Computer-aided diagnosis improves characterization of Barrett's neoplasia by general endoscopists (with video)

Jelmer B. Jukema, MD,[1] Carolus H. J. Kusters, MS,[2] Martijn R. Jong, MD,[1] Kiki N. Fockens, MD, PhD,[1] Tim Boers, MS,[2] Joost A. van der Putten, PhD,[2] Roos E. Pouw, MD, PhD,[1] Lucas C. Duits, MD, PhD,[1] Bas L. A. M. Weusten, MD, PhD,[3,4] Lorenza Alvarez Herrero, MD, PhD,[4] Martin H. M. G. Houben, MD, PhD,[5] Wouter B. Nagengast, MD, PhD,[6] Jessie Westerhof, MD, PhD,[6] Alaa Alkhalaf, MD, PhD,[7] Rosalie Mallant-Hent, MD, PhD,[8] Pieter Scholten, MD,[9] Krish Ragunath, MD, PhD,[10] Stefan Seewald, MD, PhD,[11] Peter Elbe, MD,[12,13] Francisco Baldaque Silva, MD, PhD,[12,14] Maximilien Barret, MD, PhD,[15] Jacobo Ortiz Fernández-Sordo, MD,[16] Guiomar Moral Villarejo, MD,[16] Oliver Pech, MD, PhD,[17] Torsten Beyna, MD, PhD,[18] Nahid S. M. Montazeri, PhD,[19] Fons van der Sommen, PhD,[2] Peter H. de With, PhD,[2] A. Jeroen de Groof, MD, PhD,[1] Jacques J. Bergman, MD, PhD,[1] on behalf of the BONS-AI consortium*

Amsterdam, Eindhoven, Utrecht, Den Haag, Groningen, Zwolle, Almere, the Netherlands; Perth, Australia; Zurich, Switzerland; Stockholm, Sweden; Matosinhos, Portugal; Paris, France; Nottingham, United Kingdom; Regensburg, Düsseldorf, Germany

Background and Aims: Characterization of visible abnormalities in patients with Barrett's esophagus (BE) can be challenging, especially for inexperienced endoscopists. This results in suboptimal diagnostic accuracy and poor interobserver agreement. Computer-aided diagnosis (CADx) systems may assist endoscopists. We aimed to develop, validate, and benchmark a CADx system for BE neoplasia.

Methods: The CADx system received pretraining with ImageNet and then consecutive domain-specific pretraining with GastroNet, which includes 5 million endoscopic images. It was subsequently trained and internally validated using 1758 narrow-band imaging (NBI) images of early BE neoplasia (352 patients) and 1838 NBI images of nondysplastic BE (173 patients) from 8 international centers. CADx was tested prospectively on corresponding image and video test sets with 30 cases (20 patients) of BE neoplasia and 60 cases (31 patients) of nondysplastic BE. The test set was benchmarked by 44 general endoscopists in 2 phases (phase 1, no CADx assistance; phase 2, with CADx assistance). Ten international BE experts provided additional benchmark performance.

Results: Stand-alone sensitivity and specificity of the CADx system were 100% and 98% for images and 93% and 96% for videos, respectively. CADx outperformed general endoscopists without CADx assistance in terms of sensitivity ($P = .04$). Sensitivity and specificity of general endoscopists increased from 84% to 96% and 90% to 98% with CAD assistance ($P < .001$). CADx assistance increased endoscopists' confidence in characterization ($P < .001$). CADx performance was similar to that of the BE experts.

Conclusions: CADx assistance significantly increased characterization performance of BE neoplasia by general endoscopists to the level of expert endoscopists. The use of this CADx system may thereby improve daily Barrett surveillance. (Gastrointest Endosc 2024;100:616-25.)

(footnotes appear on last page of article)

Patients with Barrett's esophagus (BE) undergo regular endoscopic surveillance to enable early detection of dysplasia or esophageal adenocarcinoma. Because progression to neoplasia is rare (<1% per patient-year[1]) and early BE neoplasia often displays only subtle endoscopic changes, most endoscopists are not familiar with the endoscopic appearance of early BE neoplasia.[2,3] Timely detection of neoplasia is, however, crucial to enable endoscopic therapy, which is associated with excellent outcomes.[4]

Endoscopic detection of BE neoplasia typically involves a 2-step process: primary detection using high-definition white-light endoscopy (WLE) in overview, followed by targeted characterization of abnormalities using optical chromoendoscopy techniques such as narrow-band imaging

(NBI). Just as endoscopists may struggle with the primary detection of early BE neoplasia, their diagnostic accuracy and interobserver agreement for subsequent NBI-based characterization are also suboptimal.[5-8] Suboptimal accuracy in NBI characterization may lead to falsely disregarding primary detections, thereby missing already detected lesions. Computer-aided diagnosis (CADx) systems have been shown to provide valuable assistance in polyp characterization[9,10] and may also be able to assist endoscopists in the characterization of Barrett's neoplasia. Moreover, a CADx system may dismiss any false-positive detections by a primary computer-aided detection (CADe) system using WLE.

Recently, the BONS-AI (Barrett's Oesophagus Imaging for Artificial Intelligence) consortium has developed an effective CADe system that has been shown to improve the primary detection of early BE neoplasia by general endoscopists.[11,12] The same consortium has now also set out to engineer an NBI-based CADx system. The aim of the current study was 3-fold: (1) to develop and validate an NBI-based CADx system for the characterization of BE neoplasia, with hardware and software specifications that allow for direct integration into current endoscopy systems; (2) to compare the performance of the NBI-based CADx system with that of BE experts and general, nonexpert endoscopists; and (3) to evaluate if the performance of general endoscopists improves when they are provided with CADx assistance.

METHODS

BONS-AI consortium

This study was conducted by the BONS-AI consortium. The consortium is led by the Department of Gastroenterology and Hepatology of the Amsterdam UMC and the Department of Electrical Engineering of Eindhoven University of Technology. The Medical Research Involving Human Subjects Act did not apply to this study; official approval was therefore waived by the medical ethics review committee of all participating centers. The study was registered at the Dutch Trial Register under the number NL8411.

Patient and public involvement

No patients were involved in determination of study design, outcome measures, or plans for dissemination of study results.

Development of the CADx system

The envisioned NBI-based CADx system is developed to work in conjunction with an WLE-based CADe system. Upon initial detection by the WLE-based CADe system, the endoscopist inspects the suspicious area in more detail using NBI. Subsequently, the CADx system can either confirm the primary CADe detection, and advise targeted biopsy acquisition, or assess the area as nondysplastic BE (NDBE), which may lead the endoscopist to disregard the primary CADe detection. The CADx system was designed to classify endoscopic images and videos as either "neoplastic" or "NDBE" using output probability scores. We also included the outcome "failure to classify" for those cases in which the system cannot reliably classify neoplasia or NDBE (Fig. 1).

The development process of the CADx system can be divided into 4 stages: (1) pretraining using both general and domain-specific data; (2) training with Barrett-specific data; (3) internal validation; and (4) external testing, including benchmarking performance.

CADx pretraining. The CADx system was initially pretrained by using ImageNet, a data set of 1.2 million images labeled into 1000 categories such as animals, houses, and cars. This process, known as generic pretraining, allows the deep learning system to learn basic features of images such as shapes and colors. The weights of the neural network are then transferred to the next phase of training. This generic pretraining on ImageNet prevents the system having to learn basic image features during the final application phase, where data are more scarce.

Pretraining, however, is likely to be more effective if the images for pretraining resemble the envisioned application (ie, domain-specific pretraining). Therefore, after generic pretraining on ImageNet, the system underwent domain-specific pretraining using GastroNet, a data set of 5,084,494 endoscopic images compiled by our consortium from the endoscopic databases of 7 Dutch hospitals. These endoscopic images were gathered from the period between 2012 and 2021 and were recorded by using endoscopy equipment of Olympus (Hamburg, Germany), Fujifilm (Tilburg, the Netherlands), and Pentax (Hamburg, Germany). In previous work, we have shown that domain-specific pretraining improves the performance of deep learning systems in GI endoscopy.[11-14]

Acquisition of Barrett-specific training data. The CADx system was subsequently trained by using NBI imagery of neoplastic and NDBE, all containing histopathologic ground truth labels. A combination of retrospective and prospective endoscopic imagery of patients with BE, collected from 8 participating centers in our consortium, was used.

Retrospectively collected images were recorded between 2012 and 2021 by using Olympus Q160Z and HQ190 endoscopes (Olympus Europa SE & Co KG, Hamburg, Germany). Endoscopic imagery was extracted from the endoscopy reporting system, after which histopathologic results were confirmed. After this, all imagery was de-identified directly at the participating center using previously developed anonymization software.[11]

Furthermore, between 2020 and 2022, endoscopic images and videos were collected prospectively following a standardized acquisition protocol. Prospective endoscopic imagery was obtained by using Olympus HQ190 and EZ1500 endoscopes and CV190 and X1 processors using

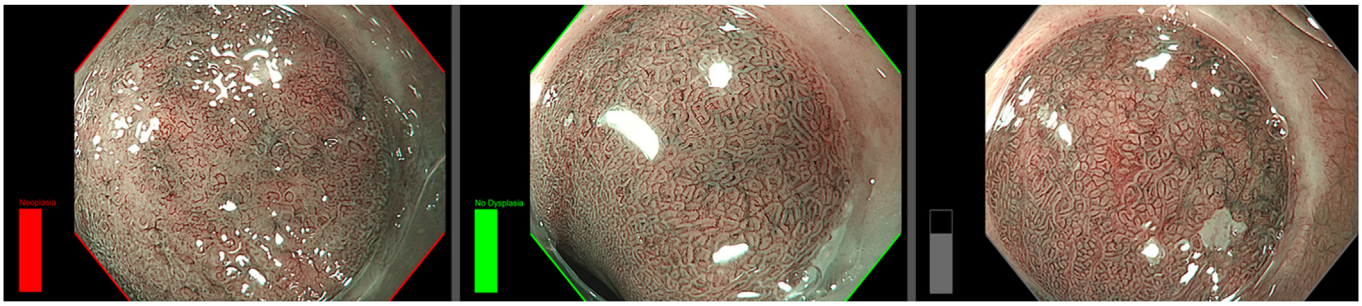


Figure 1. Output of the computer-aided diagnosis (CADx) system. Left: a neoplastic prediction by the CADx system. Middle: nondysplastic prediction by the CADx system. Right: failure to classify by the CADx system. The color bars indicate *red* for neoplastic, *green* for nondysplastic, and *gray* for failure to classify.

the NBI near focus function. For cases performed with EZ1500 endoscopes, the area was also recorded without the NBI near focus function (ie, standard focus) as this endoscope has an adjusted standard focal distance range that allows for detailed inspection without the use of the near focus mode.

Imagery was recorded by using 2 recording devices: USB300 (MediCapture, Plymouth Meeting, Penn, USA) for CV190 processors and HVO-4000ST (Sony, New York, NY, USA) for X1 processors.

All imagery was selected based on image quality, both content independent and content dependent. Content-independent quality parameters were sufficient contrast, sharpness, and absence of blur. Content-dependent quality parameters were mucosal cleanliness (in terms of absence of bubbles and blood) and illumination. Neoplastic imagery required the following: (1) endoscopically visible abnormalities; and (2) confirmation of high-grade dysplasia or esophageal adenocarcinoma in corresponding biopsy or endoscopic resection specimens. NDBE imagery required the following: (1) absence of visible lesions; and (2) no dysplasia in biopsy specimens.

Examples of retrospectively and prospectively collected data are provided in Figure 2.

Ground truth development of prospective imagery. Expert endoscopists recorded targeted areas with and without neoplasia. All areas were sampled for histopathologic examination for ground truth confirmation and were only included when pathology results showed NDBE or high-grade dysplasia/esophageal adenocarcinoma.

Training and validation of the CADx system with endoscopic images and videos. The Barrett-specific training set consisted of 1713 neoplastic images derived from 333 patients and 1755 NDBE images derived from 166 patients. To optimize the performance of the CADx system, the system was internally validated on 128 prospectively collected images. Based on this data set, the (hyper) parameters of the CADx system were optimized, and the image-based thresholds were selected. This internal validation data set consisted of 45 neoplastic images (derived from 22 patients) and 83 NDBE images (derived from 31 patients). An additional 40 neoplastic videos and 82 NDBE videos, originating from the same patients as the image validation set, were used for video-specific optimization (Supplementary Table 1, available online at www.giejournal.org).

Testing the CAD system

The image and video test sets were collected prospectively and displayed identical areas of interest from the same patients; that is, every image in the image test set had a corresponding video (with an approximate length of 5-10 seconds) in the video test set. The test sets consisted of 161 areas derived from 51 patients (Supplementary Table 1).

The test sets were isolated for a single performance test. There was no overlap in patients between the test set and the training or validation sets.

Benchmark performance of video test set

The CADx system was designed to aid general endoscopists without specific experience in BE. To provide benchmark performance and to evaluate the additive value of the CADx system, 44 endoscopists from 6 different countries reviewed the near focus video test set. A previously designed web-based module (Meducati AB, Göteborg, Sweden)[11,15,16] was adjusted for this specific study.

The benchmarking process comprised 2 phases. In the first phase, endoscopists evaluated the test set without CADx assistance. In the second phase, the same videos were assessed in a different, randomized order, and endoscopists were provided with CADx assistance. In both phases, endoscopists evaluated videos for the presence/absence of neoplasia by assigning them to 1 of 3 categories: (1) yes, likely neoplastic, indication for targeted biopsy; (2) uncertain, but I would like to take a targeted biopsy; or (3) not likely neoplastic, no indication for targeted biopsy. To ensure unbiased assessment, no feedback was given to the endoscopist in between assessment phases, and all videos were evaluated without further information on patient status. There was no washout period in between assessment phases.

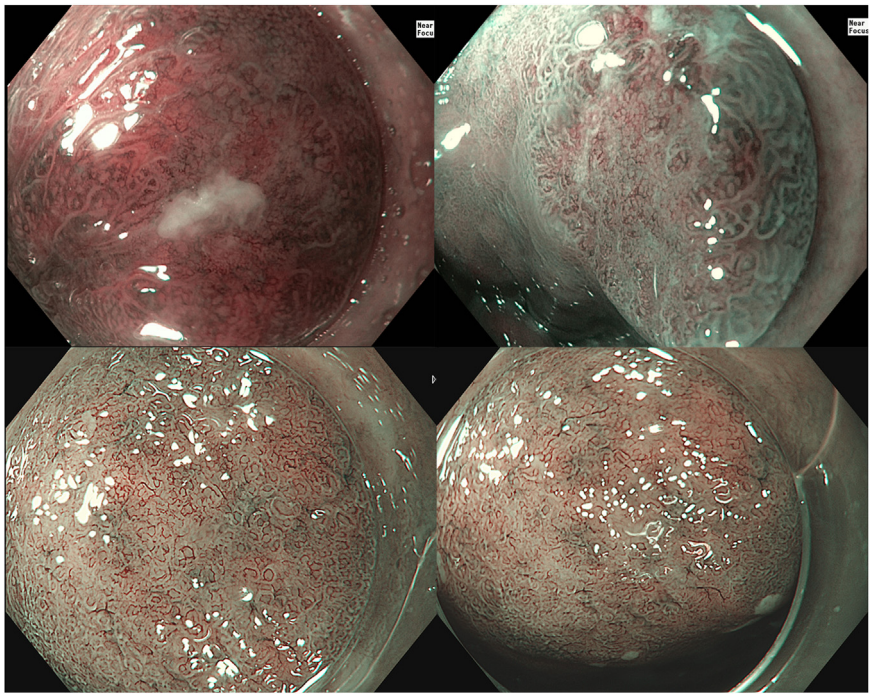


Figure 2. Examples of neoplastic Barrett images used in this study. Top row: retrospectively collected images. Bottom row: prospectively collected images, collected in accordance with acquisition protocol. The left image is in near focus, and the right image is in standard focus. A pair of corresponding videos was also recorded from every prospectively recorded area.

We also asked 10 international experts from 6 different countries with expertise in the use of optical chromoscopy techniques and treatment of early Barrett neoplasia to evaluate the video test set. The goal was to provide an expert reference for the performance of the CADx system and serve as a reference for the CADx-assisted performance of the general endoscopists.

Architecture of the CADx system

To extract relevant image information, an EfficientNet-Lite1 feature encoder[17] was used (Supplementary Fig. 1, available online at www.giejournal.org). This architecture is optimized for fast and efficient processing of real-time video imagery and can be directly implemented into current endoscopy systems. Additional technical details are described in the Appendix 1 section (available online at www.giejournal.org).

Outcomes and performance metrics of CADx and endoscopists

Primary outcomes. The primary study outcomes were (1) stand-alone classification by the CADx system, the general endoscopists, and the BE experts on images and/or videos in terms of sensitivity and specificity; and (2) the change in classification performance by general endoscopists in the video test when assisted by the CADx system.

Secondary outcome. The secondary outcome was percentage in which the CADx system failed to classify and the percentage of uncertain predictions by the general endoscopists with and without CADx assistance and the expert endoscopists. Uncertain predictions (for the endoscopists) and "failure to classify" (for the CADx system) were excluded from the classification performance.

Clinically relevant comparisons of the outcomes measurements

For the test set, we aimed for 4 clinically relevant comparisons: (1) stand-alone performance of CADx versus the performance of general endoscopists without CADx assistance; (2) performance of general endoscopists without CADx assistance versus general endoscopists with CADx assistance; (3) stand-alone performance of CADx versus the performance of the BE experts; and (4) performance of general endoscopists with CADx assistance versus the performance of BE experts.

Statistical analysis

For descriptive statistics, the median and interquartile range are used for continuous variables. Categorical variables are represented by frequencies with percentages. Performance metrics were calculated by using Python 3.8.10 (Python Software Foundation, Wilmington, Del, USA). Statistical analyses were conducted in R Studio Version 4.2.1 (RStudio, PBC, Boston, Mass, USA).

To compare sensitivity scores, a mixed-effects logistic regression model (lme4 package, R, Version 1.1-30, https://CRAN.R-project.org/package=lme4) was used on a subset

of neoplastic images/videos exclusively. Sensitivity is defined as the ratio of true positives (neoplastic images/videos detected by the specific group of interest) to all positives (ie, all neoplastic images/videos in the data set). This resulted in a conditional odds ratio for sensitivity, along with likelihood profile 95% confidence intervals. Specificity was evaluated by using a similar model but on the subset of nondysplastic images/videos. Specificity measures the ratio of true negatives (nondysplastic images/videos defined as flat BE without neoplastic lesion by the specific group of interest) to all negatives (ie, all nondysplastic images/videos in the data set).

More information on the mixed-effect logistic regressions models and noninferiority analysis is provided in the Appendix 1 (available online at www.giejournal.org).

RESULTS

Threshold selection based on validation data

An upper threshold and a lower threshold were chosen based on performance testing in the internal validation data set. For images, outputs of the CADx system <.1 were regarded as NDBE and outputs >.7 were regarded as neoplastic predictions. If the CADx system produced an output ranging from .1 to .7, the image was registered as "failure to classify." For video frames, the thresholds were set at <.3 and >.8. These thresholds were chosen in such a way that they optimized sensitivity and specificity in the internal validation set, while keeping the amount of "failures to classify" <20%.

Stand-alone CADx performance

In the image test set, CADx had a sensitivity, specificity, and failure to classify of 98%, 98%, and 22%, respectively. The distribution of the individual image classification scores is displayed in Figure 3. In the video test set, CADx had a sensitivity, specificity, and failure to classify of 96%, 94%, and 14%. In none of the cases did the CADx system provide a neoplastic and a nondysplastic prediction within the same video. Stratification of the performance results between imagery in standard focus and in near focus is presented in Table 1.

Endoscopist benchmark versus stand-alone CADx performance on video test set

In the benchmarked video test set, general endoscopists had a median sensitivity, specificity, and uncertainty of 84%, 90%, and 22%, respectively. Sensitivity of CADx was significantly higher than that of general endoscopists ($P = .04$), whereas CADx was not significantly better than general endoscopists in terms of specificity ($P = .07$) or certainty ($P = .06$). There was no statistical difference between the performance of CADx and the BE experts, who had a median sensitivity, specificity, and uncertainty of 95% ($P = .26$), 96% ($P = .38$), and 13% ($P = .71$). Individual results of general endoscopists and BE experts are shown in Figure 4, and results are summarized in Tables 2 and 3.

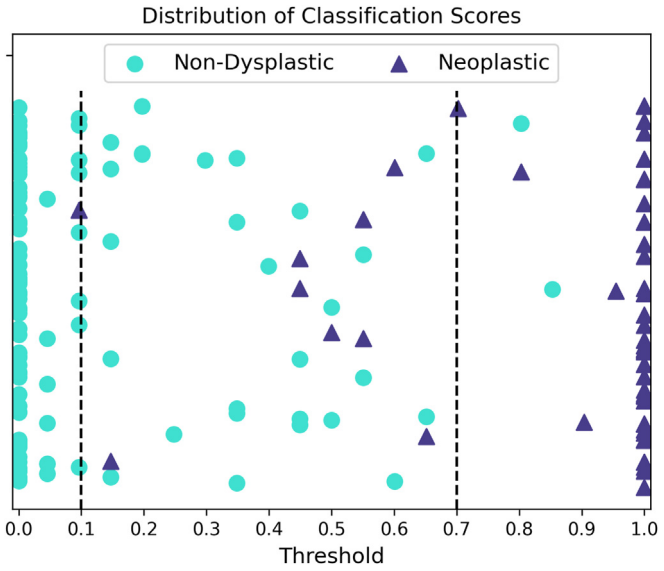


Figure 3. Distribution of classification scores for images in near focus and standard focus. The x-axis represents the classification output score of the computer-aided diagnosis (CADx) system. The y-axis does not include information and is only used to spread the individual predictions. Green dots have a nondysplastic Barrett's esophagus (NDBE) ground truth, and purple triangles have a neoplastic ground truth. Left to the .1 threshold are NDBE predictions, and right to the .7 threshold are the neoplastic predictions. In between the .1 and .7 threshold are the images that fail to classify. The most left neoplastic triangle is the lesion that was falsely predicted by the CADx system to be NDBE. This standard focus image had a probability score of .096.

Additional value of CADx assistance for general endoscopists

Sensitivity, specificity, and uncertainty all improved significantly with CADx assistance. Sensitivity increased from 84% to 96% ($P < .001$), specificity from 90% to 98% ($P < .001$), and the percentage of uncertain predictions decreased from 22% to 13% ($P < .001$) (Fig. 5). When aided by the CADx system, the performance of the general endoscopists was not statistically different from that of BE experts. Sensitivity was 96% for general endoscopists and 95% for BE experts ($P = .42$). Specificity was 98% for general endoscopists and 96% for BE experts ($P = .053$), and the percentage of uncertain predictions was 13% for both ($P = .94$). Results are summarized in Tables 2 and 3. Results stratified according to endoscopists' experience are presented in the Appendix 1.

DISCUSSION

The current study describes the development and evaluation of a CADx system for the detailed characterization of Barrett's neoplasia using NBI. The CADx system outperformed general endoscopists, and CADx assistance

TABLE 1. Stand-alone results of CADx system

Test set	Focus	Sensitivity	Specificity	Failure to classify
Images	Near focus and standard focus	98%	98%	22%
	Near focus	100%	98%	26%
	Standard focus	94%	98%	17%
Videos	Near focus and standard focus	96%	94%	14%
	Near focus	93%	96%	14%
	Standard focus	100%	91%	13%

CADx, Computer-aided diagnosis.

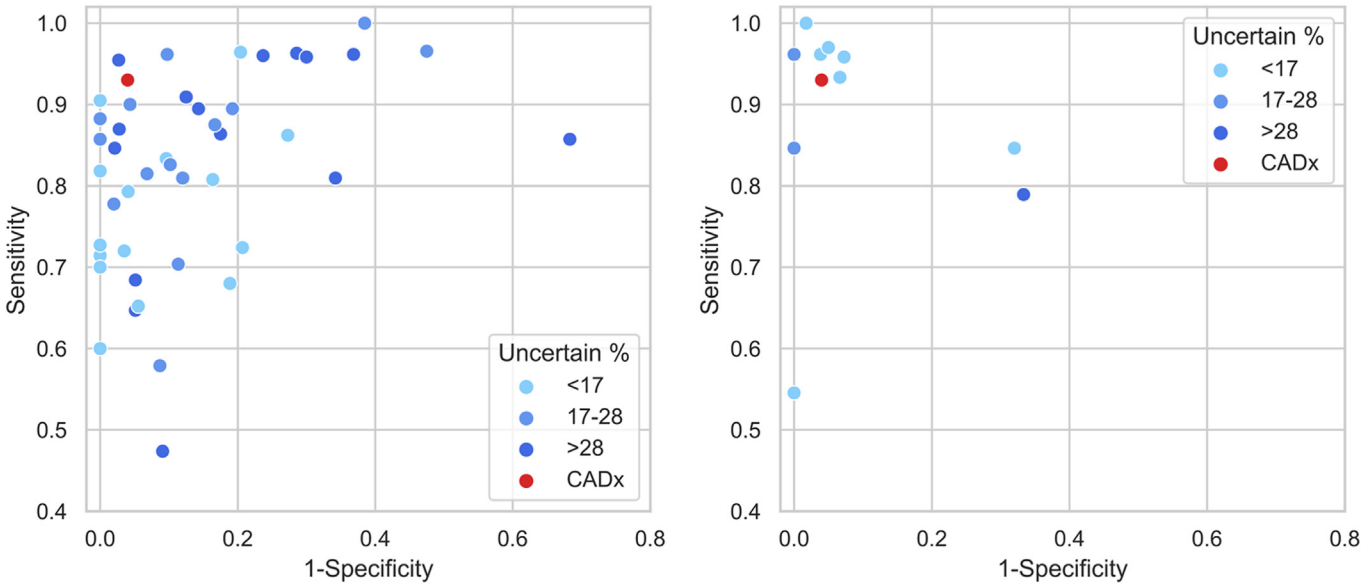


Figure 4. Individual results in terms of sensitivity, specificity, and uncertainty of general endoscopists (left) and Barrett's esophagus experts (right) on video test set. The different colors depict different amounts of uncertain predictions. It is important to consider the amount of uncertain predictions when assessing endoscopist performance as uncertain predictions affect performance calculation by being excluded from calculation. Assessing more cases as uncertain makes it easier to achieve a high performance score. *CADx*, Computer-aided diagnosis.

TABLE 2. Summary of performance on videos in near focus of the CADx system, general endoscopists without/with CADx assistance, and BE experts

Scored by	Sensitivity	Specificity	FTC/uncertain
CADx	93%	96%	14%
General endoscopists	84% (IQR, 72%-90%)	90% (IQR, 80%-97%)	22% (IQR, 14%-31%)
General endoscopists + CADx	96% (IQR, 91%-97%)	98% (IQR, 95%-100%)	13% (IQR, 9%-23%)
BE experts	95% (IQR, 93%-96%)	96% (IQR, 87%-100%)	13% (IQR, 9%-20%)

CADx, Computer-aided diagnosis; *BE*, Barrett's esophagus; *FTC*, failure to classify; *IQR*, interquartile range.

improved endoscopists' performance significantly to the level of BE experts.

The CADx system had a stand-alone sensitivity, specificity, and uncertainty of 98%, 98%, and 22% for images and 96%, 94%, and 14% for videos, respectively. When tested on a video test set, the CADx system significantly outperformed 44 general endoscopists in terms of sensitivity. In addition, the use of the CADx system significantly improved the sensitivity and specificity of general endoscopists with 12% and 8%. In addition, CADx assistance significantly reduced the number of cases in which endoscopists were uncertain about the diagnosis, with a relative reduction of 41%.

The envisioned use of the CADx system during Barrett's surveillance is to supplement a WLE-based CADe system. Upon primary detection by this system, the endoscopist

TABLE 3. Results of mixed-effect model analysis of 4 four aimed comparisons for the video test set

Comparison	Metric	Mixed-effect model	
		OR (95% CI)	*P* value
1. CADx vs general endoscopists	Sensitivity	1.75 (1.03-2.97)	.04
	Specificity	1.55 (.96-2.49)	.07
	FTC/uncertain	0.82 (.67-1.01)	.06
2. General endoscopists + CADx vs general endoscopists	Sensitivity	4.39 (3.09-6.24)	<.01
	Specificity	7.70 (5.61-10.57)	<.01
	Uncertain	.60 (.53-0.67)	<.01
3. General endoscopists + CADx vs BE experts	Sensitivity	1.57 (.53-4.72)	.42
	Specificity	3.19 (.99-10.29)	.05
	Uncertain	1.02 (.91-1.16)	.94
4. CADx vs BE experts	Sensitivity	1.66 (.69-3.99)	.26
	Specificity	1.39 (.67-2.88)	.38
	FTC/uncertain	.94 (.69-1.29)	.71

OR, Odds ratio; *CI*, confidence interval; *CADx*, computer-aided diagnosis; *FTC*, failure to classify; *BE*, Barrett's esophagus.

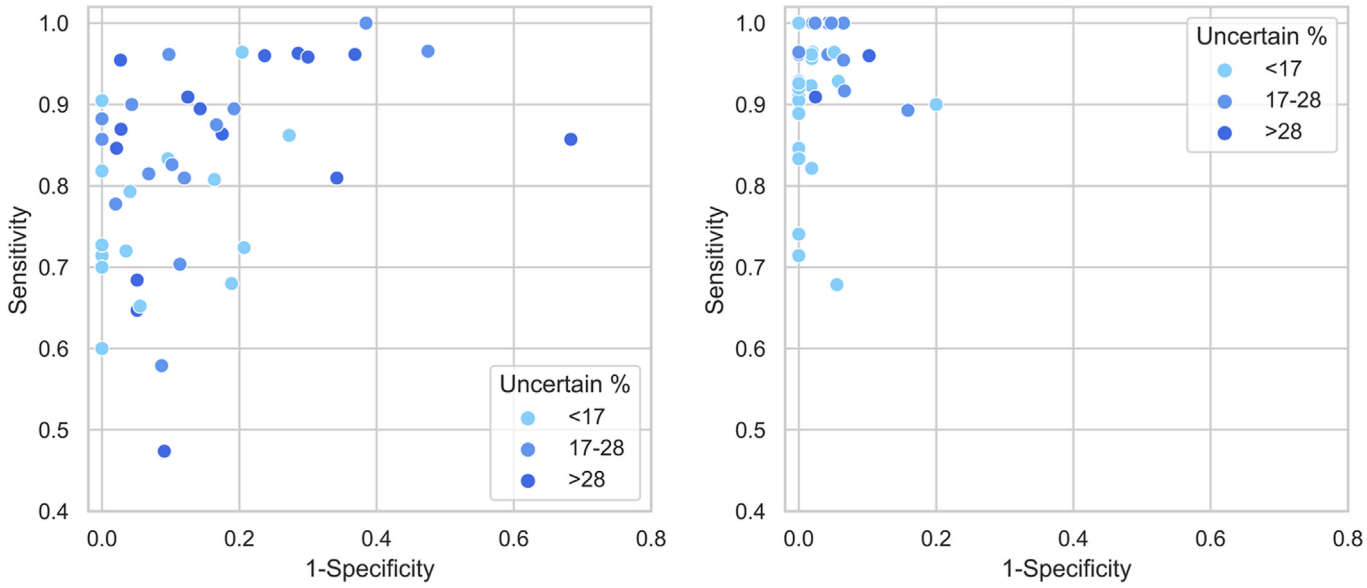


Figure 5. Change in performance of individual general endoscopists when computer-aided diagnosis (CADx) assistance was available. The left graph is without CADx assistance; the right graph is with CADx assistance.

inspects the area of interest with NBI. Subsequently, the CADx system either confirms the detection and recommends targeted biopsy or assesses the area as NDBE and advises to disregard the primary CADe detection (Supplementary Video 1). In addition, the CADx system can be used independently by endoscopists as a stand-alone tool for characterization of areas of interest detected otherwise. Besides aiding endoscopists in lesion characterization, CADx could also serve as a training tool, helping endoscopists better identify Barrett's neoplasia through CADx feedback.[18]

Although CADx systems for polyp characterization are already commercially available, only a few and relatively small-sized studies have reported on CADx systems for Barrett's neoplasia.[19-21]

In contrast to other studies, we did not force a binary outcome on the CADx prediction. If the CADx system is "not certain enough" to classify the image as either "neoplastic" or "nonneoplastic" it will output a score "failure to classify." This is a deliberate design to mitigate the potential negative consequences of incorrect predictions. In our opinion, it is preferable to refrain from predictions

rather than to provide potentially false-positive finding or, even worse, false-negative predictions. We hypothesized that a maximum failure to classify rate of 20% would be acceptable for a user-friendly CADx system. When during real-time inspection the CADx system fails to classify, it is likely that the endoscopist will alter the visualization of the region of interest (which may provide the CADx system with sufficient information to make a prediction) or that the endoscopist simply elects to perform a biopsy of the area of interest. In the current study, the use of CADx by general endoscopists resulted in 13% of areas of interest having a failure to classify score, which was found to be identical to the performance by BE experts participating in this study.

It is important to note that there was only minimal difference in performance between images and videos. This can be attributed to the fact that we included magnified imagery with relatively fixed areas of interest, as there was generally little to no significant movement of the endoscope during the video recordings. Consequently, the potential advantages of video, such as the provision of additional temporal and segmental information, may be of less relevance under these specific circumstances.

Moreover, there was no significant difference in the CADx performance between imagery captured in near focus and standard focus. This can be attributed to the adjusted focal distance of the latest version of the EZ1500 endoscope, in conjunction with the Olympus X1-processor, which enables detailed inspection of the mucosa when the endoscope tip is brought into close proximity to the area of interest without losing focus, similar to the near focus mode. However, it is worth noting that the CADx system did miss a neoplastic image in standard focus that was relatively zoomed out, thus containing a substantial amount of nondysplastic mucosa. This highlights a potential limitation of our system (Supplementary Fig. 2, available online at www.giejournal.org).

This study has several unique features. First, it is the largest study to date evaluating the use of CADx for Barrett's neoplasia. Second, the additive value of CADx was evaluated in a 2-phase benchmarking study involving 44 general endoscopists and 10 international BE experts from 6 different countries. The 2-phase design of the study allowed for evaluation of the interaction between the endoscopists and the CADx system, enhancing the clinical relevance of results. Third, the performance of our algorithm was tested on a corresponding set of images and videos, in both near focus and standard focus, providing a comprehensive understanding of the capabilities of the CADx system over multiple modalities. Finally, an important limitation of most CAD systems toward clinical implementation is their high computational complexity. One distinctive feature of our CADx system is its low computational demands, enabling real-time video-based assessment, and, most importantly, direct integration into existing endoscopy systems.

Our study also has limitations that should be acknowledged. First, the CADx system was trained, validated, and tested on high-quality data collected by expert endoscopists from 8 international centers. The performance of the system when presented with lower quality imagery has not been formally evaluated. Important parts of high-quality examinations to acquire high-quality data are inspection time, cleaning, and thorough inspection. A CAD system (and an endoscopist) can only detect what is visualized. We are therefore currently also working on a quality control algorithm to improve the procedural quality during Barrett surveillance endoscopy.

Second, the current study did not include imagery with low-grade dysplasia because of the associated uncertainty with pathologic diagnosis. Third, the system was only trained and tested on Olympus imagery. The performance on endoscopy systems of other manufacturers is unknown and should be investigated in future studies. Fourth, there was a partial overlap between our internal validation set and training set. Ideally, a strict separation of these data sets is preferred; however, to augment the volume of data for testing and internal validation purposes, we permitted a limited degree of overlap. It is crucial to note that the test set was rigorously segregated from the training and validation sets and used solely for a single performance assessment. Fifth, a learning effect between assessment phases 1 and 2 cannot be ruled out. This is inherent to the 2-phase design of the study. To limit this effect, no feedback was given after assessment round 1, and a strict washout period in between the phases was applied. Sixth, this study evaluated the CADx system as a stand-alone tool, while in clinical practice its true value may lie in conjunction with an earlier CADe system. The implementation of a supplemented CADx system has the potential to decrease the rate of false-positive results generated by the CADe system. We plan to conduct a pilot study examining the real-time interaction between the endoscopist and the CADx system. This study could explore the interaction between an initial CADe and a secondary CADx system and the effect of re-imaging areas where the system initially failed to classify.

In conclusion, we have developed and evaluated a ready-for-use CADx system for the detailed characterization of Barrett's neoplasia. The system was tested prospectively on both image and video and accurately distinguished neoplasia from NDBE. The performance of the CADx system was superior to that of general endoscopists and was comparable to that of BE experts. Moreover, CADx assistance significantly increased characterization performance of general endoscopists, indicating its additive value during Barrett surveillance endoscopies.

DISCLOSURE

The following authors disclosed financial relationships: M. Barret: research funding from Pentax Medical; consulting fees from Medtronic; medical training honorarium from Olympus; and board membership with Norgine and Ambu. J. J. Bergman: research support from Olympus Tokyo; and support for institutional review board (IRB)-approved research from Medtronic. T. Beyna: paid consultancy and lecture fees from Olympus and MicroTech; paid consultancy from Boston Scientific; and lecture fees from Fujifilm, Pentax, and Erbe. A. J. de Groof: research support from Olympus Tokyo. R. Mallant-Hent: consulting fees from Janssen. O. Pech: speaker honorarium from Boston Scientific, Medtronic, Fujifilm, Olympus, Aohua, Falk, and BMS. R. Pouw: consulting fees from Medtronic and MicroTech; and speaker fees from Pentax Medical. K. Ragunath: consulting fees from Olympus and Boston Scientific; research support from Olympus. F. van der Sommen: research support from Olympus Tokyo. P. de With: research support from Olympus Tokyo. B. Weusten: financial support for IRB-approved studies, speaker fees, and consulting fees from Pentax Medical; research support from Aqua Medical; financial support for IRB-approved studies from the St. Antonius research fund; and chair of the European Society of Gastrointestinal Endoscopy working group for the revision of the BE guideline. All other authors disclosed no financial relationships. This study received financial and logistical support from Olympus (Tokyo, Japan). Olympus did not have an active role in the design of the study, analysis of the results, or the writing of the manuscript.

ACKNOWLEDGMENTS

The authors thank the members and collaborators of the BONS-AI consortium, as listed below in alphabetical order: Alaa Alkhalaf, Isala Hospital, Zwolle, the Netherlands; Lorenza Alvarez Herrero, St. Antonius Hospital, Utrecht, the Netherlands; Francisco Baldaque-Silva, Karolinska University Hospital, Stockholm, Sweden; Maximilien Barret, Cochin Hospital, Paris, France; Jacques J. Bergman, Amsterdam UMC, Amsterdam, the Netherlands; Torsten Beyna, Evangelisches Krankenhaus Düsseldorf, Düsseldorf, Germany; Raf Bisschops, University Hospitals Leuven, Leuven, Belgium; Tim G. Boers, TU Eindhoven, Eindhoven, the Netherlands; Wouter Curvers, Catharina Hospital, Eindhoven, the Netherlands; Pierre H. Deprez, University Hospital Saint-Luc, Brussels, Belgium; Lucas C. Duits, Amsterdam UMC, Amsterdam, the Netherlands; Peter Elbe, Karolinska University Hospital & Karolinska Institute, Stockholm, Sweden; Jose Miguel Esteban, Clínico San Carlos, Madrid, Spain; Gary W. Falk, Hospital of the University of Pennsylvania, Philadelphia, Pennsylvania, USA; Kiki N. Fockens, Amsterdam UMC, Amsterdam, the Netherlands; Gregory G. Ginsberg, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, USA; Albert J de Groof, Amsterdam UMC, Amsterdam, the Netherlands; Rehan Haidry, Cleveland Clinic London and University College Hospitals NHS, London, United Kingdom; Martin H. Houben, Haga-Ziekenhuis, The Hague, the Netherlands; Anthony Infantolino, Thomas Jefferson University Hospital, Philadelphia, Pennsylvania, USA; Prasad G. Iyer, Mayo Clinic, Rochester, Minnesota, USA; Martijn Jong, Amsterdam UMC, Amsterdam, the Netherlands; Pieter-Jan de Jonge, Erasmus MC, Rotterdam, the Netherlands; Jelmer B. Jukema, Amsterdam UMC, Amsterdam the Netherlands; Arjun Koch, Erasmus MC, Rotterdam, the Netherlands; Srinadh Komanduri, Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA; Vani Konda, Baylor University Medical Center, Dallas, Texas, USA; Koen C. Kusters, TU Eindhoven, Eindhoven, the Netherlands; Philippe Leclerq, University Hospitals Leuven, Leuven, Belgium; Cadman L. Leggett, Mayo Clinic, Rochester, Minnesota, USA; Charles J. Lightdale, Columbia University Medical Center, New York, New York, USA; Rosalie C. Mallant-Hent, Flevoziekenhuis, Almere, the Netherlands; Guiomar Moral Villarejo, Nottingham University Hospitals NHS Trust, Nottingham, United Kingdom. V. Raman Muthusamy, UCLA Health, Los Angeles, California, USA; Jacobo Ortiz Fernández-Sordo, Nottingham University Hospitals NHS Trust, Nottingham, United Kingdom; Oliver Pech, St. John of God Hospital, Regensburg, Germany; Ian Penman, Royal Infirmary Edinburgh, Edinburgh, United Kingdom; Roos E. Pouw, Amsterdam UMC, Amsterdam, the Netherlands; Joost A. van der Putten, TU Eindhoven, Eindhoven, the Netherlands; Krish Ragunath, Royal Perth Hospital, Perth, Australia; Pieter Scholten, Onze Lieve Vrouwe Gasthuis, Amsterdam, the Netherlands; Stefan Seewald, Klinik Hirslanden, Zurich, Switzerland; Amritha Sethi, Columbia University Medical Center, New York, New York, USA; Michael S. Smith, Mount Sinai West & Mount Sinai Morningside Hospitals, New York, New York, USA; Fons van der Sommen, TU Eindhoven, Eindhoven, the Netherlands; Arvind Trindade, Zucker School of Medicine at Hofstra/Northwell, Long Island Jewish Medical Center, New Hyde Park, New York, New York, USA; Sachin Wani, University of Colorado Anschutz Medical Center, Aurora, Colorado, USA; Irving Waxman, Rush University Medical Center, Chicago, Illinois, USA; Jessie Westerhof, UMC Groningen, Groningen, the Netherlands; Bas L. Weusten, UMC Utrecht & St. Antonius Hospital, Utrecht, the Netherlands; Peter H. de With, TU Eindhoven, Eindhoven, the Netherlands; and Herbert C. Wolfsen, Mayo Clinic, Jacksonville, Florida, USA.

REFERENCES

1. Klaver E, Bureo Gonzalez A, Mostafavi N, et al. Barrett's esophagus surveillance in a prospective Dutch multi-center community-based cohort of 985 patients demonstrates low risk of neoplastic progression. United European Gastroenterol J 2021;9:929-37.
2. Schölvinck DW, van der Meulen K, Bergman J, et al. Detection of lesions in dysplastic Barrett's esophagus by community and expert endoscopists. Endoscopy 2017;49:113-20.
3. Beveridge CA, Mittal C, Muthusamy VR, et al. Identification of visible lesions during surveillance endoscopy for Barrett's esophagus: a video-based survey study. Gastrointest Endosc 2023;97:241-7.e242.

4. Qumseya B, Sultan S, Bain P, et al. ASGE guideline on screening and surveillance of Barrett's esophagus. Gastrointest Endosc 2019;90:335-59.e332.
5. Singh M, Bansal A, Curvers W, et al. Observer agreement in the assessment of narrowband imaging system surface patterns in Barrett's esophagus: a multicenter study. Endoscopy 2011;43:745-51.
6. Silva FB, Dinis-Ribeiro M, Vieth M, et al. Endoscopic assessment and grading of Barrett's esophagus using magnification endoscopy and narrow-band imaging: accuracy and interobserver agreement of different classification systems (with videos). Gastrointest Endosc 2011;73:7-14.
7. Curvers W, Bohmer C, Mallant-Hent R, et al. Mucosal morphology in Barrett's esophagus: interobserver agreement and role of narrow band imaging. Endoscopy 2008;40:799-805.
8. Curvers W, Baak L, Kiesslich R, et al. Chromoendoscopy and narrow-band imaging compared with high-resolution magnification endoscopy in Barrett's esophagus. Gastroenterology 2008;134:670-9.
9. Rondonotti E, Hassan C, Tamanini G, et al. Artificial intelligence-assisted optical diagnosis for the resect-and-discard strategy in clinical practice: the Artificial intelligence BLI Characterization (ABC) study. Endoscopy 2023;55:14-22.
10. Biffi C, Salvagnini P, Dinh NN, et al. A novel AI device for real-time optical characterization of colorectal polyps. NPJ Digit Med 2022;5:84.
11. Fockens KN, Jukema JB, Boers T, 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;11:324-36.
12. Fockens K, Jong M, Jukema J, et al. A deep learning system for detection of early Barrett's neoplasia: a model development and validation study. Lancet Digital Health 2023;5:e905-16.
13. Jong M, Fockens K, Jukema J, et al. Video-based computer aided detection system detects Barrett's neoplasia with high accuracy during live endoscopic procedures: a multi-center pilot and feasibility study. Endoscopy Supplement ESGE Days 2023;55:S7. Available at: https://www.esge.com/assets/downloads/pdfs/guidelines/ESGE_Days_2023.pdf.
14. de Groof AJ, Struyvenberg MR, Fockens KN, 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;91:1242-50.
15. Fockens K, de Groof J, van der Putten J, et al. Linked color imaging improves identification of early gastric cancer lesions by expert and non-expert endoscopists. Surg Endosc 2022;36:8316-25.
16. de Groof AJ, Struyvenberg MR, van der Putten J, 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;158:915-29.e914.
17. Liu R. Higher accuracy on vision models with EfficientNet-Lite. Available at: https://blog.tensorflow.org/2020/03/higher-accuracy-on-vision-models-with-efficientnet-lite.html. (2020). Accessed August 26, 2020.
18. van der Laan JJ, van der Putten JA, Zhao X, et al. Optical biopsy of dysplasia in Barrett's oesophagus assisted by artificial intelligence. Cancers 2023;15:1950.
19. Struyvenberg MR, de Groof AJ, van der Putten J, et al. A computer-assisted algorithm for narrow-band imaging-based tissue characterization in Barrett's esophagus. Gastrointest Endosc 2021;93:89-98.
20. Ebigbo A, Mendel R, Probst A, et al. Computer-aided diagnosis using deep learning in the evaluation of early oesophageal adenocarcinoma. Gut 2019;68:1143-5.
21. Hussein M, Lines D, Puyal JG, et al. Computer-aided characterization of early cancer in Barrett's esophagus on i-scan magnification imaging—multicenter international study. Gastrointest Endosc 2023;97:646-54.

Abbreviations: BE, Barrett's esophagus; CADe, computer-aided detection; CADx, computer-aided diagnosis; NBI, narrow-band imaging; NDBE, nondysplastic Barrett's esophagus; WLE, white-light endoscopy.

DIVERSITY, EQUITY, AND INCLUSION: One or more of the authors of this paper self-identifies as an under-represented gender minority in science. One or more of the authors of this paper self-identifies as an under-represented ethnic minority in science. The author list of this paper includes contributors from the location where the research was conducted who participated in the data collection, design, analysis, and/or interpretation of the work.

*All members and collaborators of the BONS-AI consortium are listed in the Supplementary Materials.

This video can be viewed directly from the *GIE* website or by using the QR code and your mobile device. Download a free QR code scanner by searching "QR Scanner" in your mobile device's app store.

0016-5107
https://doi.org/10.1016/j.gie.2024.04.013

Received September 5, 2023. Accepted April 8, 2024.

Current affiliations: Department of Gastroenterology and Hepatology, Amsterdam Gastroenterology Endocrinology Metabolism, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands (1), Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, the Netherlands (2), Department of Gastroenterology and Hepatology, UMC Utrecht, University of Utrecht, Utrecht, the Netherlands (3), Department of Gastroenterology and Hepatology, Sint Antonius Hospital, Nieuwegein, the Netherlands (4), Department of Gastroenterology and Hepatology, HagaZiekenhuis Den Haag, Den Haag, the Netherlands (5), Department of Gastroenterology and Hepatology, UMC Groningen, University of Groningen, Groningen, the Netherlands (6), Department of Gastroenterology and Hepatology, Isala Hospital Zwolle, Zwolle, the Netherlands (7), Department of Gastroenterology and Hepatology, Flevoziekenhuis Almere, Almere, the Netherlands (8), Department of Gastroenterology and Hepatology, Onze Lieve Vrouwe Gasthuis, Amsterdam, the Netherlands (9), Department of Gastroenterology and Hepatology, Royal Perth Hospital, Curtin University, Perth, WA, Australia (10), Department of Gastroenterology and Hepatology, Hirslanden Klinik, Zurich, Switzerland (11), Department of Digestive Diseases, Karolinska University Hospital, Stockholm, Sweden (12), Division of Surgery, Department of Clinical Science, Intervention and Technology, CLINTEC, Karolinska Institutet, Stockholm, Sweden (13), Center for Advanced Endoscopy Carlos Moreira da Silva, Gastroenterology Department, Pedro Hispano Hospital, ULSM Matosinhos, Portugal (14), Department of Gastroenterology and Hepatology, Cochin Hospital Paris, Paris, France (15), Department of Gastroenterology and Hepatology, Nottingham University Hospitals NHS Trust, Nottingham, United Kingdom (16), Department of Gastroenterology and Hepatology, St. John of God Hospital, Regensburg, Germany (17), Department of Gastroenterology and Hepatology, Evangalisches Krankenhaus Düsseldorf, Düsseldorf, Germany (18), Biostatistics Unit, Department of Gastroenterology and Hepatology, Amsterdam University Medical Center, location Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands (19).

Reprint requests: Jacques J. Bergman, MD, PhD, Amsterdam UMC, De Boelelaan 1117, 1081 HV Amsterdam, the Netherlands. E-mail: j.j.bergman@amsterdamumc.nl.

APPENDIX 1

Statistical analysis of 4 separate comparisons

The structure of the benchmark study allowed for 4 clinically relevant comparisons: (1) stand-alone performance of computer-aided diagnosis (CADx) versus general endoscopists; (2) general endoscopists without CADx assistance versus general endoscopists with CADx assistance; (3) stand-alone performance of CADx versus Barrett's esophagus (BE) experts; and (4) general endoscopists with CADx assistance versus BE experts.

Comparisons between endoscopists and/or CADx comprised repeated measurements (ie, with and without CADx assistance) or clustered measurements (ie, evaluation of same videos by multiple endoscopists or same endoscopist evaluation of multiple videos). All outcome parameters of these comparisons were binary variables (ie, correct prediction of either neoplasia or nondysplastic BE). Therefore, mixed-effects logistic regression models (lme4 package in R Studio) were used.

First, for the comparison between CADx and general endoscopists, the fixed effect was the evaluating entity being either an endoscopist or a CADx system. The video was considered a random effect. Second, for the comparison between general endoscopists without CADx assistance and general endoscopists with CADx assistance, the fixed effect was the assistance of CADx. The video and evaluating endoscopist were considered random effects. Third, for the comparison between the CADx system and BE experts, the fixed effect was the evaluating entity being either an endoscopist or a CADx system. The video was considered a random effect. Finally, for the comparison between general endoscopists with CADx assistance and BE experts, expertise was considered the fixed effect. The video and the evaluating endoscopist were considered random effects.

All comparisons were tested two-sided. Fixed-effect correlations with a P value $\leq .05$ were considered significant.

Technical details of the developed CADx system

Many deep neural networks in healthcare involve a considerable large number of parameters, which impose considerable hardware requirements, in terms of computation, memory, and bandwidth. This renders the direct implementation of these applications on the typical hardware platforms of current endoscopy systems unattainable. In this work, we have therefore focused on reducing the model's complexity to make real-time artificial intelligence directly feasible for endoscopic systems.

The CADx system was constructed using an EfficientNet-lite1 feature encoder (Supplementary Fig. 1), which is optimized for fast and efficient processing of real-time imagery. The EfficientNet-lite1 architecture comprises an input layer and 7 stages (Supplementary Table 2), which are constructed with Mobile Conv layers (Supplementary Fig. 3). Here, each stage has a specific kernel size, and a specific amount of repeating Mobile Conv layers. Progressing deeper into the network, the feature maps are down-scaled using strided convolutions.

To achieve real-time performance using this architecture, 4 specific optimizations were incorporated into the network: (1) the architecture is specifically optimized to achieve the best accuracy per number of operations; (2) the operations are strongly parallelizable, which means that many operations can be performed at the same time, instead of sequentially; (3) batch normalization operations are completely removed from our model by folding the batch-normalization values into the convolutional filters after the training is completed; and (4) all operations are quantized to operate in a limited integer range of 8 bits.

Because these quantized functions can only operate in a limited range, an activation function such as Rectified-Layer Unit with an unbounded activation output is not suitable. As a solution, Rectified-Layer Unit 6 is introduced to constraint the activation range by clipping the output at a value of 6. This alteration improves the conversion of the model from a higher precision float32 to integer-8 precision, as the high precision model already needs to learn to operate in a limited range.

Moving from a higher precision float value to a lower precision integer-8 value is a lossy process; hence, the network can have a significant loss of accuracy. To minimize the performance loss, the quantization-aware training (QAT) technique is introduced in the training process. This technique simulates the low precision behavior in the forward pass, while preserving high precision in the backward pass for updating the weights. For this purpose, a quantization error is induced in the total loss, which the optimizer can precisely reduce by adjusting the parameters accordingly. The result is an almost lossless model, which attains similar accuracies as high precision float32 models, and can operate on simple edge devices without vast computing capabilities.

Output stability function for video classification

Video classification was based on a stability algorithm, which provided a weighted prediction in real time. As the certainty of the CADx system increases, a gray bar located to the left of the real-time video becomes filled. Once the bar is completely filled, the bar turns red or green and the text "Neoplasia" or "No Dysplasia" is displayed (Fig. 1).

To provide a stable video classification with the proposed architecture, we propose to use a stability function on the classification output of the architecture (Supplementary Fig. 4). This stability function ensures a stable video-classification output by incorporating temporal information embedded in the frame-sequence. Real-time feedback of the combination of the algorithm and the stability function is ensured by subsampling 1 every 3 frames of 50 fps input videos, while 25 fps input videos are analyzed at the input

frame rate. The stability function consists of 5 fundamental output states (Idle, Under Inference [Neoplastic], Under Inference [Nondysplastic], Final Inference [Neoplastic], and Final Inference [Nondysplastic]) with corresponding output classes (Class 0, before classification; Class 1, under classification; Class 2, Neoplastic; and Class 3, Nondysplastic), with possible transitions in between the states. The transitions are dependent on the frame-based detected class (Neoplastic, "Failure to Classify" or Nondysplastic) of the neural network architecture, whereas the update values for each transition are dependent on the frame-based probability score (Supplementary Fig. 5).

At the beginning of each video, the stability function is initialized with cumulative value $K_{sum} = 0$ (with maximum value $K_{sum\_max} = 2^n$), which changes over time as inference proceeds (Supplementary Fig. 4). The workflow for each processed video frame consists of 4 steps. First, the frame is provided to the neural network, which produces a dysplasia probability score. Second, this probability score is compared with the predefined Lower Threshold and Upper Threshold (Supplementary Fig. 5) to obtain the frame-based detected class and corresponding update value. Third, the corresponding transition is according to the stability function. Lastly, the stability is calculated with $\text{Rounddown}([K_{sum}/K_{sum\_max}] \times 100)$. At each frame of the video, the current output class and the value of stability are shown to the user.

Multistage transfer learning

ImageNet pretraining has been an established method to learn valuable features, which have proven to also be useful even in out-of-domain data applications. Hence, using weights pretrained with natural images such as those in the ImageNet data set leads to better results compared with training from scratch with small data sets.

To increase the robustness of the model, we used an additional pretraining using a domain-specific data set, called GastroNet,[11-14] consisting of 5,084,494 images of the GI tract and a set of 3743 labeled images in the classes: esophagus, stomach, small intestines, colon, and other. Using a semi-supervised framework called Unsupervised Data Augmentation, we train a network to (1) recognize the 5 labeled classes; and (2) to predict the same output of heavily and lightly augmented images of the unlabeled data set. As shown,[11-14] domain-specific pretraining on a data set that resembles the target domain can aid in improving the performance.

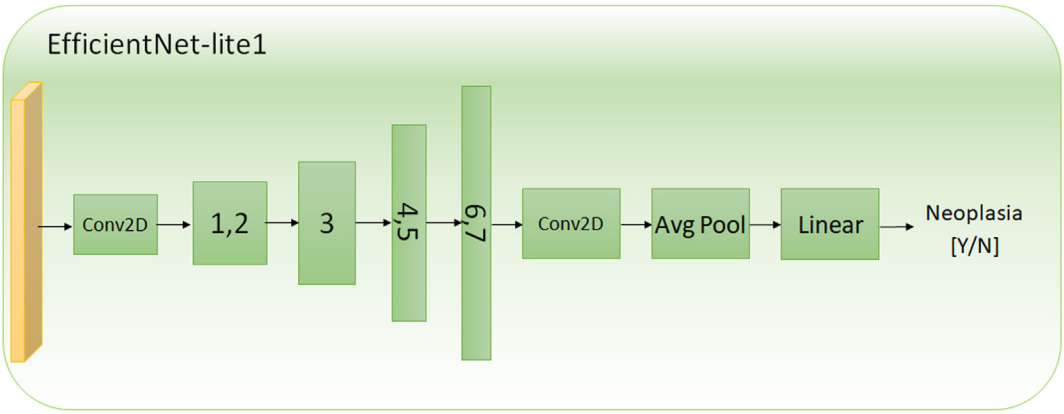
After the pretraining steps, the CADx system is fine-tuned into 2 separate stages: first, a training state in full float32 precision; and second, a QAT stage that simulates the quantized state of the model. The first step ensures that the network weights converge toward a good minimum in the loss landscape, before fine-tuning the quantized weights. Because the loss landscape is not as smooth during QAT due to the introduced rounding errors, it can be more difficult to find a good point of generalization if the training is started directly with the QAT approach.

Development infrastructure

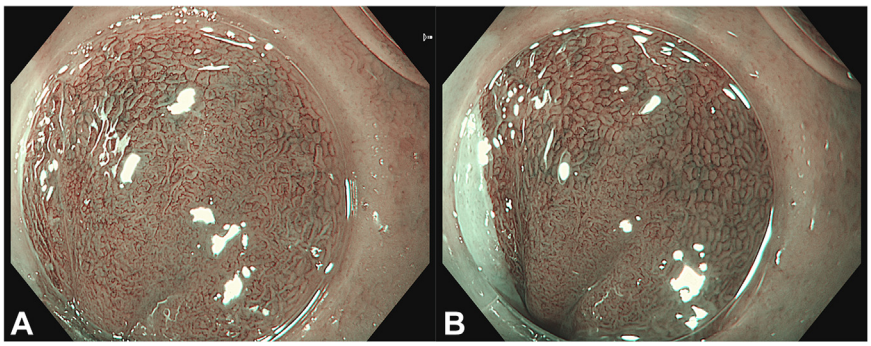
The CADx system was developed and evaluated using the Tensorflow/Keras toolkit in the Python Framework. Tensorflow/Keras was extended with the "Tensorflow Model Optimization" toolkit to quantize the model. All experiments were performed on a workstation with the following specifications: Ryzen 9 5900X CPU (AMD, Santa Clara, Calif, USA), 32 GB random-access memory, and a GeForce RTX 3080 Ti (NVIDIA, Santa Clara, Calif, USA) with 12 GB video random access memory.

False-negative CADx prediction

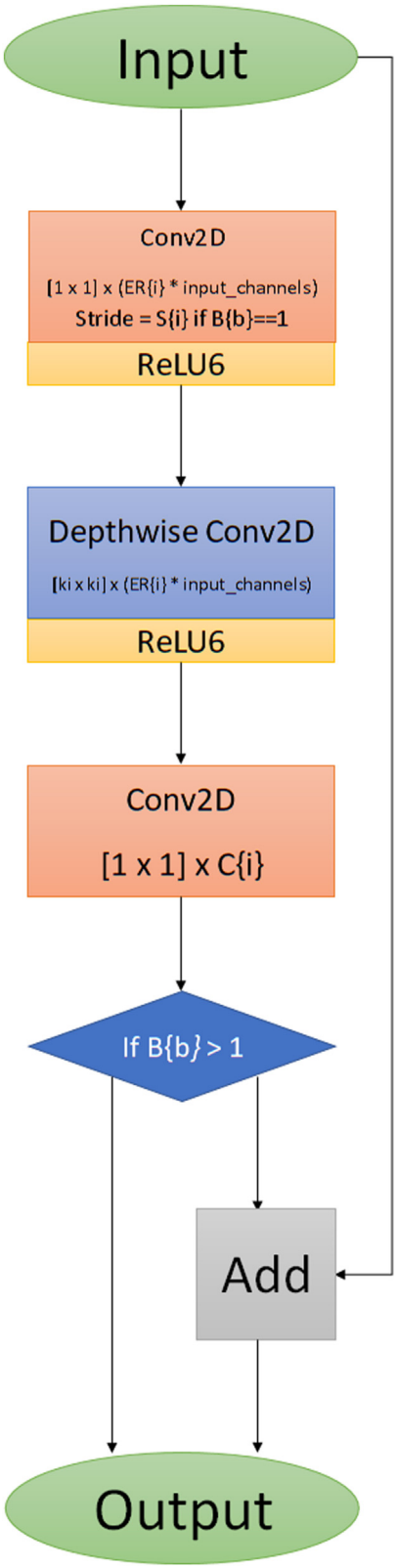
Results stratified according to endoscopists' experience. Supplementary Table 3 presents the results stratified according to the experience of the endoscopist. Supplementary Figure 6 shows the change in performance of individual general endoscopists when CADx assistance was available, stratified according to experience.



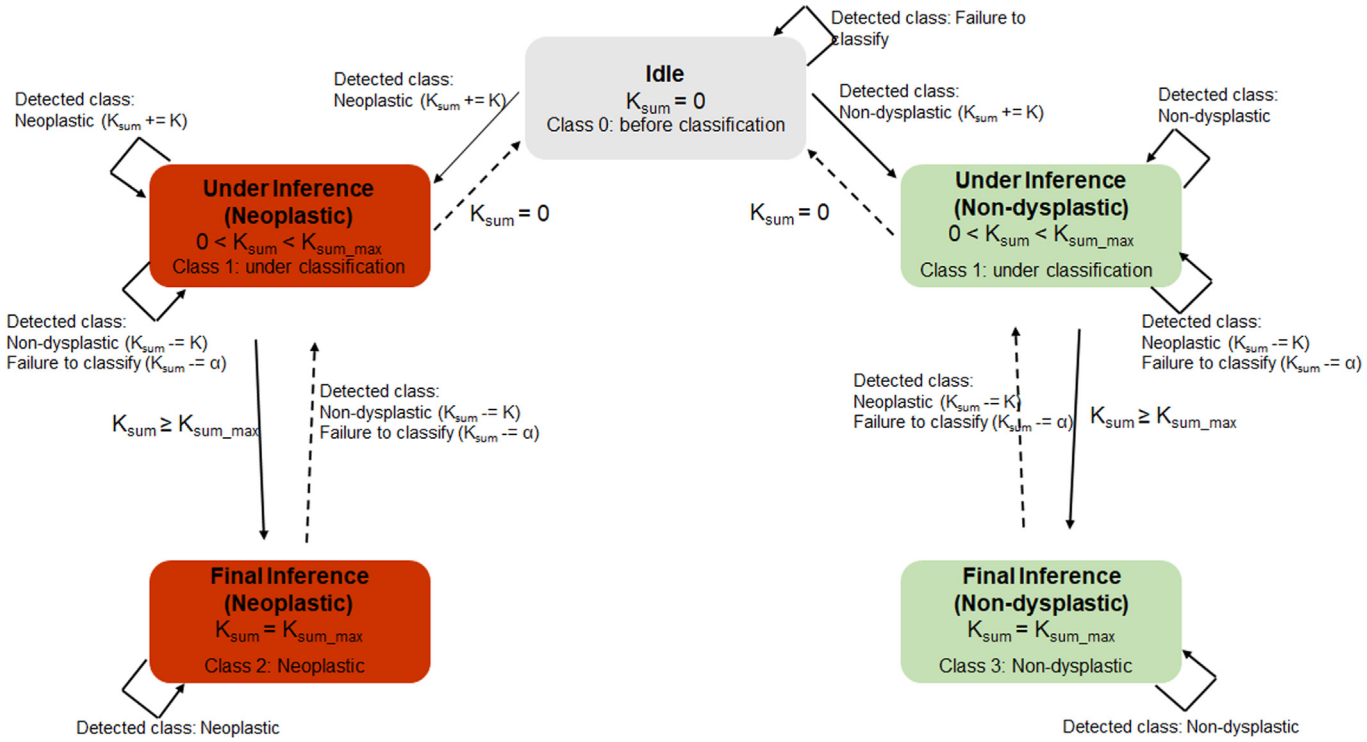
Supplementary Figure 1. Overview of the neural network design. The network uses an EfficientNet-lite1 feature encoder. Here, each independent convolutional operation is denoted with [k × k] × f, where k refers to the filter size and f to the number of filters, and is followed up with a Rectified-Layer Unit 6 activation function, except for the output layers.



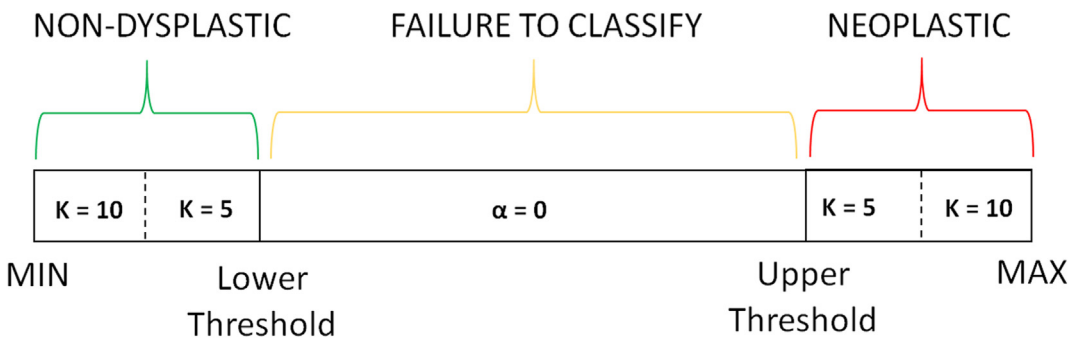
Supplementary Figure 2. Neoplastic lesion incorrectly classified by the computer-aided diagnosis system as nondysplastic in standard focus (right) while being correctly classified as neoplastic in near focus (left).



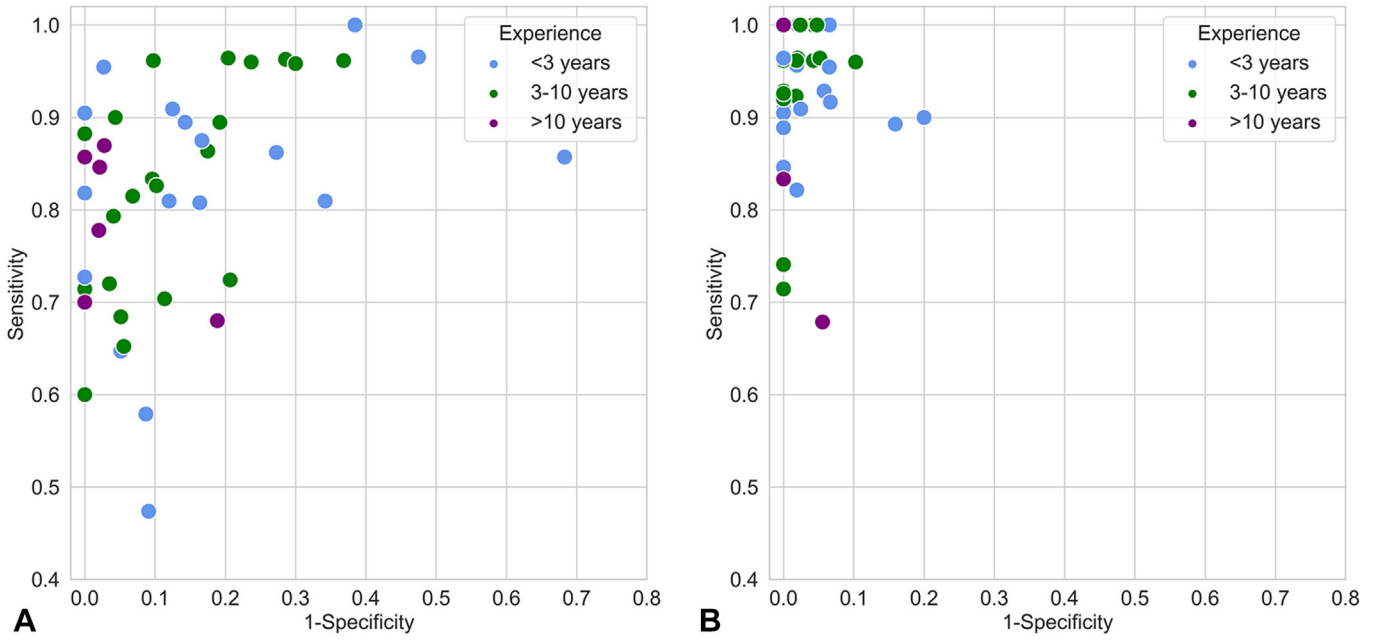
Supplementary Figure 3. Mobile Convolution block using Depthwise Separable Convolutions, which are used in the Efficient Net-lite architecture. *ReLU6*, Rectified-Layer Unit 6.



Supplementary Figure 4. Graphical diagram of the output stability function, supplementing the neural network architecture for video classification.



Supplementary Figure 5. Graphical illustration of the values for the update rules in the stability function.



Supplementary Figure 6. Change in performance of individual general endoscopists when computer-aided diagnosis (CADx) assistance was available, stratified according to experience. The left graph is without CADx assistance; the right graph is with CADx assistance.

SUPPLEMENTARY TABLE 1. Overview of data sets used for the development of the CADx system

Data set and purpose	Type of imagery	No. of images (no. of patients)		Acquisition	Type of labeling
		Neoplastic	NDBE		
1. General pretraining on nonendoscopic images	ImageNet	1,200,000 (NA)		NA	NA
2. Domain-specific pretraining on general endoscopic images	GastroNet	5,084,494 (unknown)		Retrospective acquisition	Subset: hand-labeled by 2 experts
3a. Training images	Barrett specific	801 (267)	201 (68)	Retrospective acquisition	Hand-labeled by 3 experts, corresponding pathology
3b. Training images	Barrett specific	912 (66)	1554 (98)	Prospective acquisition	Hand-labeled by 3 experts, corresponding pathology
4a. Validation images	Barrett specific	45 (22)	83 (31)	Prospective acquisition	Hand-labeled by 3 experts, corresponding pathology
4b. Validation videos	Barrett specific	40 (19)	82 (31)	Prospective acquisition	Hand-labeled by 3 experts, corresponding pathology
5a. Testing near focus images	Barrett specific	30 (20)	60 (31)	Prospective acquisition	Hand-labeled by 3 experts, corresponding pathology
5b. Testing standard focus images	Barrett specific	20 (14)	51 (26)	Prospective acquisition	Hand-labeled by 3 experts, corresponding pathology
5c. Testing near focus videos benchmark test set	Barrett specific	30 (20)	60 (31)	Prospective acquisition	Hand-labeled by 3 experts, corresponding pathology
5d. Testing standard focus videos	Barrett specific	20 (14)	51 (26)	Prospective acquisition	Hand-labeled by 3 experts, corresponding pathology

CADx, Computer-aided diagnosis; *NDBE*, nondysplastic Barrett's esophagus; *NA*, not applicable.

SUPPLEMENTARY TABLE 2. EfficientNet-lite1 detailed specifications

Stage i	Operator	Kernel size [ki × ki]	Input resolution (H{i} × W{i})	Channels (C{i})	Stride (S{i})	No. of mobile conv blocks (B{b})	Expand ratio (ER{i})
Input stage	Conv	3	256 × 256	32	2	n/a	n/a
1	MBConv	3	128 × 128	16	1	1	1
2	MBConv	3	128 × 128	24	2	3	6
3	MBConv	5	64 × 64	40	2	3	6
4	MBConv	3	32 × 32	80	2	4	6
5	MBConv	5	16 × 16	112	1	4	6
6	MBConv	5	16 × 16	192	2	5	6
7	MBConv	3	8 × 8	320	1	1	6

SUPPLEMENTARY TABLE 3. Results stratified according to experience

Experience	Phase	Sensitivity	Specificity	FTC/uncertain
<3 years (n = 17)	Phase 1	86%	88%	26%
	Phase 2	92%	98%	18%
	Change	6%	10%	–8%
3-10 years (n = 21)	Phase 1	83%	90%	21%
	Phase 2	96%	98%	13%
	Change	13%	8%	–8%
>10 years (n = 6)	Phase 1	81%	98%	22%
	Phase 2	96%	100%	9%
	Change	15%	2%	–13%

FTC, Failure to classify.