

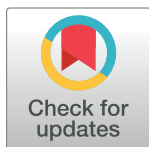
RESEARCH ARTICLE

# Accuracy of artificial intelligence-assisted endoscopy in the diagnosis of gastric intestinal metaplasia: A systematic review and meta-analysis

Na Li, Jian Yang, Xiaodong Li, Yanting Shi, Kunhong Wang \*

Department of Gastroenterology, Zibo Central Hospital, Zibo, Shandong, China

\* [kunhongwang@hotmail.com](mailto:kunhongwang@hotmail.com)



## Abstract

### OPEN ACCESS

**Citation:** Li N, Yang J, Li X, Shi Y, Wang K (2024) Accuracy of artificial intelligence-assisted endoscopy in the diagnosis of gastric intestinal metaplasia: A systematic review and meta-analysis. PLoS ONE 19(5): e0303421. <https://doi.org/10.1371/journal.pone.0303421>

**Editor:** Chih-Wei Tseng, Dalin Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, TAIWAN

**Received:** January 3, 2024

**Accepted:** April 25, 2024

**Published:** May 14, 2024

**Copyright:** © 2024 Li et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** All relevant data are within the manuscript and its [Supporting Information](#) files.

**Funding:** The author(s) received no specific funding for this work.

**Competing interests:** The authors have declared that no competing interests exist.

**Abbreviations:** AI, Artificial Intelligence; AUC, Area Under the SROC curve; BLI, Blue-Laser Imaging; CI, Confidence Interval; DOR, Diagnostic Odds

## Background and aims

Gastric intestinal metaplasia is a precancerous disease, and a timely diagnosis is essential to delay or halt cancer progression. Artificial intelligence (AI) has found widespread application in the field of disease diagnosis. This study aimed to conduct a comprehensive evaluation of AI's diagnostic accuracy in detecting gastric intestinal metaplasia in endoscopy, compare it to endoscopists' ability, and explore the main factors affecting AI's performance.

## Methods

The study followed the PRISMA-DTA guidelines, and the PubMed, Embase, Web of Science, Cochrane, and IEEE Xplore databases were searched to include relevant studies published by October 2023. We extracted the key features and experimental data of each study and combined the sensitivity and specificity metrics by meta-analysis. We then compared the diagnostic ability of the AI versus the endoscopists using the same test data.

## Results

Twelve studies with 11,173 patients were included, demonstrating AI models' efficacy in diagnosing gastric intestinal metaplasia. The meta-analysis yielded a pooled sensitivity of 94% (95% confidence interval: 0.92–0.96) and specificity of 93% (95% confidence interval: 0.89–0.95). The combined area under the receiver operating characteristics curve was 0.97. The results of meta-regression and subgroup analysis showed that factors such as study design, endoscopy type, number of training images, and algorithm had a significant effect on the diagnostic performance of AI. The AI exhibited a higher diagnostic capacity than endoscopists (sensitivity: 95% vs. 79%).

Ratio; GIM, Gastric Intestinal Metaplasia; LCI, Linked-Color Imaging; ME-NBI, Magnifying Endoscopy with NBI; NBI, Narrow-Band Imaging; NLR, Negative Likelihood Ratio; PLR, Positive Likelihood Ratio; SROC, Summary Receiver Operating Characteristic; WLI, White-Light Imaging.

## Conclusions

AI-aided diagnosis of gastric intestinal metaplasia using endoscopy showed high performance and clinical diagnostic value. However, further prospective studies are required to validate these findings.

## Introduction

Gastric cancer ranks fifth in terms of global cancer prevalence, posing a serious threat to human health [1]. Although the incidence of gastric cancer has decreased over the past three decades, the absolute number of cases continues to rise due to an aging population and a shift towards younger age groups developing gastric cancer. Hence, reducing the incidence and mortality of gastric cancer remains an urgent issue [1].

The progression of the most gastric cancers is a cascade pattern, which includes gastritis, atrophic gastritis (AG), intestinal metaplasia, heterogeneous hyperplasia, lastly culminating in cancer [2, 3]. AG and gastric intestinal metaplasia (GIM) are important intermediate- and high-risk factors for the development of gastric cancer. Early detection of these lesions is essential for delaying or halting the development of gastric cancer. In clinical practice, white light endoscopy is typically used to observe gastric lesions. However, studies have shown that the correlation between histology and general white light endoscopy diagnosis is low [4–8].

In the last decade, artificial intelligence (AI) has garnered significant attention within the scientific community, leading to considerable research being conducted on AI-related subjects, such as neural networks, machine learning, and deep learning. It has been used in various industries to provide powerful solutions to complex problems [9–11]. Computer vision is an important research area in AI. By applying various algorithms, computer vision systems can analyze and extract meaningful information from images or videos. Image classification algorithms are used to identify the category to which an image belongs, represented by VGG [12], ResNet [13], TRResNet [14], SE-ResNet [15], and EfficientNet [16]. Object detection algorithms focus on finding one or more targets in an image and framing them with rectangular boxes; typical algorithms are SSD [17], YOLO [18, 19], and R-CNN [20]. Semantic segmentation algorithms identify each pixel in the image and is capable of accurate segmentation based on the boundary of the target; typical algorithms are UNet++ [21], DeepLab [22, 23], and BiSeNet [24]. These techniques are widely used in medical imaging diagnoses [25, 26].

In gastrointestinal endoscopy, AI has been used to diagnose various diseases [27–29] and has achieved good diagnostic efficacy. Bang et al. [30] performed a meta-analysis including eight studies that specifically examined the accuracy of AI-assisted endoscopy in the diagnosis of *Helicobacter pylori* infection. Our previous study [31] conducted a meta-analysis on the accuracy of AI-assisted endoscopy in diagnosing chronic atrophic gastritis. In this study, we utilized meta-analysis to evaluate the accuracy of AI in diagnosing GIM, explored the main factors affecting AI's ability, and compared AI performance with that of endoscopists, thereby providing an objective basis for applying AI in clinical diagnosis.

## Methods

Before commencing the study, we had registered it with PROSPERO [32] (ID: CRD42022378974). This study strictly followed the PRISMA-DTA [33] guidelines. The associated checklist for PRISMA-DTA can be found in [S1 Table](#).

## Searching strategy

To obtain relevant studies, we searched the following five databases from their establishment up to October 2023: PubMed, Embase, Web of Science, Cochrane Library, and IEEE Xplore. Notably, PubMed, Embase and Cochrane Library are common medical databases, while the Web of Science is an extensive and comprehensive database. The IEEE Xplore database covers computer science, electronics, and other related fields. Related search terms include: artificial intelligence, deep learning, machine learning, computer-aided diagnosis, neural networks, gastritis, gastric precancerous, gastric tissue, and intestinal metaplasia. The detailed search strategy is shown in [S2 Table](#).

## Study selection

Inclusion criteria: (a) Studies use AI technology to analyze endoscopic images/videos to detect GIM lesions. (b) Ability to extract 2x2 table data from articles. (c) Clear presentation of diagnostic criteria. (d) A clear description of the AI algorithm and the process of diagnosing GIM. (e) The most recent studies from multiple studies on the same research group, if the AI model or study cohort was the same. Exclusion criteria: (a) Studies in which the full text was unavailable. (b) Studies in which complete four-grid table data were unavailable. (c) Reviews, meta-analyses, editorial reviews, letters to the editor, conference abstracts, and other types of literature. Two authors (J.Y. and X.L.) independently evaluated the search results, and any disagreements were resolved through discussion.

## Data extraction

The key data we extracted from each study included the first author, publication year, country/region, study design, study center, diagnostic criteria, algorithm, number of training set samples, test set type, number of test set samples, and 2x2 table data. Two authors (Y.S. and X.L.) independently extracted the data by reading the full text, and disagreements were resolved through discussion.

## Quality assessment

QUADAS-2 [34] is the widely used quality assessment tool for diagnostic accuracy studies, and includes four parts: Patient Selection, Index Testing, Reference Standards, and Flow and Timing. However, QUADAS-2 is not fully applicable to AI-centered diagnostic accuracy studies [35, 36]; therefore, we supplemented QUADAS-2 to make it more suitable for AI-centered studies. In the patient selection section, the source, size, and quality of the input data were accurately described. In index testing, whether the AI model is tested using an independent test set. In the reference standard section, whether pathological tissue biopsies were used as the “Gold Standard” is described.

## Statistical analysis

Based on a bivariate mixed-effects model, we calculated diagnostic performance indicators such as combined sensitivity, specificity, and diagnostic odds ratio (DOR). The likelihood ratio is a composite index that reflects sensitivity and specificity. The positive likelihood ratio (PLR)  $> 10$  and the negative likelihood ratio (NLR)  $< 0.1$  indicate high diagnostic performance. The area under the curve (AUC) and DOR are comprehensive measures to evaluate diagnostic accuracy.  $AUC \geq 0.9$  indicates the high accuracy of the diagnostic test. A larger DOR value indicated a better diagnostic performance.

The heterogeneity of the studies was assessed by the visual inspection of summary receiver operating characteristic and forest plots and counted by the  $I^2$  value. If heterogeneity existed, a subgroup analysis and meta-regression were performed. The clinical applicability was assessed using Fagan plots. Deek's funnel plot assessed publication bias, and when the angle between the straight line in the plot and the coordinate X-axis was closer to  $90^\circ$ , it indicated the existence of publication bias. When  $P < 0.05$  was statistically significant, publication bias was present.

Quality assessment of the included studies was performed using Review Manager 5.4 (Cochrane Collaboration, Oxford, UK). Other statistical analyses and graphing were conducted using Stata/SE16.0 (Stata, TX, USA).

## Result

### Included studies

The final search was conducted on October 12, 2023, yielding 637 papers. Among these, 228 duplicates were automatically removed through EndNote, 381 irrelevant papers were excluded by reading the titles and abstracts, two were excluded without retrieving the full text, and 14 were excluded after examining the full text. Twelve studies [37–48] (Table 1) were finally included. The flow diagram for study selection is shown in Fig 1.

**Table 1. Included studies.**

Study	Country/ Region	Endoscopy	Algorithm	Study Center	Study Design	Patient (n)	Train Image (n)	Diagnostician	Sensitivity (%)	Specificity (%)
Mu 2021 [37]	China	WLI	UNet++ ResNet	Multi	Prospective	4,587	7,326	AI	89	95
								Endoscopist	89	96
Lin 2021 [38]	China	WLI	TRResNet	Multi	Retrospective	2,741	6,489	AI	97.9	97.5
								Endoscopist	42	96
Xu 2021 [39]	China	WLI ME-NBI BLI	VGG-16	Multi	Prospective	1,384	4,138	AI	90.1	86.1
								Endoscopist	77.8	76.5
Yang 2022 [40]	China	WLI LCI	SE-ResNet	Single	Retrospective	630	17,137	AI	96.6	97.9
Yan 2020 [41]	Macau	NBI ME-NBI	EfficientNet	Single	Retrospective	416	1,880	AI	91.9	86.0
								Endoscopist	86.5	81.4
Siripoppohn 2022 [42]	Thailand	WLI NBI	Improved BiSeNet	Single	Prospective	136	642	AI	93.13	80.0
Huang 2004 [43]	Taiwan	WLI	Customized Neural Networks	Single	Retrospective	104	84	AI	83.3	91.9
Li 2021 [44]	Macau	NBI	Improved ResNet	Single	Retrospective	242	840	AI	93.16	87.1
Wong 2022 [45]	Macau	ME-NBI	Improved ResNet	Single	Retrospective	420	1372	AI	93.6	91.2
								Endoscopist	86.5	81.4
Lai 2022 [46]	Macau	WLI NBI ME-NBI	Customized Neural Networks	Single	Retrospective	513	792	AI	96.1	88.42
Li 2023 [47]	China	ESE	Customized Neural Networks	Single	Retrospective	NA	837	AI	94.39	91.81
Pornvora-phat 2023 [48]	Thailand	WLI NBI	Improved BiSeNet	Single	Retrospective	NA	1599	AI	91	96

WLI: White light imaging; LCI: Linked-color imaging; NBI, Narrow-Band imaging; ME-NBI: Magnifying endoscopy with NBI; BLI: Blue-laser imaging; ESE: Electronic staining endoscopy.

<https://doi.org/10.1371/journal.pone.0303421.t001>

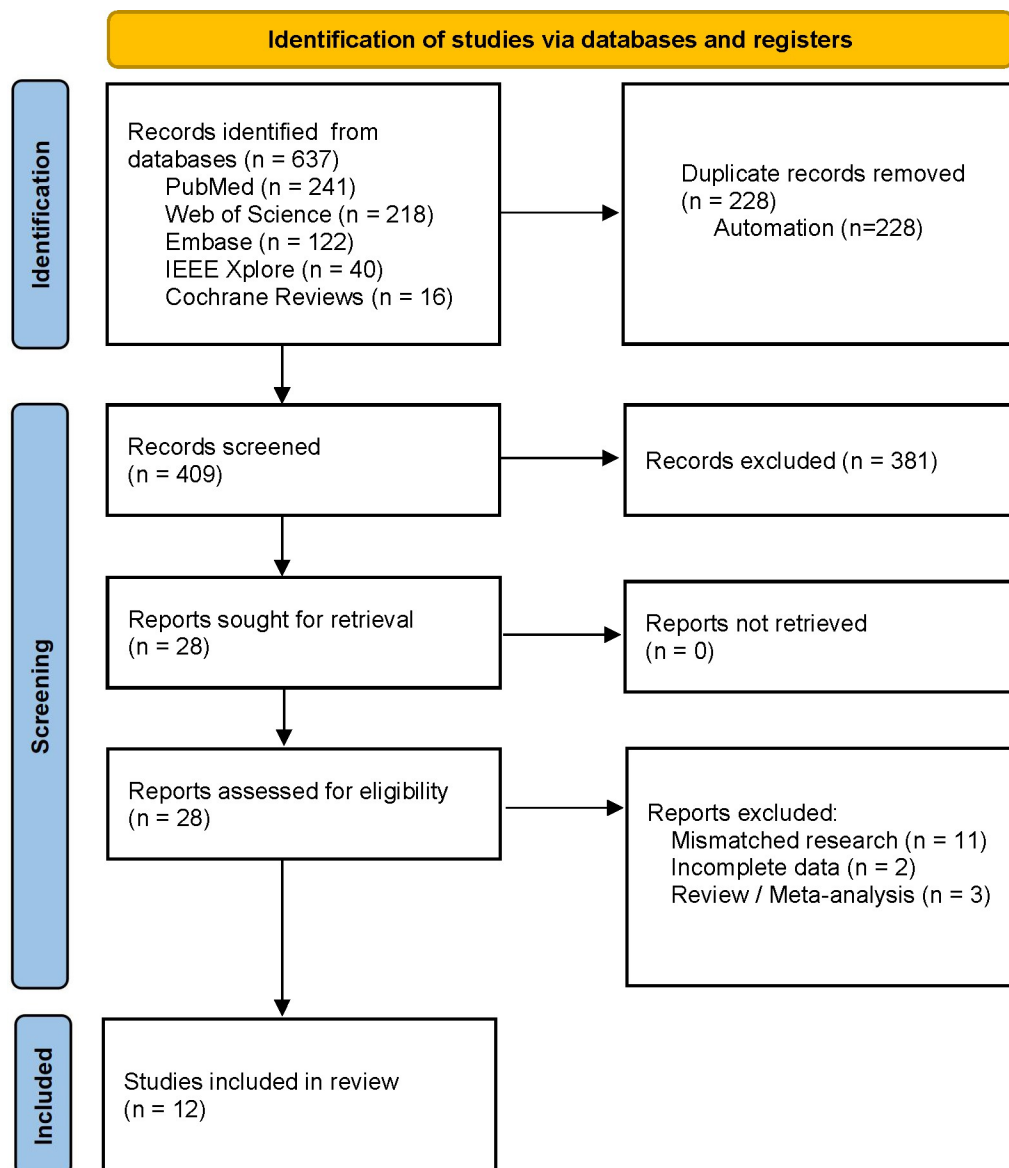


Fig 1. Flow diagram for study selection.

<https://doi.org/10.1371/journal.pone.0303421.g001>

## Study characteristics

Basic information of the 12 studies are shown in Table 1, and the participant characteristics of each study are shown in S3 Table. Among the 12 studies, three were prospective [37, 39, 42], and nine were retrospective; three studies [37–39] were multi-center and nine were single-center; three studies [44, 46, 47] used expert consensus as the diagnostic criterion and nine used pathological findings as the diagnostic criterion; three studies [37, 38, 43] used only plain white light imaging (WLI) model, and nine studies involved narrow-band imaging (NBI), magnified endoscopy with NBI (ME-NBI), blue laser imaging, and linked color imaging (LCI) model.

Mu et al. [37] developed a computer-aided system to identify non-gastritis, common gastritis, AG, and GIM. The system contains five deep learning models. ResNet was used for lesion

classification and UNet++ network was used for lesion segmentation. We extracted only the diagnostic data for GIM.

Lin et al. [38] collected 7,037 WLI images and corresponding biopsy information from 14 hospitals. The images were classified into three categories: AG, non-AG, and GIM, based on pathological findings. The AI algorithm was TResNet. The sensitivity and specificity of the AI model to diagnose GIM were 97.9% and 97.5%.

Xu et al. [39] collected WLI, ME-NBI, and blue laser imaging (BLI) images from five hospitals for model training to identify AG and GIM. The models were tested on internal, external, and prospective video test sets. The diagnostic data were collected from a randomly selected prospective video test set.

Yang et al. [40] constructed a dataset containing 21,420 WLI and LCI images to train a AI model for recognizing AG and GIM. The authors propose a dual transfer learning strategy to improve the model's performance. We extracted the data of the AI model on the WLI-independent and LCI-independent test sets and then combined them.

Yan et al. [41] collected 2,357 NBI and ME-NBI images from 416 patients for training an AI model for recognizing GIM. The sensitivity, specificity, and accuracy of the model were 91.9%, 86.0%, and 88.8%, respectively. Although the AI models performed better than the human experts, there was no significant difference between them. We combined the test results of the AI model on the NBI set and the ME-NBI set.

Siripoppohn et al. [42] implemented semantic segmentation of GIM by adding three additional preprocessing techniques to the BiSeNet network and compared it with the classical semantic segmentation algorithms, DeepLabV3+ and U-Net. Diagnostic data were extracted from the improved algorithm using a prospective video test set.

Huang et al. [43] constructed custom neural networks for identifying lesions, such as H. pylori infection, atrophy, and GIM, and extracted data relevant to identifying GIM. Although the study prospectively selected 104 patients, the model was trained and tested based on the images of these patients; therefore, we considered this to be a retrospective study.

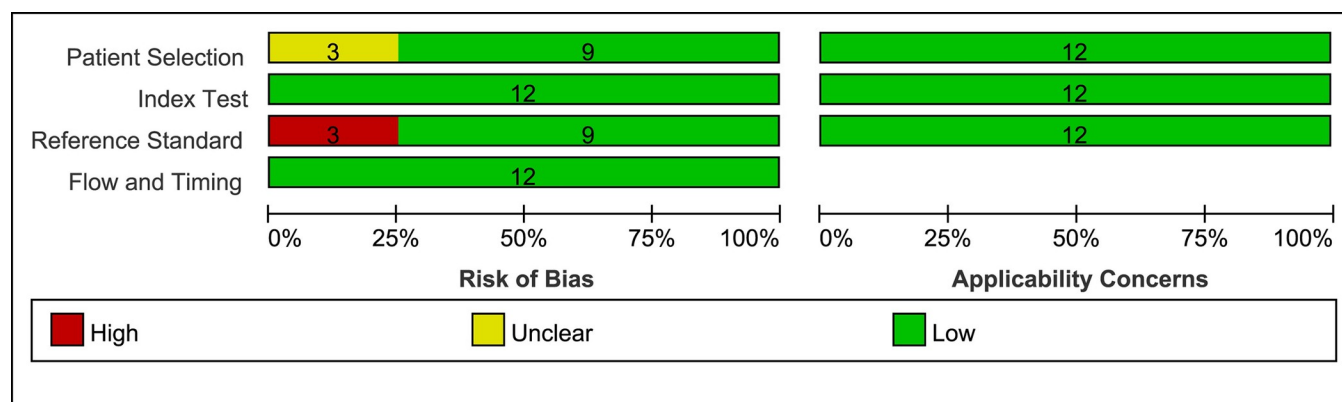
Li et al. [44] proposed a novel multi-feature fusion method to identify GIM, which extracts features from pixels, colors, and textures of endoscopic images, respectively. The authors trained and tested the model using 1,050 images and achieved an accuracy of 90.28%.

Wong et al. [45] proposed a novel broad-learning system stacking framework with multi-scale attention. This method could reliably diagnose GIM with an accuracy of 93.2%. The authors also compared the AI model's diagnostic capability to that of endoscopists, and the result showed that AI was competitive with that of skilled endoscopists.

Lai et al. [46] proposed a multi-scale multi-instance multi-feature joint-learning broad network. The network considers multiple features of each instance at multiple scales, resulting in more accurate classification. By training on a limited dataset, the network recognizes GIM with an accuracy of 85%.

Li et al. [47] proposed a combination of conventional and deep learning methods for IM lesion area localization and offset generation. The method could recognize the severity of GIM with an accuracy of 84.17%.

Pornvoraphat et al. [48] utilized AI techniques to achieve real-time segmentation of GIM. The AI algorithm is based on BiSeNet, and the authors used techniques such as negative sampling and label smoothing to improve the model's performance. The sensitivity, specificity and accuracy of the AI model were 91%, 96% and 96%, respectively.



**Fig 2. Results of the assessment of risk of bias in the included studies.**

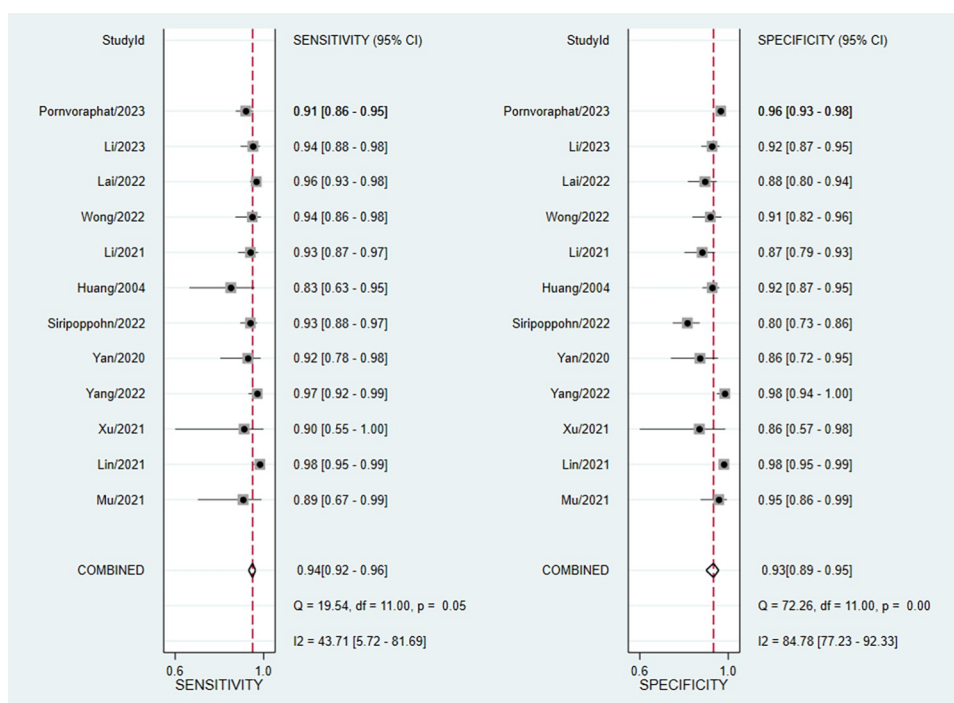
<https://doi.org/10.1371/journal.pone.0303421.g002>

## Quality assessment

The quality was assessed using the Supplemented QUADAS-2 tool (Fig 2). In the patient selection section, three studies were of unknown risk. One study [40] did not state the source of the patients, while two studies [47, 48] did not state the number of patients included. In the reference standards section, three studies [44, 46, 47] used expert consensus rather than pathological findings as the gold standard and were considered high risk.

## Meta-analysis results

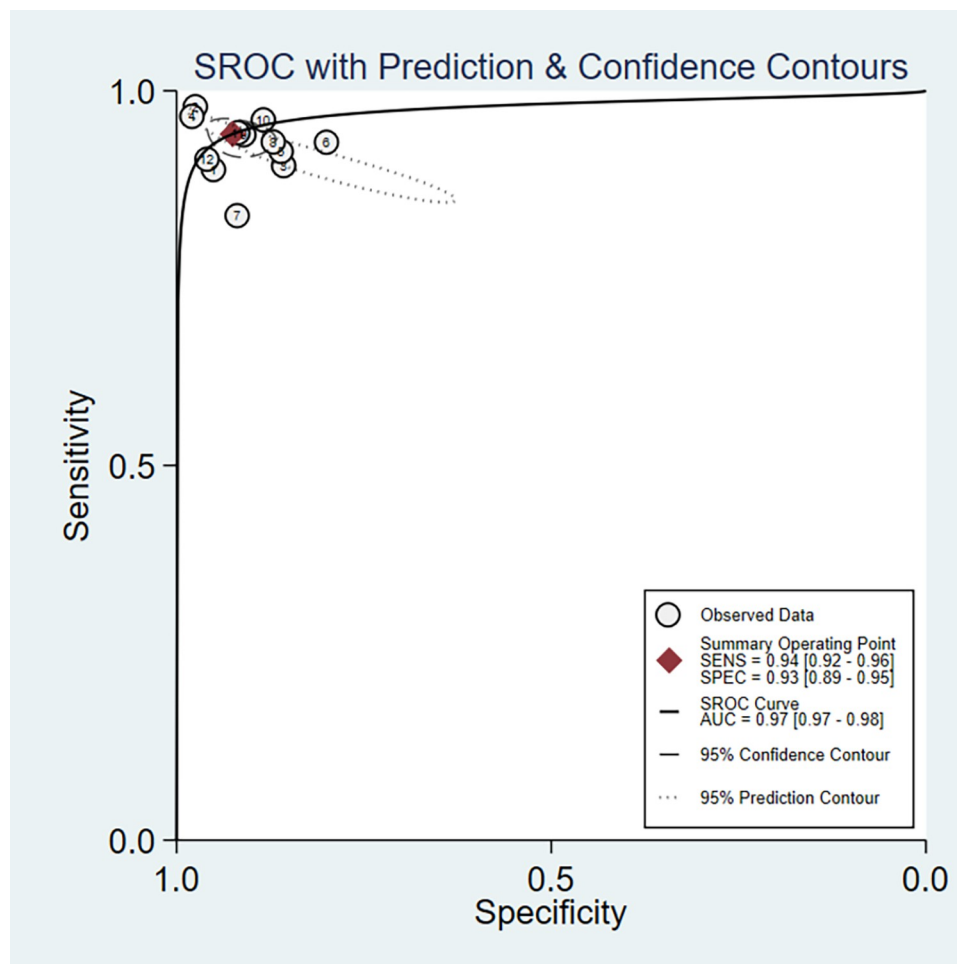
We imported data from 12 studies into Stata/SE 16.0 for meta-analysis. The pooled sensitivity and specificity of AI diagnosing GIM were 94% (95% CI: 0.92–0.96,  $I^2 = 43.71\%$ ) and 93% (95% CI: 0.89–0.95,  $I^2 = 84.78\%$ ), respectively (Fig 3). The PLR and NLR were 12.59



**Fig 3. Forest plot of sensitivity and specificity.**

<https://doi.org/10.1371/journal.pone.0303421.g003>





**Fig 4. SROC curves.**

<https://doi.org/10.1371/journal.pone.0303421.g004>

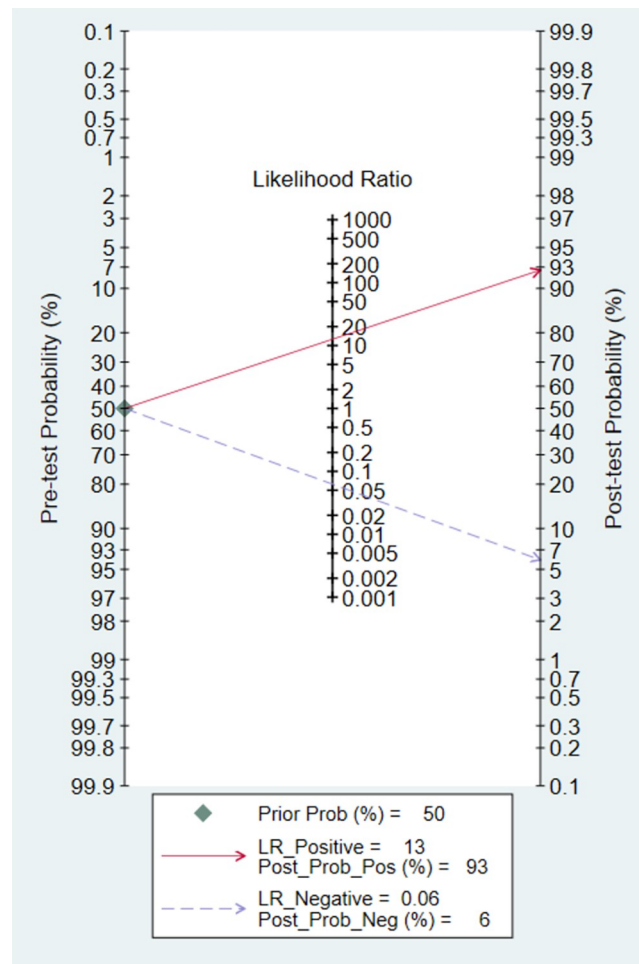
(95% CI: 0.38–18.91) and 0.06 (95% CI: 0.05–0.09), respectively (S1 Fig). The DOR (S2 Fig) and AUC (Fig 4) were 201.5 (95% CI: 110.18–368.51) and 0.97 (95% CI: 0.97–0.98), respectively. With a PLR (12.59) greater than 10, it suggested that AI had the capability to confirm the diagnosis of GIM. The NLR value ( $0.06 < 0.1$ ) indicates that AI can reliably exclude GIM. The DOR value ( $217 > 1$ ) indicated a better discriminative effect of this diagnostic test, and an AUC (0.97) closer to 1 indicated a better diagnostic effect. It is important to note that the  $I^2$  of combined sensitivity and specificity suggest a high degree of heterogeneity between studies.

Fagan plots were drawn to evaluate the clinical applicability of AI (Fig 5). With a pre-test probability set at 0.5, a positive AI diagnostic result indicated a 93% probability of the patient having GIM, while a negative result suggested a 6% likelihood, confirming or excluding the presence of Gastrointestinal Intestinal Metaplasia (GIM).

### Subgroup analysis

We conducted subgroup analyses to investigate the impact of various factors on the performance of (AI) in diagnosing GIM. The factors included study design (prospective or retrospective), study center (multi-center or single-center), endoscopy type (WLI only or others),





**Fig 5. Fagan's nomogram.**

<https://doi.org/10.1371/journal.pone.0303421.g005>

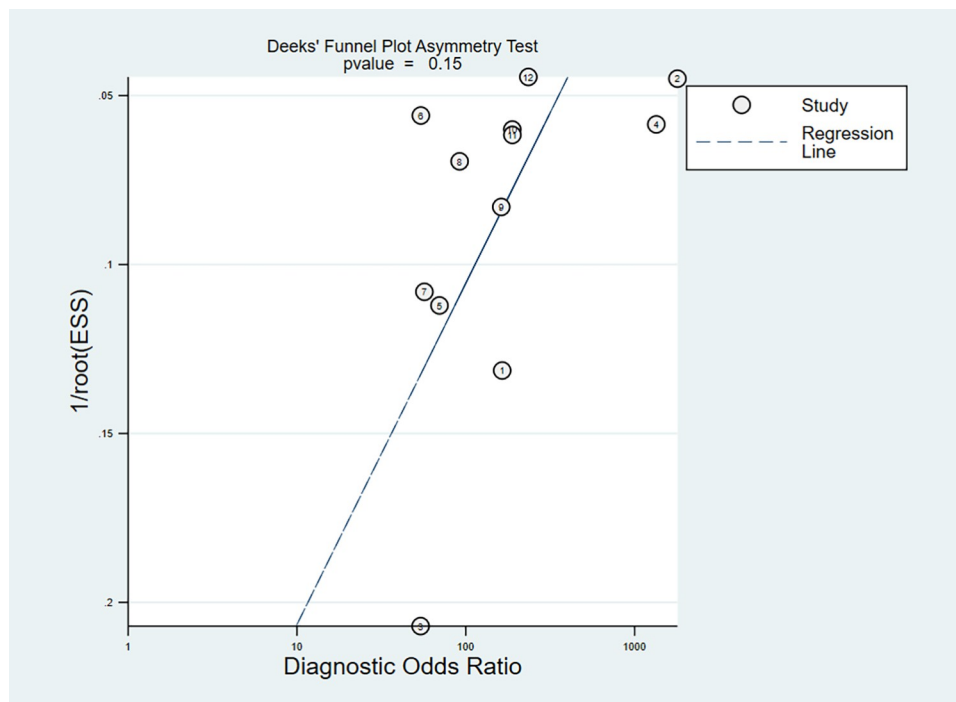
number of training images (>1500 or <1500), and algorithm type (classification algorithm or others) (Table 2).

The effect of study center on sensitivity was statistically significant, and the effect of other grouping conditions on sensitivity was extremely significant. The algorithm type had a

**Table 2. Subgroup analyses results.**

Subgroup	Condition	Studies(n)	Sensitivity(95%CI)	P	Specificity(95%CI)	P
Number of training images	> 1500	6	0.94(0.92–0.97)	<0.001	0.94(0.91–0.98)	<0.001
	< 1500	6	0.94(0.92–0.97)		0.89(0.83–0.95)	
Endoscopy type	WLI only	3	0.95(0.90–0.99)	<0.001	0.95(0.92–0.99)	0.05
	others	9	0.94(0.92–0.96)		0.90(0.86–0.94)	
Study design	prospective	3	0.93(0.88–0.98)	<0.001	0.87(0.78–0.97)	<0.001
	retrospective	9	0.94(0.93–0.96)		0.93(0.90–0.96)	
Study center	multi-center	3	0.97(0.94–0.99)	0.01	0.94(0.90–0.99)	0.18
	single-center	9	0.94(0.92–0.95)		0.91(0.87–0.95)	
Algorithm type	classification	9	0.95(0.94–0.97)	<0.001	0.92(0.88–0.96)	0.02
	others	3	0.92(0.89–0.95)		0.92(0.85–0.99)	

<https://doi.org/10.1371/journal.pone.0303421.t002>



**Fig 6. Deeks' funnel plot asymmetry test for publication.**

<https://doi.org/10.1371/journal.pone.0303421.g006>

significant effect on specificity. All of the above factors could be potential sources of heterogeneity between studies.

### Publication bias and sensitivity analysis

To assess the presence of publication bias, we performed a Deeks' funnel plot asymmetry test (Fig 6). The P value was 0.15, indicating no significant publication bias.

To delve deeper into the heterogeneity among studies, we conducted a pooled analysis by systematically excluding each study one at a time. After removing the studies by Pornvoraphat [48], the most significant changes were observed in combined sensitivity and specificity, which were found to be 95% (95% CI: 0.93–0.96,  $I^2 = 40.28\%$ ) and 92% (95% CI: 0.88–0.95,  $I^2 = 83.04\%$ ), respectively. However, this is not significantly different from the original results, indicating that the meta-analysis results were relatively stable.

### AI vs. endoscopists

To further explore the diagnostic ability of the AI, we compared it to that of the endoscopists. We collected 5 sets of data from 12 studies (Table 1) for the meta-analysis. An essential criterion for data extraction was that the test sets used by the AI and endoscopists must be identical. The comparison results are shown in Table 3. Their specificities showed no significant

**Table 3. AI vs. endoscopists.**

Performance Metrics	AI	Endoscopists	P
Sensitivity(95%CI)	0.95(0.91–0.99)	0.79(0.65–0.92)	0.39
Specificity(95%CI)	0.94(0.89–0.98)	0.90(0.83–0.96)	0.16

<https://doi.org/10.1371/journal.pone.0303421.t003>

disparity, and while AI exhibited a superior sensitivity compared to the endoscopists, this variance did not reach statistical significance.

## Discussion

Dilaghi et al. [49] conducted a meta-analysis on AI's role in the diagnosis of precancerous gastric lesions and *Helicobacter pylori* infection, with two studies involving the diagnosis of GIM. To the best of our knowledge, this is the first systematic review and meta-analysis focusing on the diagnosis of GIM using AI. This meta-analysis included 12 studies involving 11,173 relevant patients and 46,268 images/videos. In addition to pooling the diagnostic performance of AI, we explored the impact of factors such as different algorithms, varied image quantities, etc., on AI performance. Furthermore, we compared the diagnostic abilities of endoscopists with those of AI. The results demonstrate that various indicators of AI-assisted diagnosis for GIM exhibit satisfactory levels. This indicates that AI can help doctors diagnose GIM more accurately, thus reducing the rate of missed diagnoses and misdiagnoses. In addition, AI can accelerate the diagnostic process, which reduces doctors' workload and improves their efficiency.

There are still some limitations of this study: (a) The diagnostic value of AI algorithms may not be adequately assessed due to the relatively small number of studies and limited sample size. (b) The 12 included studies were conducted in Asia, and the results of the meta-analysis may not apply to a wider population. (c) Heterogeneity among the studies was very high. Although subgroup analyses were conducted, the restricted number of studies did not allow for further analysis of influencing factors, such as the type of test set (image or video) and specific endoscope type (e.g., NBI or LCI). (d) Most studies have identified only intestinal metaplasia and atrophic gastritis, and further validation is needed to determine whether other lesions affect the determination of AI. (e) Most studies were retrospective, and the test set included static images. More prospective, real-time, endoscopic-video-based studies are required to validate whether AI can be adapted to complex endoscopic environments.

Among the twelve studies, one [47] employed AI to identify GIM and assess its severity through endoscopic image analysis. This process is crucial for accurately pinpointing the most representative lesion area for biopsy, indicating a significant avenue for future research. Additionally, identifying early malignant changes from GIM remains a challenge. Previous studies have used AI to identify early gastric cancer by analyzing WLI, NBI, or ME-NBI images [50–52]. Ikenoyama et al. [51] used conventional WLI images to identify early cancers smaller than 20mm, with AI sensitivity and specificity of 58.4% and 87.3%, respectively. While these results are encouraging, AI performance still needs to be improved.

Most included studies used deep learning techniques, but none explained AI's decision-making process in detail. Due to their complexity and "black-box" nature, deep learning models often find it difficult to explain their internal working mechanisms and decision-making basis, which largely limits their clinical applications [53]. The introduction of algorithms such as LRP [54] and Grad-CAM [55] has played an important role in enhancing the explainability of existing deep learning models [56, 57]. Developing inherently explainable AI models will enable better application in clinical practice.

All the included studies were tested using their own datasets, and it was difficult to directly compare the performances of the models. High-quality publicly available datasets can be used as evaluation benchmarks to compare the performances of different algorithms. Additionally, publicly available datasets can encourage more people to participate in AI research. Currently, publicly available gastrointestinal image datasets include Hyper-Kvasir [58] and SUN-SEG [59]; however, there is a lack of large publicly available image datasets related to GIM.

It is worth noting that although AI applied to healthcare has made many technological breakthroughs, it also poses certain challenges to the current value system and legal system from the legal and ethical levels. For example, it may raise privacy and data security issues and legal liability issues when AI's decisions are made incorrectly. As AI continues to advance, collaborative efforts among governments, healthcare organizations, and AI technology companies are crucial to establishing a robust framework that ensures the responsible and fair deployment of AI in clinical settings.

## Conclusions

The pooled sensitivity of our meta-analysis was 94% (95% confidence interval: 0.92–0.96) and specificity was 93% (95% confidence interval: 0.89–0.95). Comparisons by AI vs. endoscopists showed that AI had a higher sensitivity (95% vs. 79%). The results show that AI performed excellently in diagnosing GIM, which provides an evidence-based support for the clinical application of AI. At the same time, we identified some potential limitations, such as the quality of the dataset, generalizability of the AI model, and explainable AI. The application of AI-assisted endoscopy in the medical field is promising. Future research could focus on prospective studies, improvement of the explainability of models, and adaptation to different patient characteristics.

## Supporting information

**S1 Table. PRISMA-DTA checklist.**  
(DOCX)

**S2 Table. Searching strategy to find relevant articles.**  
(DOCX)

**S3 Table. Participant characteristics and algorithmic details of included studies.**  
(DOCX)

**S1 Fig. Forest plot of PLR and NLR of AI in identifying GIM.**  
(DOCX)

**S2 Fig. Forest plot for the diagnostic odds ratio and diagnostic score after combination.**  
(DOCX)

## Author Contributions

**Conceptualization:** Na Li, Kunhong Wang.

**Data curation:** Na Li, Jian Yang, Xiaodong Li, Yanting Shi.

**Formal analysis:** Xiaodong Li, Yanting Shi.

**Investigation:** Kunhong Wang.

**Methodology:** Jian Yang, Xiaodong Li, Kunhong Wang.

**Project administration:** Kunhong Wang.

**Resources:** Xiaodong Li, Yanting Shi.

**Software:** Jian Yang, Xiaodong Li.

**Supervision:** Kunhong Wang.

**Validation:** Jian Yang, Xiaodong Li, Yanting Shi.

**Visualization:** Jian Yang, Yanting Shi.

**Writing – original draft:** Na Li.

**Writing – review & editing:** Kunhong Wang.

## References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin.* 2021; 71: 209–249. <https://doi.org/10.3322/caac.21660> PMID: 33538338
2. Correa P, Shiao YH. Phenotypic and genotypic events in gastric carcinogenesis. *Cancer Res.* 1994; 54: 1941s–1943s. PMID: 8137316
3. Correa P. Human gastric carcinogenesis: a multistep and multifactorial process—First American Cancer Society Award Lecture on Cancer Epidemiology and Prevention. *Cancer Res.* 1992; 52: 6735–6740. PMID: 1458460
4. Atkins L, Benedict EB. Correlation of gross gastroscopic findings with gastroscopic biopsy in gastritis. *N Engl J Med.* 1956; 254: 641–644. <https://doi.org/10.1056/NEJM195604052541403> PMID: 13309650
5. Carpenter HA, Talley NJ. Gastrosocopy is incomplete without biopsy: clinical relevance of distinguishing gastropathy from gastritis. *Gastroenterology.* 1995; 108: 917–924. [https://doi.org/10.1016/0016-5085\(95\)90468-9](https://doi.org/10.1016/0016-5085(95)90468-9) PMID: 7875496
6. Eshmuratov A, Nah JC, Kim N, Lee HS, Lee HE, Lee BH, et al. The correlation of endoscopic and histological diagnosis of gastric atrophy. *Dig Dis Sci.* 2010; 55: 1364–1375. <https://doi.org/10.1007/s10620-009-0891-4> PMID: 19629687
7. Red  en S, Petersson F, J  nsson K-A, Borch K. Relationship of gastroscopic features to histological findings in gastritis and *Helicobacter pylori* infection in a general population sample. *Endoscopy.* 2003; 35: 946–950. <https://doi.org/10.1055/s-2003-43479> PMID: 14606018
8. Stathopoulos G, Goldberg RD, Blackstone MO. Endoscopic diagnosis of intestinal metaplasia. *Gastrointest Endosc.* 1990; 36: 544–545. [https://doi.org/10.1016/s0016-5107\(90\)71151-0](https://doi.org/10.1016/s0016-5107(90)71151-0) PMID: 2227346
9. Zhang X, Yang J, Lin T, Ying Y. Food and agro-product quality evaluation based on spectroscopy and deep learning: A review. *Trends in Food Science & Technology.* 2021; 112: 431–441. <https://doi.org/10.1016/j.tifs.2021.04.008>
10. Mousavi SM, Beroza GC. Deep-learning seismology. *Science.* 2022; 377: eabm4470. <https://doi.org/10.1126/science.abm4470> PMID: 35951699
11. Aliramezani M, Koch CR, Shahbakhti M. Modeling, diagnostics, optimization, and control of internal combustion engines via modern machine learning techniques: A review and future directions. *Progress in Energy and Combustion Science.* 2022; 88: 100967. <https://doi.org/10.1016/j.pecs.2021.100967>
12. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv*; 2015 Apr. Report No.: arXiv:1409.1556. <https://doi.org/10.48550/arXiv.1409.1556>
13. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE; 2016. pp. 770–778. <https://doi.org/10.1109/CVPR.2016.90>
14. Ridnik T, Lawen H, Noy A, Ben E, Sharir BG, Friedman I. TRResNet: High Performance GPU-Dedicated Architecture. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). 2021. pp. 1399–1408. <https://doi.org/10.1109/WACV48630.2021.00144>
15. Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT: IEEE; 2018. pp. 7132–7141. <https://doi.org/10.1109/CVPR.2018.00745>
16. Tan M, Le Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *Proceedings of the 36th International Conference on Machine Learning.* PMLR; 2019. pp. 6105–6114.
17. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, et al. SSD: Single Shot MultiBox Detector. In: Leibe B, Matas J, Sebe N, Welling M, editors. *Computer Vision—ECCV 2016.* Cham: Springer International Publishing; 2016. pp. 21–37. [https://doi.org/10.1007/978-3-319-46448-0\\_2](https://doi.org/10.1007/978-3-319-46448-0_2)
18. Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. pp. 779–788. <https://doi.org/10.1109/CVPR.2016.91>
19. Redmon J, Farhadi A. YOLOv3: An Incremental Improvement. *arXiv*; 2018 Apr. Report No.: arXiv:1804.02767. <https://doi.org/10.48550/arXiv.1804.02767>

20. Girshick R, Donahue J, Darrell T, Malik J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE; 2014. pp. 580–587. <https://doi.org/10.1109/CVPR.2014.81>
21. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. *Deep Learn Med Image Anal Multimodal Learn Clin Decis Support* (2018). 2018; 11045: 3–11. [https://doi.org/10.1007/978-3-030-00889-5\\_1](https://doi.org/10.1007/978-3-030-00889-5_1) PMID: 32613207
22. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *arXiv*; 2016 Jun. Report No.: arXiv:1412.7062. <https://doi.org/10.48550/arXiv.1412.7062>
23. Chen L-C, Papandreou G, Schroff F, Adam H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv*; 2017 Jun. Report No.: arXiv:1706.05587. <https://doi.org/10.48550/arXiv.1706.05587>
24. Yu C, Wang J, Peng C, Gao C, Yu G, Sang N. BiSeNet: Bilateral Segmentation Network for Real-time Semantic Segmentation. *arXiv*; 2018 Aug. Report No.: arXiv:1808.00897. <https://doi.org/10.48550/arXiv.1808.00897>
25. Sutton RT, Zai Ane OR, Goebel R, Baumgart DC. Artificial intelligence enabled automated diagnosis and grading of ulcerative colitis endoscopy images. *Sci Rep*. 2022; 12: 2748. <https://doi.org/10.1038/s41598-022-06726-2> PMID: 35177717
26. Dumitrescu EA, Ungureanu BS, Cazacu IM, Florescu LM, Streba L, Croitoru VM, et al. Diagnostic Value of Artificial Intelligence-Assisted Endoscopic Ultrasound for Pancreatic Cancer: A Systematic Review and Meta-Analysis. *Diagnostics* (Basel). 2022; 12: 309. <https://doi.org/10.3390/diagnostics12020309> PMID: 35204400
27. Gehrung M, Crispin-Ortuzar M, Berman AG, O'Donovan M, Fitzgerald RC, Markowitz F. Triage-driven diagnosis of Barrett's esophagus for early detection of esophageal adenocarcinoma using deep learning. *Nat Med*. 2021; 27: 833–841. <https://doi.org/10.1038/s41591-021-01287-9> PMID: 33859411
28. Shichijo S, Nomura S, Aoyama K, Nishikawa Y, Miura M, Shinagawa T, et al. Application of Convolutional Neural Networks in the Diagnosis of Helicobacter pylori Infection Based on Endoscopic Images. *EBioMedicine*. 2017; 25: 106–111. <https://doi.org/10.1016/j.ebiom.2017.10.014> PMID: 29056541
29. Shi Y, Wei N, Wang K, Wu J, Tao T, Li N, et al. Deep learning-assisted diagnosis of chronic atrophic gastritis in endoscopy. *Front Oncol*. 2023; 13: 1122247. <https://doi.org/10.3389/fonc.2023.1122247> PMID: 36950553
30. Bang CS, Lee JJ, Baik GH. Artificial intelligence for the prediction of helicobacter pylori infection in endoscopic images: Systematic review and meta-analysis of diagnostic test accuracy. *J Med Internet Res*. 2020; 22. <https://doi.org/10.2196/21983> PMID: 32936088
31. Shi Y, Wei N, Wang K, Tao T, Yu F, Lv B. Diagnostic value of artificial intelligence-assisted endoscopy for chronic atrophic gastritis: a systematic review and meta-analysis. *Front Med* (Lausanne). 2023; 10: 1134980. <https://doi.org/10.3389/fmed.2023.1134980> PMID: 37200961
32. Booth A, Clarke M, Ghersi D, Moher D, Petticrew M, Stewart L. An international registry of systematic-review protocols. *Lancet*. 2011; 377: 108–109. [https://doi.org/10.1016/S0140-6736\(10\)60903-8](https://doi.org/10.1016/S0140-6736(10)60903-8) PMID: 20630580
33. Salameh J-P, Bossuyt PM, McGrath TA, Thombs BD, Hyde CJ, Macaskill P, et al. Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. *BMJ*. 2020; 370: m2632. <https://doi.org/10.1136/bmj.m2632> PMID: 32816740
34. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011; 155: 529–536. <https://doi.org/10.7326/0003-4819-155-8-201110180-00009> PMID: 22007046
35. Van Houwelingen HC, Zwinderman KH, Stijnen T. A bivariate approach to meta-analysis. *Statistics in Medicine*. 1993; 12: 2273–2284. <https://doi.org/10.1002/sim.4780122405> PMID: 7907813
36. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine*. 2002; 21: 589–624. <https://doi.org/10.1002/sim.1040> PMID: 11836738
37. Mu G, Zhu Y, Niu Z, Li H, Wu L, Wang J, et al. Expert-level classification of gastritis by endoscopy using deep learning: a multicenter diagnostic trial. *Endosc Int Open*. 2021; 09: E955–E964. <https://doi.org/10.1055/a-1372-2789> PMID: 34079883
38. Lin N, Yu T, Zheng W, Hu H, Xiang L, Ye G, et al. Simultaneous Recognition of Atrophic Gastritis and Intestinal Metaplasia on White Light Endoscopic Images Based on Convolutional Neural Networks: A Multicenter Study. *Clin Transl Gastroenterol*. 2021; 12: e00385. <https://doi.org/10.14309/ctg.0000000000000385> PMID: 34342293



39. Xu M, Zhou W, Wu L, Zhang J, Wang J, Mu G, et al. Artificial intelligence in the diagnosis of gastric pre-cancerous conditions by image-enhanced endoscopy: a multicenter, diagnostic study (with video). *Gastrointest Endosc*. 2021; 94: 540–548.e4. <https://doi.org/10.1016/j.gie.2021.03.013> PMID: 33722576
40. Yang J, Ou Y, Chen Z, Liao J, Sun W, Luo Y, et al. A Benchmark Dataset of Endoscopic Images and Novel Deep Learning Method to Detect Intestinal Metaplasia and Gastritis Atrophy. *IEEE J Biomed Health Inform*. 2022;PP. <https://doi.org/10.1109/JBHI.2022.3217944> PMID: 36306301
41. Yan T, Wong PK, Choi IC, Vong CM, Yu HH. Intelligent diagnosis of gastric intestinal metaplasia based on convolutional neural network and limited number of endoscopic images. *Comput Biol Med*. 2020; 126: 104026. <https://doi.org/10.1016/j.combiomed.2020.104026> PMID: 33059237
42. Siripoppohn V, Pittayanon R, Tiankanon K, Faknak N, Sanpavat A, Klaikaew N, et al. Real-time semantic segmentation of gastric intestinal metaplasia using a deep learning approach. *Clin Endosc*. 2022; 55: 390–400. <https://doi.org/10.5946/ce.2022.005> PMID: 35534933
43. Huang C-R, Sheu B-S, Chung P-C, Yang H-B. Computerized diagnosis of *Helicobacter pylori* infection and associated gastric inflammation from endoscopic images by refined feature selection using a neural network. *Endoscopy*. 2004; 36: 601–608. <https://doi.org/10.1055/s-2004-814519> PMID: 15243882
44. Li H, Vong CM, Wong PK, Ip WF, Yan T, Choi IC, et al. A multi-feature fusion method for image recognition of gastrointestinal metaplasia (GIM). *Biomedical Signal Processing and Control*. 2021; 69: 102909. <https://doi.org/10.1016/j.bspc.2021.102909>
45. Wong PK, Yao L, Yan T, Choi IC, Yu HH, Hu Y. Broad learning system stacking with multi-scale attention for the diagnosis of gastric intestinal metaplasia. *Biomedical Signal Processing and Control*. 2022; 73: 103476. <https://doi.org/10.1016/j.bspc.2021.103476>
46. Lai Q, Vong CM, Wong PK, Wang ST, Yan T, Choi IC, et al. Multi-scale Multi-instance Multi-feature Joint Learning Broad Network (M3JLBN) for gastric intestinal metaplasia subtype classification. *Knowledge-Based Systems*. 2022; 249: 108960. <https://doi.org/10.1016/j.knosys.2022.108960>
47. Li Z, Zheng X, Mu Y, Zhang M, Liu G. Color-guided deformable convolution network for intestinal metaplasia severity classification using endoscopic images. *Phys Med Biol*. 2023; 68. <https://doi.org/10.1088/1361-6560/acf3ca> PMID: 37619578
48. Pornvoraphat P, Tiankanon K, Pittayanon R, Sunthornwetchapong P, Vateekul P, Rerknimitr R. Real-time gastric intestinal metaplasia diagnosis tailored for bias and noisy-labeled data with multiple endoscopic imaging. *Comput Biol Med*. 2023; 154: 106582. <https://doi.org/10.1016/j.combiomed.2023.106582> PMID: 36738708
49. Dilaghi E, Lahner E, Annibale B, Esposito G. Systematic review and meta-analysis: Artificial intelligence for the diagnosis of gastric precancerous lesions and *Helicobacter pylori* infection. *Dig Liver Dis*. 2022; S1590-8658(22)00211–0. <https://doi.org/10.1016/j.dld.2022.03.007> PMID: 35382973
50. Ueyama H, Kato Y, Akazawa Y, Yatagai N, Komori H, Takeda T, et al. Application of artificial intelligence using a convolutional neural network for diagnosis of early gastric cancer based on magnifying endoscopy with narrow-band imaging. *J Gastroenterol Hepatol*. 2021; 36: 482–489. <https://doi.org/10.1111/jgh.15190> PMID: 32681536
51. Ikenoyama Y, Hirasawa T, Ishioka M, Namikawa K, Yoshimizu S, Horiuchi Y, et al. Detecting early gastric cancer: Comparison between the diagnostic ability of convolutional neural networks and endoscopists. *Dig Endosc*. 2021; 33: 141–150. <https://doi.org/10.1111/den.13688> PMID: 32282110
52. Li L, Chen Y, Shen Z, Zhang X, Sang J, Ding Y, et al. Convolutional neural network for the diagnosis of early gastric cancer based on magnifying narrow band imaging. *Gastric Cancer*. 2020; 23: 126–132. <https://doi.org/10.1007/s10120-019-00992-2> PMID: 31332619
53. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng*. 2018; 2: 719–731. <https://doi.org/10.1038/s41551-018-0305-z> PMID: 31015651
54. Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*. 2015; 10: e0130140. <https://doi.org/10.1371/journal.pone.0130140> PMID: 26161953
55. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 2017 IEEE International Conference on Computer Vision (ICCV). 2017. pp. 618–626. <https://doi.org/10.1109/ICCV.2017.74>
56. Rudin C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat Mach Intell*. 2019; 1: 206–215. <https://doi.org/10.1038/s42256-019-0048-x> PMID: 35603010
57. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*. 2021; 3: e745–e750. [https://doi.org/10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9) PMID: 34711379



58. Borgli H, Thambawita V, Smedsrud PH, Hicks S, Jha D, Eskeland SL, et al. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Sci Data*. 2020; 7: 283. <https://doi.org/10.1038/s41597-020-00622-y> PMID: 32859981
59. Misawa M, Kudo S, Mori Y, Hotta K, Ohtsuka K, Matsuda T, et al. Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). *Gastrointestinal Endoscopy*. 2021; 93: 960–967.e3. <https://doi.org/10.1016/j.gie.2020.07.060> PMID: 32745531