Endoscopic prediction of submucosal invasion in Barrett's cancer with the use of artificial intelligence: a pilot study

Authors

Alanna Ebigbo[*,1], Robert Mendel[*,2,3], Tobias Rückert[2], Laurin Schuster[2], Andreas Probst[1], Johannes Manzeneder[1], Friederike Prinz[1], Matthias Mende[4], Ingo Steinbrück[5], Siegbert Faiss[4], David Rauber[2,6], Luis A. de Souza[2,7], João P. Papa[7], Pierre H. Deprez[8], Tsuneo Oyama[9], Akiko Takahashi[9], Stefan Seewald[10], Prateek Sharma[11], Michael F. Byrne[12], Christoph Palm[2,3,6], Helmut Messmann[1]

Institutions

1 III Medizinische Klinik, Universitätsklinikum Augsburg, Augsburg Germany
2 Regensburg Medical Image Computing (ReMIC), Ostbayerische Technische Hochschule Regensburg (OTH Regensburg), Regensburg Germany
3 Regensburg Center of Health Sciences and Technology (RCHST), OTH Regensburg, Regensburg, Germany
4 Gastroenterology, Sana Klinikum Lichtenberg, Berlin, Germany
5 Department of Gastroenterology, Hepatology and Interventional Endoscopy, Asklepios Klinik Barmbek, Hamburg, Germany
6 Regensburg Center of Biomedical Engineering (RCBE), OTH Regensburg and Regensburg University, Regensburg, Germany
7 Department of Computing, São Paulo State University, São Paulo, Brazil
8 Cliniques Universitaires St-Luc, Université Catholique de Louvain, Brussels, Belgium
9 Saku Central Hospital Advanced Care Center, Nagano, Japan
10 GastroZentrum, Klinik Hirslanden, Zurich, Switzerland
11 Department of Gastroenterology and Hepatology, Veterans Affairs Medical Center and University of Kansas School of Medicine, Kansas City, Missouri, United States
12 Division of Gastroenterology, Vancouver General Hospital, University of British Columbia, Vancouver, British Columbia, Canada

Corresponding author

Alanna Ebigbo, MD, III Medizinische Klinik, Universitätsklinikum Augsburg, Stenglinstr. 2, 86156 Augsburg, Germany
alanna.ebigbo@gmx.de
Phone: +49 821 400 2351

submitted 1.6.2020
accepted after revision 16.11.2020
published online 16.11.2020

Bibliography
Endoscopy 2021; 53: 878–883
DOI 10.1055/a-1311-8570
ISSN 0013-726X

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

ABSTRACT

Background The accurate differentiation between T1a and T1b Barrett's-related cancer has both therapeutic and prognostic implications but is challenging even for experienced physicians. We trained an artificial intelligence (AI) system on the basis of deep artificial neural networks (deep learning) to differentiate between T1a and T1b Barrett's cancer on white-light images.

Methods Endoscopic images from three tertiary care centers in Germany were collected retrospectively. A deep learning system was trained and tested using the principles of cross validation. A total of 230 white-light endoscopic images (108 T1a and 122 T1b) were evaluated using the AI system. For comparison, the images were also classified by experts specialized in endoscopic diagnosis and treatment of Barrett's cancer.

Results The sensitivity, specificity, F1 score, and accuracy of the AI system in the differentiation between T1a and T1b cancer lesions was 0.77, 0.64, 0.74, and 0.71, respectively. There was no statistically significant difference between the performance of the AI system and that of experts, who showed sensitivity, specificity, F1, and accuracy of 0.63, 0.78, 0.67, and 0.70, respectively.

Conclusion This pilot study demonstrates the first multicenter application of an AI-based system in the prediction of submucosal invasion in endoscopic images of Barrett's cancer. AI scored equally to international experts in the field, but more work is necessary to improve the system and apply it to video sequences and real-life settings. Nevertheless, the correct prediction of submucosal invasion in Barrett's cancer remains challenging for both experts and AI.

* These authors contributed equally to this work.

Introduction

The incidence of Barrett's esophagus and Barrett's-related cancer in the West has risen significantly in the past decade [1, 2], and as this trend is expected to continue, diagnosis of Barrett's esophagus and Barrett's cancer during endoscopy must become as accurate as possible. Early diagnosis of Barrett's cancer is necessary because of its prognostic consequences [3]; however, detection and characterization pose a challenge even for experienced endoscopists with modern equipment.

In the past few years, the field of artificial intelligence (AI) has shown promising results in the diagnosis of early Barrett's cancer, especially in the detection domain [4–6]. In two initial studies, our group was able to differentiate between early Barrett's cancer/high-grade dysplasia and nondysplastic Barrett's esophagus lesions using a convolutional neural network (CNN), initially on endoscopic still images and subsequently in real time during endoscopic procedures [7, 8]. However, there are no data on the application of AI in the prediction of submucosal invasion in Barrett's cancer.

The identification of submucosal invasion (T1b) in Barrett's cancer is important because it has implications for the choice of treatment. Lesions with suspected submucosal invasion should be treated with endoscopic submucosal dissection (ESD) rather than cap-based endoscopic mucosal resection [9, 10]. In such lesions, ESD may be a valid alternative to surgery, especially if histopathological evaluation of the resected specimen fulfills the necessary criteria including submucosal invasion depth < 500 µm, well or moderate differentiation, and no lymphatic or blood vessel invasion [9, 10].

In this pilot study using endoscopic still images, we aimed to demonstrate the AI-assisted prediction of submucosal invasion in Barrett's cancer. To the best of our knowledge, this is the first report to show CNN-based differentiation between mucosal (T1a) and submucosal (T1b) invasive Barrett's cancer.

Methods

This was a retrospective, multicenter study in which endoscopic image evaluation was correlated with the results of histopathology. The primary objective of the study was to determine the diagnostic performance (sensitivity, specificity, and accuracy) of an AI system in differentiating between mucosal (T1a) and submucosal (T1b) Barrett's cancer. The secondary objective of the study was to compare the performance of the AI system with that of highly experienced Barrett's endoscopists.

Endoscopic, high-definition, white-light images of T1a and T1b Barrett's cancer were collected retrospectively in three tertiary care centers in Germany. The study was approved by the Institutional Review Board of the University Hospital Augsburg.

Images

For AI training and testing, a total of 230 white-light images (Olympus GIF-HQ190; Olympus Medical Systems, Tokyo, Japan) from 116 patients were included. For most of the patients, only one image was available; however, some patients contributed several images, with a maximum of 14 images from one patient. Overall, 108 images showed mucosal (T1a) and 122 images showed submucosal (T1b) invasive cancers. The images from the three centers varied in terms of resolution, ranging from 656 × 536 to 1350 × 1080 pixels. For our experiments, all images were downscaled to the lowest resolution.

AI system

Training and testing

The network architecture used was a 101-layer residual CNN [11]. The convolutional model, pretrained on the nonmedical ImageNet dataset [12], was mainly used as a feature extractor. Only the fully connected classifier at the end of the network was optimized with the Adam optimizer [13], a learning rate of 1e-4 with a polynomial learning policy [14], and a weight decay of 1e-4.

The network was trained for 1000 epochs and with a batch size of 32. These hyperparameters were optimized using a 5-fold cross validation approach, where the patient data were separated into disjoint sets, such that their union again resulted in the complete original dataset. Images from the same patient were not divided into different folds. For each validation fold, a separate CNN model was trained omitting the data of the validation fold. The distribution of patients to the individual validation sets was controlled by a random seed.

As the resolution of the data was nonuniform, all training images were resized such that the smaller axis had 512 pixels. Then, quadradic patches with a resolution of 512 × 512 pixels were extracted randomly along the larger axis and randomly rotated and flipped for augmentation.

Validation

For validation, which was as independent from the training as possible, again a 5-fold cross validation was performed, but with different folds from those in the training phase. However, as the dataset was of limited size, the composition of the cross validation folds set may have influenced the final result. It is possible that some subsets of images used to validate the model may have closely mirrored or completely differed from the visual properties of most of the data used for training for this fold. Additionally, the dataset also contained easy as well as difficult samples. A validation set consisting of mainly easy or mainly difficult samples would result in over- or underestimation of the model performance, respectively. To reduce these effects, multiple cross validation runs were performed using different random seeds, which were all different from the seed used for parameter optimization. To achieve the most representative result, the 5-fold cross validation scheme was run 10 times with different validation set compositions, and the individual evaluation metrics were averaged over all runs.

Histopathology

Histopathology served as the reference standard for the characterization of images. Based on the results of histopathology, endoscopic images were divided into two categories:

1. images with cancer infiltration limited to the mucosa (pT1a)
2. images with cancer infiltration into the submucosa (pT1b).

Images of lesions with infiltration deeper than the submucosa (>T1b) were excluded from the study. The depth of mucosal (m1, m2, m3, m4) or submucosal (sm1, sm2, sm3) invasion was not further evaluated. Histopathology was based on specimens resected with the ESD technique. Histopathology was confirmed by a second reference pathologist.

Image evaluation by endoscopists

The image dataset was characterized by five international expert endoscopists (A.T., T.O., P.H.D., S.S., P.S.) who were blinded to the true diagnosis of the images.

Outcome measures

The primary outcome was the sensitivity and specificity of the AI system in the prediction of T1b cancer. F1 and classification accuracy were also calculated, as follows:

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Specificity} = \frac{TN}{TN+FP}$$

$$\text{F1} = \frac{2TP}{2TP+FP+FN}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

with TP, TN, FP, FN beeing the number of true positive, true negative, false positive and false negative images, respectively

To ensure bias-free results in cross validation evaluation, these measures were calculated after totaling the confusion matrices for all folds [15].

Interobserver variation between the five experts for the differentiation between T1a and T1b cancer was calculated using Fleiss' kappa (κ) statistics for multiple raters (Microsoft Excel Version 16.0). Interpretation of kappa values was as follows: κ >0.8, almost perfect agreement; 0.8 – 0.61, substantial agreement; 0.6 – 0.41, moderate agreement; 0.4 – 0.21, fair agreement; <0.2, slight agreement; 0, agreement equal to chance; and <0 suggested disagreement [16].

Results

Performance of AI system

The sensitivity and specificity of the AI network in the differentiation between mucosal and submucosal cancer averaged over 10 runs was 0.77 (95 % confidence interval [CI] 0.75 – 0.78) and 0.64 (95 %CI 0.62 – 0.66), respectively, whereas accuracy and F1 scores showed values of 0. 71 (95 %CI 0.70 – 0.72) and 0. 74 (95 %CI 0.72 – 0.74), respectively (▶ **Table 1**).

Image evaluation and performance of expert endoscopists

The average performance of five expert endoscopists was 0.63 (95 %CI 0.53 – 0.77), 0.78 (95 %CI 0.67 – 0.89), 0.70 (95 %CI 0.67 – 0.73), and 0.67 (95 %CI 0.63 – 0.74) for sensitivity, specificity, accuracy, and F1, respectively (▶ **Table 1**). Interobserver agreement (Fleiss' kappa) was 0.49 between the five expert endoscopists.

The average performance of the AI system was similar to that of the experts who participated in the image analysis. However, there seemed to be a wider range of performance results for the expert endoscopists (▶ **Fig. 1**).

A statistical evaluation on the basis of a multivariate extension of the McNemar test revealed no statistically significant difference between the accuracy of the AI system and the mean of the expert endoscopists.

Discussion

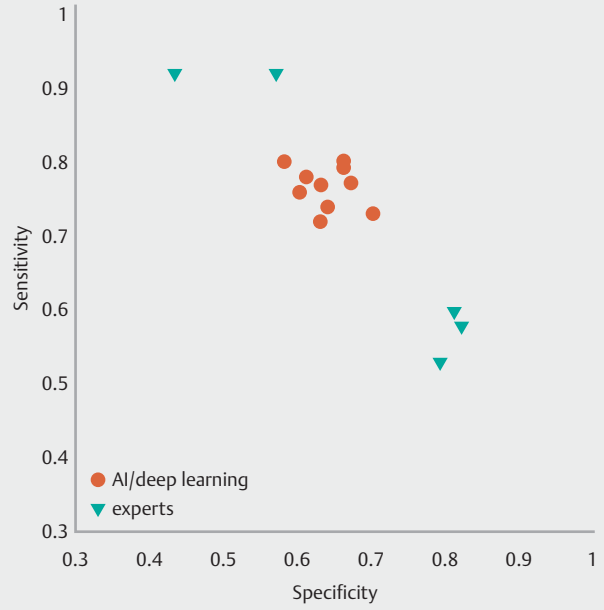
In this pilot study using white-light images, we showed that a trained AI algorithm was able to predict submucosal invasion of Barrett's-related cancer and differentiate between T1a and T1b carcinoma with a sensitivity of 77 %, specificity of 64 %, average F1 score of 74 %, and an overall accuracy of 71 % (▶ **Table 1**). These scores were comparable to the performance of five international Barrett's expert endoscopists who evaluated the same set of images with an interobserver variation of κ = 0.49.

In Barrett's cancer, preoperative differentiation between T1a and T1b cancers has relevant therapeutic and prognostic implications. Esophageal surgery for Barrett's cancer has a 30-day mortality rate of up to 30 % and a morbidity rate as high as

▶ **Table 1** Performance scores of expert endoscopists and the artificial intelligence (AI) system. The mean of the AI system is related to 10 different runs, whereas the mean of the endoscopists is related to five international expert endoscopists (interobserver agreement between 5 endoscopists, κ = 0.49).

	Endoscopists (n = 5)		AI-based results	
	Mean (95 %CI)	SD	Mean (95 %CI)	SD
F1	0.67 (0.63 – 0.74)	0.06	0.74 (0.72 – 0.74)	0.02
Accuracy	0.70 (0.67 – 0.73)	0.03	0.71 (0.70 – 0.72)	0.02
Sensitivity	0.63 (0.53 – 0.78)	0.15	0.77 (0.75 – 0.78)	0.03
Specificity	0.78 (0.67 – 0.89)	0.11	0.64 (0.62 – 0.66)	0.03

AI, artificial intelligence; CI, confidence interval; SD, standard deviation.



▶ **Fig. 1** Receiver operating characteristic curve comparing the performance of the artificial intelligence (AI) network with expert endoscopists. The AI network showed little dispersion between most measurements of different runs, whereas the experts' performance varied widely.

50 % [17]. Endoscopic resection is the method of choice for treatment of T1a lesions [17]. For lesions with suspected submucosal invasion, ESD may be a valid alternative to surgery [9, 10]. However, pretherapeutic staging to differentiate between T1a and T1b lesions is challenging, even with additional endoscopic ultrasound, which itself requires a high level of expertise for accuracy and could be associated with over- or under-staging of lesions [17, 18].

AI technology has been used to predict the invasion depth of cancers in the gastrointestinal tract [19, 20]. Horie et al. [21] demonstrated the differentiation between early (T1) and advanced (T2 – T4) cancers in the esophagus using a deep neural network, with a diagnostic accuracy of 98 %, although both squamous cell carcinomas and adenocarcinomas were included. However, the classification task of differentiating between mucosal (T1a) and submucosal (T1b) Barrett's cancer, which was done in our study, is more challenging.

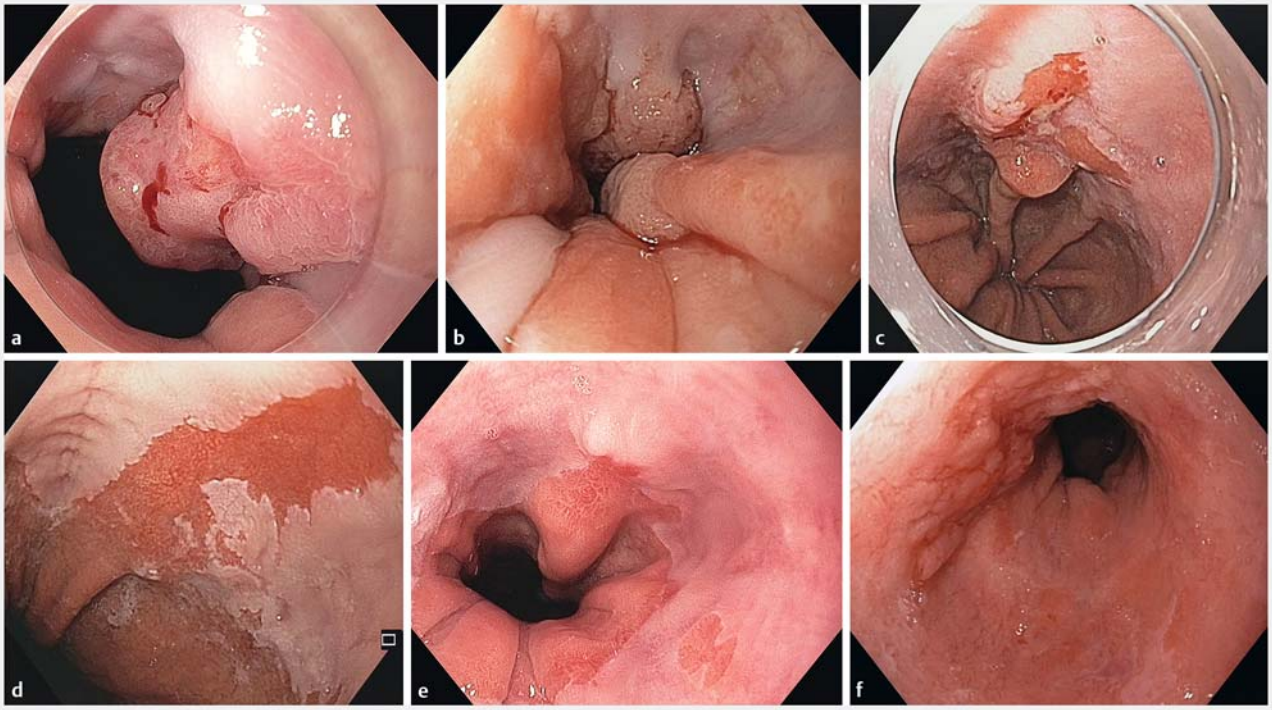
In the stomach, Zhu et al. used a CNN to predict the invasion depth of gastric cancer, with an accuracy of 89.2 %, which was significantly better than the performance of experienced endoscopists who scored an average of 77.5 % accuracy. In contrast to our study, however, Zhu et al. differentiated between mucosal/shallow submucosal cancers and deeper invasive cancers [19]. In their study, almost a third of images in the test group had invasion of the muscularis propria, subserosa or serosa (T2, T3, and T4). The interpretation of these more advanced images is less challenging than differentiating between T1a and T1b lesions, as was done in our study. This can be seen when the average performance of the endoscopists involved in both studies is considered.

In a further study in the colon, Lui et al. used AI image classifiers to differentiate between endoscopically curable and endoscopically incurable lesions, with an overall accuracy of 85.5 % and an accuracy of 94.3 % for narrow-band imaging [20]. Again, compared with the performance results in our study, these scores are clearly better. However, in the study by Lui et al., 80 % of endoscopically curable lesions were benign adenomas while 20 % were cancer lesions with submucosal invasion depths > 1000 µm. The differentiation between adenomas and deep submucosal invasive cancers is without doubt less challenging than differentiating between T1a and T1b lesions, as was done in our study. Again, when the performance of an experienced endoscopist in the study by Lui et al. is considered (accuracy of 86.4 %), then the difficulty of the images in our study can be better understood. Therefore, the performance scores of these two AI studies are not comparable to the results of our study.

We rated the performance of the AI system by comparing it with that of expert endoscopists from Japan, Europe, and the USA. The endoscopists were internationally recognized experts in the endoscopic diagnosis and treatment of early carcinomas with a focus on Barrett's esophagus. As the results of interobserver variation show, the dataset was challenging for the experts, with a Fleiss' kappa coefficient of 0.49, reflecting only moderate agreement between the experts but also the potential for using AI in predicting submucosal invasion. However, the evaluation of still images does not reflect the ideal situation in real life, where expert endoscopists will judge a lesion dynamically using features such as the movement of the esophageal wall, the softness or rigidity of the tissue around the lesion, and the behavior of the region of interest during insufflation and deflation of air. Furthermore, an expert will likely combine modalities such as white-light and virtual chromoendoscopy, as well as clean the lesion completely of all mucus before making a diagnosis.

These points also address the major limitations of our study, which include the number and quality of endoscopic images included. Data were collected retrospectively from three different centers. Some images were mere overviews of the lesion, whereas magnified endoscopic images with better details of the surface and vascular patterns made up only 12 % of the dataset (▶ **Fig. 2**). However, the fact that the results were achieved using white-light and (almost entirely) nonmagnified endoscopic images, demonstrates the high potential of the AI system. In addition, the idea of including a diverse set of images in the training of an AI system may lead to greater specificity of the network. Furthermore, a greater proportion of magnified high-quality images, as well as video sequences, may have improved the diagnostic performance of the experts and possibly also the outcome of the AI network.

The inclusion of several images from a single patient introduced statistical dependencies into the study. However, we strictly avoided splitting the images from a single patient into training and testing to ensure independent validation results. Another effect might be the over- or underestimation of per-



▶ **Fig. 2** Examples of images used in the study. Upper row shows three different examples of submucosally invasive cancer (T1b); lower row shows three different examples of mucosal cancer (T1a).

formance assuming that all images from one patient are classified the same, either correct or false. A closer look at the results revealed that no such effect occurred. This also holds true for the expert evaluations.

The validation method provided results that were not completely independent. However, using 5-fold cross validation with one seed for hyperparameter optimization and 10 different seeds for validation ensured as much independence as possible for a small dataset. The alternative of splitting the data once into training, testing, and validation is highly dependent on the distribution of the sets with a high risk of a bias due to this split. In that sense we avoided the selection bias of so-called "external validation" approaches, accepting a weak dependence of the validation data.

Endoscopic still images do not sufficiently depict the challenges the system would face in reality, which means that video recordings for validation of the network or a real-life setting would have been preferable. Finally, we did not differentiate between the depths of mucosal (m1 – m4) or submucosal (sm1 – sm3) invasion; this, however, may have been desirable as it may be almost impossible to differentiate sufficiently between a deeply mucosal (m3 /m4) invasive cancer and a shallow submucosal (sm1) invasive lesion.

Our future work will focus on improving the diagnostic ability of the system and implementing it in a real-life endoscopy setting. However, the current study may be an initial step toward developing an AI system to aid in the prediction of submucosal invasion of Barrett's cancer.

Conclusion

In this preliminary "proof of concept" study, performance scores of an AI system in the prediction of submucosal invasion in Barrett's cancer were comparable to those of expert endoscopists. The data showed that the prediction of submucosal invasion is a challenge even for Barrett's experts. However, with more training data, the diagnostic ability of the AI system can be improved considerably and then transferred to video images and to a real-life setting. Considering the difficulty this task poses to endoscopists, as well as the prognostic and therapeutic implications involved, we believe that AI has the potential to support the characterization of early Barrett's cancer in future endoscopy practice, especially for non-Barrett's experts.

Funding

Bavarian Academic Forum (BayWISS)

Acknowledgments

Hans Kiesl, Jochen Arnold, Franz Beer, Albert Beyer, Esther Endlicher, Carola Fleischmann, Simone Freund, Stephan Gölder, Frank Klebl, Johannes Manzeneder, Sandra Nagl, Friederike Prinz, Christoph Römmele, Elisabeth Schnoy.

Competing interests

The authors declare that they have no conflict of interest.

References

[1] Coleman HG, Xie SH, Lagergren J. The epidemiology of esophageal adenocarcinoma. Gastroenterology 2018; 154: 390–405

[2] Drahos J, Ricker W, Parsons R et al. Metabolic syndrome increases risk of Barrett esophagus in the absence of gastroesophageal reflux: an analysis of SEER-Medicare Data. J Clin Gastroenterol 2015; 49: 282–288

[3] Sharma P, Bergman JJ, Goda K et al. Development and validation of a classification system to identify high-grade dysplasia and esophageal adenocarcinoma in Barrett's esophagus using narrow-band imaging. Gastroenterology 2016; 150: 591–598

[4] 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: 1242–1250

[5] 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: 915–929

[6] Hashimoto R, Requa J, Tyler D et al. Artificial intelligence using convolutional neural networks for real-time detection of early esophageal neoplasia in Barrett's esophagus (with video). Gastrointest Endosc 2020; 91: 1264–1271

[7] Ebigbo A, Mendel R, Probst A et al. Real-time use of artificial intelligence in the evaluation of cancer in Barrett's oesophagus. Gut 2020; 69: 615–616

[8] Ebigbo A, Mendel R, Probst A et al. Computer-aided diagnosis using deep learning in the evaluation of early oesophageal adenocarcinoma. Gut 2019; 68: 1143–1145

[9] 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

[10] Pimentel-Nunes P, Dinis-Ribeiro M, Ponchon T et al. Endoscopic submucosal dissection: European Society of Gastrointestinal Endoscopy (ESGE) Guideline. Endoscopy 2015; 47: 829–854

[11] He K, Zhang X, Ren S et al. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. Proceedings of the IEEE conference on computer vision and pattern recognition 2016: 770–778

[12] Russakovsky O, Deng J, Su H et al. ImageNet large scale visual recognition challenge. Int J Comput Vis 2015; 115: 211–252

[13] Kingma DP, Ba J. Adam: a method for stochastic optimization. Presented at the International Conference on Learning Representations; 2015 May 7–9; San Diego, California. arXiv preprint arXiv:14126980 2014.

[14] Liu W, Rabinovich A, Berg AC. Parsenet: looking wider to see better. Presented at the International Conference on Learning Representations; 2016 May 2–4, 2016; San Juan, Puerto Rico. arXiv preprint arXiv:150604579 2015.

[15] Forman G, Scholz M. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. ACM SIGKDD Explorations Newsletter 2010; 12: 49–57

[16] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159–174

[17] Thosani N, Singh H, Kapadia A et al. Diagnostic accuracy of EUS in differentiating mucosal versus submucosal invasion of superficial esophageal cancers: a systematic review and meta-analysis. Gastrointest Endosc 2012; 75: 242–253

[18] Qumseya BJ, Brown J, Abraham M et al. Diagnostic performance of EUS in predicting advanced cancer among patients with Barrett's esophagus and high-grade dysplasia/early adenocarcinoma: systematic review and meta-analysis. Gastrointest Endosc 2015; 81: 865–874

[19] Zhu Y, Wang QC, Xu MD et al. Application of convolutional neural network in the diagnosis of the invasion depth of gastric cancer based on conventional endoscopy. Gastrointest Endosc 2019; 89: 806–815

[20] Lui TKL, Wong KKY, Mak LLY et al. Endoscopic prediction of deeply submucosal invasive carcinoma with use of artificial intelligence. Endosc Int Open 2019; 7: E514–e520

[21] Horie Y, Yoshio T, Aoyama K et al. Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks. Gastrointest Endosc 2019; 89: 25–32