GE – Portuguese Journal of Gastroenterology

Original Article

GE Port J Gastroenterol 2024;31:408–418
DOI: 10.1159/000539837

Received: February 20, 2024
Accepted: May 11, 2024
Published online: August 21, 2024

Deep Learning and Minimally Invasive Endoscopy: Panendoscopic Detection of Pleomorphic Lesions

Miguel Mascarenhas[a, b, c] Francisco Mendes[a, b] Tiago Ribeiro[a, b, c] João Afonso[a, b, c] Pedro Marílio Cardoso[a, b, c] Miguel Martins[a, b] Hélder Cardoso[a, b, c] Patrícia Andrade[a, b, c] João Ferreira[d, e] Miguel Mascarenhas Saraiva[f] Guilherme Macedo[a, b, c]

[a]Department of Gastroenterology, Precision Medicine Unit, São João University Hospital, Porto, Portugal; [b]WGO Gastroenterology and Hepatology Training Center, Porto, Portugal; [c]Faculty of Medicine of the University of Porto, Porto, Portugal; [d]Department of Mechanical Engineering, Faculty of Engineering of the University of Porto, Porto, Portugal; [e]Digestive Artificial Intelligence Development, Porto, Portugal; [f]ManopH Gastroenterology Clinic, Porto, Portugal

Keywords
Artificial intelligence · Capsule endoscopy · Deep learning · Panendoscopy

Abstract

Introduction: Capsule endoscopy (CE) is a minimally invasive exam suitable of panendoscopic evaluation of the gastrointestinal (GI) tract. Nevertheless, CE is time-consuming with suboptimal diagnostic yield in the upper GI tract. Convolutional neural networks (CNN) are human brain architecture-based models suitable for image analysis. However, there is no study about their role in capsule panendoscopy. ***Methods:*** Our group developed an artificial intelligence (AI) model for panendoscopic automatic detection of pleomorphic lesions (namely vascular lesions, protuberant lesions, hematic residues, ulcers, and erosions). 355,110 images (6,977 esophageal, 12,918 gastric, 258,443 small bowel, 76,772 colonic) from eight different CE and colon CE (CCE) devices were divided into a training and validation dataset in a patient split design. The model classification was compared to three CE experts' classification. The model's performance was evaluated by its sensitivity, specificity, accuracy, positive predictive value, negative predictive value, and area under the precision-recall curve. ***Results:*** The binary esophagus CNN had a diagnostic accuracy for pleomorphic lesions of 83.6%. The binary gastric CNN identified pleomorphic lesions with a 96.6% accuracy. The undenary small bowel CNN distinguished pleomorphic lesions with different hemorrhagic potentials with 97.6% accuracy. The trinary colonic CNN (detection and differentiation of normal mucosa, pleomorphic lesions, and hematic residues) had 94.9% global accuracy. ***Discussion/Conclusion:*** We developed the first AI model for panendoscopic automatic detection of pleomorphic lesions in both CE and CCE from multiple

Miguel Mascarenhas and Francisco Mendes contributed equally to this work.

karger@karger.com
www.karger.com/pjg

Correspondence to:
Miguel Mascarenhas, miguelmascarenhassaraiva@gmail.com

brands, solving a critical interoperability technological challenge. Deep learning-based tools may change the landscape of minimally invasive capsule panendoscopy.

Deep Learning e Endoscopia Minimamente Invasiva: Deteção panendoscópica de lesões pleomórficas

Palavras Chave
Deep learning · Endoscopia por cápsula · Inteligência artificial · Panendoscopia

Resumo
Introdução: A endoscopia por cápsula (EC) é um exame minimamente invasivo que avalia todo o trato gastrointestinal. Contudo, é morosa, com acuidade limitada no trato digestivo superior. As redes convolucionais neurais (RCN) são modelos baseados na arquitetura cerebral humana aperfeiçoados para análise de imagens. Contudo, o seu papel na panendoscopia por cápsula ainda não foi estudado. ***Métodos:*** Desenvolveu-se um modelo de inteligência artificial (IA) para deteção panendoscópica de lesões pleomórficas (nomeadamente lesões vasculares, protuberantes, resíduos hemáticos, úlceras e erosões). 355,110 imagens (6,977 esofágicas, 12,918 gástricas, 258,443 do intestino delgado e 76,772 colónicas) de oito dispositivos diferentes de enteroscopia e panendoscopia por cápsula foram divididas num *dataset* de treino e validação num desenho *patient split*. A classificação da RCN comparou-se com a de três especialistas em CE. O modelo foi avaliado através da sensibilidade, especificidade, valor preditivo positivo, valor preditivo negativo, acuidade e área sob curva *precision-recall*. ***Resultados:*** A RCN binária esofágica teve acuidade de 83.6% para lesões pleomórficas. A RCN binária para lesões gástricas pleomórficas teve acuidade de 96.6%. A RCN de 11 categorias de intestino delgado diferenciou lesões pleomórficas com diferente potencial hemorrágico com acuidade de 97.6%. A RCN trinaria colónica (mucosa normal, lesões pleomórficas e resíduos hemáticos) teve acuidade de 94.9%. ***Discussão/Conclusão:*** Desenvolveu-se o primeiro modelo de IA com elevada acuidade na deteção panendoscópica de lesões pleomórficas em dispositivos de enteroscopia e panendoscopia por cápsula, solucionando um desafio de interoperabilidade tecnológica. A utilização de modelos de deep learning pode alterar o panorama da panendoscopia por cápsula.

Introduction

Capsule endoscopy (CE) is a minimally invasive exam preconized in the study of the small bowel [1, 2], but capable of evaluating the entire gastrointestinal (GI) tract [3]. With the development of colon capsule endoscopy (CCE), CE-based panendoscopy is an evolving concept with the need to achieve a minimally invasive alternative for the evaluation of the entire GI tract [4, 5]. Whereas conventional esophagogastroduodenoscopy and colonoscopy are the current standards of care for evaluating the esophagic, gastric, and colonic mucosa, there is a need to consider the invasiveness of the exam, with a non-neglectable risk of complications like infection, bleeding, perforation and cardiopulmonary adverse events [6]. Moreover, the use of sedation techniques during the exam can increase costs related to the procedure and loss of working days by the patients [7], limiting the cost-effectiveness of both procedures in a screening setting.

Nevertheless, we must consider the intrinsic limitations of CE, especially in the upper GI tract. CE diagnostic performance for esophagic gastric lesions is still suboptimal. In fact, esophagus transit time is short and common pathologies are identified near the esophagogastric junction [8], with the scarcity of images affecting CE diagnostic yield. Regarding the stomach, the absence of insufflation in CE limits the observation of the complete structure, especially the more proximal region [9]. CE dependence of the peristaltic movements can also be a challenge concerning its diagnostic yield in the upper GI tract.

CE is a time-consuming exam, with reading times that can reach up to 120 min per exam [10]. The large number of frames produced by a single CE exam favors the use of artificial intelligence (AI) tools for image analysis. Convolutional neural networks (CNN) are a multi-layer architecture inspired by the human visual cortex, with high accuracy for imaging analysis, especially image pattern detection [11]. CNN models have been studied in several medical areas [12–14]. CE is the main focus of study for developing CNN-based technologies [15–17], augmenting its cost-effectiveness by increasing the diagnostic yield with a reduction in the reading time. Whereas there are several works about AI tools in CE for the evaluation of the small bowel [17, 18], colon [19], and even gastric mucosa [20], the role of this technology in the identification of esophageal lesions by CE is still to be explored. In this study, our group aimed to create the first worldwide AI-based model for panendoscopic (esophageal, gastric, enteric, and colonic) automatic detection of

pleomorphic lesions in a multi-device design, namely vascular lesions, hematic residues, protruding lesions, ulcers, and erosions.

Methods

Study Design

Our group aimed to develop an AI-based algorithm for the panendoscopic automatic detection of pleomorphic lesions including vascular lesions (red spots, angiectasia, and varices), xanthelasma, xanthomas, luminal blood, protruding lesions, ulcers, and erosions. This multicentric multi-device study was based on esophageal, gastric, small bowel, and colonic images obtained from eight different CE types (PillCam SB3™; PillCam SB1™; PillCam Crohn's™; PillCam Colon 1™, PillCam Colon 2™, MiroCam Capsule Endoscope™, Olympus Endocapsule™, OMOM HD Capsule Endoscopy System™) in two different centers (Centro Hospitalar Universitário São João and ManopH), comprising 5,846 CE exams in 4,372 patients between June of 2011 and December of 2022.

Our study was developed in a non-interventional fashion, respecting the Declaration of Helsinki, and was approved by the Ethics Committee of São João University Hospital/Faculty of Medicine of the University of Porto (No. CE 407/2020). Potentially identifying information of the subjects was omitted and each patient received a random number assignment in order to obtain effective data anonymization for researchers involved in the CNN network. The non-traceability of the data in conformity with general data protection regulation was ensured by a legal team with Data Protection Officer (DPO) certification (Maastricht University).

CE Protocol

CE procedures were conducted using eight different CE devices: the *PillCam* SB1™ system (Medtronic, Minneapolis, MN, USA), the *PillCam* SB3™ system (Medtronic, Minneapolis, MN, USA), the PillCam Colon 1™ (Medtronic, Minneapolis, MN, USA), the PillCam Colon 2™ (Medtronic, Minneapolis, MN, USA), the PillCam Crohn's™ (Medtronic, Minneapolis, MN, USA), the MiroCam Capsule Endoscope™ (IntroMedic, Seoul, Korea), the Olympus Endocapsule™ (Olympus, Tokyo, Japan), and the OMOM HD™ Capsule Endoscopy System (Jinshan Science & Technology Co., Chongqing, Yubei, China).

Images from PillCam SB3, PillCam SB1, PillCam Colon 2, and PillCam Crohn's CE were reviewed using the PillCam™ Software version 9 (Medtronic), whereas PillCam Colon 1 images were reviewed with an older version of PillCam™ software. The Olympus Endocapsule images were revised in the Endocapsule 10 System (Olympus). The MiroCam images are viewed in the MiroView Software (IntroMedic). The Vue Smart Software (Jinshan Science & Technology Co.) was used for reviewing the OMOM HD videos. After the removal of potential patient-identifying information, extracted frames were stored and labeled with a consecutive number.

Each patient underwent bowel preparation following previous recommendations by the European Society of Gastrointestinal Endoscopy [1]. Briefly, patients kept a clear liquid diet on the day before capsule ingestion, fasting the night before the exam. In the patients performing small bowel capsule endoscopy (SBCE), 2 L of polyethylene glycol solution was consumed before the exam. Simethicone was the chosen anti-foaming agent. 10 mg of domperidone was given to each patient as a prokinetic if the capsule remained in the stomach 1 h after ingestion (implying hourly image review on the data recorder worn by the patient). When performing CCE, a bowel preparation consisting of 4 L of polyethylene glycol solution was taken in split form (2 L in the evening before the exam and 2 L in the morning of the exam). Two boosters of 25 and 20 mL of a sodium phosphate solution were ingested when the capsule entered the small bowel and 3 h later.

Classification of Lesions

The different segments of each CE exam were reviewed for the identification of pleomorphic lesions. The pleomorphic lesions included vascular lesions (red spots, angiectasia, and varices), xanthomas, lymphangiectasias, protruding lesions, ulcers, and erosions. Our model was also evaluated for the detection of luminal blood. Classification scores used in SBCE were adapted for the definition of the different lesions [21]. Lymphangiectasias were considered white-colored points of the intestinal mucosa, while xanthomas were defined as yellowish plaque-like lesions.

Red spots were defined as flat punctuate lesions under 1 mm, with a bright red area, without vessel appearance [21]. Angiectasia consisted of reddish lesions of tortuous and dilated clustered capillaries. Varices were defined as raised serpiginous venous dilations. The subgroup of protruding lesions consisted of polyps, flat lesions, nodules and subepithelial lesions. Mucosal erosions were described as areas of minimal loss of epithelial layering with normal surrounding mucosa. Ulcers were defined as depressed loss of epithelial covering, with a whitish base and surrounding swollen mucosa, with an estimated diameter of >5 mm.

The lesions identified in the small bowel were classified into three levels of bleeding risk with the Saurin classification [22], with P0, P1, and P2 classification for absent, intermediate or high hemorrhagic risk, respectively. P0 lesions encompassed lymphangiectasia and xanthomas. P1 lesions comprised red spots, mucosal erosions, small ulcers and the majority of the protuberant lesions, whereas P2 classification encompassed angiectasia and varices, large ulcerations (>20 mm) and large (>10 mm) or ulcerated protuberant lesions. Three CE expert gastroenterologists, with an experience of over 1000 CE exams prior to the study, classified each of the extracted images.

CNN Development

The study design is displayed through a flowchart in Fig. 1. Table 1 displays the characteristics and methodological specificities of each CNN.

A total of 6,977 selected esophageal images were inserted in our CNN with transfer learning. The full esophageal dataset consisted of 3,920 images of normal mucosa and 3,057 images of esophageal lesions (namely vascular lesions, hematic residues, ulcers, erosions, and protuberant lesions). The images were divided into a training and validation dataset, in a patient split design (with all the images from a given patient allocated to the same dataset). The binary esophageal CNN (normal mucosa vs. pleomorphic lesions) was evaluated as the mean of the

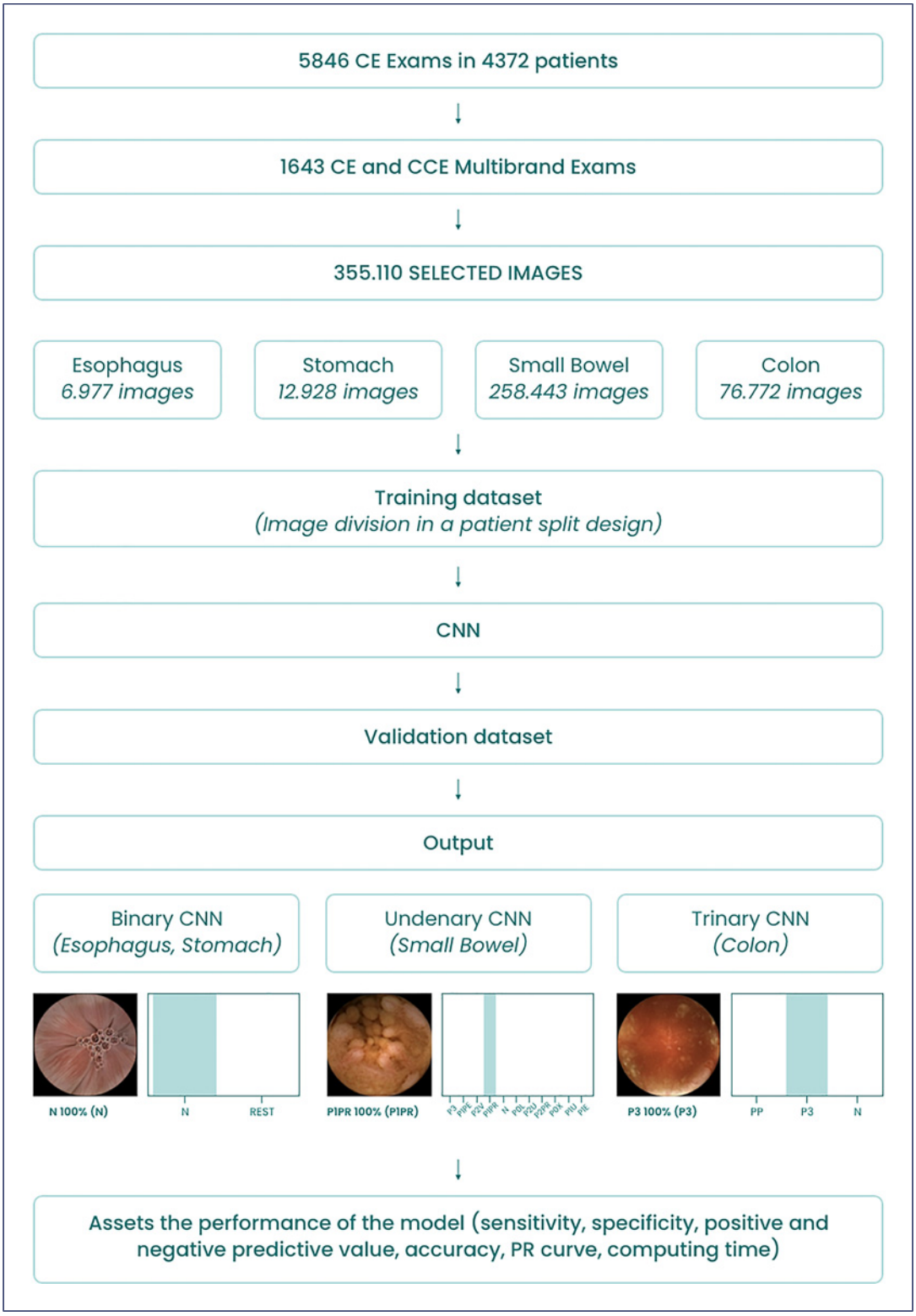

Fig. 1. Study flowchart for the training and validation phases. CE, capsule endoscopy; CCE, colon capsule endoscopy; N, normal mucosa; P3, hematic residues; P1PE, P1 red spots; P2V, P2 vascular lesions; P1PR, P1 protuberant lesions; P0L, P0 lymphangiectasia; P2U, P2 ulcers; P2PR, P2 protuberant lesions; P0X, xanthomas; P1U, P1 ulcers; P1E, P1 erosions; PP, pleomorphic lesions.

Table 1. CNN methodological characteristics in the different locations

CNN organ	No. of images	No. of exams	CE devices (No. of exams)	CNN evaluation
Esophagus	6,977	536	PillCam SB3 (270) PillCam Crohn's (207) OMOM (59)	The binary CNN (normal mucosa vs. pleomorphic lesions) was evaluated as the mean of the outcomes of three different validation dataset evaluation with different parameters
Stomach	12,918	107	PillCam SB3 (84) OMOM (14) PillCam Crohn'S (9)	The binary CNN (normal mucosa vs. pleomorphic lesions) was evaluated with the validation dataset (comprising around 10% of the total of images)
Small bowel	258,443	957	PillCam SB3 (724) OMOM (137) PillCam Crohn's (88) Colon 2 (3) MiroCam (2) PillCam SBI (2) Olympus (1)	The undenary CNN was evaluated with the validation dataset (comprising around of the total of images)
Colon	76,772	148	PillCam Crohn's (97) PillCam SB3 (25) PillCam colon 1 (17) PillCam Colon 2 (5) OMOM (4)	The trinary CNN (normal mucosa vs. pleomorphic lesions vs. hematic residues) was evaluated with the validation dataset

CNN, convolutional neural network.

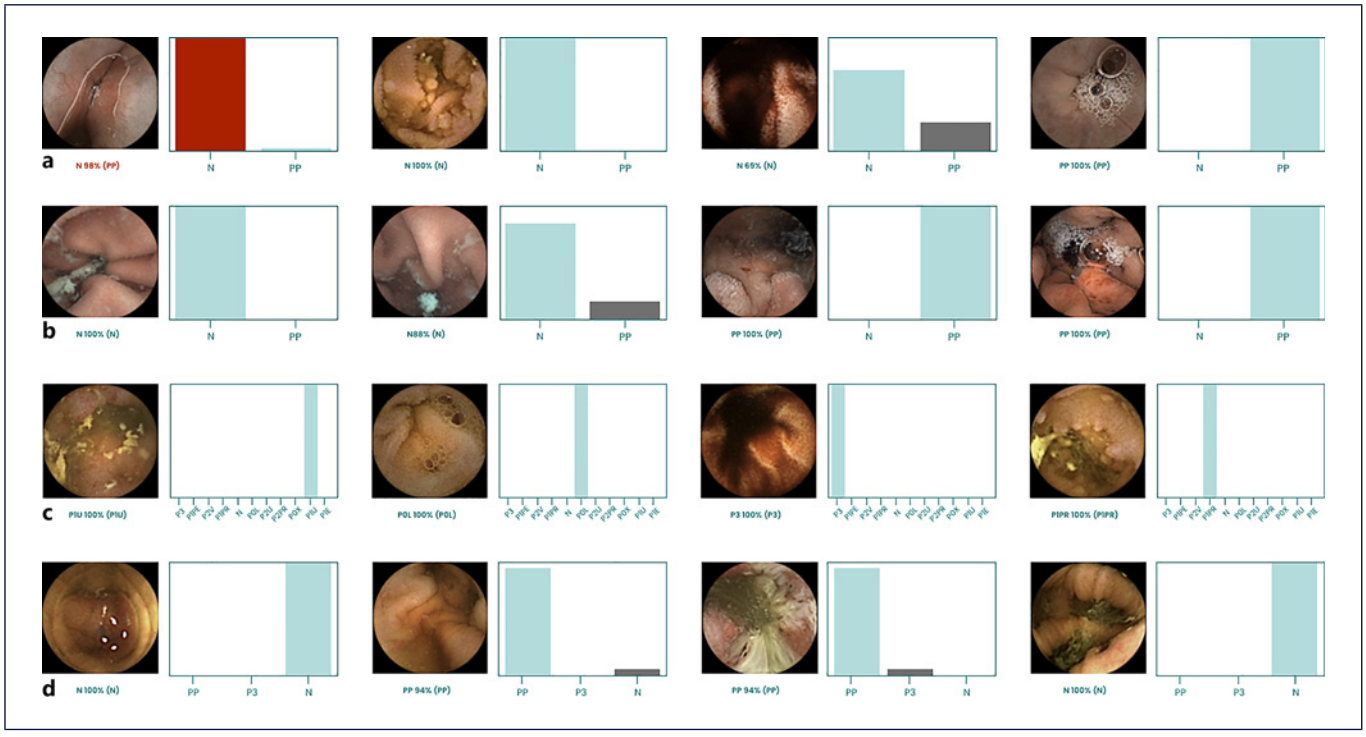

Fig. 2. Output obtained from the application of the CNN, for the pleomorphic lesions in diverse locations (esophagus [**a**]; stomach [**b**]; small bowel [**c**]; colon [**d**]). The bars represent the estimated probability by the CNN model. The finding with the highest probability was outputted as the predicted classification. The blue bars represent a correct prediction, whereas the red bars represent an incorrect prediction. N, normal mucosa; P3, hematic residues; P1PE, P1 red spots; P2V, P2 vascular lesions; P1PR, P1 protuberant lesions; P0L, P0 lymphangiectasia; P2U, P2 ulcers; P2PR, P2 protuberant lesions; P0X, xanthomas; P1U, P1 ulcers; P1E, P1 erosions; PP, pleomorphic lesions.

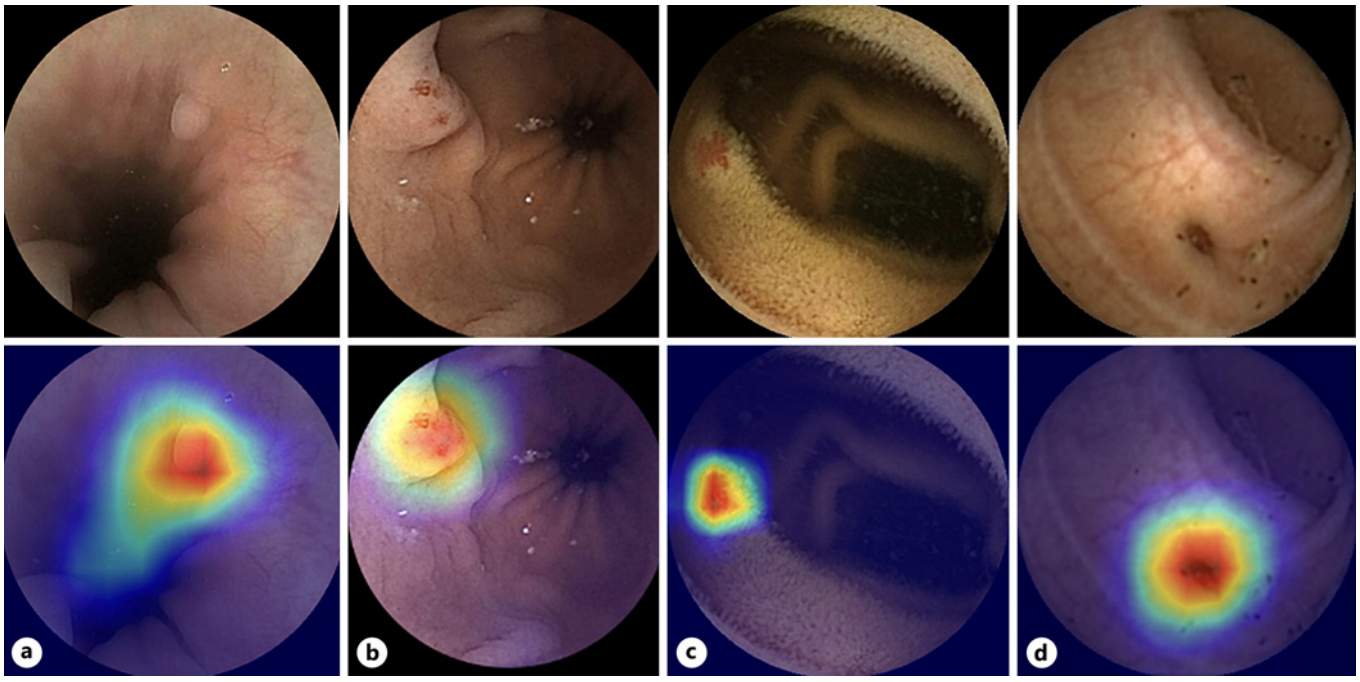

Fig. 3. Heatmaps obtained from the application of the CNN showing pleomorphic lesions in the esophagus (**a**), stomach (**b**), small bowel (**c**), and colon (**d**), as identified by the CNN.

outcomes of three different evaluations of the CNN with different model parameters.

Regarding the gastric CNN, 12,918 gastric images were used, including 6,844 normal images and 6,074 images of gastric lesions. The images were divided into a training (around 90% of the total images, n = 11,289) and validation dataset (around 10% of the total images, n = 1,629), in a patient split design. A 3-fold cross-validation was used in the development of the stomach CNN, with experimentation of different model parameters for obtaining the fittest model. The validation dataset was used to evaluate the performance of the model.

A total of 258,443 images were used for the construction of the small bowel dataset, with 62,792 normal images and 195,691 images of enteric lesions. The images were divided in a training (around 80% of the total images, n = 205,498) and validation dataset (around 20% of the total images, n = 52,945) in a patient split design. A 5-fold cross-validation was used to test the different model parameters and obtain the fittest model. The enteric CNN consisted of an undenary model, with a total of 11 categories, including normal mucosa, hematic residues and pleomorphic lesions with different hemorrhagic potential. The performance of the model was evaluated with the validation dataset.

Our colonic dataset contained 76,772 images, with 53,989 normal mucosa images, 3,918 images from hematic residues and 18,865 images from pleomorphic lesions. The images were divided into training (n = 72,438) and validation (n = 4,334) datasets, with the latter being used for the evaluation of the model. This CNN was evaluated as a trinary model, testing the CNN for distinguishing between normal mucosa, hematic residues and pleomorphic colonic lesions.

The Xception model pre-trained on ImageNet was used for the creation of the CNN. Convolutional layers of the model were kept in order to transfer this learning to our data, while the last fully connected layers were removed, and fully connected layers were attached based on the number of classes used to classify the CE images.

The model consisted of 2 blocks, comprising fully connected layers followed by a Dropout layer of 0.25 drop rate. Following these 2 blocks, a Dense layer with a size defined as the number of categories to classify was added. Our group set by trial and error a learning rate of 0.0001, batch size of 128, and the number of epochs of 20. We used Tensor-flow 2.3 and Keras libraries to prepare the data and run the model. The analyses were performed with a computer equipped with an Intel® Xeon® Gold 6130 processor (Intel, Santa Clara, CA, USA) and a NVIDIA Quadro® RTX™ 4000 graphic processing unit (NVIDIA Corporate, Santa Clara, CA, USA).

Performance Measures and Statistical Analysis

For a given image, the CNN model calculated the probability for each category (normal mucosa vs. pleomorphic lesions in the esophagus and stomach, normal mucosa vs. ten categories of lesions with different hemorrhagic potential in the small bowel, normal mucosa vs. pleomorphic lesions vs. hematic residues in the colon), with a given probability (Fig. 2), with higher probability values translating greater CNN prediction confidence. The software generated heatmaps identifying features that were the base of the prediction (Fig. 3). The CNN output was compared to the consensus classification by three CE experts', nowadays considered the gold standard for the evaluation of CE. The confusion matrix between experts and the CNN classification is presented in Table 2.

The primary performance measures included sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and accuracy (Table 3). These measures were represented with their

Table 2. Confusion matrix between experts and CNN classification

	Sn	sp	VPP	VPN	Acc	AUC-PR
Esophagus (*n* = 6,977)						
Training dataset mean	55.1 (50.4–59.4)	87.3 (85.3–89.1)	67.3 (65.0–69.5)	68.5 (67.2–69.7)	68.8 (66.7–70.8)	
Validation dataset	76.3 (71.1–80.9)	86.2 (80.5–90.9)	90.6 (86.5–93.5)	79.0 (73.7–83.2)	83.6 (79.9–86.9)	0.90
Stomach (*n* = 12,918)						
Training dataset mean	87.8 (86.9–88.7)	92.3 (91.6–93.0)	91.4 (90.6–92.0)	89.2 (88.5–89.9)	90.2 (89.6–90.7)	
Validation dataset	97.4 (96.0–98.4)	95.9 (94.4–97.1)	95.0 (93.3–96.3)	97.8 (96.7–98.6)	96.6 (95.6–97.4)	1.00
Small bowel (*n* = 258,443)						
Training dataset mean	81.5 (69.5–91.0)	98.5 (97.2–99.3)	82.6 (70.7–90.3)	98.5 (97.4–99.3)	97.5 (95.8–99.3)	
Validation dataset	78.6 (76.9–80.7)	97.6 (97.5–97.7)	72.5 (69.8–77.3)	99.2 (99.2–99.3)	97.6 (97.5–97.7)	
N versus ALL	85.2	98.4	89.9	96.6	96.5	0.95
POL versus ALL	95.2	98.0	59.5	99.8	98.0	0.91
POX versus ALL	93.1	99.8	91.6	99.9	99.7	0.97
PIPE versus ALL	87.9	99.4	82.1	99.6	99	0.93
PIPR versus ALL	96.1	99.5	99.1	97.8	98.3	1.00
PIU versus ALL	91.0	99.0	81.8	99.5	98.6	0.95
PIE versus ALL	83.1	99.5	80.0	99.6	99.1	0.91
P2V versus ALL	92.5	99.7	70.1	99.9	99.6	0.91
P2PR versus ALL	1.4	99.9	9.1	99.9	99.8	0.07
P2U versus ALL	42.3	99.8	63.0	99.6	99.5	0.55
P3 versus ALL	97.3	80.4	70.7	98.4	85.9	1.00
PO versus PI	99.1	97.5	80.6	99.9	97.6	
PO versus P2	99.9	99.5	99.9	99.7	99.8	
PO versus ALL	95.7	97.9	70.2	99.8	97.8	
PI versus P2	99.4	71.6	99.0	79.4	98.4	
PI versus ALL	95.6	96.8	96.2	96.3	96.3	
P2 versus ALL	68.0	99.5	70.9	99.5	99.0	
Colon (*n* = 76,772)						
Training dataset mean	86.5 (85.3–87.7)	93.0 (92.4–93.6)	87.6 (86.5–88.6)	94.4 (93.8–94.9)	92.5 (92.0–93.0)	
Validation dataset	85.7 (81.1–89.6)	94.0 (91.8–95.5)	87.4 (82.9–90.1)	93.5 (91.2–95.2)	94.9 (94.2–95.5)	
PP versus ALL	84.2	93.5	83.4	93.7	90.8	0.91
PP versus P3	99.4	93.6	99.6	90.7	99.1	
PP versus N	84.6	93.4	83.9	93.7	90.8	
P3 versus ALL	74.7	99.6	78.7	99.5	99.2	0.84
P3 versus N	78.7	99.6	85.5	99.4	99.1	
N versus ALL	93.1	84.4	93.2	84.1	90.4	0.98

CNN, convolutional neural network; Class, classification; N, normal mucosa; P3, hematic residues; P1PE, P1 red spots; P2V, P2 vascular lesions; P1PR, P1 protuberant lesions; P0L, P0 lymphangiectasia; P2U, P2 ulcers; P2PR, P2 protuberant lesions; P0X, xanthomas; P1U, P1 ulcers; P1E, P1 erosions; PP, pleomorphic lesions.

means and 95% confidence intervals (CI). The precision-recall (PR) curve and the area under the precision-recall curve (AUC-PR) were used to measure the performance of the model. Statistical analysis was performed using Sci-Kit learn version 0.22.2 [23].

Results

Esophagus

A total of 6,977 esophageal images from 536 CE exams in three different devices (PillCam SB3; PillCam Crohn's; OMOM HD capsule endoscopy system) were used for the development of the CNN. The esophagus CNN had a mean sensitivity of 76.3%, specificity of 86.2%, PPV of 90.6%, and NPV of 79.0%, with a mean accuracy of 83.6% and AUC-PR of 0.90. These results were achieved with an image processing time of 95 images per second.

Stomach

A total of 12,918 gastric images were obtained from a total of 107 CE exams in 3 different devices (PillCam SB3; PillCam Crohn's; OMOM HD capsule endoscopy

Table 3. CNN performance for panendoscopic automatic detection of pleomorphic lesions

	Sn	Sp	VPP	VPN	Acc	AUC-PR
Esophagus (*n* = 6,977)						
Training dataset mean	55.1 (50.4–59.4)	87.3 (85.3–89.1)	67.3 (65.0–69.5)	68.5 (67.2–69.7)	68.8 (66.7–70.8)	
Validation dataset	76.3 (71.1–80.9)	86.2 (80.5–90.9)	90.6 (86.5–93.5)	79.0 (73.7–83.2)	83.6 (79.9–86.9)	0.90
Stomach (*n* = 12,918)						
Training dataset mean	87.8 (86.9–88.7)	92.3 (91.6–93.0)	91.4 (90.6–92.0)	89.2 (88.5–89.9)	90.2 (89.6–90.7)	
Validation dataset	97.4 (96.0–98.4)	95.9 (94.4–97.1)	95.0 (93.3–96.3)	97.8 (96.7–98.6)	96.6 (95.6–97.4)	1.00
Small bowel (*n* = 258,443)						
Training dataset mean	81.5 (69.5–91.0)	98.5 (97.2–99.3)	82.6 (70.7–90.3)	98.5 (97.4–99.3)	97.5 (95.8–99.3)	
Validation dataset	78.6 (76.9–80.7)	97.6 (97.5–97.7)	72.5 (69.8–77.3)	99.2 (992–99.3)	97.6 (97.5–97.7)	
N versus ALL	85.2	98.4	89.9	96.6	96.5	0.95
POL versus ALL	952	98.0	59.5	99.8	98.0	0.91
POX versus ALL	93.1	99.8	91.6	99.9	99.7	0.97
PIPE versus ALL	87.9	99.4	82.1	99.6	99	0.93
PIPR versus ALL	9,611	99.5	99.1	97.8	98.3	1.00
PIU versus ALL	91.0	99.0	81.8	99.5	98.6	0.95
PIE versus ALL	83.1	99.5	80.0	99.6	99.1	0.91
P2V versus ALL	92.5	99.7	70.1	99.9	99.6	0.91
P2PR versus ALL	1.4	99.9	9.1	99.9	99.8	0.07
P2U versus ALL	42.3	99.8	63.0	99.6	99.5	0.55
P3 versus ALL	97.3	80.4	70.7	98.4	85.9	1.00
PO versus PI	99.1	97.5	80.6	99.9	97.6	
PO versus P2	99.9	99.5	99.9	99.7	99.8	
PO versus ALL	95.7	97.9	70.2	99.8	97.8	
PI versus P2	99.4	71.6	99.0	79.4	98.4	
PI versus ALL	95.6	96.8	96.2	96.3	96.3	
P2 versus ALL	68.0	99.5	70.9	99.5	99.0	
Colon (*n* = 76,772)						
Training dataset mean	86.5 (85.3–87.7)	93.0 (92.4–93.6)	87.6 (86.5–88.6)	94.4 (93.8–94.9)	92.5 (92.0–93.0)	
Validation dataset	85.7 (81.1–89.6)	(91.8–95.5)	87.4 (82.9–90.1)	93.5 (91.2–95.2)	94.9 (94.2–95.5)	
PP versus ALL	84.2	93.5	83.4	93.7	90.8	0.91
PP versus P3	99.4	93.6	99.6	90.7	99.1	
PP versus N	84.6	93.4	83.9	93.7	90.8	
P3 versus ALL	74.7	99.6	78.7	99.5	99.2	0.84
P3 versus N	78.7	99.6	85.5	99.4	99.1	
N versus ALL	93.1	84.4	93.2	84.1	90.4	0.98

CNN, convolutional neural network; Sn, sensitivity; Sp, specificity; PPV, positive predictive value; NPV, negative predictive value; Acc, accuracy; AUC-PR, area under precision-recall curve; N, normal mucosa; P3, hematic residues; P1PE, P1 red spots; P2V, P2 vascular lesions; P1PR, P1 protuberant lesions; P0L, P0 lymphangiectasia; P2U, P2 ulcers; P2PR, P2 protuberant lesions; P0X, xanthomas; P1U, P1 ulcers; P1E, P1 erosions; PP, pleomorphic lesions.

system). The CNN had a sensitivity of 97.4%, specificity of 95.9%, PPV of 95.0%, NPV of 97.8%, and global accuracy of 96.6%, with an AUC-PR of 1.00. The model achieved these results with an image processing time of 115 images per second.

Small Bowel

A total of 258,443 enteric images were obtained from 957 CE exams in seven different devices (PillCam SB3; PillCam SB1, PillCam Crohn's; PillCam Colon 2, OMOM HD capsule endoscopy system, MiroCam Capsule Endoscope™, Olympus Endocapsule™). The CNN revealed a global sensitivity of 78.6%, specificity of 97.6%, PPV of 72.5%, NPV of 99.2%, and accuracy of 97.6%.

When regarding the identification of specific enteric lesions, the model presented a global accuracy of 96.5% for normal mucosa, 98.0% for lymphangiectasias, and 99.7% for xanthomas. The model excelled for small protuberant lesions, ulcers, and erosions (98.3%, 98.6%, and 99.1% accuracy, respectively). The global accuracy

for vascular lesions, protuberant lesions, and ulcers with high hemorrhagic potential were 99.6%, 99.8%, and 99.5%, correspondingly. Nevertheless, the CNN presented lower sensitivities for diagnosing high-risk protuberant lesions and ulcers. These results are translated in the PR curves, with AUC-PR above 0.90 for the majority of the categories, excluding ulcers, and protuberant lesions with high hemorrhagic potential.

Finally, when considering the ability to distinguish between lesions with different hemorrhagic potentials, the CNN accurately differentiates P0 from P1 lesions (sensitivity 99.1%, specificity 97.4%, accuracy 97.6%), P0 from P2 lesions (sensitivity 99.9%, specificity 99.5%, and accuracy of 99.8%) and P1 from P2 lesions (sensitivity 99.4%, specificity 71.6%, and accuracy of 98.4%). These results were achieved with an image processing time of 282 images per second.

Colon

A total of 76,772 colonic images from 148 CE exams in five different devices (PillCam SB3; PillCam Crohn's; PillCam Colon 1, PillCam Colon 2, OMOM HD™ capsule endoscopy system) were used. The trinary CNN had a global sensitivity of 85.7%, specificity of 94.0%, PPV of 87.4%, NPV of 93.5% and global accuracy of 94.9%. The AUC-PR for normal mucosa, colonic blood and pleomorphic lesions was 0.98, 0.84, and 0.91, respectively. Furthermore, the CNN had an image processing time of 282 images per second.

Discussion

In this proof-of-concept study, our group developed the first AI model proficient in panendoscopic detection of pleomorphic lesions, in both SBCE and capsule panendoscopy devices. These results were accompanied by an image processing time that favors the clinical application of this CNN. Additionally, our group developed the first multi-device model for automatic detection of pleomorphic esophageal lesions in CE. Therefore, our group recognizes that AI-powered CE might change the landscape regarding the clinical applicability of minimally invasive capsule panendoscopy.

First, it's necessary to consider some methodologic points about the study. The division between training and validation in all the CNNs was performed in a patient split design, with all the images from a single patient included in the same dataset. This methodology significantly reduces the overfitting bias of the model (as the model would recognize similar images in the training and testing dataset). On the other side, our group preferred PR curves instead of the more common receiver operating characteristic (ROC) curves to assess the discriminating ability of the model as ROC curves reveal excessive optimism in the evaluation of model performance in cases of data imbalance [24, 25], with PR curves being less affected [26]. In our CNNs, the presence of normal mucosa images was commoner than pleomorphic lesions, thereby justifying the use of PR curves, given our objective of determining all the lesion images, instead of the commoner true negative images (implied in the ROC curve concept).

The interoperability challenge is one of the main points of interest in the discussion of the AI-based technology's role in Medicine [27, 28], with the generalization of a given technology in multiple devices as a requisite for the clinical applicability of an AI tool. Therefore, our group results in eight different CE devices, either in SBCE or capsule panendoscopy, solve the interoperability challenge with proof of diagnostic accuracy in different devices. This is, to our knowledge, not only the first panendoscopic CE CNN for detection of pleomorphic lesions but also the first capable of automatic detection in eight different CE devices, being the AI model with the largest representation of devices worldwide.

In recent years, CE-based panendoscopy has been a matter of discussion [29, 30], despite CCE is a time and resource-consuming exam, producing up to 50.000 image frames [31]. Additionally, despite numerous deep learning-based studies about small bowel and colon evaluation by CE [10, 16], there is a scarcity of studies about CNN models for esophagogastric evaluation in CE. Specific esophageal and gastric CNNs are of uttermost importance for increasing diagnostic accuracy while reducing the exam reading time and subjective bias in image evaluation by experts, which is pivotal for the implementation of minimally invasive panendoscopy.

Regardless, it is important to consider some intrinsic limitations of CE in the evaluation of the upper GI tract, which explains the different technology readiness levels (TRL) of the specific CNNs. The absence of air insufflation and dependence on abdominal peristalsis is associated with a scarcity of esophageal images and a reduction in stomach surface visualization, especially the cardia and fundus [9]. Recently, some works about CNN models for gastric evaluation have been published [32], inclusively with the use of magnetically controlled CE (MCE). However, our work is performed in much commoner CE devices (in both SBCE and CCE devices) devices and is methodologically stronger, with a patient split design that solves the overfitting problem.

On the other side, the diagnostic yield of CE in the esophagus is suboptimal, mainly because of the short transit time and scarcity of esophageal images, with a reduced number of lesion image frames [8]. The development of specific esophageal CE devices has partially overcome these limitations, without augmenting sufficiently the diagnostic yield [33]. The use of AI tools could increase the diagnostic yield of esophageal evaluation by CE. Our group developed the first multi-device CNN model for pleomorphic esophageal lesions detection, with good accuracy and image processing time.

The comprehension of the upper GI characteristics is important for the interpretation of the different CNN results. These specificities justify the lower TRLs of the esophageal and stomach CNN, with a lower number of images. However, the existence of an AI-based panendoscopy is dependent on specific gastric and esophageal models, assuring a high diagnostic yield in all GI tract locations.

This study has several limitations. First, it was performed in a retrospective manner. In the future, larger prospective multicentric studies are needed to study the clinical applicability of these technologies. Additionally, the results were based on the evaluation of still images, and studies with real-time evaluation of CE videos are needed in the future for the application of the AI model in a real-life scenario.

In conclusion, AI-based technologies might change the landscape of minimally invasive panendoscopic CE. To our knowledge, this is the first AI model capable of panendoscopic detection of pleomorphic lesions, with excellent image processing times, in both SBCE and CCE devices. This is the first study about CNN-based esophageal evaluation in CE. Additionally, this is the first study about CNN-based stomach evaluation in both SBCE and CCE devices. Furthermore, the AI model is the first to distinguish between several categories of small bowel lesions with different hemorrhagic potential in a patient split design, being also the first to excel in the diagnosis and differentiation of pleomorphic colonic lesions. The AI model was totally constructed in a patient split design, with a methodological advantage that reduces the overfitting bias of the model.

The application of these systems will improve the cost-effectiveness of a panendoscopy CE evaluation, increasing the diagnostic yield of the exam while reducing its time-consuming nature. In the future, larger real-time multicentric studies are needed for the development and application of these models.

Statement of Ethics

Our study was performed respecting the Declaration of Helsinki and was approved by the Ethics Committee of São João University Hospital/Faculty of Medicine of the University of Porto (No. CE 407/2020).

Conflict of Interest Statement

João Ferreira is a paid employee of DigestAID.

Funding Sources

The authors recognize NVIDIA support for the graphic unit acquisition.

Author Contributions

Miguel Mascarenhas and Francisco Mendes: equal contribution in study design, image extraction, drafting of the manuscript, and critical revision of the manuscript. Tiago Ribeiro and João Afonso: bibliographic review, image extraction, and critical revision of the manuscript. Pedro Marílio Cardoso and Miguel Martins: bibliographic review, image extraction, drafting of the manuscript, and critical revision of the manuscript. João Ferreira: construction and development of the CNN, statistical analysis and critical revision of the manuscript. Patrícia Andrade, Hélder Cardoso, Miguel Mascarenhas Saraiva, and Guilherme Macedo: study design and critical revision of the manuscript. All authors approved the final version of the manuscript.

Data Availability Statement

There is a limitation on data sharing due to intellectual property concerns as well as regarding non-traceability and anonymization of potentially identifying information of patients' data.

References

1 Pennazio M, Rondonotti E, Despott EJ, Dray X, Keuchel M, Moreels T, et al. Small-bowel capsule endoscopy and device-assisted enteroscopy for diagnosis and treatment of small-bowel disorders: European Society of Gastrointestinal Endoscopy (ESGE) Guideline - update 2022. Endoscopy. 2023;55(1):58–95. https://doi.org/10.1055/a-1973-3796
2 Le Berre C, Trang-Poisson C, Bourreille A. Small bowel capsule endoscopy and treat-to-target in Crohn's disease: a systematic review. World J Gastroenterol. 2019;25(31):4534–54. https://doi.org/10.3748/wjg.v25.i31.4534
3 Iddan G, Meron G, Glukhovsky A, Swain P. Wireless capsule endoscopy. Nature. 2000;405(6785):417. https://doi.org/10.1038/35013140

4 Rondonotti E, Pennazio M. Colon capsule for panendoscopy: a narrow window of opportunity. Endosc Int Open. 2021;9(12):E1860–62. https://doi.org/10.1055/a-1548-6572
5 Vuik FER, Moen S, Spaander MCW. Colon capsule endoscopy as panendoscopy: using current knowledge to enhance possibilities. Endosc Int Open. 2022;10(5):E584. https://doi.org/10.1055/a-1785-4810
6 Levy I, Gralnek IM. Complications of diagnostic colonoscopy, upper endoscopy, and enteroscopy. Best Pract Res Clin Gastroenterol. 2016;30(5):705–18. https://doi.org/10.1016/j.bpg.2016.09.005
7 Helmers RA, Dilling JA, Chaffee CR, Larson MV, Narr BJ, Haas DA, et al. Overall cost comparison of gastrointestinal endoscopic procedures with endoscopist- or anesthesia-supported sedation by activity-based costing techniques. Mayo Clin Proc Innov Qual Outcomes. 2017;1(3):234–41. https://doi.org/10.1016/j.mayocpiqo.2017.10.002
8 Park J, Cho YK, Kim JH. Current and future use of esophageal capsule endoscopy. Clin Endosc. 2018;51(4):317–22. https://doi.org/10.5946/ce.2018.101
9 Kim JH, Nam SJ. Capsule endoscopy for gastric evaluation. Diagnostics. 2021;11(10):1792. https://doi.org/10.3390/diagnostics11101792
10 Piccirelli S, Mussetto A, Bellumat A, Cannizzaro R, Pennazio M, Pezzoli A, et al. New generation express view: an artificial intelligence software effectively reduces capsule endoscopy reading times. Diagnostics. 2022;12(8):1783. https://doi.org/10.3390/diagnostics12081783
11 Richards BA, Lillicrap TP, Beaudoin P, Bengio Y, Bogacz R, Christensen A, et al. A deep learning framework for neuroscience. Nat Neurosci. 2019;22(11):1761–70. https://doi.org/10.1038/s41593-019-0520-2
12 Khurshid S, Friedman S, Reeder C, Di Achille P, Diamant N, Singh P, et al. ECG-based deep learning and clinical risk factors to predict atrial fibrillation. Circulation. 2022;145(2):122–33. https://doi.org/10.1161/CIRCULATIONAHA.121.057480
13 Islam MM, Poly TN, Walther BA, Yeh CY, Seyed-Abdul S, Li YJ, et al. Deep learning for the diagnosis of esophageal cancer in endoscopic images: a systematic review and meta-analysis. Cancers. 2022;14(23):5996. https://doi.org/10.3390/cancers14235996
14 Wu X, Chen D. Convolutional neural network in microsurgery treatment of spontaneous intracerebral hemorrhage. Comput Math Methods Med. 2022;2022:9701702. https://doi.org/10.1155/2022/9701702
15 Mascarenhas M, Afonso J, Andrade P, Cardoso H, Macedo G. Artificial intelligence and capsule endoscopy: unravelling the future. Ann Gastroenterol. 2021;34(3):300–9. https://doi.org/10.20524/aog.2021.0606
16 Soffer S, Klang E, Shimon O, Nachmias N, Eliakim R, Ben-Horin S, et al. Deep learning for wireless capsule endoscopy: a systematic review and meta-analysis. Gastrointest Endosc. 2020;92(4):831–9.e8. https://doi.org/10.1016/j.gie.2020.04.039
17 Chu Y, Huang F, Gao M, Zou DW, Zhong J, Wu W, et al. Convolutional neural network-based segmentation network applied to image recognition of angiodysplasias lesion under capsule endoscopy. World J Gastroenterol. 2023;29(5):879–89. https://doi.org/10.3748/wjg.v29.i5.879
18 Mascarenhas Saraiva MJ, Afonso J, Ribeiro T, Ferreira J, Cardoso H, Andrade AP, et al. Deep learning and capsule endoscopy: automatic identification and differentiation of small bowel lesions with distinct haemorrhagic potential using a convolutional neural network. BMJ Open Gastroenterol. 2021;8(1):e000753. https://doi.org/10.1136/bmjgast-2021-000753
19 Mascarenhas M, Ribeiro T, Afonso J, Ferreira JPS, Cardoso H, Andrade P, et al. Deep learning and colon capsule endoscopy: automatic detection of blood and colonic mucosal lesions using a convolutional neural network. Endosc Int Open. 2022;10(2):E171–7. https://doi.org/10.1055/a-1675-1941
20 Xia J, Xia T, Pan J, Gao F, Wang S, Qian YY, et al. Use of artificial intelligence for detection of gastric lesions by magnetically controlled capsule endoscopy. Gastrointest Endosc. 2021;93(1):133–9.e4. https://doi.org/10.1016/j.gie.2020.05.027
21 Leenhardt R, Li C, Koulaouzidis A, Cavallaro F, Cholet F, Eliakim R, et al. Nomenclature and semantic description of vascular lesions in small bowel capsule endoscopy: an international Delphi consensus statement. Endosc Int Open. 2019;7(3):E372–9. https://doi.org/10.1055/a-0761-9742
22 Saurin JC, Delvaux M, Gaudin JL, Fassler I, Villarejo J, Vahedi K, et al. Diagnostic value of endoscopic capsule in patients with obscure digestive bleeding: blinded comparison with video push-enteroscopy. Endoscopy. 2003;35(7):576–84. https://doi.org/10.1055/s-2003-40244
23 Pedregosa FVG, Gramfort A, Mi- chel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12 2825–30.
24 Movahedi F, Padman R, Antaki JF. Limitations of receiver operating characteristic curve on imbalanced data: assist device mortality risk scores. J Thorac Cardiovasc Surg. 2023;165(4):1433–42.e2. https://doi.org/10.1016/j.jtcvs.2021.07.041
25 Halligan S, Altman DG, Mallett S. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. Eur Radiol. 2015;25(4):932–9. https://doi.org/10.1007/s00330-014-3487-0
26 Fu GH, Yi LZ, Pan J. Tuning model parameters in class-imbalanced learning with precision-recall curve. Biom J. 2019;61(3):652–64. https://doi.org/10.1002/bimj.201800148
27 Bazoukis G, Hall J, Loscalzo J, Antman EM, Fuster V, Armoundas AA. The inclusion of augmented intelligence in medicine: a framework for successful implementation. Cell Rep Med. 2022;3(1):100485. https://doi.org/10.1016/j.xcrm.2021.100485
28 He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med. 2019;25(1):30–6. https://doi.org/10.1038/s41591-018-0307-0
29 Cortegoso Valdivia P, Elosua A, Houdeville C, Pennazio M, Fernandez-Urien I, Dray X, et al. Clinical feasibility of panintestinal (or panenteric) capsule endoscopy: a systematic review. Eur J Gastroenterol Hepatol. 2021;33(7):949–55. https://doi.org/10.1097/MEG.0000000000002200
30 Chetcuti Zammit S, Sidhu R. Capsule endoscopy - recent developments and future directions. Expert Rev Gastroenterol Hepatol. 2021;15(2):127–37. https://doi.org/10.1080/17474124.2021.1840351
31 Eliakim R, Yassin K, Niv Y, Metzger Y, Lachter J, Gal E, et al. Prospective multicenter performance evaluation of the second-generation colon capsule compared with colonoscopy. Endoscopy. 2009;41(12):1026–31. https://doi.org/10.1055/s-0029-1215360
32 Afonso J, Mascarenhas M, Ribeiro T, Cardoso P, Andrade A, Ferreira J, et al. S594 development and validation of a convolutional neural network for the automatic detection of multiple gastric lesions in multi-brand capsule endoscopy videos: a pilot study. Am J Gastroenterol. 2022;117(10S):e418–e419. https://doi.org/10.14309/01.ajg.0000859016.17459.c6
33 Duvvuri A, Desai M, Vennelaganti S, Higbee A, Gorrepati VS, Dasari C, et al. Diagnostic accuracy of a novel third generation esophageal capsule as a non-invasive detection method for Barrett's esophagus: a pilot study. J Gastroenterol Hepatol. 2021;36(5):1222–5. https://doi.org/10.1111/jgh.15283