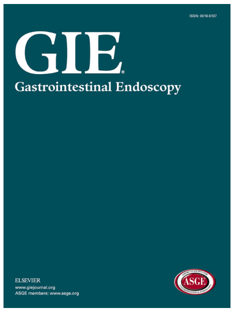

ISSN: 0016-5107
GIE®
Gastrointestinal Endoscopy
ELSEVIER
www.giejournal.org
ASGE members: www.asge.org
AMERICAN SOCIETY FOR
ASGE
GASTROINTESTINAL ENDOSCOPY

Detection of duodenal villous atrophy on endoscopic images using a deep learning algorithm

Markus W. Scheppach, M.D., David Rauber, M.Sc., Dr. Johannes Stallhofer, M.D., Anna Muzalyova, Ph.D., Vera Otten, M.D., Carolin Manzeneder, M.D., Tanja Schwamberger, M.D., Julia Wanzl, M.D., Jakob Schlottmann, M.D., Vidan Tadic, M.D., Andreas Probst, M.D., Elisabeth Schnoy, M.D., Christoph Römmele, M.D., Carola Fleischmann, M.D., Michael Meinikheim, M.D., Silvia Miller, M.D., Bruno Märkl, M.D., Andreas Stallmach, M.D., Christoph Palm, Ph.D., Helmut Messmann, M.D., Alanna Ebigbo, M.D.

PII: S0016-5107(23)00009-3

DOI: https://doi.org/10.1016/j.gie.2023.01.006

Reference: YMGE 13521

To appear in: *Gastrointestinal Endoscopy*

Received Date: 28 September 2022

Revised Date: 16 December 2022

Accepted Date: 1 January 2023

Detection of duodenal villous atrophy on endoscopic images using a deep learning algorithm

Authors:

Markus W. Scheppach *[1] M.D., David Rauber *[2,3] M.Sc., Dr. Johannes Stallhofer[4] M.D., Anna Muzalyova[1] Ph.D., Vera Otten[1] M.D., Carolin Manzeneder[1] M.D., Tanja Schwamberger[1] M.D., Julia Wanzl[1] M.D., Jakob Schlottmann[1] M.D., Vidan Tadic[1] M.D., Andreas Probst[1] M.D., Elisabeth Schnoy[1] M.D., Christoph Römmele[1] M.D., Carola Fleischmann[1] M.D., Michael Meinikheim[1] M.D., Silvia Miller[5] M.D., Bruno Märkl[5] M.D., Andreas Stallmach[4] M.D., Christoph Palm[2,3] Ph.D., Helmut Messmann[1] M.D., Alanna Ebigbo[1] M.D.

Author affiliations:

1. Internal Medicine III – Gastroenterology, University Hospital of Augsburg, Augsburg, Germany
2. Regensburg Medical Image Computing (ReMIC), Ostbayerische Technische Hochschule Regensburg (OTH Regensburg), Regensburg, Germany
3. Regensburg Center of Biomedical Engineering (RCBE), OTH Regensburg, Germany
4. Department of Internal Medicine IV (Gastroenterology, Hepatology, and Infectious Diseases), Jena University Hospital, Jena, Germany
5. Pathology, University Hospital of Augsburg, Augsburg, Germany

* Both authors contributed equally to this work.

Corresponding author:

Markus W. Scheppach M.D., Internal Medicine III – Gastroenterology, University Hospital of Augsburg, Stenglinstrasse 2, 86156 Augsburg, Germany

Detection of duodenal villous atrophy on endoscopic images using a deep learning algorithm

Authors:

Markus W. Scheppach *[1], David Rauber *[2,3], Johannes Stallhofer[4], Anna Muzalyova[1], Vera Otten[1], Carolin Manzeneder[1], Tanja Schwamberger[1], Julia Wanzl[1], Jakob Schlottmann[1], Vidan Tadic[1], Andreas Probst[1], Elisabeth Schnoy[1], Christoph Römmele[1], Carola Fleischmann[1], Michael Meinikheim[1], Silvia Miller[5], Bruno Märkl[5], Andreas Stallmach[4], Christoph Palm[2,3], Helmut Messmann[1], Alanna Ebigbo[1]

Author affiliations:

1. Internal Medicine III – Gastroenterology, University Hospital of Augsburg, Augsburg, Germany
2. Regensburg Medical Image Computing (ReMIC), Ostbayerische Technische Hochschule Regensburg (OTH Regensburg), Regensburg, Germany
3. Regensburg Center of Biomedical Engineering (RCBE), OTH Regensburg, Germany
4. Department of Internal Medicine IV (Gastroenterology, Hepatology, and Infectious Diseases), Jena University Hospital, Jena, Germany
5. Pathology, University Hospital of Augsburg, Augsburg, Germany

* Both authors contributed equally to this work.

Abstract

Background and aims:
Celiac disease with its endoscopic manifestation of villous atrophy is underdiagnosed worldwide. The application of artificial intelligence (AI) for the macroscopic detection of villous atrophy at routine esophagogastroduodenoscopy may improve diagnostic performance.

Methods:
A dataset of 858 endoscopic images of 182 patients with villous atrophy and 846 images from 323 patients with normal duodenal mucosa was collected and used to train a ResNet 18 deep learning model to detect villous atrophy. An external data set was used to test the algorithm, in addition to six fellows and four board certified gastroenterologists. Fellows could consult the AI algorithm's result during the test. From their consultation distribution, a stratification of test images into "easy" and "difficult" was performed and used for classified performance measurement.

Results:
External validation of the AI algorithm yielded values of 90 %, 76 %, and 84 % for sensitivity, specificity, and accuracy, respectively. Fellows scored values of 63 %, 72 % and 67 %, while the corresponding values in experts were 72 %, 69 % and 71 %, respectively. AI consultation significantly improved all trainee performance statistics. While fellows and experts showed significantly lower performance for "difficult" images, the performance of the AI algorithm was stable.

Conclusion:
In this study, an AI algorithm outperformed endoscopy fellows and experts in the detection of villous atrophy on endoscopic still images. AI decision support significantly improved the performance of non-expert endoscopists. The stable performance on "difficult" images suggests a further positive add-on effect in challenging cases.

Key words (5): celiac disease, villous atrophy, endoscopy, detection, artificial intelligence,

Abbreviations: AI, artificial intelligence; AI-CDSS, artificial intelligence clinical decision support solution; EGD, esophagogastroduodenoscopy; VA, villous atrophy;

Conflict of interest disclosure:

All authors declare no conflict of interest.

Introduction

Celiac disease, a disorder caused by an inflammatory reaction of the small intestinal mucosa to ingested gluten in genetically susceptible persons, has a worldwide prevalence of 1.4% [1]. While the prevalence is reported to be rising, the disease continues to be underreported [2-4] and more than 50% of cases are undiagnosed worldwide. This seems to be due to its unspecific symptoms [5], as well as the endoscopic manifestation (small intestinal villous atrophy), which is often subtle and easily overlooked at inspection [6]. Villous atrophy is most often caused by celiac disease, but can also occur in other disorders, such as tropical sprue or Whipple's disease [7]. Endoscopic markers of villous atrophy include a mosaic pattern and deep groves of the mucosa, scalloping and, in severe cases, loss of duodenal folds, as well as visible submucosal vessels and duodenal erosions [8]. At least 23% of histologically and serologically confirmed cases of villous atrophy and celiac disease showed no macroscopic signs of villous atrophy during conventional endoscopic examination [9]. The histological examination shows villous effacement, crypt hypertrophy and an accumulation of lymphocytes in the mucosa and is most often classified by the Marsh-Oberhuber classification [10, 11]. As 70% of patients are diagnosed as adults [12] and the time between the onset of symptoms and the definitive diagnosis amounts to 11 years on average [13], there is an apparent need for scientific innovation into the diagnostic yield of this disease. Blood serology can make the diagnosis with high accuracy [5], however this test is only applied if celiac disease is considered by the clinician. Esophagogastroduodenoscopy (EGD) on the other hand is a diagnostic tool that is performed frequently for upper GI conditions unrelated to celiac disease. It stands to reason that villous atrophy may be present concomitantly in a relevant percentage of these examinations. To improve macroscopic detection in these cases by modern techniques of image analysis may be scientifically and clinically interesting.

Deep learning algorithms have been developed with great success for the recognition of colorectal polyps during colonoscopy [14, 15] as well as other gastrointestinal disorders [16]. We therefore aimed to design a deep learning algorithm for the detection of villous atrophy on images of the duodenum and jejunum.

Various national guidelines give recommendations concerning the training of endoscopists in their respective countries. In Great Britain independent performance of EGD is permitted after 250 cases under supervision if certain criteria are met [17]. Considering the prevalence of celiac disease, it is likely that the visual diagnosis of villous atrophy is often made without supervision for the first time. This fact further suggests that an artificial intelligence clinical decision support solution (AI-CDSS) for the detection of villous atrophy and celiac disease may have a potential clinical benefit, especially for gastroenterology fellows in training.

Methods

The main objective of this study was to demonstrate that an AI algorithm detects villous atrophy with higher sensitivity than trainees in endoscopy. Sensitivities of 85% and 70% were assumed for the AI algorithm and trainees, respectively. To show this difference with a power of 80% and a p-value of < 5 % a sample size of greater than 131 test images per group was calculated.

858 still images of the duodenum or jejunum from 182 patients with histologically confirmed villous atrophy (VA) (Marsh-classification grade III) [10] were retrospectively extracted from Augsburg University Hospital database for the years 2010 to 2021. 846 further images from 323 patients with macroscopically and histologically confirmed non-atrophic small intestinal mucosa (controls) were extracted for the same period. Patients with known celiac disease under gluten free diet were excluded from the control data set. Images were recorded during routine clinical practice using Olympus gastroscopes (GIF-HQ190, GIF-HQ-180, GIF-HQ1500; Olympus Medical Systems, Tokyo, Japan). At least one image and at most 69 images were included per patient. Characteristics of the VA and control datasets are shown in **Table 1**.

The training dataset was split into five equal-sized subsets. Splitting the images from one patient into multiple subsets was avoided. To classify these images, a Convolutional Neural Network (CNN) was used as a model. This type of network consists of a sequence of convolutional and non-linear layers. In this case the ResNet architecture was employed [18]. The model uses so-called skip-connections, which allow to propagate low-level features. For this project, a ResNet with 18 layers (ResNet18) was chosen [19]. This model was trained with the images of four subsets and then validated internally with the remaining subset (five-fold cross-validation). This process was repeated for each subset, such that each subset was validated once. An additional external test data set was obtained from Jena University Hospital, Jena, Germany. Following the same rules of inclusion as for the training data, the test set comprised 194 VA images and 155 control images. Indications for EGDs in adults in descending order of frequency included abdominal pain, diarrhea, anemia, Crohn's disease, non-cardiac chest pain and suspected mastocytosis. In children, EGD was only performed for the clinical suspicion of celiac disease, which included abdominal discomfort, diarrhea, anemia and failure to thrive, as well as positive serology. Further details of this dataset are shown in **Table 2**. Images were recorded during clinical practice using Olympus gastroscopes (GIF-HQ190, GIF-HQ185, GIF-HQ1500; Olympus Medical Systems, Tokyo, Japan).

The trained AI algorithm, as well as four board certified gastroenterologists (experts) with > 1000 EGDs and six gastroenterology fellows (trainees) with an experience of 100 – 1000 EGDs were tested on the external test data set. The mean endoscopic experience of trainees ± standard deviation was 278 ± 173 examinations at the time of the study. A binary decision for a macroscopic suspicion of villous atrophy and subsequent indication for duodenal biopsy was asked for each image. The trainees were given access to the results of the AI algorithm. This means that after documentation of their suspected diagnosis, trainees were allowed to consult the AI algorithm, whenever they were unsure or in doubt. Consultation of the AI algorithm was documented for each test image. Finally, a definitive diagnosis was documented if the AI algorithm was consulted. This group was informed about the sensitivity and specificity of the AI algorithm on the external data set in advance. For evaluation, trainees were regarded as two groups, once before the AI result could be consulted, once after optional consultation of the AI algorithm's result for all test questions. The test images were divided into two subcategories: "Easy" images were defined as images for which zero or one trainee consulted the AI, "difficult" images were defined as images, for which two or more trainees consulted the AI algorithm.

The categorical variables are expressed as absolute numbers and percentages. Pooled sensitivity, specificity and accuracy of each group were determined and are presented as percentages. These quality criteria / performance indices were compared between gastroenterologists' experience levels as well as depending on images' difficulty within each experience level. The different experience levels were compared using the McNemar test [20]. The difficulty within each experience level was tested using Fisher's Exact Test. Correction for multiple comparisons was performed by the Bonferroni method. A p-value of less than or equal to 0.05 was considered statistically significant. Ethics approval was obtained for the entire study, from the Ethics Committee of Ludwig-Maximilians-University, Munich (Project Nr: 21-1215) and for the external data set from the Ethics Committee of Jena University Hospital (Registration Nr. 2021-2297). The approval included data acquisition, data processing for the development of an AI algorithm and preclinical evaluation of this algorithm.

Results

The internal cross-validation yielded values of 82 %, 85 % and 84 % for sensitivity, specificity and accuracy for the AI algorithm. On the external test data, the AI algorithm achieved values of 90 %, 76 %, and 84 % for sensitivity, specificity, and accuracy, respectively. Sensitivities, specificities and accuracies of the different groups of endoscopists and the AI algorithm for the external test data set are shown in **Figure 1**. All differences reached statistical significance except for the comparisons of specificities of trainees vs. experts and trainees with AI support vs. AI alone.

Within the group trainee with AI support, the AI algorithm's finding was consulted in 21 % (N=438) of overall pooled test questions (N=2094). In 42 % (N=185) of these cases, the AI algorithm disagreed with the test subject. In cases of disagreement with the AI finding, the trainees changed their final diagnosis in 81 % (N=149). In 92 % (N=139) of these cases, the decision change led to the correct diagnosis. In cases of agreement of primary diagnosis and AI finding (58%, N=253), the decision was kept in 97% (N=246) of cases and agreement with the AI finding led to the right diagnosis in 79% (N=200) of cases.

Of all 349 test images, 30% (105 images) triggered no consultation of the AI by any of the six trainees. For 29 % of the tests (101 images) one trainee consulted the AI, two trainees consulted the AI in 28 % of the tests (99 images), three trainees consulted the AI in 11 % (37 images) and four consulted the AI in 2 % of the tests (7 images). There were no test images, for which five or all six of trainees consulted the AI algorithm. Hence, the test images were divided into two subcategories: "Easy" images were defined as images for which zero or one trainee consulted the AI, "difficult" images were defined as images, for which two or more trainees consulted the AI algorithm. Sensitivities, specificities and accuracies for all groups after this subdivision are shown in **Figure 2**.

Discussion

Detection of villous atrophy, in the vast majority of cases due to celiac disease, by artificial intelligence has been attempted by different groups. Gadermayr et al. [21] achieved an accuracy of 94 % to 100 % for the detection of villous atrophy during EGD using a combination of multi-resolution local binary patterns, improved Fisher vectors and a multi-fractal spectrum with expert knowledge. However, this technique requires water immersion of the duodenum and the study was conducted in children. Villous atrophy can also be detected on capsule endoscopy images with a high accuracy of over 90 % using different forms of artificial intelligence [22, 23]. These studies were done in the setting of a high pre-test probability or the clinical suspicion of celiac disease. Water immersion of the duodenum, as well as capsule endoscopy are not routine examinations, and reserved for particular cases. The aim of the current study was the development of an application for routine EGD and for supporting the endoscopist in making the incidental diagnosis of villous atrophy. Since celiac disease causes unspecific symptoms, false diagnoses such as gastritis or even IBS may be made, because the differential diagnosis of celiac disease was not considered.

Celiac disease reportedly has a rising prevalence of at least 1 % worldwide, of which more than 50 % are undiagnosed [2-4, 24]. According to large epidemiologic studies, patients may often present without gastrointestinal symptoms [12] and therefore are difficult to detect clinically. In this setting of low pre-test probability for celiac disease, serology testing is rarely performed by clinicians. This suggests a potential benefit of an AI-CDSS for the detection of villous atrophy and, consequently celiac disease during routine EGD, i.e. in cases, where celiac disease is not a probable differential diagnosis before the intervention. The reduction of lag time between the onset of symptoms and the final diagnosis by means of an AI application may prove valuable: It may reduce the burden of advanced disease und may thus be cost effective.

This study was designed to show a superiority of an AI algorithm over trainees in the detection of villous atrophy, which was indeed demonstrated. An improvement of trainee performance by AI support was a secondary outcome parameter. A superiority of the AI algorithm over experts or a benefit of AI support for this group were considered unlikely, which is why these questions were not addressed in the study. The measured difference between AI and experts was an unexpected finding, which may generate hypotheses for further research.

~~Fellows consulted the AI tool in 21 % of pooled test questions. A subdivision into “easy” and “difficult” test images was done according to the frequency of AI consultation by the test subjects for a specific test image.~~ The results show a clinically relevant and statistically significant difference between “easy” and “difficult” images in all performance parameters and for all groups, except for the AI algorithm. ~~With regard to the sensitivity in VA detection by AI, there was no statistically significant difference between “easy” and “difficult” images (91 % vs. 89 %).~~ It classified images, which were easy or difficult for endoscopists to assess, with stable performance. Consequently, there may be parameters in the endoscopic image, which cannot be detected by the human eye but can be used for diagnosis by an AI algorithm. These results suggest a clinical benefit in the detection of villous atrophy (and, thereby, celiac disease) by the application of the AI algorithm, especially for endoscopy fellows in training and in macroscopically challenging cases.

This study may have several limitations. While the dataset is comparably large considering the rarity of the disease, only cases with a high degree of histologic alterations in the duodenal mucosa were included (Marsh III). An increase of mucosal lymphocytes (Marsh I) and the proliferation of crypts (Marsh II) were not included, since they are not visible on the macroscopic endoscopic image and were rarely found on biopsy in our population (data not shown). Mild cases of celiac disease might therefore be missed by the AI algorithm. Furthermore, the test decision was based on the inspection of a single duodenal image. This practice gives less visual information to endoscopists than they would obtain in a clinical setting; their diagnostic capability might be diminished simply due to this circumstance. However, also AI performance may be improved upon application to video data. The low number of test subjects calls into question if the results can be generalized. In order to circumvent this problem, we used statistical methods for low subject numbers (McNemar test) and a large test data set (349 test images) for a more accurate measurement of the subjects’ performance. A further limitation is the retrospective nature of the data set, which might entail a lower image quality, than is standard today, as well as a lack of standardization of image collection. However, non-conformity of images provides a more realistic data set, reduces the risk of overfitting and improves the robustness of the resulting algorithm.

The composition of the test dataset with an approximately 50:50 split of VA to control patients does not reflect real life, where the true prevalence of celiac disease is 1.4 % [1]. A theoretical test dataset with a spit of 1.4 to 98.6 % and a sufficient number of VA images for statistical testing would have required over 10,000 images in the control group. This setting would have been impractical for human testing. Therefore, a high prevalence of VA images was tolerated in the test.

Furthermore, the test dataset was created according to the relevant test parameters of microscopic villous atrophy and physiological mucosa, resulting in a non-matched dataset with a difference in mean age between the groups. Since the study was focused solely on the detection of villous atrophy on the endoscopic image, it is unlikely that age disparity impaired test validity.

It could be argued that the disclosure of the AI algorithm's performance on the test data set to endoscopists might have introduced a bias. However, disclosure of accurate information on AI performance was considered critical to establishing realistic testing conditions. To minimize a possible confounding effect, subjects were left unaware of the fact that the disclosed performance was derived from the test data. Furthermore, since results from internal cross validation and external validation were similar, a relevant confounding effect was unlikely.

In summary, AI significantly outperformed endoscopy fellows and experts in the detection of villous atrophy and showed stable diagnostic ability in images which were difficult for humans to assess. Further clinical studies are needed to evaluate this new technology in real life.

Acknowledgements:

This work was not supported by external funding.

References

[1] Singh P, Arora A, Strand TA, Leffler DA, Catassi C, Green PH, et al. Global Prevalence of Celiac Disease: Systematic Review and Meta-analysis. Clin Gastroenterol Hepatol. 2018;16:823-36 e2.
[2] Rubio-Tapia A, Ludvigsson JF, Brantner TL, Murray JA, Everhart JE. The prevalence of celiac disease in the United States. Am J Gastroenterol. 2012;107:1538-44; quiz 7, 45.
[3] Ludvigsson JF, Rubio-Tapia A, van Dyke CT, Melton LJ, 3rd, Zinsmeister AR, Lahr BD, et al. Increasing incidence of celiac disease in a North American population. Am J Gastroenterol. 2013;108:818-24.
[4] Ludvigsson JF, Murray JA. Epidemiology of Celiac Disease. Gastroenterol Clin North Am. 2019;48:1-18.
[5] Felber J, Bläker H, Fischbach W, Koletzko S, Laaß MW, Lachmann N, Lorenz P, Lynen P, Reese I, Scherf K, Schuppan D, Schumann M. Aktualisierte S2k-Leitlinie Zöliakie der Deutschen Gesellschaft für Gastroenterologie, Verdauungs- und Stoffwechselkrankheiten (DGVS). In: Deutsche Gesellschaft für Gastroenterologie V-uSD, editor.2021. p. 1-135.
[6] Barada K, Habib RH, Malli A, Hashash JG, Halawi H, Maasri K, et al. Prediction of celiac disease at endoscopy. Endoscopy. 2014;46:110-9.
[7] Schiepatti A, Cincotta M, Biagi F, Sanders DS. Enteropathies with villous atrophy but negative coeliac serology in adults: current issues. BMJ Open Gastroenterol. 2021;8.
[8] Dickey W. Endoscopic markers for celiac disease. Nat Clin Pract Gastroenterol Hepatol. 2006;3:546-51.
[9] Dickey W, Hughes D. Disappointing sensitivity of endoscopic markers for villous atrophy in a high-risk population: implications for celiac disease diagnosis during routine endoscopy. Am J Gastroenterol. 2001;96:2126-8.
[10] Marsh MN. Grains of truth: evolutionary changes in small intestinal mucosa in response to environmental antigen challenge. Gut. 1990;31:111-4.

[11] Oberhuber G, Granditsch G, Vogelsang H. The histopathology of coeliac disease: time for a standardized report scheme for pathologists. Eur J Gastroenterol Hepatol. 1999;11:1185-94.

[12] Fasano A, Berti I, Gerarduzzi T, Not T, Colletti RB, Drago S, et al. Prevalence of celiac disease in at-risk and not-at-risk groups in the United States: a large multicenter study. Arch Intern Med. 2003;163:286-92.

[13] Green PHR, Stavropoulos SN, Panagi SG, Goldstein SL, McMahon DJ, Absan H, et al. Characteristics of adult celiac disease in the USA: results of a national survey. Am J Gastroenterol. 2001;96:126-31.

[14] Hassan C, Wallace MB, Sharma P, Maselli R, Craviotto V, Spadaccini M, et al. New artificial intelligence system: first validation study versus experienced endoscopists for colorectal polyp detection. Gut. 2020;69:799-800.

[15] Shahidi N, Rex DK, Kaltenbach T, Rastogi A, Ghalehjegh SH, Byrne MF. Use of Endoscopic Impression, Artificial Intelligence, and Pathologist Interpretation to Resolve Discrepancies Between Endoscopy and Pathology Analyses of Diminutive Colorectal Polyps. Gastroenterology. 2020;158:783-5 e1.

[16] Ebigbo A, Mendel R, Probst A, Manzeneder J, Souza LA, Jr., Papa JP, et al. Computer-aided diagnosis using deep learning in the evaluation of early oesophageal adenocarcinoma. Gut. 2019;68:1143-5.

[17] Siau K, Beales ILP, Haycock A, Alzoubaidi D, Follows R, Haidry R, et al. JAG consensus statements for training and certification in oesophagogastroduodenoscopy. Frontline Gastroenterol. 2022;13:193-205.

[18] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[19] Rauber D, Mendel R, Scheppach M, Ebigbo A, Messmann H, Palm C. Analysis of Celiac Disease with Multimodal Deep Learning. In: Maier-Hein K, Deserno TM, Handels H, Maier A, Palm C, Tolxdorff T, editors. Bildverarbeitung f \ u r die Medizin 2022: Springer Fachmedien Wiesbaden; 2022. p. 115--20.

[20] Hawass NE. Comparing the sensitivities and specificities of two diagnostic procedures performed on the same group of patients. Br J Radiol. 1997;70:360-6.

[21] Gadermayr M, Kogler H, Karla M, Merhof D, Uhl A, Vecsei A. Computer-aided texture analysis combined with experts' knowledge: Improving endoscopic celiac disease diagnosis. World J Gastroenterol. 2016;22:7124-34.

[22] Wang X, Qian H, Ciaccio EJ, Lewis SK, Bhagat G, Green PH, et al. Celiac disease diagnosis from videocapsule endoscopy images with residual learning and deep feature extraction. Comput Methods Programs Biomed. 2020;187:105236.

[23] Stoleru CA, Dulf EH, Ciobanu L. Automated detection of celiac disease using Machine Learning Algorithms. Sci Rep. 2022;12:4071.

[24] Lohi S, Mustalahti K, Kaukinen K, Laurila K, Collin P, Rissanen H, et al. Increasing prevalence of coeliac disease over time. Aliment Pharmacol Ther. 2007;26:1217-25.

Tables and Table and Figure legends:

Table 1	Villous Atrophy (VA) Set		Control Set	
	Patients (N=182)	Images (N=858)	Patients (N=323)	Images (N=846)
< 18 yrs	119 (65.4%)	401 (46.7%)	34 (10.5%)	58 (6.9%)
> 18 yrs	63 (34.6%)	457 (53.3%)	289 (89.5%)	788 (93.1%)
Male	71 (39.0%)	319 (37.2%)	155 (48.0%)	419 (49.5%)
Female	111 (61.0)	539 (62.8%)	168 (52.0%)	427 50.5%)
WLI mode	751 (87.5%)		764 (90.3%)	
NBI mode	107 (12.5%)		82 (9.7%)	
Near focus mode	43 (5.0%)		52 (6.1%)	
Indigo-carmine staining	123 (14.3%)		19 (2.2%)	

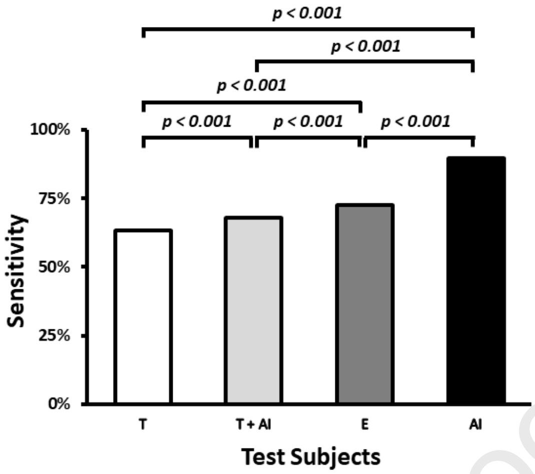
Table 1, Legend: Training data set: Characteristics of included patients and images; WLI, white light imaging; NBI, narrow band imaging; percentages are given based on the subsets (VA and control); in total 505 patients and 1704 images.

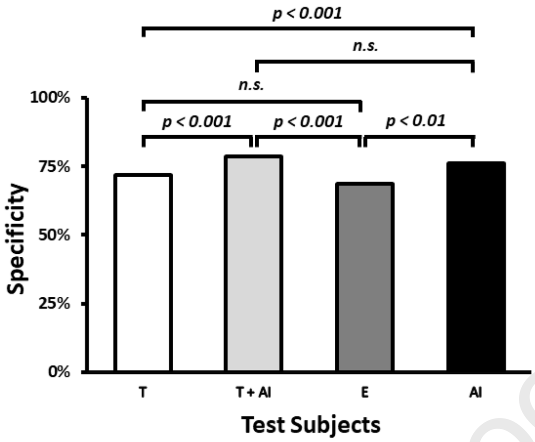
Table 2	Villous Atrophy (VA) Set		Control Set	
	Patients (N=63)	Images (N=194	Patients (N=65)	Images (N=155)
< 18 yrs	32 (50.8%)	89 (45.9%)	2 (3.1%)	9 (5.8%)
≥ 18 yrs	31 (49.2%)	105 (54.1%)	63 (96.9%)	146 (94.2%)
Mean age ± SD	28.4±23.8		46.4±19.1	
Median age	17		42	
Male	22 (34.9%)	68 (35.1%)	21 (32.3%)	49 (31.6%)
Female	41 (65.1%)	126 (64.9%)	44 (67.7%)	106 (68.4%)
WLI mode	190 (97.9%)		152 (98.1%)	
NBI mode	4 (2.1%)		3 (1.9%)	
Near focus mode	9 (4.6%)		6 (3.9%)	
Indigo-carmine staining	0 (0%)		0 (0%)	

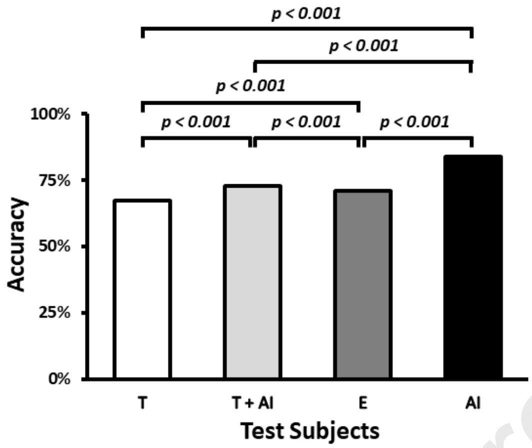
Table 2, Legend: External test data set: Characteristics of included patients and images; WLI, white light imaging; NBI, narrow band imaging; SD = standard deviation; percentages are given based on the subsets (VA and control), in total 128 patients and 349 images.

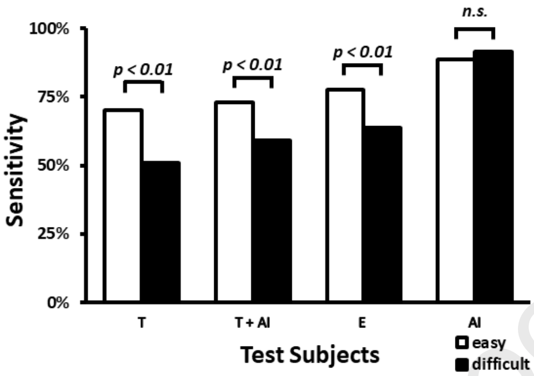
Legend of figure 1: Sensitivities (A), specificities (B) and accuracies (C) of the different groups in the evaluation by the external test set; T: Trainees; T + AI: Trainees with AI support, pooled result for all final diagnoses of all test images; E: Experts; AI: result of the AI algorithm.

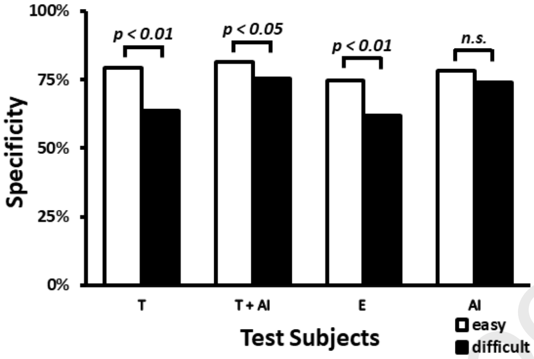
Legend of figure 2: Sensitivities (A), specificities (B) and accuracies (C) of the different groups and for the two subdivisions into “easy” and “difficult” images; easy: white columns; difficult: black columns; T: Trainees; T + AI: Trainees with AI support, pooled result for all final diagnoses of all test questions; E: Experts; AI: result of the AI algorithm.

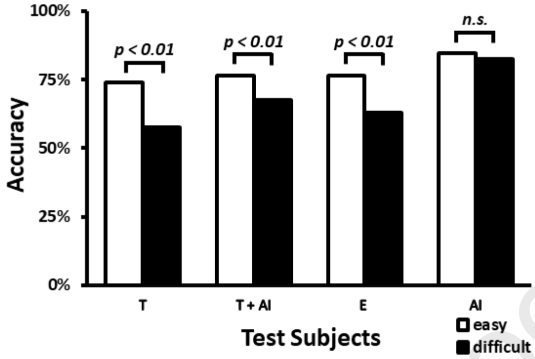

p < 0.001
p < 0.001
p < 0.001
p < 0.001
p < 0.001
p < 0.001
100%
75%
50%
25%
0%
Sensitivity
T
T + AI
E
AI
Test Subjects


p < 0.001
n.s.
n.s.
p < 0.001
p < 0.001
p < 0.01
100%
75%
50%
25%
0%
Specificity
T
T + AI
E
AI
Test Subjects


p < 0.001
p < 0.001
p < 0.001
p < 0.001
p < 0.001
p < 0.001
100%
75%
50%
25%
0%
Accuracy
T
T + AI
E
AI
Test Subjects


100%
75%
50%
25%
0%
Sensitivity
p < 0.01
p < 0.01
p < 0.01
n.s.
T
T + AI
E
AI
Test Subjects
□ easy
■ difficult


100%
75%
50%
25%
0%
Specificity
p < 0.01
p < 0.05
p < 0.01
n.s.
T
T + AI
E
AI
Test Subjects
□ easy
■ difficult


100%
75%
50%
25%
0%
Accuracy
p < 0.01
p < 0.01
p < 0.01
n.s.
T
T + AI
E
AI
Test Subjects
□ easy
■ difficult

Abbreviations: AI, artificial intelligence; AI-CDSS, artificial intelligence clinical decision support solution; EGD, esophagogastroduodenoscopy; VA, villous atrophy;