

Towards a robust and compact deep learning system for primary detection of early Barrett's neoplasia: Initial image-based results of training on a multi-center retrospectively collected data set

Kiki N. Fockens¹ | Jelmer B. Jukema¹ | Tim Boers² | Martijn R. Jong¹ | Joost A. van der Putten² | Roos E. Pouw¹  | Bas L. A. M. Weusten^{3,4} | Lorenza Alvarez Herrero⁴ | Martin H. M. G. Houben⁵ | Wouter B. Nagengast⁶  | Jessie Westerhof⁶ | Alaa Alkhalaf⁷ | Rosalie Mallant⁸ | Krish Ragunath⁹ | Stefan Seewald¹⁰ | Peter Elbe^{11,12} | Maximilien Barret¹³ | Jacobo Ortiz Fernández-Sordo¹⁴ | Oliver Pech¹⁵ | Torsten Beyna¹⁶ | Fons van der Sommen² | Peter H. de With² | A. Jeroen de Groof¹  | Jacques J. Bergman¹

Correspondence

A. Jeroen de Groof, Amsterdam UMC, location AMC Meibegdreef 9 Amsterdam, 1105 AZ, The Netherlands.
Email: a.j.degroof@amsterdamumc.nl

Funding information

Olympus Tokyo, Japan

Abstract

Introduction: Endoscopic detection of early neoplasia in Barrett's esophagus is difficult. Computer Aided Detection (CADe) systems may assist in neoplasia detection. The aim of this study was to report the first steps in the development of a CADe system for Barrett's neoplasia and to evaluate its performance when compared with endoscopists.

Methods: This CADe system was developed by a consortium, consisting of the Amsterdam University Medical Center, Eindhoven University of Technology, and 15 international hospitals. After pretraining, the system was trained and validated using 1.713 neoplastic (564 patients) and 2.707 non-dysplastic Barrett's esophagus (NDBE; 665 patients) images. Neoplastic lesions were delineated by 14 experts. The performance of the CADe system was tested on three independent test sets. Test set 1 (50 neoplastic and 150 NDBE images) contained subtle neoplastic lesions representing challenging cases and was benchmarked by 52 general endoscopists. Test set 2 (50 neoplastic and 50 NDBE images) contained a heterogeneous case-mix of neoplastic lesions, representing distribution in clinical practice. Test set 3 (50 neoplastic and 150 NDBE images) contained prospectively collected imagery. The main outcome was correct classification of the images in terms of sensitivity.

Kiki N. Fockens and Jelmer B. Jukema both authors contributed equally to this article

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. United European Gastroenterology Journal published by Wiley Periodicals LLC on behalf of United European Gastroenterology.

Results: The sensitivity of the CAdE system on test set 1 was 84%. For general endoscopists, sensitivity was 63%, corresponding to a neoplasia miss-rate of one-third of neoplastic lesions and a potential relative increase in neoplasia detection of 33% for CAdE-assisted detection. The sensitivity of the CAdE system on test sets 2 and 3 was 100% and 88%, respectively. The specificity of the CAdE system varied for the three test sets between 64% and 66%.

Conclusion: This study describes the first steps towards the establishment of an unprecedented data infrastructure for using machine learning to improve the endoscopic detection of Barrett's neoplasia. The CAdE system detected neoplasia reliably and outperformed a large group of endoscopists in terms of sensitivity.

KEYWORDS

artificial intelligence, Barrett's esophagus, Barrett's neoplasia, computer aided detection, endoscopy, machine learning

INTRODUCTION

Barrett's esophagus (BE) is a well-known precursor for the development of esophageal adenocarcinoma (EAC). BE patients undergo regular endoscopic surveillance to detect neoplasia at an early stage.^{1,2} Early neoplastic lesions may be difficult to detect, given their sometimes-subtle endoscopic appearance.^{3,4} Current state-of-the-art endoscopes enable visualization of nearly all subtle endoscopic changes. However, these subtle lesions are not always recognized since most endoscopists are unfamiliar with their appearance. A tool assisting the endoscopist in the recognition of such subtle changes may improve the quality of BE surveillance.

Computer aided detection (CAdE) systems for gastrointestinal applications largely focus on colonic polyp detection and characterization.⁵ CAdE systems for the detection of Barrett's neoplasia are still under development.^{6–8} For successful clinical implementation, such systems require robust training and efficient software to enable real-time application. Most CAdE systems described in the literature are trained with relatively small data sets originating from a single center. This limits their generalizability and results in suboptimal performance when applied in daily practice.⁹ In addition, most of the currently developed CAdE systems require significant computational resources, which limits real-time video-based application and hampers efficient integration in existing endoscopy systems.

Our consortium envisions to develop a robust and 'ready-for-use' CAdE system for the detection of early Barrett's neoplasia in a stepwise manner, using efficient and generic software that enables easy integration in the endoscopy suite. In this paper, we aim to describe the research infrastructure of our consortium and to report the first results of CAdE performance on still images after training the algorithm on retrospectively collected imagery.

METHODS

We aimed to develop a CAdE system for the primary detection of Barrett's neoplasia on white-light endoscopy (WLE) images in

Key summary

1. Summarise the established knowledge on this subject

- Endoscopic detection of Barrett's neoplasia is difficult
- Computer Aided Detection (CAdE) systems can assist the endoscopist in the detection of neoplasia

2. What are the significant and/or new findings of this study?

- This study describes the rigorous development of a CAdE system for Barrett's neoplasia, which detected neoplasia with high accuracy
- Our CAdE system outperforms the vast majority of general endoscopists in terms of the detection of Barrett's neoplasia

overview. It classifies images as either neoplastic or non-dysplastic, followed by the localization of neoplasia (if present) by the projection of a green bounding box around the lesion, thereby guiding the endoscopist to the region of interest.

BONS-AI consortium

This study was performed by the Department of Gastroenterology and Hepatology of the Amsterdam University Medical Centers, the Netherlands, a tertiary referral center for BE neoplasia, and the Department of Electrical Engineering of the Eindhoven University of Technology, the Netherlands. The Barrett's OesophaguS imaging for Artificial Intelligence (BONS-AI) consortium consists of 15 participating medical centers from 7 countries, all expert centers in the field of Barrett endoscopy: University Medical Center Utrecht, the Netherlands, Sint Antonius hospital Nieuwegein, the Netherlands, University Medical Center Groningen, the Netherlands, Isala hospital Zwolle, the Netherlands, Haga Teaching Hospital, the

Netherlands, Flevohospital Almere, the Netherlands, Onze Lieve Vrouwe Gasthuis hospital Amsterdam, the Netherlands, Cochin Hospital Paris, France, Hirslanden Klinik Zürich, Switzerland, Karolinska University Hospital Stockholm, Sweden, Evangelisches Krankenhaus Düsseldorf, Germany, Krankenhaus Barmherzige Brüder Regensburg, Germany, Nottingham University Hospital, United Kingdom, Royal Perth Hospital, Australia. Centers participated in the consortium by collecting data and providing ground truth delineations (see below). For the contribution per center, please see Supplementary Table 1.

The Medical Research Involving Human Subjects Act did not apply to this study. Official approval for this study was therefore waived by the medical ethics review committee of participating centers. This study was registered at the Dutch Trial Register under the number NL8411.

Data collection

To effectively train our CAdE system, we envisioned to create the largest Barrett imagery data set that is currently described. To this end, both retrospective and prospective data were collected by collaborative partners in our consortium. Retrospectively recorded endoscopic images were collected from the endoscopic databases of the participating centers that included patients under surveillance for their non-dysplastic Barrett's esophagus (NDBE) or undergoing endoscopic treatment of BE neoplasia. Endoscopic images recorded with Olympus 100-series endoscopes (H180 and HQ190) and processors (CV-180 and CV-190; Olympus, Tokyo, Japan) between 2012 and 2021 were selected. The images were automatically extracted from the system and de-identified using proprietary software specifically designed for this project. The de-identification software was developed to detect text (*i.e.*, patient identifiers) within imagery and subsequently place a black box over the text, followed by automatic overwriting of the original image. Any meta-data was automatically removed as well.

The majority of the participating centers also collected prospective imagery following a standardized image acquisition protocol. In this protocol, endoscopic images are collected in WLE in both NDBE and neoplastic patients. Endoscopic images were recorded for each 2 cm of the Barrett's segment. All imagery was obtained in overview without specific focus on areas of interest such as potential neoplastic lesions, if present. For an extensive description of the prospective data acquisition protocol, please see the supplementary materials.

All endoscopic imagery was manually labeled as either NDBE or neoplasia according to the histopathology definitions (see below). To guarantee patient privacy, endoscopic imagery was saved and de-identified. Prospective imagery was recorded completely anonymously in participating centers. Images were stored in lossless PNG, BMP, or TIFF format; videos were saved as MP4 files. All imagery was stored on a secured server to which only the researchers of this consortium had access.

Definitions and selection of neoplastic and NDBE images

For our data sets, we included WLE images from treatment naïve patients. Images were selected based on image quality. Parameters for inclusion included both content-independent quality (*e.g.*, sufficient contrast and sharpness and absence of blur) and content-dependent quality (*e.g.*, adequate esophageal expansion, illumination and mucosal cleanliness in terms of absence of blood and bubbles, thereby enabling circumferential evaluation of mucosal surface).

For neoplastic images, we employed the following additional selection criteria: (1) a visible neoplastic lesion within the image, and (2) high-grade dysplasia (HGD) or early adenocarcinoma (EAC) in biopsies or endoscopic resection specimens.

For NDBE images, we imposed the following additional selection criteria: (1) no visible abnormalities within the endoscopic image, and (2) absence of any degree of dysplasia in all tissue samples.

Images were excluded in case of (1) presence of tools within the image (*e.g.*, clips or biopsy forceps), (2) collapsed esophagus, (3) presence of blood or extensive amounts of mucous or bubbles, and (4) biopsies showing low-grade dysplasia in pathology results.

Creation of ground truth

All images were reviewed by three experts (KF, JJ, and MJ) for quality assessment and for the absence of any visible abnormalities (NDBE images) or the presence of a visible neoplastic lesion (neoplastic images).

To indicate the location of the neoplastic lesion within each neoplastic image, all lesions were delineated by at least two expert endoscopists from a group of 14 expert endoscopists from the participating centers (RP, BW, MH, WN, JW, LA, AA, KR, MB, JO, OP, TB, SS, and JB) using an online module (Meducati AB, Göteborg, Sweden). Experts were asked to provide two delineations for each image. First, they were asked to delineate the outer peripheral extent of the neoplastic lesion. This delineation contained the more subtle mucosal and vascular changes of the lesion. Subsequently, the experts were asked to delineate only the area where the neoplastic lesion was most profound. This area was considered to represent the highest likelihood of neoplasia (Figure 1). After delineation, the experts were asked to score: (1) the most prominent macroscopic component of the lesion (Paris classification 0-I, 0-II, or 0-III); (2) location of the lesion within the image (based on distance to endoscope, angle of imaging, insufflation/desufflation), on a scale of 1–3: 1 representing a poor location, 2 representing a moderate location, and 3 representing a good location; (3) quality of the image (based on the image resolution, overexposure of light and shadowing artifacts), on a scale of 1–3, where 1 represents poor quality and 3 good quality; (4) level of cleaning of the mucosa, on a scale of 1–3 with 1 representing poor cleaning and 3 good cleaning; (5) the subtlety of the neoplastic lesion on the image, incorporating all the four relevant features into a single visual analogue scale score ranging from 1 (very subtle) to 100

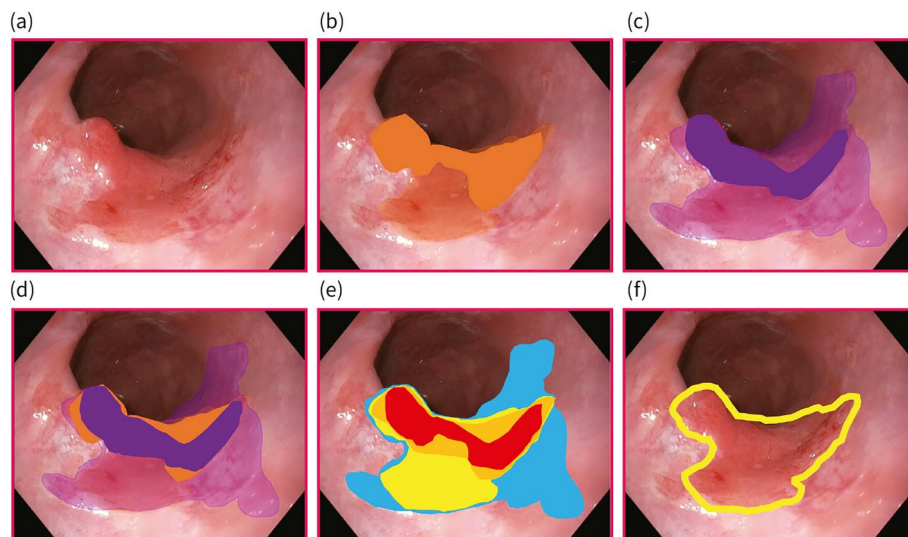


FIGURE 1 Creation of ground truth using expert delineations, an example of one case. (a) original image, (b) delineations of expert 1; (c) delineations of expert 2; (d) combined delineations of the two experts; (e) heatmap based on expert delineations; (f) eventual delineation used for ground truth.

(very obvious). Prior to the delineations, all experts received extensive instructions on the delineation process and rating criteria with endoscopic examples per category and these examples were available during the rating process.

Each neoplastic image was delineated by two experts after which delineation overlap was evaluated. In case of insufficient overlap (defined as a Dice score¹⁰ of overlapping delineations <0.3), a third expert delineated that particular image. The delineations of the two experts with the highest Dice score were subsequently used as ground truth for neoplasia. If it was not possible to reach consensus (Dice scores persistently <0.3), the image was excluded.

Figure 1 shows the combination of expert delineations in an exemplary case. The area containing both higher likelihood delineations and the overlapping area of the lower likelihood delineation of both experts is displayed in yellow and was used as ground truth for both training and testing of the CADe system. This area was considered to contain the most valuable information about true neoplasia.

For an extensive description of the process of creating the ground truth for training and testing the CADe system, please see Supplementary Video 1.

Data sets for the development of the Computer Aided Detection system

For the development of this CADe system, three different data sets of increasing proximity to the final application were used for stepwise training (Table 1):

First, a publicly available data set of random images (ImageNet 1K, <https://www.image-net.org/download.php>) was used for general pretraining.¹¹ This data set contains 1.200.000 general color images labeled into 1.000 different categories, none of them specifically related to endoscopy. During such pretraining, a deep learning

system learns basic features of images, such as edges and shapes. This training method eliminates the need to learn these basic features from images corresponding to the final application (i.e., Barrett's imagery), which are generally more scarcely available.

Second, GastroNet was used for domain-specific pretraining. This data set has been described before and contains 494.364 general endoscopic images obtained from the endoscopic archives of the Amsterdam University Medical Center, location Academic Medical Center.¹² A subset of 3.743 images were manually labeled by two experts into five categories: the esophagus, stomach, small intestines, colon, and others. The remaining images were classified into the same five categories as part of this pretraining method.¹³ The rationale for domain-specific pretraining was to familiarize the CADe system with the basic features of endoscopic imagery, thereby enhancing the pretraining process.

After pretraining, a third data set was used for domain-specific training of the CADe system. This data set contained 4.420 Barrett-specific WLE images derived from 1.229 patients (1.713 neoplastic images from 564 patients and 2.707 NDBE images from 665 patients), which were all retrospectively collected in 10 participating centers. All images were obtained using Olympus CV-190 processors using GIF-H180 and GIF-HQ190 endoscopes.

A subset of this training data set was subtracted prior to training, and therefore not used in the training data set. This subset was used for validation of the CADe system. Based on the results of this subset, the (hyper)parameters of the trained CADe system were optimized and the threshold for differentiating neoplastic from NDBE images was determined. This subset contained 233 neoplastic images (129 unique patients) and 374 NDBE images (73 unique patients). The images were carefully selected based on the subtlety score of the neoplastic images to create a subset comparable to the test set (see below) and to optimize the CADe system toward its intended application.

TABLE 1 Performance in terms of sensitivity and specificity on validation data set to determine optimal threshold for Computer Aided Detection (CADe) performance.

Data set and purpose	Type of imagery	Number of images (# patients)		Acquisition	Type of labeling
		Neoplastic	NDBE		
1. General pretraining on non-endoscopic images	ImageNet	1.200.000 (n.a.)		n.a.	n.a.
2. Domain-specific pretraining on general endoscopic images	GastroNet	494.364 (15.286)		Retrospective acquisition	Subset: Hand-labeled by 2 experts
3. Training	Barrett specific	1.480 (435)	2.333 (592)	Retrospective acquisition	Hand-labeled by 3 experts, correlating pathology, delineated by ≥ 2 experts
4. Validation	Barrett specific	233 (129)	374 (73)	Retrospective acquisition	Hand-labeled by 3 experts, correlating pathology, delineated by ≥ 2 experts
5. Testing	Barrett specific	50 (50)	150 (150)	Retrospective acquisition	Hand-labeled by 3 experts, correlating pathology, delineated by ≥ 2 experts
	Barrett specific	50 (50)	50 (50)	Retrospective acquisition	Hand-labeled by 3 experts, correlating pathology, delineated by ≥ 2 experts
	Barrett specific	50 (39)	150 (74)	Prospective acquisition	Hand-labeled by 3 experts, correlating pathology

Data sets for testing the Computer Aided Detection system

To test the performance of the CADe system, three independent test sets were created (Table 1, Figure 2). The images in these test sets were not used during pretraining, training, or validation of the CADe system and a patient-split was maintained between all data sets.

Test set 1 contained 200 retrospectively collected images: 50 neoplastic images derived from 50 patients and 150 NDBE images derived from 150 patients. A ratio of 1:3 neoplasia/NDBE was used since this better reflects clinical practice than a 1:1 split as was used in our previous studies.^{12,14} This first test set was artificially enriched with subtle neoplastic lesions (based on preselection of cases with a likelihood-of-detection-score <50 followed by independent further selection by two experts), mimicking a clinical setting during surveillance endoscopy where recognition of early neoplasia might be at stake.

Test set 2 contained 100 retrospectively collected images (50 neoplastic and 50 NDBE images, 1 image per patient). This test set contained a wide variety of neoplastic lesions in terms of subtlety, representing the distribution of neoplastic lesions encountered in clinical practice.

Test set 3 contained prospectively collected images recorded with the latest Olympus X1 endoscopy system using either HQ190 or EZ1500 endoscopes. This set contained 200 images: 50 neoplastic images derived from 39 patients and 150 NDBE images derived from 74 patients. Although our CADe system was not trained on either prospectively collected imagery or next-generation endoscopes, we wanted to have an estimate of its performance once under state-of-the-art circumstances.

Benchmark performance by general endoscopists on test set 1

The first test set was benchmarked by 52 general endoscopists originating from 3 countries to provide a reference for CADe performance. A previously designed web-based module (Meducati AB, Göteborg, Sweden) was used and adjusted for this specific study.^{15,16} For each endoscopic image, the endoscopists indicated if they detected neoplasia and, where applicable, placed a biopsy mark on the most abnormal part of the lesion. This mark represents the location where they would have taken a targeted biopsy during real-time endoscopic examination.

Architecture of Computer Aided Detection system

The CADe system was constructed using an EfficientNet-Lite1 encoder¹⁷ to extract the relevant image information and a MobileNetV2 DeepLabV3+ decoder¹⁸ to generate an output segmentation (Supplementary Figure 2). Both architectures are optimized for fast and efficient processing of real-time imagery and can be directly implemented into current endoscopy systems.

After the two-step pretraining process and transfer of the corresponding learned features, the encoder and decoder branch of the system were trained simultaneously using data set 3. This enabled optimal use of the classification labels and expert delineations of neoplastic images for both branches. Furthermore, the neoplastic images without expert delineations could be efficiently leveraged to improve the training of the classification branch. During validation, using only the segmentation branch (decoder) achieved the best performance for both classification and segmentation. For this

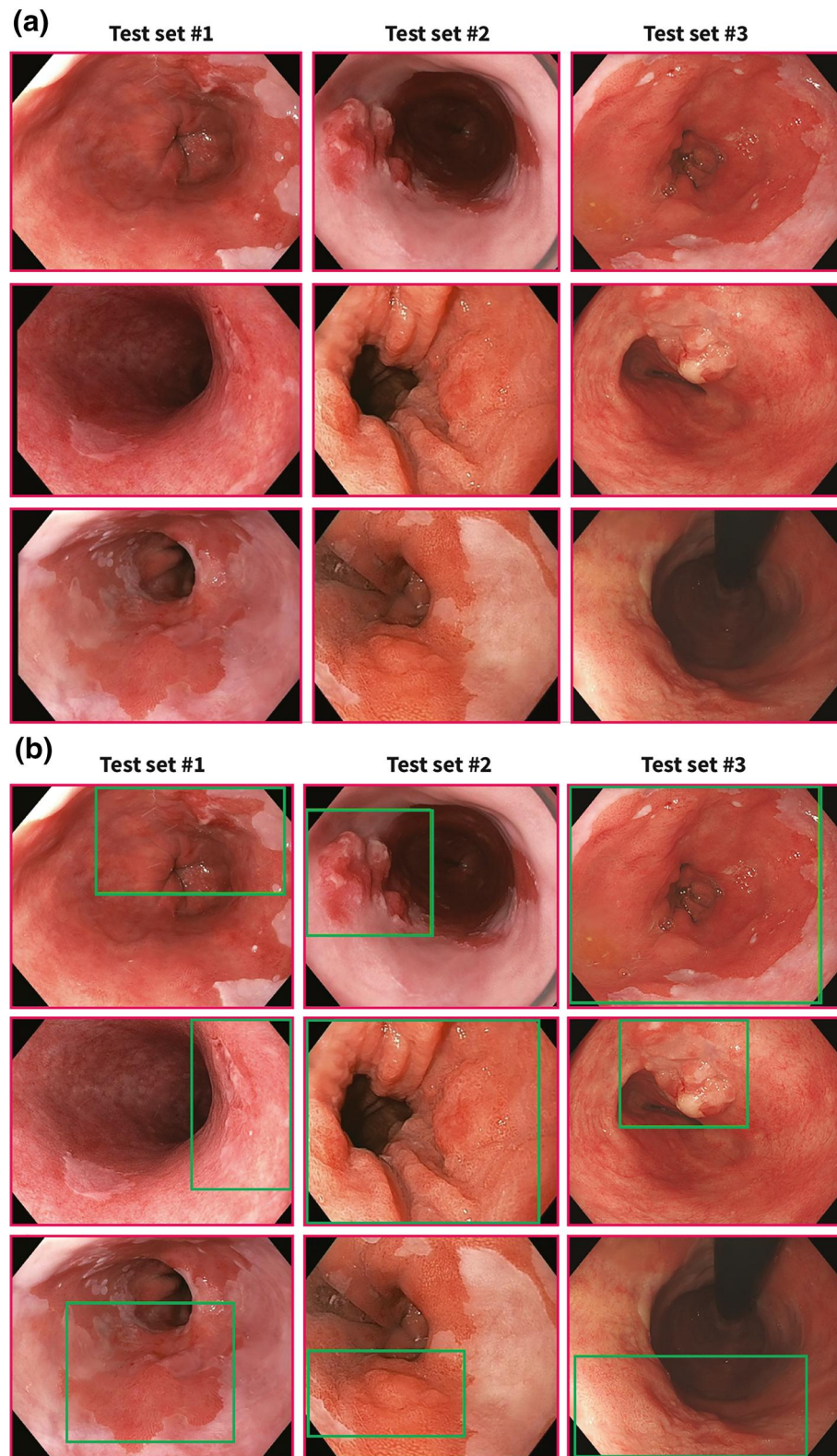


FIGURE 2 (a) Example of neoplastic images per test set; (b) corresponding bounding boxes based on Computer Aided Detection (CAdE) prediction. Test set 1: retrospective test set enriched with more subtle lesions; test set 2: retrospective test set with less subtle lesions; test set 3: prospective test set with different kinds of lesions.

reason, positive detection on the test set was defined as any prediction for which one or more pixels of the image exceeded the customizable threshold.

Additional technical details are described in the supplementary materials of this manuscript.

Outcome measurements

Primary outcome measurements:

- Correct classification of neoplastic images, reported in terms of sensitivity for the CAdE system and for general endoscopists on test set 1.
- Correct classification of neoplastic images, reported in terms of sensitivity for the CAdE system on test sets 2 and 3.

Secondary outcome measurements:

- Correct classification of NDBE images, reported in terms of specificity for the CAdE system and for general endoscopists on test set 1.
- Correct localization of neoplasia for test set 1, defined as overlap of the bounding box (CAdE system) or biopsy mark (endoscopists) with experts' ground truth (Figure 3).

- Correct classification of NDBE images reported in terms of specificity for the CAdE system on test sets 2 and 3.
- Processing speed of endoscopic images for the CAdE system.

Statistical analysis

Statistical analyses were performed using Python 3.8.10 (Python Software Foundation). Diagnostic accuracy per image was displayed using sensitivity and specificity. Due to the 25/75 split of neoplastic and non-dysplastic data, it was chosen not to display the accuracy. Localization performance was evaluated for the images that were correctly classified as neoplastic.

RESULTS

Internal validation results

During the training phase, the CAdE system's performance was evaluated on the validation set (Table 2, Figure 4) to optimize the (hyper)parameters and to determine the threshold for neoplasia detection afterward. This threshold is the optimal cut-off point value in terms of sensitivity while maintaining acceptable specificity. A threshold of 0.25 (with a corresponding sensitivity of 88.0% and

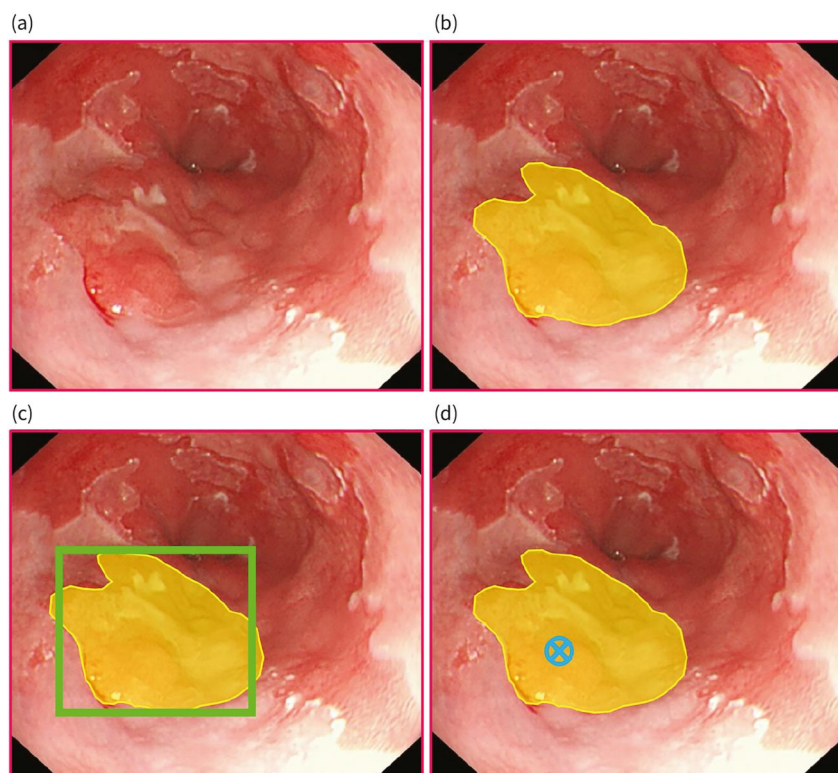
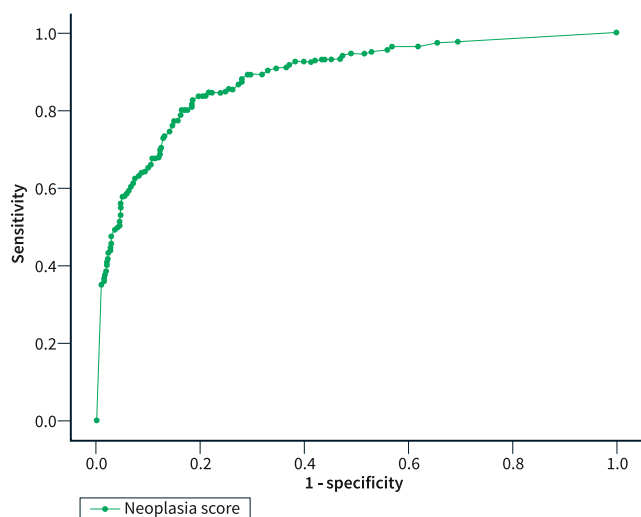


FIGURE 3 Difference in localization for Computer Aided Detection (CAdE) system and endoscopists: (a) original image; (b) ground truth based on expert delineations; (c) green bounding box of the CAdE system indicating the localization of the neoplastic lesion; (d) the biopsy mark of the endoscopist indicating the localization of the neoplastic lesion.

TABLE 2 Summary of performance on all three test sets for the Computer Aided Detection (CAdE) system and general endoscopists.

Threshold	Sensitivity	Specificity
0.15	92.7	60.0
0.20	89.7	68.2
0.25	88.0	72.2
0.30	85.0	76.2
0.35	84.1	80.5
0.40	80.7	82.6
0.45	75.1	85.8

**FIGURE 4** The ROC curve for internal validation. The vertical axis represents sensitivity, and the horizontal axis represents the inverse of specificity. The best performance on this data set is located closest to the left upper corner. ROC, receiver operating curve.

specificity of 72.2% on the validation set) was chosen for further use on the three independent test sets. This threshold indicates that if the prediction of the CAdE system, ranging between 0 and 1, is ≥ 0.25 , this image will be classified as neoplastic. A prediction <0.25 will classify the image as non-dysplastic.

Performance on test set 1

The CAdE system correctly classified 42/50 neoplastic images as neoplastic and 99/150 NDBE images as non-dysplastic, corresponding to sensitivity and specificity of 84% and 66%, respectively. The results are summarized in Table 3.

In 41 of 42 (97%) correctly identified neoplastic images, the bounding box of the CAdE system overlapped with the ground truth of experts.

Benchmark performance of general endoscopists

Fifty-two general endoscopists originating from France, the United Kingdom, and the Netherlands completed the web-based module. The median sensitivity for the general endoscopists was 63% (IQR 50%–78%) and median specificity was 87% (IQR 79%–94%; Figure 5). The general endoscopists placed their biopsy mark within the experts' ground truth in 96% (IQR 94%–100%) of the correctly classified neoplastic images. The CAdE system outperformed 88% of the endoscopists in terms of sensitivity. The median absolute difference between the performance of the CAdE system and the general endoscopists was 21%, resulting in a relative increase in neoplasia detection of 33%. The lowest scoring 25% of the endoscopists (median sensitivity of 44% (IQR 44%–48%)) would benefit most from the assistance of the CAdE system with a potential relative increase of 63% in their neoplasia detection (absolute increase 40%).

Performance on test set 2

The CAdE system classified all 50 neoplastic images correctly (sensitivity 100%) and 33/50 NDBE images (specificity 66%; Table 3).

Performance on test set 3

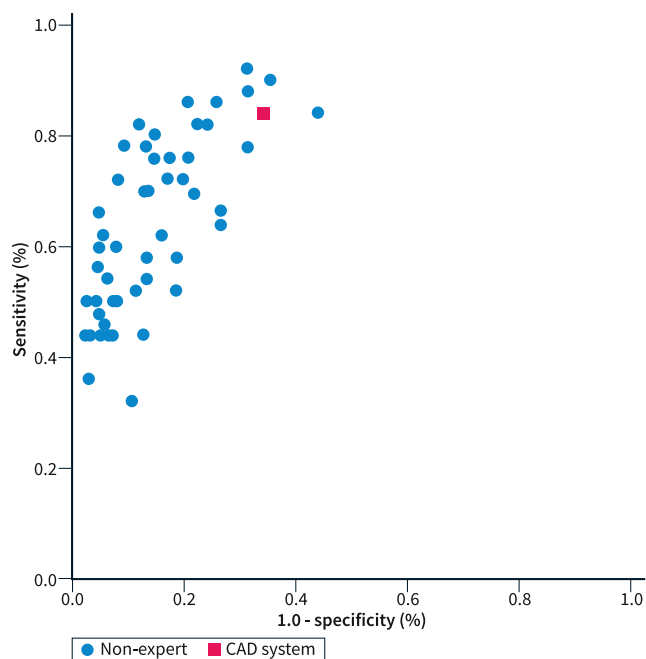
In this prospectively collected test set, the CAdE system correctly classified 44/50 neoplastic images and 96/150 NDBE images (sensitivity 88%, specificity 64%; Table 3).

Processing speed on endoscopic images

The CAdE system classified an endoscopic image in 0.029 seconds corresponding to an analysis speed of 35 frames per second, which is sufficient for real-time application during endoscopic procedures.

TABLE 3 Overview of data sets used for the development of the Computer Aided Detection (CAdE) system.

Data set	Scored by	Classification		Localization	
		Sensitivity	Specificity	Performance	Method
Retrospective test set 1	CAdE system	84%	66%	97%	Bounding box
	General endoscopists	63% (IQR 50%–78%)	87% (IQR 79%–94%)	96% (IQR 94%–100%)	Biopsy mark
Retrospective test set 2	CAdE system	100%	66%	n.a.	n.a.
Prospective test set 3	CAdE system	88%	64%	n.a.	n.a.

**FIGURE 5** Classification performance of the Computer Aided Detection (CAdE) system (red square) and individual general endoscopists (blue dot) on test set 1.

DISCUSSION

We describe the first step in the development of a robust CAdE system for Barrett's neoplasia. In this study, we aimed to describe the infrastructure of our consortium and report the first results based on retrospectively collected endoscopic images. This CAdE system was trained with an unprecedented number of images in the field of BE and outperformed the vast majority of general endoscopists in terms of neoplasia detection.

For our primary analyses, we constructed a test set enriched for subtle neoplastic lesions, representing challenging cases in which endoscopists would benefit most from CAdE assistance. This test set was benchmarked by 52 general endoscopists who detected 63% of neoplastic lesions versus 84% for CAdE (Figure 5). A sensitivity of 63% for the general endoscopists corresponds to a miss-rate of approximately one third of the neoplastic lesions. Approximately 25% of the endoscopists missed over 50% of the neoplastic lesions.

Most endoscopists with a sensitivity comparable to that of the CAdE system (Figure 5) had a significantly lower specificity than those who performed poorly in detecting neoplasia. This inverse relationship between sensitivity and specificity was also reflected in the CAdE performance. The CAdE system misclassified 51/150 NDBE images as neoplastic, which corresponds to a specificity of 66% versus a median specificity of 87% for the general endoscopists.

In the light of the ongoing development of the CAdE system, we wanted to understand the false-positive predictions of the CAdE system. During this evaluation, we noticed that the bounding box in some cases included evident false-positive detections (*i.e.*, the esophageal lumen or the endoscope itself) or subtle mucosal abnormalities, even though the corresponding pathology showed no dysplasia (Figure 6). We reasoned that the obvious false-positive detections would be rejected immediately by the endoscopist. However, the CAdE system should detect all visible abnormalities, regardless of corresponding pathology, to adhere to current guidelines. Therefore, we performed a post-hoc analysis in which 2 expert endoscopists independently assessed the false-positive detections of the system.

For test set 1, 24/51 false-positive detections, the NDBE images indeed contained subtle visible abnormalities as assessed post-hoc by two expert endoscopists (Figure 6a). Based on the more focused endoscopic inspection and clinical information, the endoscopist may then decide to dismiss the detection, interrogate the area in detail, and/or obtain targeted biopsy. The potential negative clinical consequences of false-positive CAdE detection (*i.e.*, obtaining an unnecessary additional targeted biopsy against the background of having to take multiple random biopsies anyway) are clearly of minor importance compared to the potential clinical consequences of false-negative detection, in which a neoplastic lesion may be left undetected and a 3–5 years surveillance interval may follow.

Ten of 51 false-positive predictions were considered clear flaws of the CAdE system, for example, detecting the endoscope, lumen, or bubbles as abnormal (Figure 6b). We anticipate that endoscopists will easily dismiss such positive detections. In only 17/150 predictions, the CAdE system labeled normal mucosa as being neoplastic (Figure 6c). In our opinion, these are “true” false-positive detections, corresponding to an adjusted specificity of 89% (Figure 7). To compare the adjusted specificity fairly with the assessors, the “subtle abnormality” false-positive predictions were disregarded for the assessors and specificity was recalculated. The median adjusted specificity for assessors was 88% (original specificity 87%).

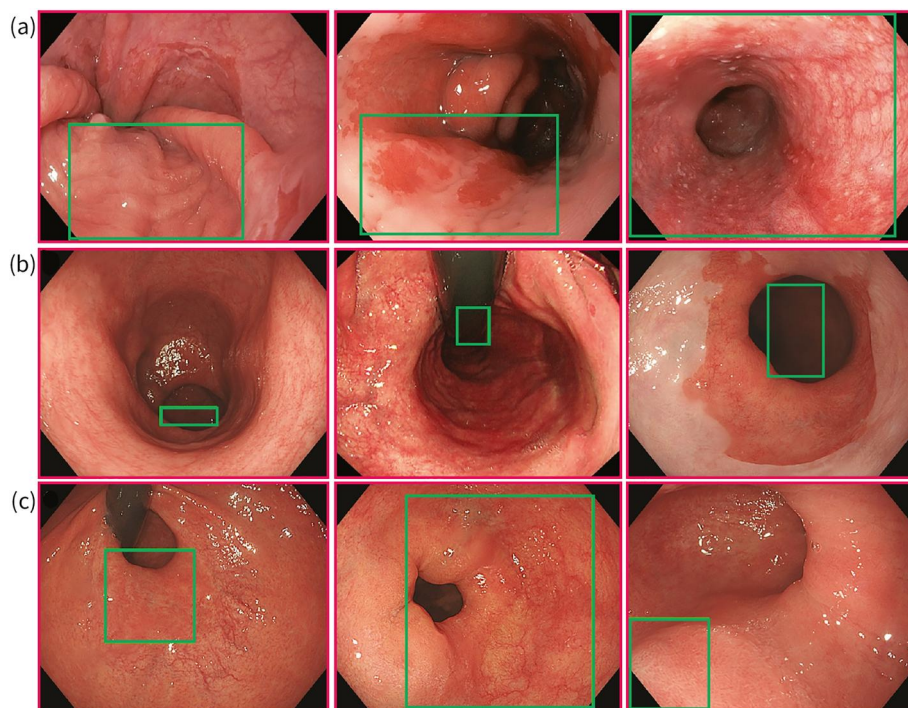


FIGURE 6 Examples of false positive Computer Aided Detection (CAdE) predictions: (a) subtle visible abnormalities; (b) clear flaws of the CAdE system; and (c) “true” false-positive detections.

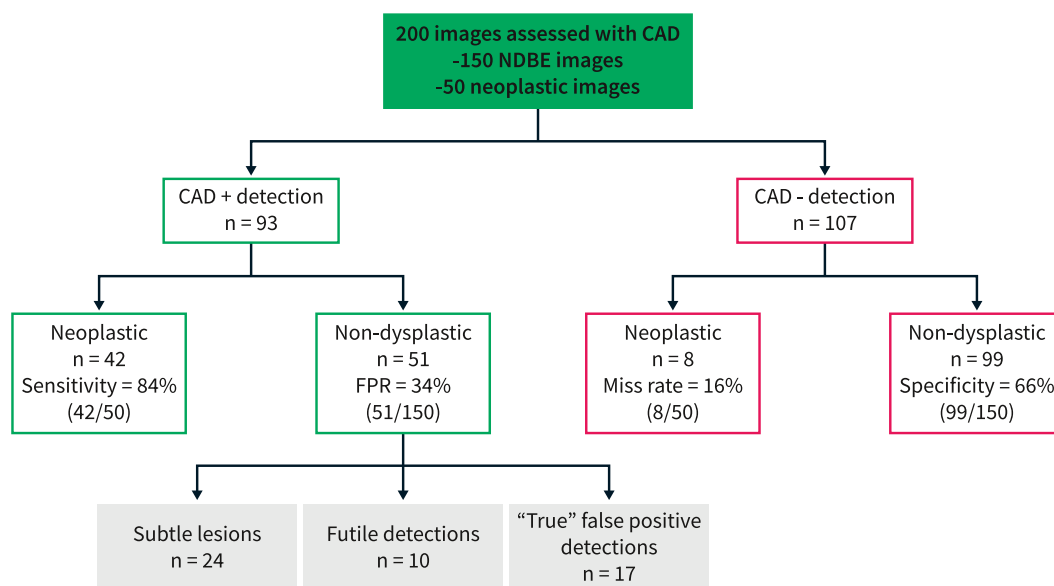


FIGURE 7 Flow-chart of Computer Aided Detection (CAdE) performance on test set 1.

The post hoc analyses of false-positive findings in test sets 2 and 3 showed similar findings with adjusted specificity of 88% and 87%, respectively. We anticipate that a significant proportion of false-positive detections on a flat-type mucosa of NDBE will be easily dismissed after more detailed inspection with optical chromoscopy techniques. We have recently reported on a computer-assisted characterization algorithm using narrow-band imaging (NBI-CADx)¹³ that could be used for such a purpose.

In test set 2, all neoplastic lesions were detected (sensitivity 100%). This test set was created to represent the normal variety of neoplastic lesions encountered in daily practice and was not artificially enriched with subtle neoplasia, explaining the difference in sensitivity in test sets 1 and 2. These results suggest that in daily practice, the CAdE system should be able to detect virtually all early neoplastic lesions encountered during Barrett surveillance endoscopy.

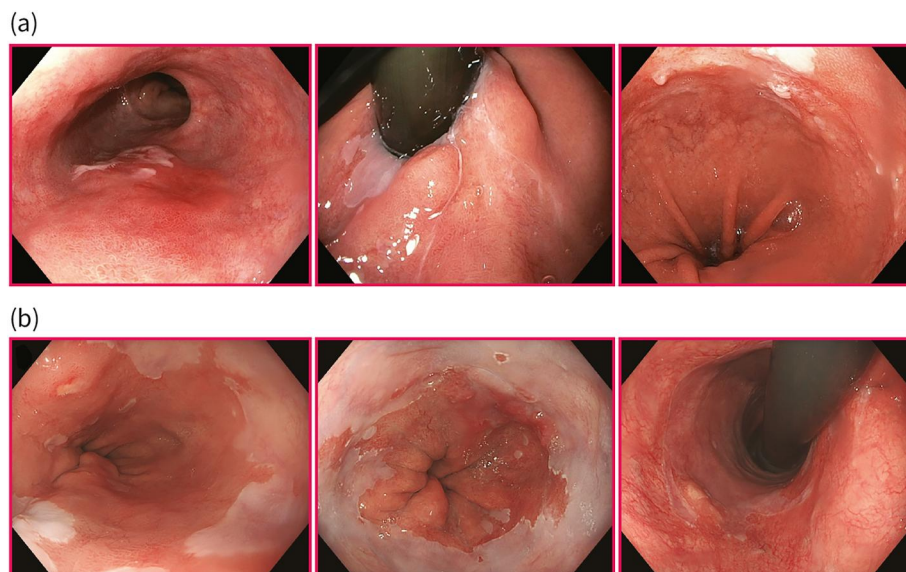


FIGURE 8 Representing the difference in focus for neoplastic lesions: (a) retrospective images focused on the neoplastic lesion and (b) prospectively recorded images in overview without specific focus on the neoplastic lesion.

In test set 3, the CADe system was tested on images obtained using the Olympus X1 processor and the latest generation EZ1500-gastrosopes. The results of this prospective test set are compared to those of test set 1 (sensitivity 88% vs. 84%, specificity 64% vs. 66%). The results of test set 3 suggest that the current CADe system is pretty robust, even without actual training on imagery acquired with this set-up. It is important to point out that in our prospective image acquisition protocol images are obtained in overview *without specific focus on any lesion* (i.e., mimicking the situation where lesions are overlooked). Training and testing CADe algorithms using imagery collected in overview eliminate an important hidden bias that is inherent to the use of retrospectively collected imagery of early neoplasia: in the latter – almost per definition – the available image has been acquired *because the lesion has been detected* by the endoscopist. Therefore, *retrospectively* collected imagery of neoplastic lesions is biased toward easily detectable lesions and shows the neoplastic lesion in a different endoscopic configuration (i.e. relatively more often focused onto the lesion) than during the envisioned endoscopic application of CADe in Barrett surveillance. In the eventual application, the endoscopist provides an overview of the Barrett's segment and CADe might provide clinical benefit for those neoplastic lesions that remain unrecognized by the endoscopist. This is an important source of systemic bias in developing CADe algorithms on retrospective imagery (Figure 8). In our ongoing prospective image acquisition, all imagery is obtained in overview without specific focus on imaging lesions. This provides a more realistic imagery for training a CADe system and ensures that imagery of NDBE and neoplastic cases are recorded under the same circumstances.

The ratio of neoplastic and non-dysplastic images in our test sets does not reflect the real-time prevalence of neoplasia in a general surveillance setting. If we extrapolate our findings to an estimated prevalence of visible neoplastic lesions of 1/200 surveillance

endoscopy, our current algorithm will result in more false-positive detections than true-positive detections (i.e., the positive predictive value would be 1%, vs. a negative predictive value of 100%). As mentioned above, we anticipate that many false-positive detections will be dismissed by the endoscopist and that the remaining detections justify targeted biopsies based on current guideline recommendations.

There are several opportunities to further improve the current CADe performance on still images. First, we will expand the pre-training GastroNet data set tenfold to include 5.000.000 general endoscopic images. Second, we will increase the number of training data by adding more prospectively collected data. Third, we will expand the number of subtle neoplastic images in the training set: since we artificially enriched our test set with subtle neoplastic lesions, our training set may have been relatively depleted for these types of lesions and therefore may have been underpowered for optimal training here. Fourth, we will further curate the training set by excluding NDBE images with subtle visible abnormalities to avoid that the CADe system is trained with images of subtle abnormalities which are labeled as non-neoplastic. These two measures should increase the sensitivity for detection of subtle neoplastic lesions. Fifth, we strive to exploit the ambiguity contained in the ground truth delineations by the expert endoscopists for better training. Conceptually, the overlap area of two experts for the most profound part of the neoplastic lesion has different information than an area which was delineated by only one expert as being subtle in neoplastic appearance (Figure 1).

This study has several unique features. First, due to the low demands on computational resources and generic architecture of the current CADe system, it is suitable for direct clinical implementation in current endoscopy systems. Second, our CADe system was trained with the largest reported number of Barrett's images to date. The total number of images (close to 500.000 endoscopic images and 4.920 Barrett-specific images derived from 1.642 unique patients)

were obtained from 11 participating centers in Europe. The heterogeneity of the data sets used in the development and evaluation of the CAdE system increases the robustness of the results. All neoplastic images were delineated by at least two out of 14 international expert endoscopists to create the optimal ground truth for training and testing. This approach ensures improved heterogeneity and robustness over our previous CAdE system.^{12,14}

This study also has some limitations. First, only retrospective data were used for the development of this CAdE system. However, the comparable performance on test set 3 suggests robustness for prospective data. Second, due to the setup of the current benchmarking study, it is not possible to report the additive value of the CAdE system. This would require a two-phase assessment in which assessors first assess images without CAdE assistance, followed by a second assessment in which the same images are evaluated with CAdE assistance. Third, all images used in this study were recorded by expert endoscopists, resulting in high-quality images with proper mucosal cleaning and a well-expanded esophagus. This might differ from the eventual application in the current surveillance setting. In future studies, we are planning to include more heterogeneous and prospective data to increase the robustness of the CAdE system. Finally, the localization of the CAdE system was considered correct if the bounding box overlapped the experts' ground truth with a minimum of one pixel, which could theoretically increase localization performance. However, when analyzing CAdE predictions, the bounding box virtually always included the major part of the neoplastic lesion, if not the complete lesion.

In future studies, to improve the performance of the current CAdE system, our consortium aims to expand the training set, in particular by the prospectively collected imagery, with a set goal on 15.000 images in total derived from 2.000 patients. Second, we will work on the transition from an image-based CAdE system toward a video-based CAdE system to work toward a real-time application during endoscopic procedures. Third, to evaluate the robustness and generalizability of the CAdE system, we aim to expand the number of independent test sets for images and videos. Furthermore, we will include expert endoscopists performance as a reference performance and we want to test the performance of general endoscopists without and with the assistance of the CAdE system in sequential benchmarking studies to investigate the additive value of the CAdE system when used by general endoscopists. Finally, we aim to incorporate an NBI-CADx system to reduce false-positive CAdE detections.

In conclusion, we report the preliminary results of a robust CAdE system for Barrett's neoplasia with low computational demands allowing real-time applications. The CAdE system detected neoplasia with high accuracy and near-perfect localization. The accuracy for detecting neoplasia of the CAdE system was higher compared to general endoscopists, suggesting improved neoplasia detection by the use of CAdE assistance. Future studies will focus on the acquisition of prospective training data and prospective testing with 'endoscopist plus CAdE' performance evaluations, followed by live testing in the endoscopy suite.

AFFILIATIONS

- ¹Department of Gastroenterology and Hepatology, Amsterdam Gastroenterology, Endocrinology and Metabolism, University of Amsterdam, Amsterdam, the Netherlands,
- ²Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, the Netherlands
- ³Department of Gastroenterology and Hepatology, UMC Utrecht, University of Utrecht, Utrecht, the Netherlands
- ⁴Department of Gastroenterology and Hepatology, Sint Antonius Hospital, Nieuwegein, the Netherlands
- ⁵Department of Gastroenterology and Hepatology, Haga Teaching Hospital, Den Haag, the Netherlands
- ⁶Department of Gastroenterology and Hepatology, University of Groningen, Groningen, the Netherlands
- ⁷Department of Gastroenterology and Hepatology, Isala Hospital Zwolle, Zwolle, the Netherlands
- ⁸Department of Gastroenterology and Hepatology, Flevoziekenhuis Almere, Almere, the Netherlands
- ⁹Department of Gastroenterology and Hepatology, Royal Perth Hospital, Perth, Australia
- ¹⁰Department of Gastroenterology and Hepatology, Hirslanden Klinik, Zurich, Switzerland
- ¹¹Department of Digestive Diseases, Karolinska University Hospital, Stockholm, Sweden
- ¹²Division of Surgery, Department of Clinical Science, Intervention and Technology, CLINTEC, Karolinska Institutet, Stockholm, Sweden
- ¹³Department of Gastroenterology and Hepatology, Cochin Hospital Paris, Paris, France
- ¹⁴Department of Gastroenterology and Hepatology, Nottingham University Hospital, Nottingham, UK
- ¹⁵Department of Gastroenterology and Hepatology, Krankenhaus Barmherzige Brüder Regensburg, Regensburg, Germany
- ¹⁶Department of Gastroenterology and Hepatology, Evangelische Klinik Düsseldorf, Düsseldorf, Germany

ACKNOWLEDGMENT

The authors are grateful to the 52 general endoscopists who participated in this study using benchmarking test set 1. This has provided an important reference for the performance of the CAdE system.

CONFLICT OF INTEREST STATEMENT

This project was financially supported by Olympus Tokyo, Japan.

DATA AVAILABILITY STATEMENT

Research data are not shared.

ORCID

Roos E. Pouw  <https://orcid.org/0000-0002-4707-1186>

Wouter B. Nagengast  <https://orcid.org/0000-0002-6164-1536>

A. Jeroen de Groof  <https://orcid.org/0000-0003-2334-0043>

REFERENCES

1. American Gastroenterological Association, Spechler SJ, Sharma P, Inadomi JM, Shaheen NJ, American gastroenterological association medical position statement on the management of Barrett's

- esophagus. *Gastroenterology*. 2011;140(3). <https://doi.org/10.1053/j.gastro.2011.01.030>
2. Weusten BLAM, Bisschops R, Coron E, Dinis-Ribeiro M, Dumonceau JM, Esteban JM, et al. Endoscopic management of Barrett's esophagus: European Society of Gastrointestinal Endoscopy (ESGE) position statement. *Endoscopy*. 2017;49(2):191–8. <https://doi.org/10.1055/s-0042-122140>
 3. Schölvinc DW, van der Meulen K, Bergman JJGHM, Weusten BLAM. Detection of lesions in dysplastic Barrett's esophagus by community and expert endoscopists. *Endoscopy*. 2017;49(2):113–20. <https://doi.org/10.1055/s-0042-118312>
 4. Bergman JJGHM, de Groof AJ, Pech O, Ragunath K, Armstrong D, Mostafavi N, et al. An interactive web-based educational tool improves detection and delineation of Barrett's esophagus-related neoplasia. *Gastroenterology*. 2019;156(5):1299–308.e3. <https://doi.org/10.1053/j.gastro.2018.12.021>
 5. Byrne MF, Chapados N, Soudan F, Oertel C, Linares Perez M, Kelly R, et al. Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. *Gut*. 2019;68(1):94–100. <https://doi.org/10.1136/gutjnl-2017-314547>
 6. Ebigbo A, Mendel R, Probst A, Manzeneder J, Souza Jr LA, Papa JP, et al. Computer-aided diagnosis using deep learning in the evaluation of early oesophageal adenocarcinoma. *Gut*. 2019;68(7):1143–5. <https://doi.org/10.1136/gutjnl-2018-317573>
 7. Hashimoto R, Requa J, Dao T, Ninh A, Tran E, Mai D, et al. Artificial intelligence using convolutional neural networks for real-time detection of early esophageal neoplasia in Barrett's esophagus (with video). *Gastrointest Endosc*. 2020;91(6):1264–71.e1. <https://doi.org/10.1016/j.gie.2019.12.049>
 8. Ghatwary N, Zolgharni M, Ye X. Early esophageal adenocarcinoma detection using deep learning methods. *Int J Comput Assist Radiol Surg*. 2019;14(4):611–21. <https://doi.org/10.1007/s11548-019-01914-4>
 9. van der Sommen F, de Groof J, Struyvenberg M, van der Putten J, Boers T, Fockens K, et al. Machine learning in GI endoscopy: practical guidance in how to interpret a novel field. *Gut*. 2020;69(11):2035–45. <https://doi.org/10.1136/gutjnl-2019-320466>
 10. Milletari F, Navab N, Ahmadi SA, V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: *Proceedings - 2016 4th international conference on 3D vision, 3DV 2016*; 2016. <https://doi.org/10.1109/3DV.2016.79>
 11. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: *Institute of electrical and electronics engineers (IEEE)*; 2010. p. 248–55. <https://doi.org/10.1109/cvpr.2009.5206848>
 12. de Groof AJ, Struyvenberg MR, van der Putten J, van der Sommen F, Fockens KN, Curvers WL, et al. Deep-learning system detects neoplasia in patients with Barrett's esophagus with higher accuracy than endoscopists in a multistep training and validation study with benchmarking. *Gastroenterology*. 2020;158(4):915–29.e4. <https://doi.org/10.1053/j.gastro.2019.11.030>
 13. van der Putten J, de Groof J, van der Sommen F, Struyvenberg M, Zinger S, Curvers W, et al. Pseudo-labeled bootstrapping and multi-stage transfer learning for the classification and localization of dysplasia in Barrett's esophagus. In: *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*. LNCS; 2019: 11861. https://doi.org/10.1007/978-3-030-32692-0_20
 14. de Groof AJ, Struyvenberg MR, Fockens KN, van der Putten J, van der Sommen F, Boers TG, et al. Deep learning algorithm detection of Barrett's neoplasia with high accuracy during live endoscopic procedures: a pilot study (with video). *Gastrointest Endosc*. 2020;91(6):1242–50. <https://doi.org/10.1016/j.gie.2019.12.048>
 15. de Groof AJ, Fockens KN, Struyvenberg MR, Pouw RE, Weusten BL, Schoon EJ, et al. Blue-light imaging and linked-color imaging improve visualization of Barrett's neoplasia by nonexpert endoscopists. *Gastrointest Endosc*. 2020;91(5):1050–7. <https://doi.org/10.1016/j.gie.2019.12.037>
 16. Fockens K, de Groof J, van der Putten J, Khurelbaatar T, Fukuda H, Takezawa T, et al. Linked color imaging improves identification of early gastric cancer lesions by expert and non-expert endoscopists. *Surg Endosc*. 2022;36(11):8316–8325 Published online. <https://doi.org/10.1007/s00464-022-09280-0>
 17. Liu R. Higher accuracy on vision models with EfficientNet-Lite. <https://blog.tensorflow.org/2020/03/higher-accuracy-on-vision-models-with-efficientnet-lite.html>
 18. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: inverted residuals and linear bottlenecks. In: *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*; 2018. <https://doi.org/10.1109/CVPR.2018.00474>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Fockens KN, Jukema JB, Boers T, Jong MR, van der Putten JA, Pouw RE, et al. Towards a robust and compact deep learning system for primary detection of early Barrett's neoplasia: Initial image-based results of training on a multi-center retrospectively collected data set. *United European Gastroenterol J*. 2023;11(4):324–36. <https://doi.org/10.1002/ueg2.12363>