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

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

Augsburg, Regensburg, Jena, Germany

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

Methods: A dataset of 858 endoscopic images of 182 patients with VA and 846 images from 323 patients with normal duodenal mucosa was collected and used to train a ResNet18 deep learning model to detect VA. An external dataset was used to test the algorithm, in addition to 6 fellows and 4 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 corresponding values of 63%, 72%, and 67% and experts scored 72%, 69%, and 71%, respectively. AI consultation significantly improved all trainee performance statistics. Although fellows and experts showed significantly lower performance for difficult images, the performance of the AI algorithm was stable.

Conclusions: In this study, an AI algorithm outperformed endoscopy fellows and experts in the detection of VA on endoscopic still images. AI decision support significantly improved the performance of nonexpert endoscopists. The stable performance on difficult images suggests a further positive add-on effect in challenging cases. (Gastrointest Endosc 2023;97:911-6.)

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] Although the prevalence is reported to be rising, the disease continues to be under-reported,[2-4] and more than 50% of cases are undiagnosed worldwide. This seems to be because of its unspecific symptoms[5] and the endoscopic manifestation (small intestinal villous atrophy [VA]), which is often subtle and easily overlooked at inspection.[6]

VA is most often caused by celiac disease but can also occur in other disorders, such as tropical sprue or Whipple's

Abbreviations: AI, artificial intelligence; VA, villous atrophy.

DISCLOSURE: *All authors disclosed no financial relationships.*

*Drs Scheppach and Rauber contributed equally to this article.

0016-5107
https://doi.org/10.1016/j.gie.2023.01.006

Received September 28, 2022. Accepted January 1, 2023.

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

Reprint requests: Markus W. Scheppach, MD, Internal Medicine III–Gastroenterology, University Hospital of Augsburg, Stenglinstrasse 2, 86156 Augsburg, Germany.

disease.[7] Endoscopic markers of VA include a mosaic pattern and deep groves of the mucosa, scalloping, and, in severe cases, loss of duodenal folds and visible submucosal vessels and duodenal erosions.[8] At least 23% of histologically and serologically confirmed cases of VA and celiac disease showed no macroscopic signs of VA 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]

Because 70% of patients are diagnosed as adults[12] and the time between onset of symptoms and 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. 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 VA 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 GI disorders.[16] We therefore aimed to design a deep learning algorithm for the detection of VA on images of the duodenum and jejunum.

Various national guidelines give recommendations concerning the training of endoscopists in their respective countries. In Great Britain, the 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 VA is often made without supervision for the first time. This fact further suggests that an artificial intelligence (AI) clinical decision support solution for the detection of VA 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 VA 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.

From 2010 to 2021, 858 still images of the duodenum or jejunum from 182 patients with histologically confirmed VA (Marsh classification grade III)[10] were retrospectively extracted from the Augsburg University Hospital database. A further 846 images from 323 patients with macroscopically and histologically confirmed nonatrophic small-intestinal mucosa (control subjects) were extracted for the same period. Patients with known celiac disease on a gluten-free diet were excluded from the control dataset. Images were recorded during routine clinical practice using gastroscopes (GIF-HQ190, GIF-HQ180, or GIF-HQ1500; Olympus Medical Systems, Tokyo, Japan). At least 1 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 5 equal-sized subsets. Splitting the images from 1 patient into multiple subsets was avoided. To classify these images, a convolutional neural network was used as a model. This type of network consists of a sequence of convolutional and nonlinear layers. In this case the ResNet architecture was used.[18] The model uses so-called skip connections that allow the propagation of low-level features. For this project, a ResNet with 18 layers (ResNet18) was chosen.[19] This model was trained with the images of 4 subsets and then validated internally with the remaining subset (5-fold cross-validation). This process was repeated for each subset such that each subset was validated once. An additional external test dataset 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, noncardiac 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 gastroscopes (GIF-HQ190, GIF-HQ185, and GIF-HQ1500; Olympus Medical Systems).

The trained AI algorithm and 4 board-certified gastroenterologists (experts) with >1000 EGDs and 6 gastroenterology fellows (trainees) with an experience of 100 to 1000 EGDs were tested on the external test dataset. The mean endoscopic experience of trainees was 278 ± 173 examinations at the time of the study. A binary decision for a macroscopic suspicion of VA and subsequent indication for duodenal biopsy sampling was asked for each image. Trainees were given access to the results of the AI algorithm in which after documentation of their suspected diagnosis, trainees were allowed to consult the AI algorithm when 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 dataset in advance. For evaluation, trainees were regarded as 2 groups, once before the AI result could be consulted and once after optional consultation of the AI algorithm's result for all test questions. The test images were divided into 2 subcategories: "Easy" images were defined as images for

TABLE 1. Training dataset: characteristics of included patients and images

Characteristics	Villous atrophy set		Control set	
	Patients (n = 182)	Images (n = 858)	Patients (n = 323)	Images (n = 846)
Age <18 y	119 (65.4)	401 (46.7)	34 (10.5)	58 (6.9)
Age >18 y	63 (34.6)	457 (53.3)	289 (89.5)	788 (93.1)
Male sex	71 (39.0)	319 (37.2)	155 (48.0)	419 (49.5)
Female sex	111 (61.0)	539 (62.8)	168 (52.0)	427 50.5)
White-light imaging mode	751 (87.5)		764 (90.3)	
Narrow-band imaging 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)	

Values are n (%). Percentages are given based on the subsets (villous atrophy and control).

TABLE 2. External test dataset: characteristics of included patients and images

Characteristics	Villous atrophy set		Control set	
	Patients (n = 63)	Images (n = 194)	Patients (n = 65)	Images (n = 155)
Age <18 y	32 (50.8)	89 (45.9)	2 (3.1)	9 (5.8)
Age ≥18 y	31 (49.2)	105 (54.1)	63 (96.9)	146 (94.2)
Mean age ± standard deviation, y	28.4 ± 23.8		46.4 ± 19.1	
Median age, y	17		42	
Male sex	22 (34.9)	68 (35.1)	21 (32.3)	49 (31.6)
Female sex	41 (65.1)	126 (64.9)	44 (67.7)	106 (68.4)
White-light imaging mode	190 (97.9)		152 (98.1)	
Narrow-band imaging mode	4 (2.1)		3 (1.9)	
Near-focus mode	9 (4.6)		6 (3.9)	
Indigo carmine staining	0 (0)		0 (0)	

Values are n (%) unless otherwise defined. Percentages are given based on the subsets (villous atrophy and control).

which no or 1 trainee consulted the AI, whereas "difficult" images were defined as images for which ≥2 trainees consulted the AI algorithm.

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 and performance indices were compared between gastroenterologist experience levels and the 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 the Fisher exact test. Correction for multiple comparisons was performed by the Bonferroni method. A $P \leq .05$ was considered as statistically significant.

Ethics approval was obtained for the entire study from the Ethics Committee of Ludwig-Maximilians-University, Munich (project no. 21-1215) and for the external dataset from the Ethics Committee of Jena University Hospital (registration no. 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, respectively, 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 dataset are shown in Figure 1. All differences reached statistical significance except for the comparisons of specificities of trainees versus experts and trainees with AI support versus 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% of these cases (n = 185), 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% of these cases (n = 139), the decision change led to the correct diagnosis. In cases of agreement of the

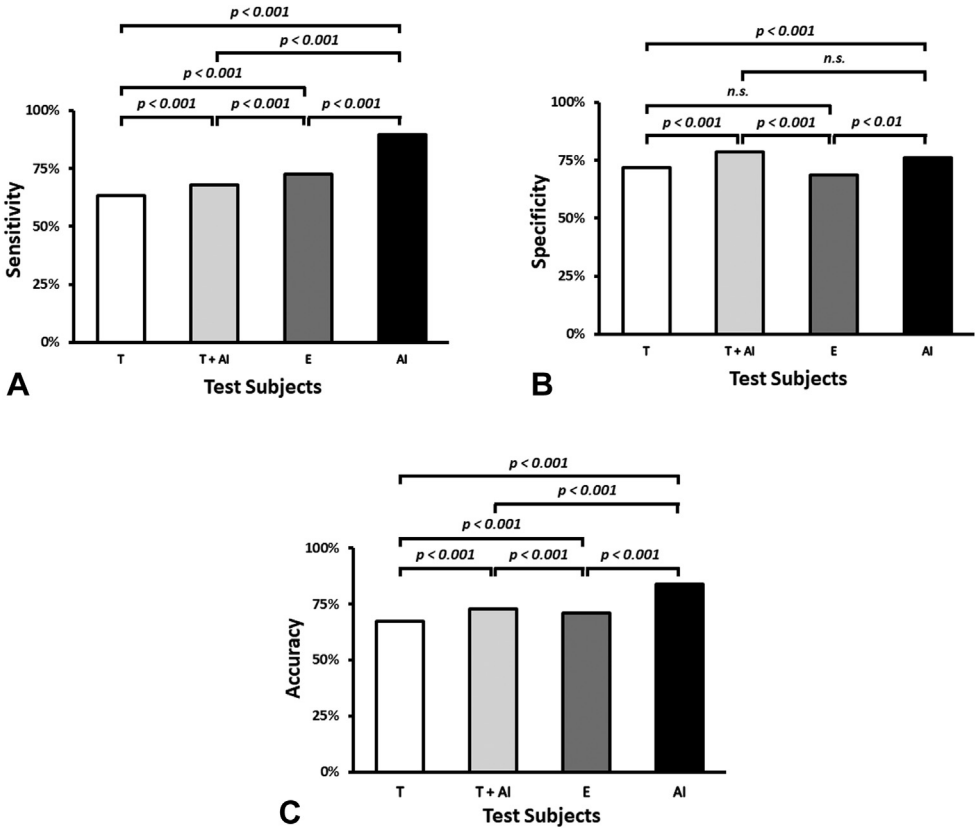

Figure 1. Sensitivities (**A**), specificities (**B**), and accuracies (**C**) of the different groups in the evaluation by the external test set. *T*, Trainees; *T + AI*, trainees with artificial intelligence support, pooled result for all final diagnoses of all test images; *E*, experts; *AI*, result of the artificial intelligence algorithm.

primary diagnosis and AI finding (58%, n = 253), the decision was kept in 97% of cases (n = 246), and agreement with the AI finding led to the right diagnosis in 79% of cases (n = 200).

Of all 349 test images, 30% (105 images) triggered no consultation of the AI by any of the 6 trainees. One trainee consulted the AI in 29% of the tests (101 images), 2 trainees consulted the AI in 28% of the tests (99 images), 3 trainees consulted the AI in 11% (37 images), and 4 consulted the AI in 2% of the tests (7 images). In no test image did 5 or all 6 trainees consult the AI algorithm. Hence, the test images were divided into 2 subcategories, easy and difficult, as defined above in Methods. Sensitivities, specificities, and accuracies for all groups after this subdivision are shown in Figure 2.

DISCUSSION

Detection of VA, in most cases of celiac disease, by AI has been attempted by different groups. Gadermayr et al[21] achieved an accuracy of 94% to 100% for the detection of VA during EGD using a combination of multiresolution local binary patterns, improved Fisher vectors, and a multifractal spectrum with expert knowledge. However, this technique requires water immersion of the duodenum, and the study was conducted in children. VA can also be detected on capsule endoscopy images with a high accuracy of over 90% using different forms of AI.[22,23] These studies were done in the setting of a high pretest probability or the clinical suspicion of celiac disease. Water immersion of the duodenum and capsule endoscopy are not routine examinations and are reserved for particular cases. The aim of the current study was to develop an application for routine EGD to support the endoscopist in making the incidental diagnosis of VA. Because celiac disease causes unspecific symptoms, false diagnoses such as gastritis or even irritable bowel syndrome 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 GI symptoms[12] and therefore are difficult to detect clinically. In this setting of low pretest probability for celiac disease, serology testing is rarely performed by clinicians. This suggests a potential benefit of an AI clinical decision support solution for the detection of VA and, consequently, celiac disease during routine EGD (ie, in cases where celiac disease is not a

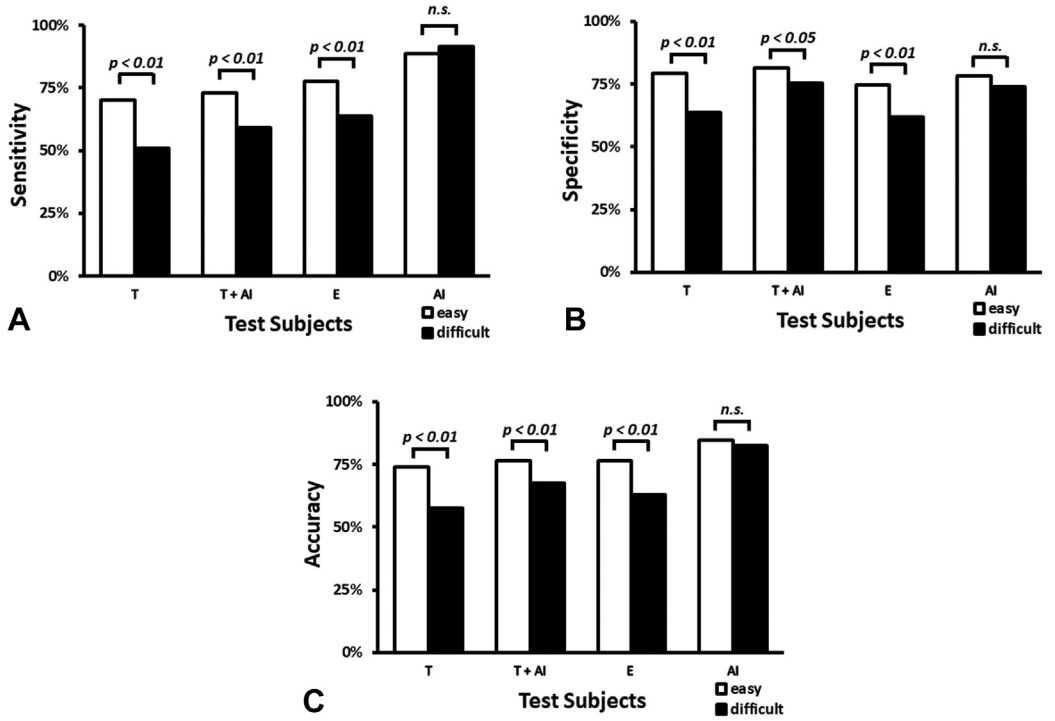

Figure 2. Sensitivities (**A**), specificities (**B**), and accuracies (**C**) of the different groups and for the 2 subdivisions into "easy" (*white columns*) and "difficult" (*black columns*) images. *T*, Trainees; *T + AI*, trainees with artificial intelligence support, pooled result for all final diagnoses of all test questions; *E*, experts; *AI*, result of the artificial intelligence algorithm.

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 by reducing the burden of advanced disease and may thus be cost-effective.

This study was designed to show a superiority of an AI algorithm over trainees in the detection of VA, 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 was 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.

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. AI classified images that were easy or difficult for endoscopists to assess with stable performance. Consequently, there may be parameters in the endoscopic image that 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 VA (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. Although the dataset was 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, because they are not visible on the macroscopic endoscopic image and were rarely found on biopsy samples 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 because of this circumstance. However, AI performance may also be improved on application to video data. The low number of test subjects calls into question if the results can be generalized. To circumvent this problem, we used statistical methods for low subject numbers (McNemar test) and a large test dataset (349 test images) for a more accurate measurement of the subjects' performance. A further limitation is the retrospective nature of the dataset, which might entail a lower image quality than is standard today, as well as a lack of standardization of image collection. However, nonconformity of images provides a more realistic dataset, 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 VA and physiologic mucosa, resulting in a nonmatched dataset with a difference in mean age between the groups. Because the study was focused solely on the detection of VA 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 dataset 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, because 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 VA and showed stable diagnostic ability in images that were difficult for humans to assess. Further clinical studies are needed to evaluate this new technology in real life.

ACKNOWLEDGMENT

Johannes Stallhofer was supported by the Interdisciplinary Center of Clinical Research (IZKF) of the Medical Faculty Jena (Advanced Clinician Scientist Program ACSP 05).

REFERENCES

1. Singh P, Arora A, Strand TA, et al. Global prevalence of celiac disease: systematic review and meta-analysis. Clin Gastroenterol Hepatol 2018;16:823-36.
2. Rubio-Tapia A, Ludvigsson JF, Brantner TL, et al. 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, 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, et al. Aktualisierte S2k-Leitlinie Zöliakie der Deutschen Gesellschaft für Gastroenterologie, Verdauungs- und Stoffwechselerkrankungen (DGVS). Z Gastroenterol 2022;60:790-856.
6. Barada K, Habib RH, Malli A, et al. Prediction of celiac disease at endoscopy. Endoscopy 2014;46:110-9.
7. Schiepatti A, Cincotta M, Biagi F, et al. Enteropathies with villous atrophy but negative coeliac serology in adults: current issues. BMJ Open Gastroenterol 2021;8:e000630.
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, 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, 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, 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, et al. 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.
16. 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.
17. Siau K, Beales ILP, Haycock A, et al. JAG consensus statements for training and certification in oesophagogastroduodenoscopy. Frontline Gastroenterol 2022;13:193-205.
18. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA, 2016. p. 770-8. Available at: https://ieeexplore.ieee.org/document/7780459. Accessed March 16, 2023.
19. Rauber D, Mendel R, Scheppach M, et al. Analysis of celiac disease with multimodal deep learning. In: Maier-Hein K, Deserno TM, Handels H, et al, eds. Bildverarbeitung fur 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, et al. 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, et al. Celiac disease diagnosis from video-capsule 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, et al. Increasing prevalence of coeliac disease over time. Aliment Pharmacol Ther 2007;26:1217-25.