandari P, Antonelli G, Repici A. Artificial intelligence for non-polypoid colorectal neoplasms. *Dig Endosc* 2021; **33**: 285–9.
4 Ahmad OF, Mori Y, Misawa M *et al*. Establishing key research questions for the implementation of artificial intelligence in colonoscopy: A modified Delphi method. *Endoscopy* 2021; **53**: 893–901.
5 Vinsard DG, Mori Y, Misawa M *et al*. Quality assurance of computer-aided detection and diagnosis in colonoscopy. *Gastrointest Endosc* 2019; **90**: 55–63.
6 Vázquez D, Bernal J, Sánchez FJ *et al*. A benchmark for endoluminal scene segmentation of colonoscopy images. *J Healthc Eng* 2017; **2017**: 1–9.
7 Bernal J, Sánchez J, Vilariño F. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognit* 2012; **45**: 3166–82.
8 Bernal J, Sánchez FJ, Fernández-Esparrach G, Gil D, Rodríguez C, Vilariño F. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Comput Med Imaging Graph* 2015; **43**: 99–111.
9 Bernal J, Tajkbaksh N, Sanchez FJ *et al*. Comparative validation of polyp detection methods in video colonoscopy: Results from the MICCAI 2015 endoscopic vision challenge. *IEEE Trans Med Imaging* 2017; **36**: 1231–49.
10 Silva J, Histace A, Romain O, Dray X, Granado B. Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. *Int J Comput Assist Radiol Surg* 2014; **9**: 283–93.
11 Hassan C, Badalamenti M, Maselli R *et al*. Computer-aided detection-assisted colonoscopy: Classification and relevance of false positives. *Gastrointest Endosc* 2020; **92**: 900–4.e4.
12 Misawa M, Kudo S, Mori Y *et al*. Artificial intelligence-assisted polyp detection for colonoscopy: Initial experience. *Gastroenterology* 2018; **154**: 2027–9.e3.
13 Misawa M, Kudo S, Mori Y *et al*. Development of a computer-aided detection system for colonoscopy and a publicly

accessible large colonoscopy video database (with video). *Gastrointest Endosc* 2021; **93**: 960–7.e3.
14 Wang P, Xiao X, Glissen Brown JR *et al*. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. *Nat Biomed Eng* 2018; **2**: 741–8.
15 Yamada M, Saito Y, Imaoka H *et al*. Development of a real-time endoscopic image diagnosis support system using deep learning technology in colonoscopy. *Sci Rep* 2019; **9**: 1–9.
16 Hassan C, Wallace MB, Sharma P *et al*. New artificial intelligence system: First validation study versus experienced endoscopists for colorectal polyp detection. *Gut* 2020; **69**: 799–800.
17 Zhou G, Xiao X, Tu M *et al*. Computer aided detection for laterally spreading tumors and sessile serrated adenomas during colonoscopy. *PLoS One* 2020; **15**: e0231880.
18 Wang P, Liu X, Berzin TM *et al*. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): A double-blind randomised study. *Lancet Gastroenterol Hepatol* 2020; **5**: 343–51.
19 Livovsky DM, Veikherman D, Golany T *et al*. Detection of elusive polyps using a large-scale artificial intelligence system (with videos). *Gastrointest Endosc* 2021; **94**: 1099–109.E10.
20 Guo Z, Nemoto D, Zhu X *et al*. Polyp detection algorithm can detect small polyps: *Ex vivo* reading test compared with endoscopists. *Dig Endosc* 2021; **33**: 162–9.
21 Mori Y, Bretthauer M, Kalager M. Hopes and hypes for artificial intelligence in colorectal cancer screening. *Gastroenterology* 2021; **161**: 774–7.

SUPPORTING INFORMATION

ADDITIONAL SUPPORTING INFORMATION may be found in the online version of this article at the publisher's web site.

Table S1 False positive classifications.

Table S2 Algorithm subgroup analysis for flat lesions, sessile serrated lesions, advanced colorectal lesions and laterally spreading tumors.

Table S3 Algorithm subgroup analysis on all testing datasets combined based on polyp histopathology and morphology.

Appendix S1 Algorithm development, evaluation metrics, annotation methods and endoscopist video evaluation on UCL-subtle dataset.

Video S1 Videos demonstrating subtle polyp detections by the CNN, including advanced adenomas, laterally spreading tumors and sessile serrated lesions.