

Deep-learning based detection of gastric precancerous conditions

Pedro Guimarães,¹ Andreas Keller,¹ Tobias Fehlmann,¹ Frank Lammert,² Markus Casper ²

► Additional material is published online only. To view please visit the journal online (<http://dx.doi.org/10.1136/gutjnl-2019-319347>).

¹Chair for Clinical Bioinformatics, Saarland University, Saarbrücken, Germany

²Department of Medicine II, Saarland University Medical Center, Saarland University, Homburg, Germany

Correspondence to

Dr Markus Casper, Department of Medicine II, Saarland University Medical Center, Saarland University, 66421 Homburg, Germany; markus.casper@uks.eu

Received 22 June 2019

Revised 17 July 2019

Accepted 18 July 2019

Published Online First

2 August 2019

MESSAGE

Conventional white-light endoscopy has high interobserver variability for the diagnosis of gastric precancerous conditions. Here we present a deep-learning (DL) approach for the diagnosis of atrophic gastritis developed and trained using real-world endoscopic images from the proximal stomach. The model achieved an accuracy of 93% (area under the curve (AUC): 0.98; *F*-score 0.93) in an independent data set, outperforming expert endoscopists. DL may overcome conventional appraisal of white-light endoscopy and support human decision making. The algorithm is available free of charge via a web-based interface (<https://www.ccb.uni-saarland.de/atrophy>).

IN MORE DETAIL

Introduction

Chronic inflammation of the gastric mucosa induces a cascade of precancerous conditions (chronic atrophic gastritis, intestinal metaplasia) and lesions (dysplasia) that may result in the development of intestinal-type gastric cancer.¹ Infection with *Helicobacter pylori* and autoimmune gastritis are the most relevant factors initiating these mechanisms. Conventional white-light endoscopy has moderate sensitivity and specificity, as well as a high interobserver variability, and is therefore not sufficient to reliably diagnose gastric atrophy or intestinal metaplasia.^{2,3} Thus, especially in Western countries, histology-based diagnosis of precancerous conditions using standardised biopsy protocols is favoured. Advanced endoscopic techniques (eg, virtual or conventional chromoendoscopy, magnification endoscopy, confocal laser endomicroscopy) are often hindered by technical availability and costs.

DL has demonstrated potential in medical imaging, including GI endoscopy.⁴ In this field, DL has been used to diagnose focal pathologies (in particular colorectal polyps and oesophageal adenocarcinoma), and only occasionally for diseases diffusely affecting the GI mucosa (eg, *H. pylori*-associated gastritis).⁴⁻⁷ Here, for the first time, we present a DL approach that overcomes the limitations of white-light endoscopy in diagnosing atrophic gastritis.

Patients and methods

For a first data set, we identified 200 real-world images from patients with and without histology-proven atrophic gastritis (100 each) from subjects undergoing routine oesophagogastroduodenoscopy between 2008 and 2018 (data set DS1). Endoscopies were performed with various generations of

Olympus scopes (GIF-Q160, GIF-Q160Z, GIF-1TQ160, GIF-Q165, GIF-H180, GIF-H190; Olympus Europe, Hamburg, Germany). Images were unaltered white-light images anonymised and exported as Digital Imaging and Communications in Medicine (DICOMs). Non-standardised images (eg, various scope positions, distances, angles and illumination; bile, food and mucus contaminations) were taken from the non-overinflated proximal stomach (gastric corpus and fundus). All images were cropped, resized and normalised to have a set average and SD.

An independent second data set (data set DS2) of 70 images (30 with atrophy; 40 without) was used for independent testing and evaluation by six endoscopists with less than 1500 and more than 1500 esophagogastroduodenoscopy (EGDs). Since the two groups (three each) did not differ ($p > 0.05$), their ratings were combined. Table 1 summarises the patient characteristics. Patients included in the study had no evidence of persisting *H. pylori* infection. Histopathological assessment of H&E-stained slices was carried out by seven board-certified academic pathologists, with at least two pathologists evaluating each specimen (non-blinded) using the updated Sydney system.⁸

With traditional machine learning, handcrafted features are fed to a model for classification. With DL these are computed incrementally by the model without expert intervention. Thus, there is no theoretical limit that prevents it from learning any feature representation. Convolutional neural networks (CNNs) are the gold standard for image analysis. CNNs take advantage of the local structural relationships in the image and create progressively more complex abstract representations from layer to layer. However, this requires a large amount of training data.

To overcome this limitation, we used a fine-tuned, pretrained CNN; that is, we used pretrained weights to initialise the network, thus improving the stability and performance of our model (figure 1). The training data were artificially augmented by image rotation, mirroring and scaling. First, we assessed the best architecture (pretrained models on ImageNet).⁹ We performed 10-fold stratified cross-validation on DS1. For each round, data were split into training, tuning and testing sets (80%/10%/10%). The test set was classified using the best performing hyperparameter combination, as assessed in the tuning set (early stop grid search to select dropout, learning rate, momentum). All images were used for testing once only. In a



© Author(s) (or their employer(s)) 2020. No commercial re-use. See rights and permissions. Published by BMJ.

To cite: Guimarães P, Keller A, Fehlmann T, *et al.* *Gut* 2020;**69**:4-6.

Table 1 Data set characteristics

	Data set 1	Data set 2
Patients with atrophy	n=37	n=13
Age (range)	69±13 (39–91)	70±13 (47–83)
Gender (female/male)	22/15 (59%/41%)	7/6 (54/46%)
Autoimmune gastritis	28 (76%)	8 (62%)
Severe atrophy	11 (30%)	4 (31%)
Intestinal metaplasia	27 (73%)	9 (70%)
Images/patient (range)	2.7 (1–15)	2.3 (1–5)
Images fundus	48 (48%)	19 (63%)
Images corpus	52 (52%)	11 (37%)
Patients without atrophy	n=64	n=22
Age (years)	64±15 (18–86)	66±17 (26–83)
Gender (female/male)	29/35 (45%/55%)	11/11 (50%/50%)
Normal mucosa	32 (50%)	11 (50%)
Chronic gastritis	32 (50%)	11 (50%)
Images/patient, median (range)	1.5 (1–4)	1.8 (1–5)
Images fundus	62 (62%)	20 (50%)
Images corpus	38 (38%)	20 (50%)

For data sets 1 and 2, baseline characteristics are given. For patient age, mean and SD as well as range (in parentheses) are presented. Medians and ranges of the numbers of images used per patient are shown.

second stage, DS1 was used for training and tuning (90%/10% split), whereas DS2 was used for testing. Model architecture was chosen from the first-stage results. Hyperparameters were

assessed in the tuning set. Online supplementary file 1 provides an indepth description of the methods.

For all models and expert evaluations, accuracy, balanced accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and *F*-score were computed. In addition, NPV, PPV and accuracy were computed for prevalence rates between 20% and 50% in steps of 1%. Statistical differences between the expert evaluations and DL were assessed with Wilcoxon signed-rank test. Further, we computed the receiver operating characteristic (ROC) curves and the AUC.

Results

The best performing pretrained DL model for diagnosis of atrophic gastritis, as assessed by cross-validation, was VGG16.¹⁰ The algorithm yielded results for all images. Table 2 summarises the results.

Accuracy, balanced accuracy and *F*-score were significantly lower for the endoscopists when compared with the DL-based approach ($p=0.03$). There was no significant difference between the endoscopy experts and the model for the remaining performance metrics. Online supplementary figure 1A,B shows the ROC curves.

DISCUSSION

We present a DL approach capable of surpassing expert assessment for endoscopic diagnosis of atrophic gastritis. Despite the low number of images available, our model achieved a diagnostic

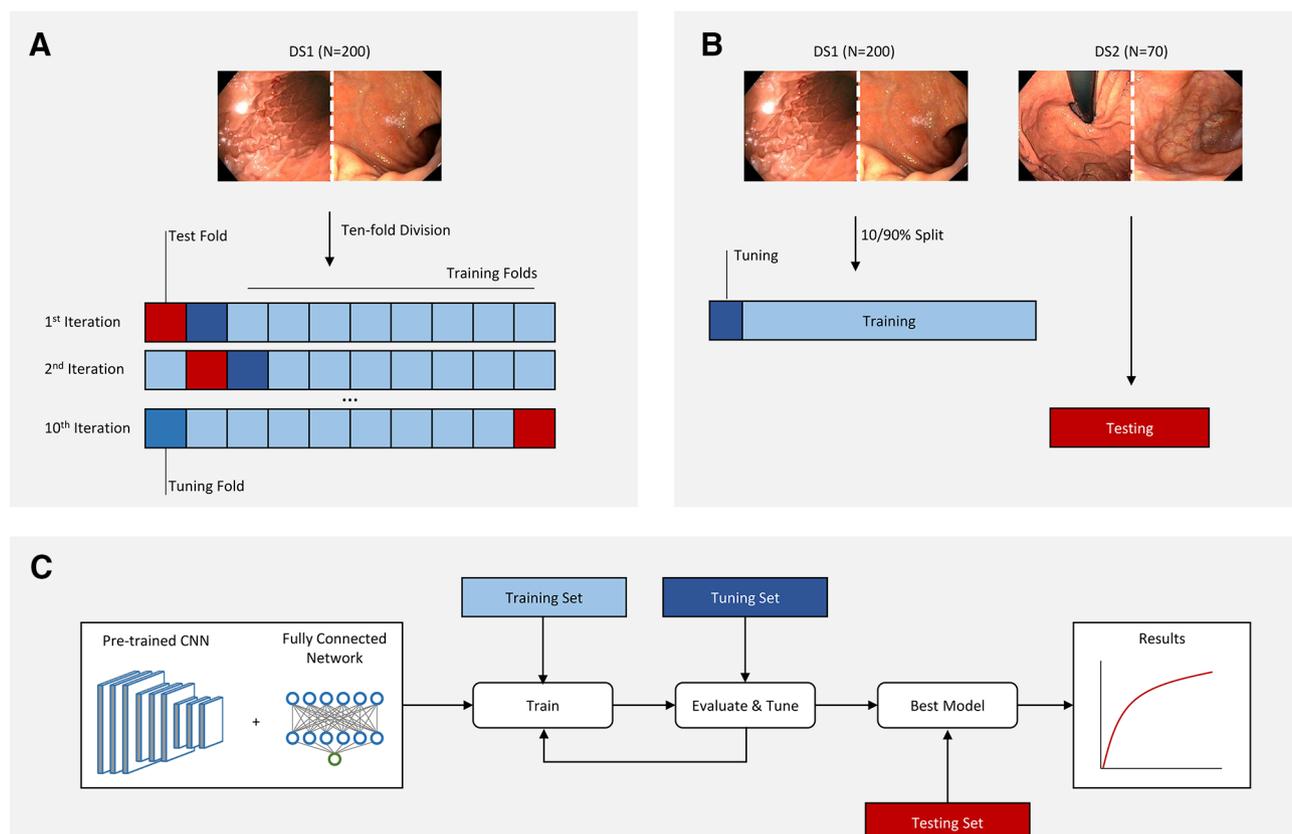


Figure 1 Study workflow. DS1 was used to perform 10-fold stratified cross-validation (A). At each iteration, one fold was used for testing, one was used for tuning, and the remainder were used for training. In a second step, DS1 was used for tuning and training (10/90% split), whereas DS2 was used for hold-out testing (B). The workflow for each training, tuning and testing cycle was performed as presented in C. The testing set (DS2) was only used to derive results and never to train the model or to tune hyperparameters, remaining completely independent. CNN, convolutional neural network; DS1, data set 1; DS2, data set 2.

Table 2 Results of the DL-based algorithm in comparison with evaluation by endoscopists

	Data set 1		Data set 2		
	DL-based algorithm		Endoscopists		
	Cross-validation	Independent validation	Combined (n=6)	Less experienced (n=3)	More experienced (n=3)
Accuracy	0.935	0.929	0.800 (0.07)	0.814 (0.11)	0.786 (0.05)
Balanced accuracy	0.935	0.938	0.800 (0.09)	0.832 (0.10)	0.768 (0.08)
Sensitivity	0.930	1.000	0.800 (0.26)	0.956 (0.08)	0.644 (0.30)
Specificity	0.940	0.875	0.800 (0.17)	0.708 (0.17)	0.892 (0.15)
PPV	0.939	0.857	0.799 (0.15)	0.724 (0.14)	0.875 (0.14)
NPV	0.931	1.000	0.879 (0.14)	0.956 (0.08)	0.802 (0.15)
F-score	0.935	0.933	0.758 (0.13)	0.820 (0.10)	0.695 (0.15)
AUC	0.984	0.981	–	–	–

For the evaluation by groups of endoscopists, SD is given in parentheses. AUC, area under the curve; DL, deep learning; NPV, negative predictive value; PPV, positive predictive value.

accuracy of 93%, which was significantly better than the combined results of endoscopists working at a tertiary referral centre.

Endoscopic surveillance is generally advised for patients with extensive atrophy or intestinal metaplasia, but not in case of precancerous conditions restricted to the antrum.² Thus, we decided to focus on the proximal stomach (gastric corpus and fundus). Histopathology was used as gold standard. This method is, especially in initial or patchy disease, prone to sampling error. Thus, a false-positive rate of 12.5% is acceptable, because false-negative results of histopathology (at least two biopsies in the proximal stomach) cannot be ruled out. Our algorithm cannot sharply discriminate simple atrophy from metaplastic atrophic gastritis, since most patients in both cohorts suffered from atrophic gastritis with intestinal metaplasia, which is the most reliable histological marker of atrophy.²

The strength of our approach is that we used real-world images for training, tuning and evaluation. Thus, our algorithm has the capability to work reliably under these conditions and is not dependent on high-quality, ideal images. Nevertheless, the generalisation of these results needs to be taken cautiously since the size of the training data set was limited. The prevalence of atrophic gastritis varies in different parts of the world,¹¹ and affected patients are more likely to be present in endoscopy-based cohorts. Therefore, we extrapolated the performance metrics for the reported prevalence range from 20% to 50%.¹¹ Notwithstanding that the algorithm performs adequately across these real-world prevalence rates, as shown in online supplementary figure 1C, further prospective evaluation in additional cohorts is inevitable before standard implementation.

To provide worldwide direct access for a broad group of users, we developed a web-based software tool where image files can be uploaded for analysis by the DL-based algorithm (available free of charge at <https://www.ccb.uni-saarland.de/atrophy>). Moreover, uploaded images from different settings may lead to more robust algorithms in the future, overcoming the limitations associated with one training data set.

In conclusion, DL can support human decision making in complex settings of GI endoscopy and is a promising tool for clinically relevant endoscopy applications.

Acknowledgements The authors thank Thomas Adams, MD, Dr Bettina Friesenhahn-Ochs, MD, Dr Katharina Grotemeyer, MD, Dr Oliver Linn, MD, Dr Matthias Reichert, MD, and Simone Zimmermann, MD, for blinded evaluation

of endoscopy images, and the team of Professor Dr Rainer Bohle, MD, for histopathological evaluation.

Contributors PG: programming of the deep-learning algorithm, image analysis, statistics, manuscript preparation. AK: revision and editing of the manuscript, supervision of the artificial intelligence part, idea for the study. TF: Programming of the web-based software tool. FL: revision and editing of the manuscript; supervision of the clinical part, idea for the study. MC: manuscript preparation, patient identification and coordination of image evaluation by endoscopists.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval The study was approved by the ethics committee of Ärztekammer des Saarlandes (Saarbrücken, Germany; #36/19).

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iD

Markus Casper <http://orcid.org/0000-0002-1146-288X>

REFERENCES

- Correa P, Piazuelo MB. The gastric precancerous cascade. *J Dig Dis* 2012;13:2–9.
- Pimentel-Nunes P, Libânio D, Marcos-Pinto R, *et al.* Management of epithelial precancerous conditions and lesions in the stomach (maps II): European Society of endoscopy (ESGE), European Helicobacter and microbiota Study Group (EHMSG), European Society of pathology (ESP), and Sociedade Portuguesa de Endoscopia Digestiva (SPED) guideline update 2019. *Endoscopy* 2019;51:365–88.
- Redéen S, Petersson F, Jönsson KA, *et al.* Relationship of gastroscopic features to histological findings in gastritis and Helicobacter pylori infection in a general population sample. *Endoscopy* 2003;35:946–50.
- Shichijo S, Nomura S, Aoyama K, *et al.* Application of Convolutional neural networks in the diagnosis of Helicobacter pylori infection based on endoscopic images. *EBioMedicine* 2017;25:106–11.
- Mori Y, Kudo SE, Mohmed HEN, *et al.* Artificial intelligence and upper gastrointestinal endoscopy: current status and future perspective. *Dig Endosc* 2018.
- 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.
- Ebigbo A, Mendel R, Probst A, *et al.* Computer-Aided diagnosis using deep learning in the evaluation of early oesophageal adenocarcinoma. *Gut* 2019;68:1143–5.
- Dixon MF, Genta RM, Yardley JH, *et al.* Classification and grading of gastritis. The updated Sydney system. International workshop on the histopathology of gastritis, Houston 1994. *Am J Surg Pathol* 1996;20:1161–81.
- Russakovsky O, Deng J, Su H, *et al.* ImageNet large scale visual recognition challenge. *Int J Comput Vis* 2015;115:211–52.
- Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition, 2014. Available: <http://arxiv.org/abs/1409.1556>
- Weck MN, Brenner H. Prevalence of chronic atrophic gastritis in different parts of the world. *Cancer Epidemiol Biomarkers Prev* 2006;15:1083–94.