Deep-Learning System Detects Neoplasia in Patients With Barrett's Esophagus With Higher Accuracy Than Endoscopists in a Multi-Step Training and Validation Study with Benchmarking

A.J. de Groof, M.R. Struyvenberg, J. van der Putten, F. van der Sommen, K.N. Fockens, W.L. Curvers, S. Zinger, R.E. Pouw, E. Coron, F. Baldaque-Silva, O. Pech, B. Weusten, A. Meining, H. Neuhaus, R. Bisschops, J. Dent, E.J. Schoon, P.H. de With, J.J. Bergman

PII: S0016-5085(19)41586-2
DOI: <https://doi.org/10.1053/j.gastro.2019.11.030>
Reference: YGAST 63029

To appear in: *Gastroenterology*
Accepted Date: 18 November 2019

Please cite this article as: de Groof AJ, Struyvenberg MR, van der Putten J, van der Sommen F, Fockens KN, Curvers WL, Zinger S, Pouw RE, Coron E, Baldaque-Silva F, Pech O, Weusten B, Meining A, Neuhaus H, Bisschops R, Dent J, Schoon EJ, de With PH, Bergman JJ, Deep-Learning System Detects Neoplasia in Patients With Barrett's Esophagus With Higher Accuracy Than Endoscopists in a Multi-Step Training and Validation Study with Benchmarking, *Gastroenterology* (2019), doi: <https://doi.org/10.1053/j.gastro.2019.11.030>.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2019 by the AGA Institute

Deep learning system detects Barrett's Esophagus neoplasia with high accuracy when compared to endoscopists: a multi-step training and external validation study with benchmarking

Primary study outcome:
Classification performance

Neoplasia
present

Neoplasia
not present

Secondary study outcome:
Localization performance

Gastroenterology

Deep-Learning System Detects Neoplasia in Patients With Barrett's Esophagus With Higher Accuracy Than Endoscopists in a Multi-Step Training and Validation Study with Benchmarking

A.J. de Groof¹, M.R. Struyvenberg¹, J. van der Putten², F. van der Sommen², K.N. Fockens¹, W.L. Curvers³, S. Zinger², R.E. Pouw¹, E. Coron⁴, F. Baldaque-Silva⁵, O. Pech⁶, B. Weusten⁷, A. Meining⁸, H. Neuhaus⁹, R. Bisschops¹⁰, J. Dent¹¹, E.J. Schoon³, P.H. de With², J.J. Bergman¹

¹ Department of Gastroenterology and Hepatology, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands.

² Department of Electrical Engineering, VCA group, Eindhoven University of Technology, Eindhoven, the Netherlands

³ Department of Gastroenterology and Hepatology, Catharina Hospital Eindhoven, Eindhoven, the Netherlands

⁴ Institut des Maladies de l'Appareil Digestif, University Hospital of Nantes place Alexis Ricordeau, Nantes, France

⁵ Department of Digestive Diseases, Karolinska University Hospital and Karolinska Institute, Stockholm, Sweden

⁶ Gastroenterology and interventional Endoscopy, Krankenhaus Barmherzige Brüder, Regensburg, Germany

⁷ Department of Gastroenterology and Hepatology, St. Antonius Hospital, Nieuwegein, the Netherlands

⁸ Center of Internal Medicine, Ulm University, Ulm, Germany

⁹ Internal Medicine, Evangelisches Krankenhaus Düsseldorf, Düsseldorf, Germany

¹⁰ Department of Gastroenterology and Hepatology, University Hospitals Leuven, Leuven, Belgium

¹¹ Department of Medicine, University of Adelaide and Royal Adelaide Hospital, Adelaide, South Australia

The Netherlands National Trials Registry (NTR) number: NTR7072

This research is supported by the Dutch Cancer Society and Technology Foundation STW, as part of their joint strategic research program 'Technology for Oncology'

Corresponding author:

J.J. Bergman, MD PhD

Professor of Gastrointestinal Endoscopy

Director of Endoscopy

Amsterdam UMC, location Academic Medical Center

Meibergdreef 9

1105 AZ Amsterdam

The Netherlands

Phone number: +31 (0)20 5663556

Fax number: +31 (0)20 6917033

Author contribution:

Study design: A.J. de Groof, M.R. Struyvenberg, J. van der Putten, F. van der Sommen, S. Zinger, R. Bisschops, W.L. Curvers, E.J. Schoon, P.H. de With, J.J. Bergman

Patient recruitment and data collection: A.J. de Groof, M.R. Struyvenberg, K.N. Fockens, F. Baldaque Silva, W.L. Curvers, R. Bisschops, E.J. Schoon, R.E. Pouw, B. Weusten, E. Coron, J.J. Bergman

Image delineation: R.E. Pouw, B. Weusten, O. Pech, A. Meining, H. Neuhaus, R. Bisschops, E.J. Schoon, J.J. Bergman, A.J. de Groof, M.R. Struyvenberg, E. Coron

CAD system development: J. van der Putten, F. van der Sommen, S. Zinger, P.H. de With

Data analysis and interpretation: J. van der Putten, F. van der Sommen, A.J. de Groof, M.R. Struyvenberg, W.L. Curvers, J.J. Bergman

Writing of manuscript: A.J. de Groof, M.R. Struyvenberg, J. van der Putten, F. van der Sommen, J. Dent, W.L. Curvers, P.H. de With, J.J. Bergman

All authors reviewed the final manuscript for important intellectual content and agreed to submit.

Disclosures:

A.J. de Groof: *None declared*

M.R. Struyvenberg: *None declared*

J. van der Putten: *None declared*

F. van der Sommen: *None declared*

K.N. Fockens: *None declared*

W.L. Curvers: *None declared*

S. Zinger: *None declared*

R.E. Pouw: *None declared*

E. Coron: *Abbvie, Alfa Sigma, Olympus (speaking fees), Cook, Mayoly Spindler (conference visits), FUJIFILM, Medtronic (consultant)*

F. Baldaque-Silva: *Boston Scientific*

O. Pech: *Medtronic, FUJIFILM, OLYMPUS, Boston Scientific, Cook (speaker fees)*

B. Weusten: *Pentax Medical (research support, speakers fees)*

A. Meining: *OVESCO (consulting fees)*

H. Neuhaus: *Boston Scientific, CDx Diagnostics, Cook Medical, Demcom, Erbe, Falk Foundation, Fujifilm, Medtronic, Olympus (consulting, advisory, speaking fees); Boston Scientific, CDx Diagnostics, Demcom, Erbe, Fujifilm, Olympus, Pentax Medical (research support)*

R. Bisschops: *Fujifilm (research support, consulting fees, speakers fees)*

J. Dent: *None declared*

E.J. Schoon: *None declared*

P.H. de With: *None declared*

J.J. Bergman: *Fujifilm, NinePoint Medical (research support); Fujifilm (speaking fees)*

ABSTRACT

Background & Aims: We aimed to develop and validate a deep-learning computer-aided detection (CAD) system, suitable for use in real time in clinical practice, to improve endoscopic detection of early neoplasia in patients with Barrett's esophagus (BE).

Methods: We developed a hybrid ResNet-UNet model CAD system using 5 independent endoscopy datasets. We performed pre-training using 494,364 labelled endoscopic images collected from all intestinal segments. Then, we used 1704 unique esophageal high-resolution images of rigorously confirmed early-stage neoplasia from patients with BE, and non-dysplastic BE, derived from 669 patients. System performance was assessed using datasets 4 and 5. Dataset 5 was also scored by 53 general endoscopists with a wide range of experience from 4 countries to benchmark CAD system performance. Coupled with histopathology findings, scoring of images that contained early-stage neoplasia in datasets 2–5 were delineated in detail for neoplasm position and extent by multiple experts whose evaluations served as the ground truth for segmentation.

Results: The CAD system classified images as containing neoplasms or non-dysplastic BE with 89% accuracy, 90% sensitivity, and 88% specificity (dataset 4, 80 patients and images). In dataset 5 (80 patients and images) values for the CAD system vs those of the general endoscopists were 88% vs 73% accuracy, 93% vs 72% sensitivity, and 83% vs 74% specificity.

The CAD system achieved higher accuracy than any of the individual 53 non-expert endoscopists, with comparable delineation performance. CAD delineations of the area of neoplasm overlapped with those from the BE experts in all detected neoplasia in datasets 4 and 5. The CAD system identified the optimal site for biopsy of detected neoplasia in 97% and 92% of cases (dataset 4 and 5 respectively).

Conclusions: We developed, validated, and benchmarked a deep-learning computer-aided system for primary detection of neoplasia in patients with BE. The system detected neoplasia with high accuracy and near-perfect delineation performance.

KEY WORDS: Barrett surveillance, machine learning, artificial intelligence, esophageal cancer

INTRODUCTION

Endoscopic surveillance is now accepted as the routine management of the significantly increased risk for development of esophageal adenocarcinoma (EAC) in Barrett's Esophagus (BE), because most surveillance-detected EAC is cured by low morbidity endoscopic surgery¹⁻³.

There is now general agreement that high-definition white light endoscopy (HD-WLE) is the best option for initial recognition of BE neoplasia by surveillance^{1,2}. The image quality of high-definition endoscopes is now so good that even the very subtle mucosal surface abnormalities seen in some cases of high-grade dysplasia (HGD) and EAC (collectively "early neoplasia") are recognized reliably by expert endoscopists whose practice is highly specialized on management of BE⁴.

Despite these major improvements, endoscopic detection of early BE neoplasia fails not infrequently, when surveillance is carried out by general endoscopists who must undertake the bulk of surveillance⁴. This problem arises because general endoscopists are infrequently confronted with early neoplasia in BE, since progression to neoplasm is rare (<1% per patient year)⁵. General endoscopists thus have limited familiarity with the endoscopic appearances of early BE neoplasia. This issue has been addressed recently by development and validation of an internet-delivered interactive training program for improvement of identification of early neoplasia by endoscopists, based on stored HD-WLE imagery⁶.

Another possible option to improve the accuracy of BE surveillance is to use *artificial intelligence* or *machine learning*, given its potential to recognize informative patterns in imagery. The recent success of deep neural networks for a wide variety of applications has greatly leveraged this potential⁷. To date, most studies have focused on colon polyp classification, with machine learning systems developed to *classify* dedicated images or videos⁷⁻¹⁵. The development of machine learning tools for *primary detection* of lesions is however more challenging as these do not only classify imagery, but also localize lesions within imagery. These tools are usually referred to as *computer aided detection* (CAD) systems.

The ARGOS-consortium (see Title Page) has been working to develop a clinically useful CAD system for detection of early BE neoplasia. Two previous ARGOS studies have shown that earlier "clinically inspired" versions of CAD system can recognize early neoplasia with good accuracy^{16,17}. Despite this success, the functioning of a clinically inspired CAD system is potentially restricted by the human perceptual bias in preselecting the features the system is supposed to use for detecting neoplasia (i.e. color and texture). Recent developments in CAD have shown promising results using deep learning techniques, a less strictly supervised approach not hampered by human perceptual bias. In contrast to *clinically inspired* methods,

deep learning techniques autonomously learn to distinguish features with increasing abstraction through the use of multiple processing layers. These techniques may thereby identify alternative discriminative features for neoplasia⁷. Also, importantly, in contrast to the clinically inspired system, a CAD system developed with deep learning techniques has the potential to achieve superior clinical utility by completing such rapid image analysis that it would be suitable for use in real-time, to support the recognition and targeted biopsy of possible BE neoplasia during surveillance endoscopy.

The aim of this study was to develop a state-of-the-art deep learning CAD system for endoscopic images of BE, followed by internal- and external validation, and to subsequently compare CAD performance with the performance of a group of international endoscopist assessors.

METHODS

Setting and design

This study was performed at the departments of Gastroenterology and Hepatology of 4 tertiary clinical referral centers for BE neoplasia (the Amsterdam University Medical Centers, location Academic Medical Center, the Catharina Hospital Eindhoven, the University Hospital Nantes, and the University Hospital Leuven), in partnership with the Department of Electrical Engineering of the Eindhoven University of Technology. Since only anonymized images were used, the “Medical Research Involving Human Subjects Act” did not apply to this study. Official approval of this study was therefore waived by the Medical Ethics Review Committees of the participating centers.

Deep learning CAD system design

The primary goal of our CAD system is to serve as a detection tool, by highlighting areas within the Barrett segment that may contain early neoplasia, thereby assisting endoscopists to detect areas of interest during surveillance endoscopy. Therefore, the CAD system was developed to first *classify* an image as “neoplastic” or “not neoplastic”. The CAD system then directly produces a “heatmap” and corresponding delineation in any neoplastic image, encircling the region suspicious for neoplasia (Figure 1). After delineating this region, the CAD system subsequently “red-flags” what it determines to be the most abnormal part of the lesion, by placing a *crosshair* in the image (Figure 1), thus marking the preferred primary biopsy site. In clinical practice, the actual delineation of lesions that follows upon their primary detection (i.e. prior to endoscopic resection) is usually performed with the use of optical chromoscopy in magnified view by endoscopists in expert centers. This is not the envisioned application of this CAD system and therefore pixel-precise delineation metrics are less relevant and not evaluated in this study. Primary detection is achieved by a combination of the CAD system delineation and the red-flag indicator crosshair, indicating the most abnormal area. In our opinion, this is a more clinical relevant application than a detailed CAD system to assess the true demarcation line of a lesion.

Datasets used for training and validation of CAD system

Five datasets were used for stepwise training, validation and benchmarking the deep learning CAD system (Table 1): 1) a dataset of 494,364 endoscopic images, containing a variety of endoscopic imagery, and labelled as either esophagus, gastric, small bowel or colon; 2) 1247 images, of early BE neoplasia (with the neoplastic lesion delineated by 2 experts) and non-dysplastic BE (NDBE); 3) 297 images of early BE neoplasia (with the neoplastic lesion delineated by 3 experts), and NDBE for refinement training and internal validation of the CAD system; 4) 40 images of early BE neoplasia, delineated by 3 experts, and 40

images labelled as NDBE, for primary external validation of the CAD system; 5) Finally, 40 images of subtle BE neoplasia, delineated by 6 experts, and 40 images labelled as NDBE, for secondary external validation of the CAD system. This fifth dataset was also evaluated by 53 international endoscopist assessors to benchmark CAD performance.

Pre-training

Generally, deep learning networks are pre-trained on computer-vision datasets, such as ImageNet¹⁸. Such datasets contain a large number of hand-labelled images of random objects such as boats, cars, and dogs, that enable deep learning networks to learn basic discriminative image features¹⁹. The acquired knowledge from pre-training are stored in so called “weights”. These weights are subsequently carried forward to the next level of refinement training.

On the basis of our rationale that an endoscopic CAD system would gain more discriminative knowledge from a specific, endoscopy-driven pre-training, dataset 1 was developed. This dataset consisted of 494,364 labelled endoscopic images taken from the esophagus, stomach, small intestine and colon of 15,286 patients (Table 1). These images were collected retrospectively from the endoscopic database in the Amsterdam University Medical Centers. Images were automatically de-identified, resized and stored directly after extraction from the database resized at a resolution of 256x256 pixels. All images were recorded with Olympus HQ190, HQ180, and HQ290 endoscopes, and FUJIFILM 700 series endoscopes. Subsequently, each individual image was labeled by organ system, using a separate computer algorithm we developed specifically for this project. Details of this separate semi-supervised labelling algorithm are described in the supplementary materials. Given its potential to pre-train deep learning algorithms of various endoscopic CAD systems, we refer to this dataset as *GastroNet*. The weights obtained from *GastroNet* were further refined by the subsequent datasets described below, using both transfer learning- and ensemble learning techniques.

To enable a first evaluation whether pre-training with *GastroNet* resulted in improved outcomes, we did multiple experiments in which we compared our CAD system performance when pre-trained with *GastroNet*, versus performance of our CAD system when pre-trained with ImageNet, versus no pre-training¹⁹. In these so-called *ablation experiments*, we furthermore assessed the additive value of multi-stage transfer learning. Given the importance of these analyses and their potential impact on development of deep learning architectures in endoscopy, these experiments are reported as secondary outcome measures in Results, but details of the methodology of these experiments are described in the Supplementary Materials, due to word count constraints.

Refinement training

Dataset 2, used for refinement training of the CAD system, consisted of 1247 retrospectively collected WLE images from 414 BE patients (557 NDBE images and 690 early neoplasia images, Table 1), obtained with Olympus HQ190, HQ180, and HQ260 endoscopes. This dataset was collected in the Amsterdam UMC. The NDBE images were reviewed by two experts (AdG, MS) for absence of visible neoplasia. The neoplastic images in this dataset were delineated independently by the same experts, using an online software module specifically designed for this project. This software allowed endoscopic images to be delineated on a computer screen, mapping the assessed extent of the neoplasia on the endoscopic image, and subsequently allowed for extraction of the delineations as black-and white masks to serve as training input for the CAD system. After review and delineation, all images were resized and stored at a resolution of 256x256 pixels, to enable input into the CAD system.

Training and internal validation

Dataset 3 contained 129 images of early BE neoplasia derived from 54 patients, and 168 images from 41 NDBE patients (Table 1), that were prospectively collected at 2 tertiary referral centers (Amsterdam UMC and Catharina Hospital Eindhoven) by a standardized image acquisition protocol, as described in previous publications by our group^{16, 18}. All of these HD-WLE images were gathered with the ELUXEO™ 7000 endoscopy system (FUJIFILM, Tokyo, Japan). All lesions on the neoplastic images were resected by either endoscopic mucosal resection or endoscopic submucosal dissection, and were shown to contain early neoplasia (defined as either HGD or EAC). All NDBE images were derived from patients who had no dysplasia in any of their previous endoscopies *and* no dysplasia in any of the random biopsies of the endoscopic procedure in which the imagery used for our study was obtained. The NDBE images were reviewed by two experts (AdG, WC) for absence of visible neoplasia. The neoplastic images were delineated by three other experts (OP, BW, RP), using the online software module described for “Refinement training”. Images were resized and stored at a resolution of 256x256 pixels. The CAD system was internally validated on this dataset, using 4-fold cross validation. In this validation approach, the dataset is split into 4 equal parts, after which one part is used as a validation set, while the other 3 parts are used for training. This was repeated 4 times, each time with a different validation set, after which the scores of the 4 experiments were pooled into a point estimate, to reduce any influence from variations of the mix of image contents among the 4 parts of the dataset.

Double external validation

The CAD system was then tested on 2 separately gathered external image datasets, to test for generalizability of CAD results to new, unseen, data.

Dataset 4 was collected retrospectively from the Amsterdam UMC, Catharina Hospital Eindhoven, University Hospital Leuven, University Hospital Nantes. It contained 40 images of early neoplastic lesions and 40 NDBE images visualized with the ELUXEO endoscopy system in 80 patients (Table 1). All images had corresponding histopathology reports. The neoplastic cases were selected purely on the basis of their histopathology and image quality, thereby creating a heterogeneous dataset in terms of lesion subtlety. These were considered representative of the variability of early lesions presenting in a general endoscopic setting. None of these images were included in any other dataset. Lesions on the neoplastic images were independently delineated by 3 experts (RP, EC, JB), using the software module described above. All NDBE images were reviewed by two experts (AdG, JB) for absence of visible abnormalities on the images.

Dataset 5 provided material for the second external validation (Table 1). These images were collected prospectively for another project, with the same protocol and endoscope used to gather dataset 3 and 4. This dataset also contained 40 images of early BE neoplasia and 40 NDBE images from 80 patients. In contrast to dataset 4 however, this dataset mostly contained subtle neoplastic lesions, representing a greater detection and delineation challenge than those in dataset 4. None of the dataset 5 images were included in any of the other datasets and all had correlating pathology reports. The NDBE images were reviewed by 2 experts (AdG, WC). Lesions on the neoplastic images were independently delineated by 6 experts (JB, RB, OP, ES, AM, HN), with the software described above.

Ground truth development

All images in dataset 3, 4, and 5 had a corresponding histopathological diagnosis of either early neoplasia or NDBE. The independent delineations of neoplastic lesions in datasets 2, 3, 4 and 5 by multiple BE expert contributors to this study developed a ground truth for both training and validation of the CAD system. The entire area delineated as being neoplastic by at least one expert, was labelled as the *soft spot*. This area was used to train the CAD system to recognize early BE neoplasia. The *soft spot* was also used to assess the CAD system's segmentation performance (i.e. the ability to *localize* neoplastic lesions in images) in both the internal validation (dataset 3) and external validation (datasets 4 and 5). The area of neoplastic change that the majority of experts (i.e. ≥ 2 experts in datasets 2, 3 and 4, and ≥ 4 experts in

dataset 5) delineated, was labelled as the *sweet spot*. This area was used to further assess segmentation performance during both internal- and external validation.

Model description

The CAD system was constructed using a custom-made, fully residual, hybrid ResNet/U-Net model. ResNet models are commonly used to *classify* images, while U-Net models are used for segmentation predictions within an image. Generally, these two models are trained separately. In our custom-made approach, however, classification and segmentation tasks were trained jointly. This was achieved by simultaneously optimizing the model using multi-task learning. The binary cross entropy loss was calculated at the end of the encoding branch and a weighted dice loss was calculated at the end of the decoder. The model weights associated with their respective losses were updated during a single backward pass through the network. This allowed pre-training on classification with GastroNet, before increasing complexity of the model by training with segmentation masks in datasets two and three. Our model used fully-residual skip connections instead of the standard concatenations²⁰. This approach reduces the amount of tunable parameters and thereby decreases the risk of overfitting, a frequently occurring problem in machine learning. To further decrease overfitting, batch normalization and data augmentation were applied.

The CAD algorithm encompassed five separate pseudo-models, which were first trained using different subsets of the GastroNet dataset. These five pseudo models were further fine-tuned with five bootstraps of data set 2, resulting in 25 different pre-trained models²¹. These 25 models underwent a final fine-tuning phase with data set 3 using ensemble learning, an effective method to reduce the variance of the output predictions of multiple models. The 10 best performing models (as evaluated on the internal validation data set 3) were selected, and subsequently the combined, pooled input of these 10 models was used to produce the final predictions in the external validation sets. An overview of the model is shown in Figure 2. An extensive, technical description of the model can be found in the supplementary materials and Figure 5.

All experiments were performed on a desktop PC with the following specifications: an Intel Xeon CPU E5-1650 v4 @3.60 GHz, 32 GB RAM, NVIDIA Titan Xp 12GB GPU.

Benchmark performance by non-expert endoscopist assessors

To relate the performance of the CAD system to that of endoscopists, all 80 images from dataset 5 were also evaluated by 53 international endoscopist assessors, none of whom was a BE endoscopic expert.

Assessors were classified into 4 groups (Table 2): 'Novice' (GI research fellows with no endoscopic expertise), 'fellow' (endoscopy fellows in a GI training program), 'junior' (board certified endoscopists with <3 years of endoscopic experience), and 'senior' (board certified endoscopists with >5 years of endoscopic experience). Assessors were recruited from the Netherlands, Belgium, France, and Sweden. An online scoring- and delineation module was developed specifically for this assessment. In this module, assessors were able to view all 80 images which were delivered via the internet with no loss of image quality. Assessors first classified an image to be neoplastic or NDBE. When a neoplastic lesion was diagnosed, assessors then had to indicate their preferred biopsy location, and delineated the borders of the neoplastic lesion. Assessors were required to complete their evaluations of all cases in a single session. Interpretations of images were locked after being scored to prevent assessors going back to revise their earlier evaluations. The order in which images were presented for evaluation was randomized among assessors.

Outcome measurements

Primary outcome measurements

- Classification scores in the internal validation set (dataset 3), in terms of diagnostic accuracy, using 4-fold cross-validation methodology. Classification was considered correct when the CAD system correctly labelled an image as neoplastic or NDBE.
- Classification scores in the two external validation sets (datasets 4 and 5), in terms of diagnostic accuracy, based on per-image analyses.

Secondary outcome measurements

- Segmentation performance: Delineation scores for the internal- and external validation (datasets 3, 4 and 5) were defined as the proportion of neoplastic images in which the CAD's delineation overlapped the ground truth (i.e. the experts' soft- and sweet spot);
- Segmentation performance: Biopsy site scores for the internal- and external validation (datasets 3, 4 and 5) were defined as the proportion of neoplastic images where the CAD system positioned the preferred biopsy location in the experts' ground truth;

- Benchmarking to general endoscopists: The CAD system's performance was compared with that of the 53 general endoscopist assessors in assessing dataset 5, in terms of diagnostic accuracy of classification, and segmentation performance (i.e. delineation scores and biopsy site scores);
- Time required for analyzing an image in dataset 5 by the CAD system and endoscopist assessors.

Post-hoc analysis

To assess the benefit of pre-training with GastroNet and multi-stage transfer learning, we evaluated the classification performance of the CAD system on the 2 external validation databases when: 1) the CAD system received no pre-training nor multi-stage transfer learning; 2) the CAD system received no pre-training, yet was trained with multi-stage transfer learning; 3) the CAD system was pre-trained by ImageNet and multi-stage transfer learning; and 4) the CAD system was pre-trained by GastroNet and multi-stage transfer learning. Extensive details of these experiments are listed in the Supplementary Materials.

Statistical analysis

The CAD system was developed and evaluated in the PyTorch framework and statistical tests were performed with Software package Matlab 2018a (Mathworks, Inc., Natick, Massachusetts, USA), and with SPSS Statistical software package for Windows (version 25, SPSS Inc., Chicago). Since this was the first study in its kind which developed and validated a deep learning CAD system for early BE neoplasia, no formal sample size calculation was made. Diagnostic accuracy per image was displayed in terms of accuracy, sensitivity, and specificity. To evaluate the operating points and establish optimal system accuracy, the F1 score was computed as the harmonic mean between the sensitivity and specificity²².

RESULTS

Internal validation on dataset 3

After training, the CAD system was first validated internally on dataset 3. The rationale for internal validation is to optimize (hyper-)parameters and gauge system performance. Four-fold per-patient cross validation was performed to ensure that images were not used for both training and testing. The results for internal validation are summarized in Table 2.

Classification performance

The CAD system correctly classified 113/129 images of early neoplasia, and 149/168 images of NDBE. Therefore, the primary outcome measures of the accuracy, sensitivity, and specificity for the classification score were 88%, 88% and 89%, respectively. Misclassified images are considered in the Discussion. Based on these results, the optimal cut-off threshold for the internal validation set was based on the F1 score and established to be at $T=0.6$. An image was therefore labeled as “neoplastic” when the classification score was ≥ 0.6 .

Segmentation performance

The CAD system only enables the secondary outcome measures of segmentation predictions in images that are assessed as containing neoplasia. In all 113 images where the CAD system correctly classified the image as “neoplastic”, the CAD delineation overlapped with the experts’ ground truth, resulting in a delineation score of 88% (113/129). In 83% (107/129) cases, the CAD delineation also recognized the area *the majority of* experts assessed to be neoplastic (*the sweet spot*). When only evaluating the correctly classified images, and thus purely focusing on the segmentation functionality of the system, the delineation scores were 100% (113/113) and 95% (107/113) for soft spot and sweet spot recognition, respectively.

The biopsy site score, indicating the preferred biopsy location, was positioned in the experts’ soft spot in 102/129 images (79%), and in 92/129 images (71%) in the sweet spot. When focusing on the 113 correctly classified neoplastic images, the biopsy site scores were 90% and 81% for soft- and sweet spot, respectively.

External validation on dataset 4

After internal validation, the CAD system was validated externally on images from 80 new patients that were not included in dataset 1-3. This external validation was executed without any alteration of the

CAD system settings, to enable true validation. The results for external validation are summarized in Table 2.

Classification performance

The CAD system correctly classified 36/40 images that contained neoplasia, and 35/40 of the NDBE images. The primary outcome measures of accuracy, sensitivity, and specificity for the classification score were 89%, 90%, and 88%, respectively (Table 2).

Segmentation performance

In all images correctly classified as containing neoplasia, the CAD delineation overlapped with both the expert soft- and sweet spots. The segmentation functionality of the CAD system was therefore 100%. Taking into account the 4 undetected lesions, the delineation score was 90%. The CAD system indicated the preferred biopsy site within the soft spot in all detected cases (100%), and the sweet spot in 35/36 cases (97%). Focusing on the undetected lesions, the biopsy site scores were 90% and 88%, respectively (Table 2).

External validation on dataset 5

The CAD system was then validated externally on images from 80 new patients not included in dataset 1-4, that represented the more challenging cases during surveillance endoscopies.

Classification performance

The CAD system correctly classified 37/40 of the images that contained neoplasia, and 33/40 of the NDBE images. The primary outcome measures of accuracy, sensitivity, and specificity for the classification score were 88%, 93%, and 83%, respectively (Table 2).

Segmentation performance

For the 36 correctly classified neoplastic images, the CAD delineation overlapped with the experts' soft- and sweet spots (100%). Taking into account the *undetected* lesions, the delineation score was therefore 93% (37/40).

The biopsy site score in the external validation set was 90% (36/40). In 85% (34/40), the preferred biopsy location was positioned in the sweet spot. Focusing on the segmentation functionality, the biopsy site scores were 97% and 92% for soft- and sweet spot recognition, respectively.

Benchmarking of CAD performance to that of endoscopist assessors

Dataset 5 was also evaluated by 53 international endoscopist assessors, from the Netherlands (30),

Belgium (10), France (7), and Sweden (6). This group consisted of 10 novices, 18 fellows, 8 junior endoscopists and 17 senior endoscopists (Table 2).

Classification performance

The primary outcome measure of mean scores for classification accuracy, sensitivity, and specificity for all 53 general endoscopists were 73%, 72%, and 74%, respectively.

Mean scores for assessors in the 4 groups of endoscopic experience are given in Table 2. These were all substantially inferior to the CAD system scores. The plot of the classification scores of the CAD system and the individual general endoscopists in Figure 3 shows that the CAD system achieved higher accuracy than any of the individual endoscopists in terms of accuracy, and all but one endoscopist in terms of sensitivity: the specificity and accuracy of this assessor were however low at 35% and 68%, respectively.

Segmentation performance

In nearly all correctly classified neoplastic images, assessors' delineations overlapped with the expert ground truth (98% for both soft- and sweet spot delineation scores). Taking into account the undetected lesions, the mean assessors' delineation score was 71%. Furthermore, when focusing on segmentation ability, assessors located their preferred biopsy location within the soft- and sweet spot in 96% and 94%, respectively.

Computation time

The mean time for classification per image in dataset 5 was 0.111 seconds (SD 0.011). The mean time for segmentation (e.g. analyzation and producing delineation and red-flag indicator) was 0.124 seconds (SD 0.007). The mean time per image for both classification and delineation (if applicable) by assessors was 46.3 seconds (SD 135).

Additive value of pre-training with GastroNet and multi-stage transfer learning

Results listed in Supplementary Table 1 show that the classification performance of the CAD system improved for datasets 4 and 5 when pre-training had been used. In both dataset 4 and 5, the CAD system performed better when it was pre-trained with GastroNet, compared to pre-training with ImageNet. This was most prominent in dataset 5. Classification performance increased when multi-stage transfer learning was used.

DISCUSSION

This paper describes the rigorous development, validation, and benchmarking of a novel deep learning computer aided detection system that has been shown to achieve a higher accuracy than non-expert general endoscopists in detection of early BE neoplasia.

The rapid rise of computer aided detection in endoscopy offers great opportunities to overcome endoscopist-dependent limitations of neoplasia detection. The exponential increase in computational power of modern-day computers, and increasing amounts of available data, now enable deep learning techniques which offer a promising alternative for conventional machine learning methods⁷. The potential of deep learning in endoscopy has recently increased even further by the introduction of transfer learning, a technique that re-uses networks trained on large datasets from other domains, such as ImageNet^{7, 19}. Deep learning techniques have already shown promising results in endoscopy, mostly in the field of colonic polyps^{8-15, 23}. Studies evaluating CAD systems for BE neoplasia are however scarce^{16, 17, 24-28}.

The external validation of our CAD system with dataset 4 tested it with a relatively wide variation of lesion subtlety covered the natural variability of early lesions likely to be encountered by a general endoscopist. The high diagnostic accuracy that the system achieved suggests that the CAD system should perform well in a general clinical setting, outside major referral centers that specialize in BE assessment.

The CAD system was subsequently tested on dataset 5 which mainly contained subtle lesions, representative of diagnostically more challenging surveillance endoscopy cases. In our opinion, this is the type of lesion for which our BE CAD system should have add greatest value.

The benchmarking of the CAD system with general endoscopists provides an important calibration. This showed that the CAD system was considerably more accurate than the non-expert endoscopist assessors, for all 4 levels of expertise, with a difference in sensitivity of approximately 20%. These results suggest that our CAD system may improve the endoscopic detection of early BE neoplasia by non-expert endoscopists. It is important to note though, that the 28 fellows and novices within the group of 53 non-expert endoscopic assessors should be regarded as either insufficiently trained or virtually untrained in the recognition of early BE neoplasms. This possible reservation is allayed by the still relatively high failure rate of the 25 more senior endoscopists.

The 3 neoplastic images that were incorrectly labelled as NDBE by the CAD system in dataset 5, contained flat and subtle lesions, as shown in Figure 4. In 51%, non-expert endoscopist assessors also incorrectly classified these images as NDBE. Further training of the CAD system with even larger datasets should decrease the rate of missed BE neoplasm by the system. The CAD system incorrectly classified 7 of the 40 NDBE images (18%) as containing neoplasm. In clinical practice, the consequences of false-positive detections will be relatively limited, since these may be adequately down-staged by subsequent detailed endoscopic inspection (*e.g.* by incorporation of optical chromoscopy techniques) and by targeted biopsies.

The computational speed of 0.24 seconds for classification and delineation of the endoscopic images is relatively slow for deep learning systems, but still fast enough for the system to be used in real time during endoscopic surveillance. This “slow-speed” is mainly due to the limited number of graphics processing units (GPUs) in the computer we used for our experiments. A dedicated software implementation on more powerful GPUs will certainly reduce processing time enough to enable real-time processing when working towards video applications.

Our study has 6 major strengths. Firstly, it used five well-defined independent datasets to train and validate our CAD system. The size of these datasets is unprecedented in CAD applications for BE neoplasia: in total, 1704 images of early BE neoplasia and NDBE, derived from 4 different centers and delineated by >10 experts, were available for training and testing. The heterogeneity and size of our datasets contribute to the robustness of our CAD system. We could have easily tripled the number of images included in the external datasets ($n=160$), yet we insisted on including only one image per patient to prevent overfitting results at patient-level. In comparison, our previous publications used only a total of 60 and 100 images for both training and testing^{16 17}. Other publications on CAD systems for BE neoplasia are also hampered by small datasets, never using more than 250 images. Also, these studies have used the same database for training and validation, making external testing of the CAD systems impossible²⁷.

Secondly, the availability of histological correlations in datasets 3-5, and the use of a representative panel of ≥ 10 BE experts to delineate neoplasia in datasets 2-5, enabled robust development of the ground truth for all 899 images that contained neoplasia. This ensured high-quality training input for the CAD system. Based on previous experience, we decided that a minimum of 2 experts was needed to establish ground truth⁶. For neoplastic lesions, also at least 2 experts were required to enable calculation of the sweet spot.

Thirdly, state-of-the art machine learning techniques, such as transfer learning and ensemble learning, were used to train the system. Transfer learning is a widely used method in deep learning to prevent overfitting on small data sets and to exploit the power of deep neural networks for better training on image interpretation with scarce training data²⁹. In this study, the CAD system was first presented with an abundance of data, and then trained with more computational budget. For this, we used a unique custom-made, fully residual, hybrid ResNet-UNet model. The use of this multi-task model increases training efficiency, by using expert delineations not only for segmentation performance but also to increase classification performance, and simultaneously decreases the risk of overfitting due to better regularization.

Fourthly, we used a new approach for pre-training our deep learning algorithm. Pre-training with large numbers of images of random objects such as boats, cars, and dogs is a widely used concept in deep learning to learn the system basic discriminative image features¹⁹. We developed GastroNet in the hope that a pre-training database consisting purely of gastrointestinal endoscopic images would enhance the training process. The training superiority conferred by GastroNet was greatest in dataset 5, which contained the more challenging neoplasia cases. These post-hoc analyses of the value of GastroNet we report in the current paper are only the first experiments in the evaluation of GastroNet. We are currently working on expanding the number of images in GastroNet from 494,364 to 5,000,000. Although the concept of endoscopy driven pre-training appears logical and the results of these post-hoc analyses results indicate an additive value of GastroNet, at this stage, we do not take this as definitive proof. Direct comparison of different forms of pre-training is technically challenging, and prone to bias, as described in the Supplementary Materials. Ongoing studies should provide better data on the value of GastroNet in deep learning algorithms in endoscopy.

Fifthly, this is the first study to validate a Barrett CAD system on an external dataset of images, which had not been seen previously by the algorithm. The use of two independent datasets increased the heterogeneity of the images and demonstrated the robustness of our CAD system's performance.

Finally, our study is the first to benchmark the performance of our primary CAD system for BE against the performance of a large group of general endoscopist assessors. These were grouped by years of endoscopic expertise, based on previous studies by our group⁶. We acknowledge that this might not directly relate to familiarity with Barrett's neoplasia. However, the CAD system achieved higher accuracy than any of the endoscopists, regardless of level of endoscopic expertise. Delineation performance of the

CAD system was comparable to that of the endoscopists. Although we have not formally assessed any statistical differences, these data provide a useful calibration of the performance of our CAD system.

Our study has possible limitations to its generalizability.

Firstly, the images in datasets 3-5 were generally of high quality and were collected by expert endoscopists. This may have resulted in selection bias. In a less specialized setting, imagery could be of lower quality, e.g. by the presence of blur or water and mucus on the lens, and may reduce the performance of the CAD system.

Secondly, although the pre-training and refinement training was performed with a variety of endoscopic imagery, the CAD system was only specifically trained and validated on FUJIFILM ELUXEO imagery.

Thirdly, our CAD system is specifically designed to detect early BE neoplasia in overview with white light endoscopy, without the use of magnification or optical chromoscopy techniques. In our opinion, these techniques are more useful for detailed *characterization* of lesions after their primary detection with HD-WLE in overview³⁰. A dedicated computer aided diagnosis (CADx) system for detailed characterization after primary detection is under development by our group³¹. This CADx system could potentially be used for optical biopsy, by the use of optical chromoscopy in magnification and advise endoscopists on the necessity of obtaining biopsies.

Fourthly, our CAD system is still image-based whereas in general video-based CAD systems are considered state-of-the-art. The rationales for the development of a still image-based CAD system are that an effective image-based CAD system is a crucial step towards the development of a video-based CAD system, since such a system will analyze videos frame-by-frame, resembling a still image-based approach operating at high processing speed. Frame-by-frame video imagery, however, contains additional spatial and temporal information, which may lead to superior performance over single-image based CAD. However, a still image-based CAD system with an image processing time as short as our CAD system may be easier to use in a real-time clinical setting. An average Barrett segment can be visualized in approximately 5 overview images, and the acquisition of high quality images is generally easier than high quality videos, e.g. due to esophageal movement caused by motility, breathing and patient movement. Endoscopists can obtain still images of the Barrett segment in overview every 1-2 centimeter, which are then immediately analyzed by the image-based CAD system and depicted in real-time. A pilot study is underway interrogating this functionality, while our group continues its work towards a video-based CAD system.

In conclusion, we report the development of the first externally validated deep learning CAD system for early BE neoplasia. The system detected neoplasia with high accuracy and near-perfect localization. The CAD system achieved higher accuracy than a panel of non-expert endoscopist assessors, strongly suggesting that CAD may improve the accuracy of surveillance-detected early BE neoplasia by general endoscopists. The use of a pre-training dataset of a large number of gastrointestinal endoscopic images for deep learning in endoscopy is the first of its kind. We anticipate that the system's performance will further improve by expanding the number of images in the different database layers of our deep learning system and by including video recordings.

REFERENCES

1. Spechler SJ, Sharma P, Souza RF, et al. American Gastroenterological Association medical position statement on the management of Barrett's esophagus. *Gastroenterology* 2011;140:1084-91.
2. Weusten B, Bisschops R, Coron E, et al. Endoscopic management of Barrett's esophagus: European Society of Gastrointestinal Endoscopy (ESGE) Position Statement. *Endoscopy* 2017;49:191-198.
3. Pech O, May A, Manner H, et al. Long-term efficacy and safety of endoscopic resection for patients with mucosal adenocarcinoma of the esophagus. *Gastroenterology* 2014;146:652-660.e1.
4. Scholvinck DW, van der Meulen K, Bergman JJ, et al. Detection of lesions in dysplastic Barrett's esophagus by community and expert endoscopists. *Endoscopy* 2017;49:113-120.
5. Hvid-Jensen F, Pedersen L, Drewes AM, et al. Incidence of adenocarcinoma among patients with Barrett's esophagus. *N Engl J Med* 2011;365:1375-83.
6. Bergman J, de Groof AJ, Pech O, et al. An Interactive Web-Based Educational Tool Improves Detection and Delineation of Barrett's Esophagus-Related Neoplasia. *Gastroenterology* 2019;156:1299-1308.e3.
7. van der Sommen F, Curvers WL, Nagengast WB. Novel Developments in Endoscopic Mucosal Imaging. *Gastroenterology* 2018;154:1876-1886.
8. Iakovidis DK, Maroulis DE, Karkanis SA. An intelligent system for automatic detection of gastrointestinal adenomas in video endoscopy. *Comput Biol Med* 2006;36:1084-103.
9. Maroulis DE, Iakovidis DK, Karkanis SA, et al. CoLD: a versatile detection system for colorectal lesions in endoscopy video-frames. *Comput Methods Programs Biomed* 2003;70:151-66.
10. Takemura Y, Yoshida S, Tanaka S, et al. Computer-aided system for predicting the histology of colorectal tumors by using narrow-band imaging magnifying colonoscopy (with video). *Gastrointest Endosc* 2012;75:179-85.
11. Kominami Y, Yoshida S, Tanaka S, et al. Computer-aided diagnosis of colorectal polyp histology by using a real-time image recognition system and narrow-band imaging magnifying colonoscopy. *Gastrointest Endosc* 2016;83:643-9.
12. Chen PJ, Lin MC, Lai MJ, et al. Accurate Classification of Diminutive Colorectal Polyps Using Computer-Aided Analysis. *Gastroenterology* 2018;154:568-575.
13. Tischendorf JJ, Gross S, Winograd R, et al. Computer-aided classification of colorectal polyps based on vascular patterns: a pilot study. *Endoscopy* 2010;42:203-7.
14. Misawa M, Kudo SE, Mori Y, et al. Characterization of Colorectal Lesions Using a Computer-Aided Diagnostic System for Narrow-Band Imaging Endocytoscopy. *Gastroenterology* 2016;150:1531-1532.e3.
15. Urban G, Tripathi P, Alkayali T, et al. Deep Learning Localizes and Identifies Polyps in Real Time With 96% Accuracy in Screening Colonoscopy. *Gastroenterology* 2018;155:1069-1078.e8.
16. Groof Jd, van der Sommen F, van der Putten J, et al. The Argos project: The development of a computer-aided detection system to improve detection of Barrett's neoplasia on white light endoscopy. *United European Gastroenterology Journal* 2019;2050640619837443.
17. van der Sommen F, Zinger S, Curvers WL, et al. Computer-aided detection of early neoplastic lesions in Barrett's esophagus. *Endoscopy* 2016;48:617-24.
18. de Groof AJ, Swager A, Pouw RE, et al. Blue-light imaging has an additional value to white-light endoscopy in visualization of early Barrett's neoplasia. an international multicenter cohort study. *Gastrointest Endosc* 2018.

19. Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database. IEEE Conf Comput Vis Pattern Recognit 2009 2–9.
20. Shelhamer E, Long J, Darrell T. Fully Convolutional Networks for Semantic Segmentation. IEEE Trans Pattern Anal Mach Intell 2017;39:640-651.
21. Mooney. Bootstrapping: A nonparametric approach to statistical inference In: D DR, ed. Newbury Park, CA: Sage University Paper series on Quantative applicatinos in the social sciences, 1993.
22. Powers D. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation, 2008.
23. Byrne MF, Chapados N, Soudan F, 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;68:94-100.
24. Mendel R. EA, Probst A., Messmann H., Palm C. Barrett's Esophagus Analysis Using Convolutional Neural Networks: Springer Vieweg, Berlin, Heidelberg, 2017.
25. Ebigbo A, Mendel R, Probst A, et al. Computer-aided diagnosis using deep learning in the evaluation of early oesophageal adenocarcinoma. Gut 2018.
26. Ghatwary N, Zolgharni M, Ye X. Early esophageal adenocarcinoma detection using deep learning methods. Int J Comput Assist Radiol Surg 2019;14:611-621.
27. de Souza LA, Jr., Palm C, Mendel R, et al. A survey on Barrett's esophagus analysis using machine learning. Comput Biol Med 2018;96:203-213.
28. Ghatwary N, Ahmed A, Ye X. Automated Detection of Barrett's Esophagus Using Endoscopic Images: A Survey, Cham, Springer International Publishing, 2017.
29. SJ P. A survey on transfer learning. In: Q Y, ed: IEEE Transactions on knowledge and data engineering, 2010:1345-59.
30. Boerwinkel DF, Swager A, Curvers WL, et al. The clinical consequences of advanced imaging techniques in Barrett's esophagus. Gastroenterology 2014;146:622-629 e4.
31. Struyvenberg MR, de Groof J, van der Putten J, et al. Deep Learning Algorithm for Characterization of Barrett's Neoplasia Demonstrates High Accuracy on Nbi-Zoom Images. Gastroenterology 2019;156:S-58.

Figure 1. Example of two neoplastic lesions (A) with the “heatmap” visualization by the CAD system (B) with its corresponding delineation and biopsy site indicator (C). Ground truth is established by expert delineations (D).

Figure 2. Visualization of CAD model.

Figure 3. Classification performance of all individual endoscopist assessors and the CAD system on dataset 5.

Figure 4. False negative cases in external validation set (dataset 5).

Figure 5. Graphical overview of neural network design, visualizing both classification mode (ResNet inspired; top) and segmentation mode (U-Net inspired; bottom).

TABLES

Table 1. Overview of datasets used for training and validation of the CAD system

	Number of images	Number of patients	Data acquisition	Endoscopy system	Type of labelling	Purpose of dataset
Dataset 1 (GastroNet)	494,364	15,286	Amsterdam UMC; retrospective acquisition	Olympus HQ190, HQ180, HQ260, FUJIFILM 700 series	Hand-labelling of subset by experts, followed by implementation of pseudo- labelling software	Pre-training
Dataset 2	1247	414	Amsterdam UMC; retrospective acquisition	Olympus HQ190, HQ180, HQ260		Refinement training
Neoplastic images	690	198			Delineated by 2 experts	
NDBE images	557	216			Hand-labelled by 2 experts	
Dataset 3	297	95	Amsterdam UMC, Catharina Hospital Eindhoven; prospective acquisition	FUJIFILM 700 series		Refinement training and internal validation
Neoplastic images	129	54			Delineated by 3 experts + Correlating pathology	
NDBE images	168	41			Hand-labelled by 2 experts + Correlating pathology	
Dataset 4	80	80	Amsterdam UMC, Catharina Hospital Eindhoven, University Hospital Leuven, University Hospital Nantes; retrospective acquisition	FUJIFILM 700 series		External validation
Neoplastic images	40	40			Delineated by 3 experts + Correlating pathology	
NDBE images	40	40			Hand-labelled by 2 experts + Correlating pathology	
Dataset 5	80	80	Amsterdam UMC, Catharina Hospital Eindhoven, University Hospital Leuven; prospective acquisition	FUJIFILM 700 series		External validation
Neoplastic images	40	40			Delineated by 6 experts + Correlating pathology	
NDBE images	40	40			Hand-labelled by 2 experts + Correlating pathology	

Table 2. CAD system performance in internal- and external validation, and endoscopist assessor performance in external validation dataset 5.

		Classification Performance				Segmentation Performance			
		Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score	Delineation score soft spot (%)	Delineation score sweet spot (%)	Biopsy site score soft spot (%)	Biopsy site score sweet spot (%)
CAD system	Internal validation Dataset 3	88.2 (262/297)	87.6 (113/129)	88.6 (149/168)	88.4	100 (113/113)	94.6 (107/113)	90.3 (102/113)	81.4 (92/113)
	External validation Dataset 4	88.8 (71/80)	90.0 (36/40)	87.5 (35/40)	88.7	100 (36/36)	100 (36/36)	100 (36/36)	97.2 (35/36)
	External validation Dataset 5	87.5 (70/80)	92.5 (37/40)	82.5 (33/40)	88.1	100 (37/37)	100 (37/37)	97.3 (36/37)	91.9 (34/37)
		Accuracy (% , 95% CI)	Sensitivity (% , 95% CI)	Specificity (% , 95% CI)		Delineation score soft spot (% , 95% CI)	Delineation score sweet spot (% , 95% CI)	Biopsy site score soft spot (% , 95% CI)	Biopsy site score sweet spot (% , 95% CI)
Benchmark assessment dataset 5 by endoscopist assessors	All assessors (n=53)	73.0 (72-74)	71.8 (68-75)	74.3 (70-78)		98.2 (97-99)	97.7 (97-99)	96.2 (94-99)	94.3 (91-98)
	Senior endoscopists (n=17)	74.8 (72-77)	76.5 (72-81)	73.1 (66-80)		98.9 (97-100)	98.6 (97-100)	98.6 (97-100)	97.8 (96-100)
	Junior endoscopists (n=8)	77.7 (74-81)	79.1 (73-85)	76.3 (66-86)		99.7 (99-100)	99.3 (98-100)	99.3 (98-100)	97.8 (96-100)
	Fellows (n=18)	73.3 (71-76)	70.6 (64-77)	76.1 (71-81)		97.9 (96-100)	97.4 (96-99)	96.3 (94-99)	93.9 (90-97)
	Novices (n=10)	65.9 (63-69)	60.3 (50-70)	71.5 (54-89)		96.6 (93-100)	95.3 (91-100)	90.0 (76-100)	86.4 (69-100)

5 classes - stomach, esophagus,
colon, duodenum, other. (n=494,364)

Stage 1: pre-training

GastroNet architecture
Semi supervised training

prediction

Calculate
classification loss

2 classes with annotations for training
Dysplastic (n=690), non-dysplastic (n=557)

Stage 2: training

GastroNet architecture
Supervised training with segmentations

Copy trained weights

Prediction Delineation

Calculate
classification +
segmentation loss

2 classes with annotations for fine-tuning
Dysplastic (n=129), non-dysplastic (n=168)

Stage 3: fine-tuning

GastroNet architecture
Supervised training with segmentations

Copy trained weights

Prediction Delineation

Calculate
classification +
segmentation loss

2 classes with annotations for testing
in two separate test sets, each containing
Dysplastic (n=40), non-dysplastic (n=40)

Stage 4: testing

GastroNet architecture
Test on new images

Copy trained weights

Image class label prediction

Output delineation

SUPPLEMENTARY MATERIALS

Neural network design

The CAD system was constructed using a custom-made, fully residual, hybrid ResNet/U-Net model. The encoder-decoder network consists of an encoder section very similar to ResNet18. All convolutional layers have 3x3 filters with stride 1 save for the very first layer, which has 7x7 filters and stride 2. This operation effectively reduces the processing time by reducing the dimensionality of the input in the first operation. Standard ReLUs are used as the activation function in all layers. After the 2x2 average pooling operators, the number of filters is doubled to maintain computational complexity. Additionally, the base ResNet18 model is extended with a fully connected convolutional layer with 32 nodes. This layer is included to enable CNN-codes, which is where the features extracted from this fully connected layer are used to train traditional classification models such as support vector machines (SVM's) or Random Forests instead of end-to-end learning usually employed with deep learning. Deep learning usually requires specific hardware which is not always available for all research groups. With the addition of this fully connected layer, the use of deep learning by CNN-codes is still possible for researchers with limited computational resources. Connected to the above-mentioned fully connected layer, a standard classification layer with a sigmoid activation function is added.

The decoder section of the network has the same structure as the encoder (up until the fully connected layer). Learnable transposed convolutional layers are used to up sample the encoded features into a segmentation map. In the U-Net architecture, the output of the encoder features in the same depth level are concatenated with the up sampled feature map. For this network, a residual can be calculated from the encoder features and the decoder features as they have the same dimensions. Fully residual auto-encoders have the benefit of fewer parameters without a decrease in complexity which reduces the risk of overfitting. The output of the decoder is a probability map which is facilitated by a sigmoidal activation function.

Most CAD models are either classification or segmentation models. The addition of the classification layer in the bridge of the encoder-decoder network has two advantages: (1) Expert delineations can be used to increase the classification performance by using a dual loss. The primary loss is calculated at the end of the decoder using binary cross entropy. While a second segmentation loss can be calculated using a weighted DICE loss function. The final loss is calculated by taking the sum of the two losses. This approach solves a classification problem directly with guidance of the expert annotations instead of thresholding the predicted delineations of a segmentation problem to find classifications in post-

processing; (2) some datasets do not have expert annotations as these are generally laborious to obtain. With this network design, a mix of samples with segmentations and samples with only a label can be used. When training for a classification problem, such as during pre-training with the *GastroNet* dataset, the encoding branch is initialized with a set of useful features for classification. This initialization is based on a large variety of endoscopic images and can later be leveraged during training with the full network when segmentations are also available (dataset 2 and 3).

A graphical overview of the model is shown in Supplementary Figure 5.

Development of the GastroNet data set

The images from the *GastroNet* data set were extracted from the endoscopic database (ENDOBASE software, Olympus Europe, Hamburg, Germany) of the Amsterdam University Medical Centers and automatically de-identified directly after extraction from the database. For this we used a random coded search in ENDOBASE for the years 2012-2018, selecting gastrointestinal interventions. All labels and patient identifiers were automatically removed after extraction.

First, a subset of 3.743 images was created in which images were labeled manually by 2 experts (JG and MS). Labels consisted of ‘esophagus’ (n=1,305), ‘stomach’ (n=876), ‘duodenum’ (n=339), ‘colon’ (n=679), and ‘other’ (n=544). After training on this subset, the remaining 490,621 images were then automatically classified in these 5 categories by using a pseudo-labeling algorithm¹.

Then, a base classification model was trained from the smaller subset of supervised samples, after which pseudo-labels were predicted for the remaining unlabeled data. Second, the network is trained until convergence using all supervised + pseudo-labeled samples. Third, N random subsets are sampled from the pseudo-labeled set and a percentage p of the most certain samples are extracted. Finally, N separate models are fine-tuned with the different pseudo-labeled subset + labeled samples until convergence.

This approach results in N models which have all been trained with all data from the *GastroNet* data set. However, each model is still diverse as a different subset of the entire data set is subsequently used to fine-tune the models. In our experiments we chose N = 5 where each subset consisted of 50.000 randomly selected samples from the unlabeled images and probability $p = 0.1$.

Multi-stage transfer learning

Recently, transfer learning techniques have been applied successfully in many real-world applications². Transfer learning is particularly effective in domains where image and ground truth acquisition is

expensive, e.g. cancer detection. Many image classification algorithms that use transfer learning employ a network pre-trained on the ImageNet data set as a starting point for training^{3,4}. Using weights pre-trained with natural images such as those in the ImageNet data set leads to better results compared to training from scratch with small data sets. However, the main reason the natural image domain is often used is because of accessibility and 'cheap' labeling, and not because pre-training with natural images is optimal. Datasets which are closely related to the target domain are preferred for transfer learning when the order of images is comparable. In this work, *GastroNet* is employed as an alternative to ImageNet. Since the images in the *GastroNet* dataset resemble the target domain (endoscopic imagery of Barrett's esophagus) to a high extent, we hypothesized that the learned features are more representative for both domains. To further fine-tune the final algorithm, we employ two additional data sets (dataset 2 and 3). Both data sets contain WLE overview images of Barrett's neoplasia as well as non-dysplastic BE. However, the data is captured with different brands of endoscopes. After pre-training with *GastroNet*, the algorithm is first further fine-tuned with dataset 2 and subsequently with dataset 3. With each step, the total training time and learning rate is reduced since the network parameters get closer to the target domain with each step.

Post-hoc analyses of the value of pre-training with GastroNet and multi-stage transfer learning

To assess the individual contributions of both our *GastroNet*- and multi-stage transfer learning approach, an ablation study was performed. Ablation studies are typically performed to evaluate performance when specific components of the CAD system are removed ("ablated") or altered. In this ablation study, we hypothesized that pre-training with *GastroNet*, would be non-inferior to pre-training with ImageNet.

Ablation experiment description

For the ablation study, 4 additional experiments were performed.

In the first experiment, no pre-training and no multi-stage transfer learning is employed. The network was initialized with random weights and only trained with data set 3. Performance was then evaluated on dataset 4 and dataset 5.

In the second experiment, again no pre-training was used, however in this experiment dataset 2 and dataset 3 were sequentially used with the multi-stage transfer learning methodology. Performance was evaluated on dataset 4 and dataset 5.

In the third experiment, ImageNet was used to pre-train the network prior to using multi-stage transfer learning with dataset 2 and dataset 3. Again, performance was evaluated on dataset 4 and dataset 5.

In the fourth experiment, only the labeled portion of the GastroNet data set was used to pre-train the network. Subsequently, the multi-stage transfer learning methodology was employed using datasets 2 and 3. The classification performance was measured on data set 4 and 5 as well.

In the fifth experiment, the entirety of GastroNet was used with a semi-supervised algorithm to pre-train the network, followed by multi-stage transfer learning with dataset 2 and 3 and evaluation of performance on dataset 4 and dataset 5.

During all training steps for all experiments, the training procedure was terminated when the validation error stopped increasing, the models with the best performance on the validation sets were selected for training in the following stages. For our primary result, *GastroNet* was used to create an ensemble of models. To facilitate a fair comparison, 5-fold cross validation was employed with dataset 3. This resulted in 5 models trained on different subsets of dataset 3. An ensemble of these models was then used to generate the classification scores.

Ablation experiment results and discussion

An overview of the results of the ablation study is listed in Supplementary Table 1. It is clear that using multi-stage transfer learning increases the performance on the 2 test sets. This result is not surprising as approximately four times the amount of data is used during training with multi-stage transfer learning. Most CAD systems employ ImageNet to initialize the network parameters, to allow deep learning with relatively small datasets. Our ablation experiments show that pre-training with ImageNet in combination with multi-stage transfer learning leads to comparable results with our approach on data set 4 in terms of accuracy. However, on dataset 5 - which is considered more difficult due to the subtlety of lesions – our approach appears to outperform all other models, including the model pre-trained with ImageNet. The results from these ablation experiments indicate that pre-training with *GastroNet* has an important effect on the generalizability of the model, in particular when the network is pre-trained with the proposed semi-supervised learning approach. When pretraining with only the labeled GastroNet samples, a performance similar to ImageNet pre-training is obtained. However, much less data is used to attain these results. It should be noted that this is only a first experiment, performed to enable a first assessment of the additive value of *GastroNet*. Although results are promising, there are 2 potential limitations: First, *GastroNet* pre-training was specifically designed to create a diverse ensemble. The ensembles in the other experiments are created with a valid approach which is often used in literature. However, the difference in these results might therefore partly stem from the ensembling approach, instead of an effect that solely stems from methods of pre-training. As a counterpoint, the ensembling

method used is unique to how *GastroNet* was used and can therefore not be effectively employed for the other experiments. Second, the hyperparameters used for the ablation experiments were identical to the *GastroNet* experiments. While these hyperparameter values are a good approximation, they might not have been optimal for each individual ablation experiment. Since, in general, ablation studies become less meaningful when experiment details vary for each ablation experiment, we elected to equalize hyperparameters in our ablation experiments, to enable assessment of additive value solely of pre-training and multi-stage transfer learning.

These results therefore provide the first insights into pre-training with an alternative to ImageNet. A comprehensive study on the merits of *GastroNet* will be performed in future work. This study will contain evaluations on multiple gastrointestinal sets with multiple standard neural network architectures to assess the predictive power of gastro-specific pre-training compared to ImageNet pre-training.

Implementation details of the algorithm

The algorithm was trained using Adam and AMS-grad with a weight decay of 10^{-5} ⁵. A cyclic cosine learning rate scheduler was used to control the learning rate⁶. Additionally, batch normalization was used to regularize the network⁷. The model was further regularized with data augmentation. Images were randomly rotated with $\theta \in \{0, 90, 180, 270\}$ degrees and random flipped along the x- and y-axis with probability $p = 0.5$. Additionally, random permutations were made to the color, contrast and brightness of the images. Furthermore, images were randomly sheared by up to 8 degrees and randomly translated by up to 10% of image width. Lastly, random crops of various sizes were taken from the rescaled images without resizing to preserve feature size consistency.

The source code and weights of the trained network will be shared through an open online repository upon publication of this paper.

Assessor recognition

The authors are most grateful to the 53 volunteer endoscopist assessors listed in Supplementary Table 2 who evaluated dataset 5 and thereby provided benchmark results for the performance of the CAD system.

SUPPLEMENTARY TABLES*Supplementary Table 1. Ablation experiment results.*

Pre-training	Multi-stage transfer learning	Dataset 4			Dataset 5		
		Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)
No	No	67.5	82.5	52.5	62.5	97.5	27.5
No	Yes	77.5	75.0	80.0	68.8	70.0	67.5
ImageNet	Yes	85.0	82.5	87.5	72.5	82.5	62.5
GastroNet (labelled only)	Yes	86.3	90.0	82.5	75.0	80.0	70.0
GastroNet (all)	Yes	88.8	90.0	87.5	87.5	92.5	82.5

Supplementary Table 2. List of volunteer endoscopist assessors.

Clarisse Bohmer	Liselotte Zwager
Hanke Brandse	Andre Blomberg
Warner Bruins Slot	Georgios Katinios
Tim Grool	Ioannis Rouvelas
Joep Grootjans	Malin Ljunggren
Jeroen Jansen	Morteza Shafazand
Joep IJspeert	Tomasz Radomski
Sjoerd Kuiken	Robin Erken
Johan Kuyvenhoven	Bertrand Brieau
Selma Lekkerkerker	Marie-Anne Guillaumot
Petula Nijeboer	Einas Abouali
Sanne Rosekrans	Diane Lorenzo
Bouke Salden	Marion Chartier
Pieter Scholten	Clement Breteau
Meike van der Ree	Felix Corre
Brechje van Eijck	Hanneke Beaumont
Ellert van Soest	Sofie de Meulder
Wytske Westra	Alia Hadeffi
Tessa Uiterwaal	Ana-Maria Bucalau

Willemijn de Klaver	Filip van Herpe
Willemijn de Rooij	Filip Janssens
Myrte Gorris	Philip Caenepeel
Laura Haasnoot	Alexander van den Bulcke
Renske Oude Nijhuis	Klara Wyffels
Anne Snijder	Jao Sabino
Devica Umans	Alessia Settesoldi
Sara van Gennep	

REFERENCES SUPPLEMENTARY MATERIALS

1. Lee D-H. Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks, 2013.
2. Pan S, Yang Q. A Survey on Transfer Learning, 2010.
3. Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database. IEEE Conf Comput Vis Pattern Recognit 2009 2–9.
4. Urban G, Tripathi P, Alkayali T, et al. Deep Learning Localizes and Identifies Polyps in Real Time With 96% Accuracy in Screening Colonoscopy. Gastroenterology 2018;155:1069-1078.e8.
5. Kingma D, Ba J. Adam: A Method for Stochastic Optimization. In: International Conference on Learning Representations (ICLR). vol. 5 (2015), 2014.
6. Smith LN. Cyclical Learning Rates for Training Neural Networks, In 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), 24-31 March 2017, 2017.
7. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In: Francis B, David B, eds. Proceedings of the 32nd International Conference on Machine Learning. Volume 37. Proceedings of Machine Learning Research: PMLR, 2015:448–456.

Manuscript Number: GASTRO 19-01851

What you need to know:

BACKGROUND AND CONTEXT: A deep-learning computer-aided detection system might be used to improve endoscopic evaluation of patients with Barrett's esophagus (BE).

NEW FINDINGS: The authors developed, validated, and benchmarked a deep-learning computer-aided system for primary detection of neoplasia in patients with BE. The system detected neoplasia with higher accuracy than endoscopists and near-perfect delineation.

LIMITATIONS: Further studies are needed in separate populations.

IMPACT: This system could increase detection rates of esophageal neoplasia in patients with BE.

Lay Summary: The authors developed and tested a computer-aided system to analyze endoscopy images from patients with BE. The system accurately identified neoplastic lesions of the esophagus.