Annals of Gastroenterology (2024) **37,** 1-9

Use of artificial intelligence for the detection of *Helicobacter pylori* infection from upper gastrointestinal endoscopy images: an updated systematic review and meta-analysis

Om Parkash[a], Abhishek Lal[a], Tushar Subash[b], Ujala Sultan[b], Hasan Nawaz Tahir[c], Zahra Hoodbhoy[d], Shiyam Sundar[c], Jai Kumar Das[d]

The Aga Khan University, Karachi, Pakistan

Abstract

Background *Helicobacter pylori* (*H. pylori*) infection is associated with various gastrointestinal diseases and may lead to gastric cancer. Currently, endoscopy is the gold standard modality used for diagnosing *H. pylori* infection, but it lacks objective indicators and requires expert interpretation. In the past few years, the use of artificial intelligence (AI) for diagnosing gastrointestinal pathologies has increased tremendously and may improve the diagnostic accuracy of endoscopy for *H. pylori* infection. This study aimed to evaluate the diagnostic accuracy of AI algorithms for detecting *H. pylori* infection using endoscopic images.

Methods Three investigators searched the PubMed, CINHAL and Cochrane databases for studies that compared AI algorithms with endoscopic histopathology for diagnosing *H. pylori* infection using endoscopic images. We assessed the methodological quality of studies using the QUADAS-2 tool and performed a meta-analysis to estimate the pooled sensitivity, specificity, and accuracy of AI for detecting *H. pylori* infection.

Results A total of 11 studies were identified that met our inclusion criteria. All were conducted in different countries based in Asia. Our meta-analysis showed that AI had high sensitivity (0.93, 95% confidence interval [CI] 0.90-0.95), specificity (0.92, 95%CI 0.89-0.94), and accuracy (0.92, 95%CI 0.90-0.94) for detecting *H. pylori* infection using endoscopic images. However, there was also high heterogeneity among the studies (Tau^2=0.87, I^2=76.10% for generalized effect size; Tau^2=1.53, I^2=80.72% for sensitivity; Tau^2=0.57, I^2=70.86% for specificity).

Conclusion This systematic review and meta-analysis showed that AI had high diagnostic accuracy for detecting *H. pylori* infection using endoscopic images.

Keywords Artificial intelligence, deep learning, machine learning, *Helicobacter pylori*, endoscopy

Ann Gastroenterol 2024; 37 (XX): 1-9

[a]Section of Gastroenterology, Department of Medicine (Om Parkash, Abhishek Lal, Tushar); [b]Medical College (Subash, Ujala Sultan); [c]Department of Community Health Sciences (Hasan Nawaz Tahir, Shiyam Sundar); [d]Department of Paediatrics and Child Health (Zahra Hoodbhoy, Jai Kumar Das), The Aga Khan University, Karachi, Pakistan

Conflict of Interest: None

Correspondence to: Om Parkash, The Aga Khan University, Faculty Office Building, National Stadium Road, The Aga Khan University, Karachi, 75500 Pakistan, e-mail: om.parkash@aku.edu

Received 22 April 2024; accepted 11 July 2024;
published online 20 October 2024

DOI: https://doi.org/10.20524/aog.2024.0913

Introduction

Helicobacter pylori (*H. pylori*) infection is among the most common gastrointestinal pathologies and affects more than half of the global population [1]. *H. pylori* is associated with various gastrointestinal diseases, such as peptic ulcer disease, gastritis, mucosa-associated lymphoid tissue (MALT) lymphoma, and gastric adenocarcinoma [2]. If left untreated, *H. pylori* infection can lead to gastric metaplasia, chronic gastric atrophy, and gastric carcinomas. According to the International Agency for Cancer Research, *H. pylori* has been categorized as a class 1 carcinogen [3]. Hence, the eradication of *H. pylori* holds vital significance in terms of mitigating the risk of gastric cancer. However, the diagnosis of *H. pylori* infection is challenging, as it requires invasive or noninvasive methods that have limitations in terms of accuracy, cost, availability, and patient compliance.

Invasive methods include biopsy and a urease test, which involve taking tissue samples from the stomach and testing them for the presence of *H. pylori* or its urease enzyme. These methods are considered the gold standard for the diagnosis of *H. pylori* infection, but they are expensive, time-consuming, and may cause complications such as bleeding and perforation [4]. Endoscopy involves visual inspection of the gastric mucosal lesions in doubt, during which biopsy samples are taken and undergo histopathological analysis for definitive diagnosis. Endoscopic features associated with *H. pylori* infection include erythema, atrophy, mucosal folds, ulcerations, and nodularity [5]. Noninvasive methods include serology and urea breath test, which detect the antibodies or the urea metabolites of *H. pylori* in the patients' blood or breath samples. These methods are simple, convenient and widely available, but they have drawbacks that include low specificity, cross-reactivity, and inability to distinguish between current and past infection.

Endoscopic features used to identify potential cases of *H. pylori* infection lack objective indicators, with possible variance in terms of interobserver and intraobserver reliability for visual inspection of endoscopic images. Expert gastroenterologists and hepatologists may accurately recognize endoscopic images for *H. pylori* infection; however, amateur specialists need a considerable amount of time to execute this task precisely. Moreover, there are no uniform features associated with the detection of *H. pylori* infection, and hence no established modalities for diagnosing *H. pylori* infection through endoscopic examination [6]. The final diagnosis of the lesions is based on a histopathological analysis of the biopsy sample. To overcome such challenges, there has been a recent surge in interest in utilizing artificial intelligence (AI).

Recently, AI has emerged as a promising tool to assist endoscopists in the detection and characterization of gastrointestinal lesions. AI can analyze endoscopic images using deep learning algorithms that can learn from large datasets and perform tasks such as classification, segmentation and localization. AI-based image analysis has shown promise in terms of its diagnostic capacities, including the evaluation of endoscopic images for the detection of *H. pylori* infection. As a result, its use in patients suspected of suffering from such infection has increased.

Several studies have reported the application of AI for the detection of *H. pylori* infection from upper gastrointestinal endoscopy images using various modalities, such as white-light endoscopy, narrow-band imaging, and magnifying endoscopy [7,8]. These studies have shown that AI can achieve high accuracy and sensitivity in identifying *H. pylori* infection from endoscopic images, and can also differentiate between active and inactive infection. Moreover, these studies have demonstrated that AI can reduce interobserver variability and improve endoscopists' diagnostic confidence. Keeping this in mind, this systematic review and meta-analysis aimed to evaluate the accuracy of AI in the diagnosis of *H. pylori* infection using endoscopic images. The findings of this research could have a significant impact on the diagnosis and treatment of *H. pylori* infection.

Material and methods

This systematic review has been registered with PROSPERO (CRD42023437688). We followed the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) guidelines for diagnostic test accuracy for reporting in this study [9].

Eligibility criteria and search strategy

In this systematic review, we included all observational studies that aimed to detect *H. pylori* infection based on upper gastrointestinal endoscopy utilizing AI algorithms, compared to a reference standard (specialist opinion, and laboratory investigations). For the literature search, no restrictions were applied to the type of algorithms used, or the age at which *H. pylori* infection was diagnosed. Studies that diagnosed gastrointestinal pathologies other than *H. pylori* infection, or were published in languages other than English, were excluded from this review. Review studies, letters to the editor, conference proceedings, scientific reports and opinions were also excluded, as were studies conducted on animal or non-human subjects.

We used PubMed, CINHAL, and Cochrane as databases for our literature search, to identify articles published up to August 2023. The search terms used were "Artificial Intelligence", "Algorithms, "Machine Learning", "Deep Learning", Supervised Machine Learning, "Unsupervised Machine Learning", "*Helicobacter Pylori*", "*H. Pylori*", "Endoscopy", "Gastro*", "Peptic Ulcer", "Cancer", "Carcinoma", "Endoscopy", "Diagnosis", "Sensitivity", "Specificity", "Accuracy", and "Area Under Curve". The complete search strategy used is available in the supplementary section. The initial list of articles was imported to EndNote and duplicates were removed.

Screening and data extraction

Three authors (AL, TS, and US) independently screened the search results for their titles and abstracts to assess their potential eligibility in this review. Full texts of the articles were then reviewed by the 3 authors to ensure the selection of the articles relevant to this review. The bibliographies and citations of the included studies were also reviewed to include any further studies that might have been missed during the electronic search. The entire screening process was performed independently by the 3 authors and conflicts were resolved by the fourth and fifth authors (HNT and OP).

After the final screening process, data from the included studies were entered into a preformed data extraction form in MS Word. The data entered included title, name of journal, country of publication, study design, study setting, sample size, patient characteristics, type of AI algorithms, reference standard used, data analysis, and performance metrics (specificity, sensitivity, and accuracy), validation of the model, and subgroups if mentioned. We included studies in this

review where the types of AI systems and training data were mentioned.

Risk of bias assessment

The risk of bias was assessed by 3 authors (AL, TS, and SS) independently using Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2), which evaluates 4 domains: patient selection, index test, reference standard, and flow and timing. Each domain is rated as low risk, high risk, or unclear risk of bias or applicability concern. Any discrepancy between the reviewers was resolved by discussion or consultation with a fourth reviewer (OP).

Statistical analysis

When possible, we constructed 2×2 contingency tables, with values such as true positive, false positive, true negative and false negative, and used them to calculate sensitivity, specificity and accuracy. Data from all the included studies were entered into STATA (version 17.0) software, where we constructed forest plots and receiver operating characteristic (ROC) curves, using the sensitivity, specificity and accuracy of the studies. The subgroup analyses were conducted based on the Quality of Study, Study Format, Number of Patients, Published Year, and Type of AI.

Results

Our search strategy yielded an initial count of 4118 research papers returned by title/abstract searches. Of these 4118 articles, 127 duplicates were removed and a further 3991 research papers were excluded. A total of 11 studies were included in this systematic review and meta-analysis, as presented in Fig. 1.

Of the 11 studies [7,8,10-18], 6 were retrospective cohort studies [8,11,13,15,17,18], 4 were prospective cohort studies [7,10,14,16], and 1 study was a clinical trial [12] (Table 1). Most of the studies were from high-income countries, mainly Japan (n=6), followed by China (n=2), Malaysia (n=1), Korea (n=1), and Taiwan (n=1). All of the included studies performed a diagnostic analysis of AI's accuracy in detecting *H. pylori* infection using endoscopic images (n=11). The results of the AI-based diagnostic models were compared with histopathological analysis via biopsy.

Methodological quality of studies

Of the 11 studies included in the final analysis, 10 studies showed a low risk of bias; however, 1 study was found to have a high risk of bias (Fig. 2).

Meta-analysis

The meta-analysis entailed the computation of sensitivity, specificity and accuracy, based on the data

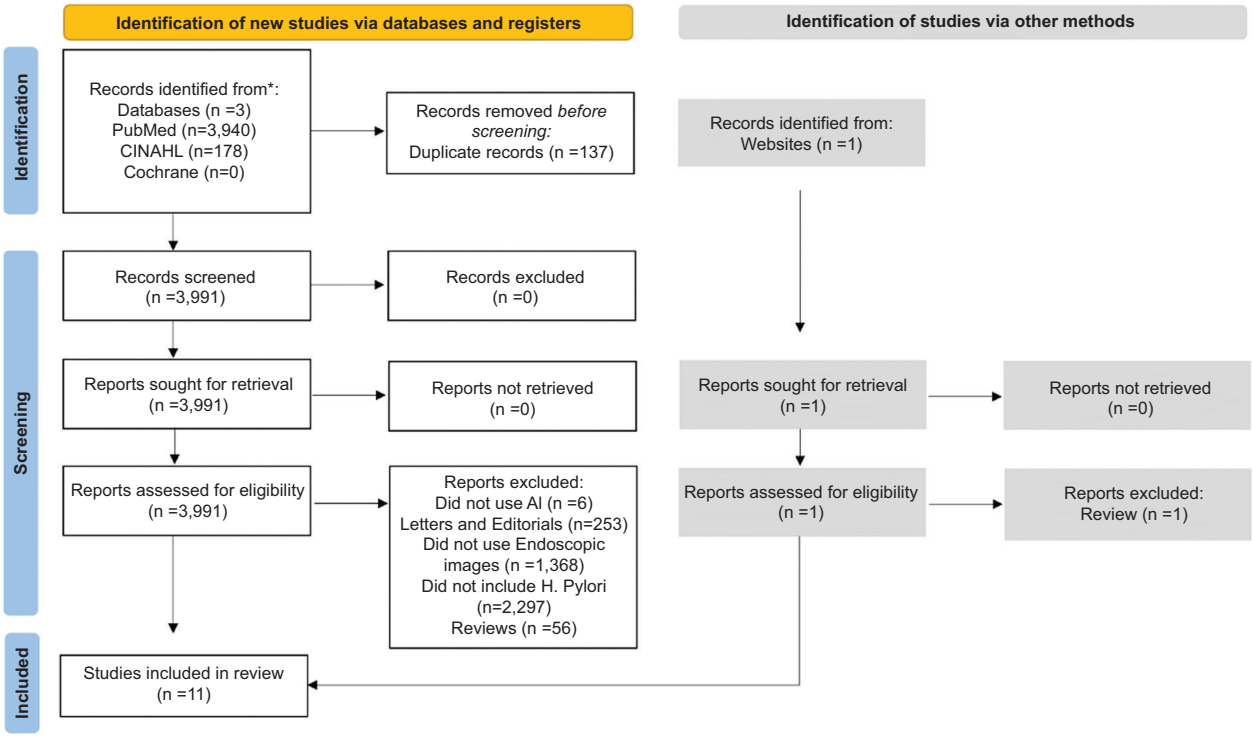


Figure 1 PRISMA flowchart

Table 1 Summary of characteristics of studies included (n=11)

Author, Year [ref.]	Country	Study Design	AI algorithm	Sample Size (Patients Images)	Endoscope System	TP	FP	TN	FN	PPV	NPV	Sensitivity	Specificity
Zhang *et al*, 2023 [17]	China	Retrospective Cohort	ResNet-50 and LSTM	47,239 images from 1,826 patients	Standard endoscope (GIF-HQ290, GIF-H260; Olympus, Tokyo, Japan; EG-L590ZW; Fujifilm, Tokyo,Japan)	56 78	4 7	58 77	6 6	93.3 91.8	90.6 92.8	90.3%	93.6%
Itoh *et al*, 2018 [10]	Japan	Prospective Cohort	GoogLeNet DCNN[d]	179 images from 139 patients	EG-L580NW endoscope (Fujifilm, Tokyo, Japan)	56	10	64	9	84.9	87.7	86.7%	86.7%
Nakashima *et al*, 2018 [7]	Japan	Prospective Pilot Cohort	GoogLeNet as a pretrained DCNN[d] model	2124 images from 222 patients	EG-L580NW instrument (FUJIFILM Co., Japan)	29 29 20	4 5 12	26 25 18	1 1 10	87.9 85.3 62.5	96.3 96.2 64.3	96.7%	86.7%
Nakashima *et al*, 2020 [12]	Japan	Prospective, single-center clinical study	Deep convolutional neural networks (DCNNd)	395 patients	EG-L580NW and EG-6400 N (Fujifilm Co., Japan)	25 24	6 11	74 69	15 16	80.6 68.6	83.2 81.2	62.5%	92.5%
Shichijo *et al*, 2017 [14]	Japan	Retrospective Cohort	Deep neural network architecture, GoogLeNet	32,208 images from 1768 in training set and 11,481 images from 397 patients in testing set	EGD[c] (EVIS GIF-XP290N, GIF-XP260, GIF-XP260NS, GIF-N260; Olympus Medical Systems, Co., Ltd., Tokyo, Japan)	59 64	54 41	271 284	13 8	52.1 60.9	95.4 97.3	88.9%	87.4%
Zheng *et al*, 2019 [18]	China	Retrospective Cohort	ResNet-50 (CNN[e])	11,729 images (training data set) and 3755 images (validation data set)	Standard endoscope (GIF-Q260J; Olympus, Tokyo, Japan)	252 284	14 2	128 140	58 26	94.7 99.3	68.8 84.3	81.4%	90.1%
Yoshi *et al*, 2020 [16]	Japan	Prospective Cohort	NM[a]	498 patients	Olympus H260 and Xp260NS	42	22	393	28	65.6	93.4	91.6%	88.6%
Yasuda *et al*, 2020 [15]	Japan	Retrospective Cohort	Support Vector Machine (Machine Learning)	525	LCI[b] based Upper gastrointestinal endoscope (EG-L590ZW or EGL600ZW) (Fujifilm Co.)	38	9	54	4	80.9	93.1	90.4%	85.7%
Yacob *et al*, 2023 [8]	Malaysia	Retrospective Cohort	Pre-trained CNN[e]: ShuffleNet version 1	20 patients	NM[a]	65	1	47	1	98.5	97.9	98.5%	97.9%

(*Contd...*)

Table 1 (*Continued*)

Author, Year [ref.]	Country	Study Design	AI algorithm	Sample Size (Patients Images)	Endoscope System	TP	FP	TN	FN	PPV	NPV	Sensitivity	Specificity
Seo *et al*, 2023 [13]	Korea	Retrospective Cohort	Inception-v3 using TensorFlow (CNN[e])	13,403 images from 952 patients (Training Dataset) 5,636 images from 411 patients (Test Dataset – Internal Validation | Koreans) 2,812 images from 131 patients (Test Dataset – Internal Validation | Non-Koreans) 1,338 images from 160 patients (External validation dataset)	Conventional white-light videoscopes (GIF-H260 and GIF-H290; Olympus Co, Ltd, Tokyo, Japan)	115 53 35	14 5 14	269 59 93	13 14 18	89.2 91.4 71.4	95.4 80.8 83.8	96%	90%
Lin *et al*, 2023 [11]	Taiwan	Retrospective Cohort	Convolutional Neural Network (CNN[e]) and Concurrent Spatial and Channel Squeeze and Excitation (scSE[f]) network, combined with different classification models for deep learning	302 patients (584 images in derivation dataset and 375 images in validation dataset)	Standard endoscope (GIF-Q260J; Olympus, Tokyo, Japan)	65	10	45	0	86.7	100	93%	81%

NM[a], not mentioned; LCI[b], linked color imaging; EGD[c], esophagogastroduodenoscopy; DCNN[d], deep convolutional neural networks; CNN[e], convolutional neural network; scSE[f], spatial and channel squeeze and excitation

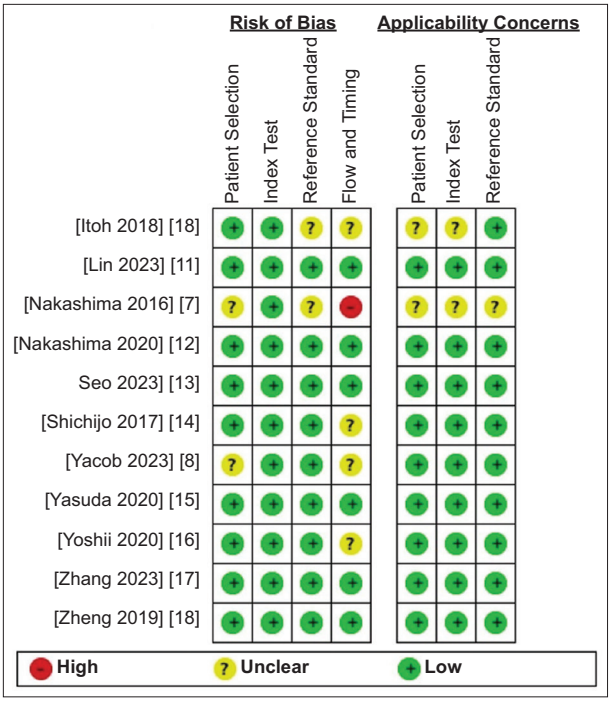


Figure 2 QUADAS 2 tool for risk of bias analysis of the included studies (n=11)

extracted from the included studies. The pooled estimates for the sensitivity and specificity of AI for diagnosing *H. pylori* infection via endoscopic images were found to be 0.90 (95% confidence interval [CI] 0.80-0.95), and 0.92 (95%CI 0.88-0.95).

Heterogeneity

The assessment of heterogeneity among the included studies was conducted using between-study heterogeneity statistics, found to be as: generalized heterogeneity: Tau^2=0.87, I^2=76.10%, sensitivity heterogeneity: Tau^2=1.53, I^2=80.72%, specificity heterogeneity: Tau^2=0.57, I^2=70.86%. These statistics indicate moderate to high heterogeneity across the studies concerning diagnostic accuracy, as illustrated by the forest plot in Fig. 3. The summary ROC curve of the included studies is presented in Fig. 4.

Subgroup analysis

Quality of study

The investigation into between-study heterogeneity statistics revealed a moderate level of heterogeneity across the studies, as indicated by a generalized Tau^2 of 0.87 and an I^2 value of 76.10%. The sensitivity and specificity analyses demonstrate high values of 0.80 and 0.71, respectively. The LR test comparing the random-effects model with the fixed-effects model yielded a statistically significant chi-squared value of 162.87, with 3 degrees of freedom and a P<0.001, suggesting the superiority of the random-effects model for this meta-analysis.

The study-specific test accuracy was further stratified based on the level of disease severity. Gastritis includes early-stage infection, mild gastritis and cases without significant gastric lesions, while peptic ulcer, atrophic gastritis, MALT lymphoma or cancer were considered as more severe. For studies focused on gastritis [8,10-18], the summary estimates reported a sensitivity of 0.90 (95%CI 0.80-0.95) and specificity of 0.92 (95%CI 0.88-0.95). In contrast, Nakashima *et al* 2018 [7], representing peptic ulcers, atrophic gastritis, MALT lymphoma, or cancer, exhibited a sensitivity of 0.97 (95%CI 0.83-1.00) and a specificity of 0.87 (95%CI 0.69-0.96). The overall summary estimates for both categories were consistent, with a sensitivity of 0.90 (95%CI 0.80-0.95) and specificity of 0.92 (95%CI 0.88-0.95). This analysis suggests that the diagnostic accuracy of the test remains robust across different disease categories (Supplementary Fig. 1).

Study format

The examination of between-study heterogeneity statistics revealed a moderate level of heterogeneity among the studies, with a generalized Tau^2 of 0.87 and an I^2 value of 76.10%. Sensitivity and specificity were notably high at 0.81 and 0.70, respectively. The LR test comparing the random-effects model with the fixed-effects model yielded a significant chi-squared value of 162.87, 3 degrees of freedom, and a P<0.001, supporting the preference for the random-effects model in this meta-analysis.

The study-specific test accuracy was further dissected based on the study design, distinguishing between prospective and retrospective studies. In the prospective studies [7,10,12,16], the summary estimates report a sensitivity of 0.90 (95%CI 0.80-0.95) and specificity of 0.92 (95%CI 0.88-0.95). For retrospective studies [8,11,13-15,17,18], similar summary estimates of sensitivity (0.90, 95%CI 0.80-0.95) and specificity (0.92, 95%CI 0.88-0.95) were observed (Supplementary Fig. 2). The overall summary estimates for both prospective and retrospective studies were consistent, further affirming the robustness of the diagnostic accuracy across different study designs.

Number of patients

The study-specific test accuracy analysis, focusing on absolute measures, provides insights into sensitivity and specificity across different subgroups based on the number of patients. For studies with more than 500 patients, Itoh *et al* 2018 [10] reported an estimated sensitivity of 0.86 (95%CI 0.75-0.93) and specificity of 0.86 (95%CI 0.77-0.93). Similarly, Seo *et al* 2023 [13], Zhang *et al* 2023 [17], and Zheng *et al* 2019 [18] found varying but favorable sensitivity and specificity estimates. The

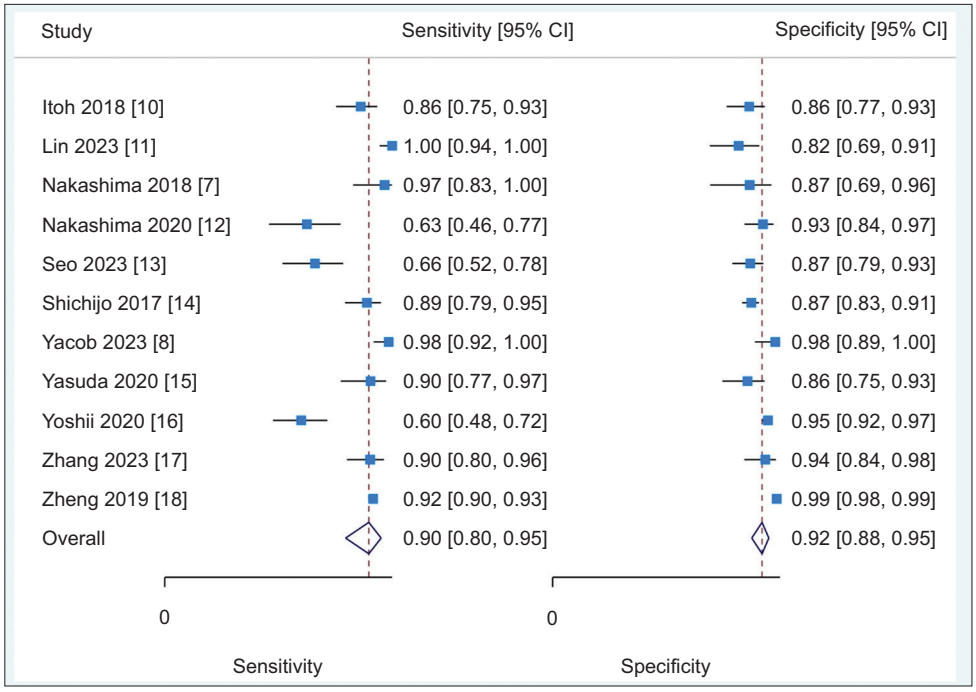


Figure 3 Forest plot of sensitivity and specificity of different studies using artificial intelligence for the detection of *Helicobacter pylori* infection
CI, confidence interval

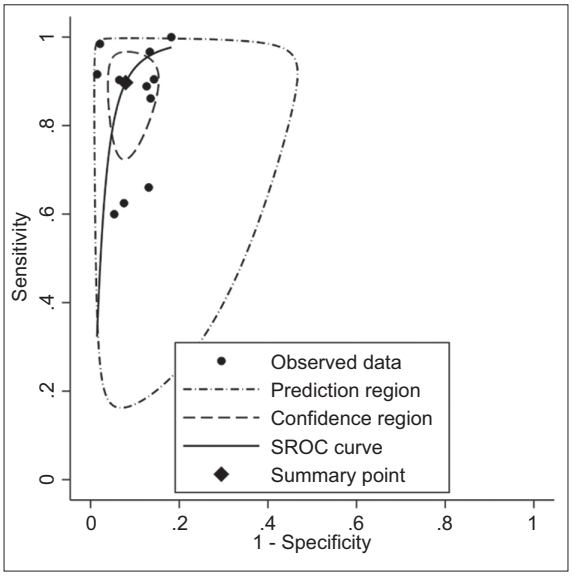


Figure 4 Summary receiver operating characteristic (SROC) curve with 95% confidential interval for predictions of artificial intelligence models for detecting *Helicobacter pylori* infection

summary for this subgroup indicates an overall sensitivity of 0.90 (95%CI 0.80-0.95) and specificity of 0.92 (95%CI 0.88-0.95) (Supplementary Fig. 3).

In the subgroup with fewer than 500 patients, Lin *et al* 2023 [11], Nakashima *et al* 2018 [7], Nakashima *et al* 2020 [12], Shichijo *et al* 2017 [14], Yacob *et al* 2023 [8], Yasuda *et al* 2020 [15], and Yoshii *et al* 2020 [16] present diverse sensitivity and specificity estimates. Notably, Lin *et al* 2023 [11] showed perfect sensitivity (1.00) and high specificity (0.91). The summary for this subgroup reports an overall sensitivity of 0.90 (95%CI 0.80-0.95) and specificity of 0.92 (95%CI 0.88-0.95).

The overall summary across all studies, irrespective of the number of patients, yields an estimated sensitivity of 0.90 (95%CI 0.80-0.95) and specificity of 0.92 (95%CI 0.88-0.95). These findings underscore the consistency of test accuracy measures across subgroups, reinforcing the robustness of the meta-analysis results in evaluating diagnostic performance within varying patient cohorts.

Published year

The examination of between-study heterogeneity statistics reveals a moderate level of heterogeneity across the included studies, as indicated by a generalized Tau^2 of 0.87 and an I^2 value of 76.10%. The sensitivity and specificity analyses show high values of 0.80 and 0.70, respectively. The LR test comparing the random-effects model with the fixed-effects model yielded a statistically significant chi-squared value of 162.87, with 3 degrees of freedom and a P<0.001, suggesting that the random-effects model is more appropriate for this meta-analysis.

Further exploration into study-specific test accuracy, categorized by the publication year, reveals varying estimates for sensitivity and specificity. Studies conducted before 2020 [7,10,14,18], exhibit favorable sensitivity and specificity values, contributing to an overall summary estimate of sensitivity at 0.90 (95%CI 0.80-0.95) and specificity at 0.92 (95%CI 0.88-0.95). Studies conducted after 2020 [8,11-13,15-17], present varying but generally high sensitivity and specificity estimates. The summary for this subgroup also reports an

overall sensitivity of 0.90 (95%CI 0.80-0.95) and specificity of 0.92 (95%CI 0.88-0.95) (Supplementary Fig. 4). The analysis by the publication year suggests consistent test accuracy measures across studies, emphasizing the robustness of the findings across different timeframes.

Type of AI

The examination of between-study heterogeneity statistics reveals a moderate level of heterogeneity across the included studies, as indicated by a generalized Tau^2 of 0.87 and an I^2 value of 76.10%. The sensitivity and specificity analyses show high values of 0.81 and 0.70, respectively. The LR test comparing the random-effects model with the fixed-effects model yields a statistically significant chi-squared value of 162.87, with 3 degrees of freedom and a P<0.001, suggesting that the random-effects model is more appropriate for this meta-analysis.

Further exploration into study-specific test accuracy, categorized by the machine learning algorithm used, reveals varying estimates for sensitivity and specificity. Studies utilizing convolutional neural networks (CNNs) [7,10-14,17,18], exhibit favorable sensitivity and specificity values. The summary estimate for this subgroup reports an overall sensitivity of 0.90 (95%CI 0.80-0.95) and specificity of 0.92 (95%CI 0.88-0.95). For studies employing a support vector machine (SVM), Yasuda *et al* 2020 [15] reported sensitivity and specificity values within the acceptable range, contributing to the overall summary estimate of sensitivity at 0.90 (95%CI 0.80-0.95) and specificity at 0.92 (95%CI 0.88-0.95). Additionally, Yoshii *et al* 2020 [16], using an unspecified machine learning algorithm, reports sensitivity at 0.60 (95%CI 0.48-0.72) and specificity at 0.95 (95%CI 0.92-0.97) (Supplementary Fig. 5). The analysis by machine learning algorithm suggests consistent test accuracy measures across different algorithms, emphasizing the robustness of the findings.

Discussion

This systematic review and meta-analysis aimed to provide evidence regarding the comparison of endoscopic and histopathological diagnosis versus AI-based diagnosis of the endoscopic images. The results of this systematic review and meta-analysis confirmed the high performance of AI-based methods for *H. pylori* detection from upper gastrointestinal endoscopy images. The pooled estimates for sensitivity and specificity were 0.90 (95%CI 0.80-0.95) and 0.92 (95%CI 0.88-0.95), respectively, with no significant heterogeneity among the studies. The area under the summary ROC curve was 0.97 (95%CI 0.96-0.99), indicating excellent diagnostic accuracy.

AI is an emerging technology that can analyze complex and large-scale data, such as endoscopy images, and provide automated and objective diagnosis [19]. AI-based methods for *H. pylori* detection mainly use deep learning techniques, such as CNNs, to extract features and classify endoscopy images into positive or negative for *H. pylori* infection. However, in our review, most of the studies employed deep learning algorithms. These methods can achieve high accuracy, sensitivity and specificity, comparable or superior to human experts [11]. Moreover, AI-based methods can reduce the workload and variability of endoscopists and provide real-time and noninvasive diagnosis of *H. pylori* infection. However, it is currently not clear how endoscopists and gastroenterologists would react and interact with diagnoses suggested by AI. Therefore, further studies aiming to assess AI's implementation into clinical practice can be of significant importance and value.

Gastroenterologists are clinicians who use various diagnostic techniques, such as endoscopies and colonoscopies, to reach a definitive diagnosis. AI-based models have proven to possess great specificities and sensitivities towards diagnosing various gastric pathologies including gastric polyps, Barrett's esophagus, celiac disease, and inflammatory bowel disease [20]. In our review, we found that most of the studies used deep learning AI algorithms with pooled sensitivities and specificities of 0.90 and 0.92. A similar systematic review and meta-analysis conducted by Bang *et al* found slightly lower pooled sensitivity and specificity (0.87 and 0.86) [21]. One reason for such differences could be that the AI models used are being continuously improved compared with previous versions; therefore, better outcomes are being generated.

Given that endoscopic biopsy is an invasive procedure, a significant percentage of patients may require fewer unnecessary biopsies if a highly accurate AI algorithm is applied during endoscopic inspection. However, at the same time, current AI models cannot completely eliminate the need for the expert opinion of endoscopists and gastroenterologists to make the final decision [22]. Presently, AI is rarely tasked to assist clinicians in making better clinical decisions and our results prove that this needs to be changed.

Although this systematic review and meta-analysis thoroughly assessed endoscopic histopathological diagnosis versus AI diagnosis for *H. pylori* infection, our analysis had some limitations. Firstly, the studies used different AI algorithms, endoscope systems and reference standards, which may have introduced heterogeneity and variability in the performance and quality of the AI methods. Secondly, the studies had different study designs, such as retrospective cohort, prospective cohort and clinical trial, which may have different levels of validity and reliability and could affect the risk of bias and confounding factors. Thirdly, some studies did not report some important information, such as the patient characteristics, the AI training and testing methods, the endoscopic image quality and resolution, and the potential sources of error and uncertainty. Fourthly, in this meta-analysis, we compared the sensitivity, specificity, and accuracy of several AI algorithms for identifying *H. pylori* infection. While this technique provides wide clinical insights, it has severe technological constraints. Different AI algorithms, such as CNNs and SVMs, have been created and evaluated in a variety of scenarios. The variability in datasets, preparation methodologies and implementation specifics used in research makes direct performance comparisons difficult. As a result, while our pooled estimates provide relevant clinical

information, they should be treated with caution, particularly when considering technical evaluations. Lastly, a notable constraint of our analysis is the geographic concentration of the included studies, which were all carried out in Asian nations. Given the large frequency of *H. pylori* infection in Asia, it is reasonable to expect considerable study from this region. However, this limits the generalizability of our results to non-Asian groups. Genetic, environmental, and behavioral differences across populations may influence the performance and application of AI systems. To ensure their worldwide applicability and efficacy, these AI models must be validated in a variety of demographic and geographical scenarios.

In conclusion, this systematic review and meta-analysis showed that AI had a high diagnostic accuracy for detecting *H. pylori* infection using endoscopic images. However, there was also a high heterogeneity among the studies, due to various factors that may affect the performance of AI. Therefore, more rigorous and consistent studies are needed to confirm and improve the reliability and validity of AI for diagnosing *H. pylori* infection in clinical practice.

Summary Box

What is already known:

- *Helicobacter pylori* (*H. pylori*) infection is responsible for various gastrointestinal (GI) pathologies with biopsy being the gold standard diagnostic modality

What the new findings are:

- Use of artificial intelligence to detect *H. pylori* infection by analyzing upper GI endoscopic images may facilitate diagnosis, decreasing the number of biopsies performed and reducing patient cost
- Such technologies can be used as an adjunct to decisions taken by doctors to treat patients suffering from *H. pylori* infection

References

1. Hooi JKY, Lai WY, Ng WK, et al. Global prevalence of *Helicobacter pylori* infection: systematic review and meta-analysis. *Gastroenterology* 2017;**153**:420-429.
2. Bang CS, Lee JJ, Baik GH. The most influential articles in Helicobacter pylori research: a bibliometric analysis. *Helicobacter* 2019;**24**:e12589.
3. Ahn HJ, Lee DS. *Helicobacter pylori* in gastric carcinogenesis. *World J Gastrointest Oncol* 2015;**7**:455-465.
4. Patel SK, Pratap CB, Jain AK, Gulati AK, Nath G. Diagnosis of *Helicobacter pylori*: what should be the gold standard? *World J Gastroenterol* 2014;**20**:12847-12859.
5. Glover B, Teare J, Patel N. A systematic review of the role of non-magnified endoscopy for the assessment of *H. pylori* infection. *Endosc Int Open* 2020;**8**:E105-E114.
6. Kim JW. Usefulness of narrow-band imaging in endoscopic submucosal dissection of the stomach. *Clin Endosc* 2018;**51**: 527-533.
7. Nakashima H, Kawahira H, Kawachi H, Sakaki N. Artificial intelligence diagnosis of *Helicobacter pylori* infection using blue laser imaging-bright and linked color imaging: a single-center prospective study. *Ann Gastroenterol* 2018;**31**:462-468.
8. Yacob YM, Alquran H, Mustafa WA, Alsalatie M, Sakim HAM, Lola MS. *H. pylori* related atrophic gastritis detection using enhanced convolution neural network (CNN) learner. *Diagnostics (Basel)* 2023;**13**:336.
9. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;**372**:n71.
10. Itoh T, Kawahira H, Nakashima H, Yata N. Deep learning analyzes *Helicobacter pylori* infection by upper gastrointestinal endoscopy images. *Endosc Int Open* 2018;**6**:E139-E144.
11. Lin CH, Hsu PI, Tseng CD, et al. Application of artificial intelligence in endoscopic image analysis for the diagnosis of a gastric cancer pathogen-*Helicobacter pylori* infection. *Sci Rep* 2023;**13**:13380.
12. Nakashima H, Kawahira H, Kawachi H, Sakaki N. Endoscopic three-categorical diagnosis of *Helicobacter pylori* infection using linked color imaging and deep learning: a single-center prospective study (with video). *Gastric Cancer* 2020;**23**:1033-1040.
13. Seo JY, Hong H, Ryu WS, Kim D, Chun J, Kwak MS. Development and validation of a convolutional neural network model for diagnosing *Helicobacter pylori* infections with endoscopic images: a multicenter study. *Gastrointest Endosc* 2023;**97**:880-888.
14. Shichijo S, Nomura S, Aoyama K, et al. Application of convolutional neural networks in the diagnosis of *Helicobacter pylori* infection based on endoscopic images. *EBioMedicine* 2017;**25**:106-111.
15. Yasuda T, Hiroyasu T, Hiwa S, et al. Potential of automatic diagnosis system with linked color imaging for diagnosis of *Helicobacter pylori* infection. *Dig Endosc* 2020;**32**:373-381.
16. Yoshii S, Mabe K, Watano K, et al. Validity of endoscopic features for the diagnosis of *Helicobacter pylori* infection status based on the Kyoto classification of gastritis. *Dig Endosc* 2020;**32**:74-83.
17. Zhang M, Pan J, Lin J, et al. An explainable artificial intelligence system for diagnosing *Helicobacter pylori* infection under endoscopy: a case-control study. *Therap Adv Gastroenterol* 2023;**16**:17562848231155023.
18. Zheng W, Zhang X, Kim JJ, et al. High accuracy of convolutional neural network for evaluation of *Helicobacter pylori* infection based on endoscopic images: preliminary experience. *Clin Transl Gastroenterol* 2019;**10**:e00109.
19. El Hajjar A, Rey JF. Artificial intelligence in gastrointestinal endoscopy: general overview. *Chin Med J (Engl)* 2020;**133**:326-334.
20. Visaggi P, de Bortoli N, Barberio B, et al. Artificial intelligence in the diagnosis of upper gastrointestinal diseases. *J Clin Gastroenterol* 2022;**56**:23-35.
21. 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**:e21983.
22. Song YQ, Mao XL, Zhou XB, et al. Use of artificial intelligence to improve the quality control of gastrointestinal endoscopy. *Front Med (Lausanne)* 2021;**8**:709347.

Supplementary material

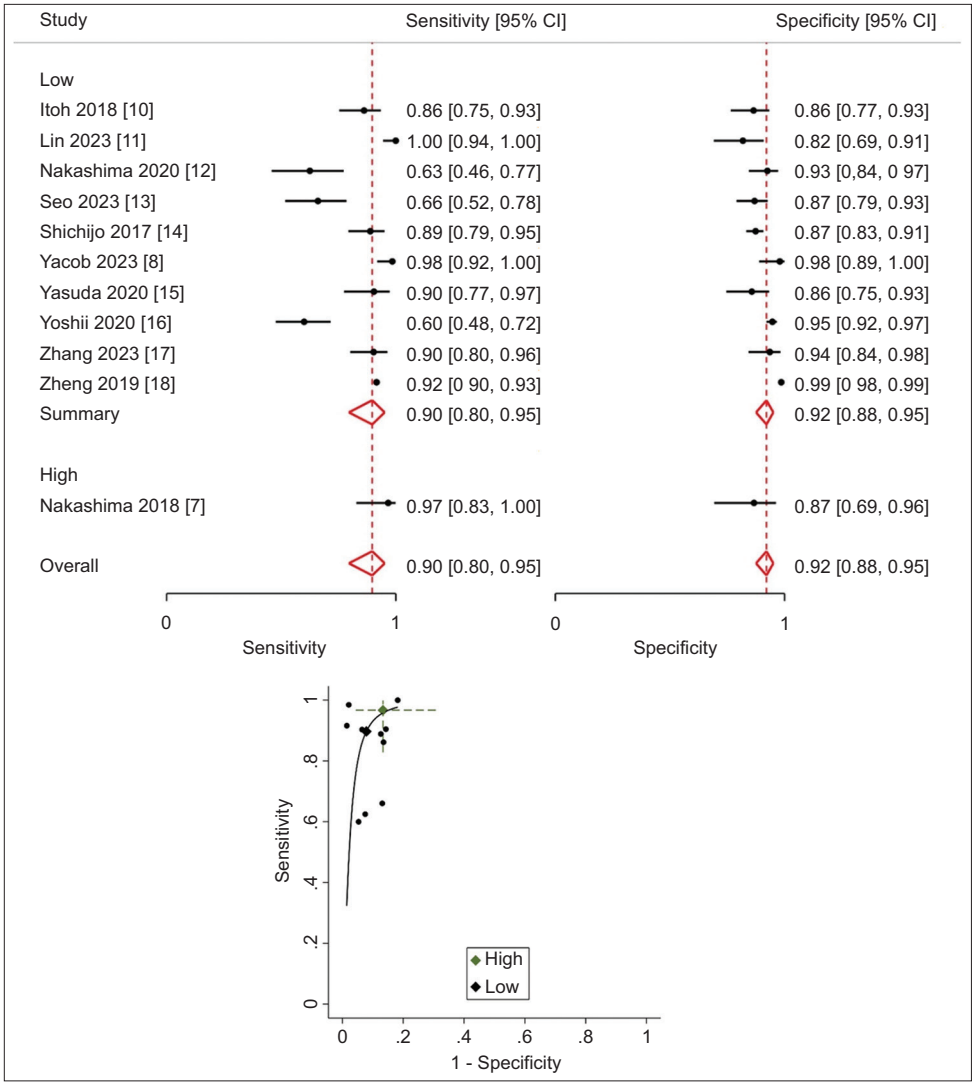
Supplementary Table 1 PRISMA checklist

Section and Topic	Item #	Checklist item	Location where item is reported
		TITLE	
Title	1	Identify the report as a systematic review.	1
		ABSTRACT	
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	3-4
		INTRODUCTION	
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	6-7
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	7
		METHODS	
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	7
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	8
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	8
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	8
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	8-9
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	8-9
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	8-9
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	9
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	-
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	10
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	10
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	10
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	10-15
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	10-15
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	-
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	10

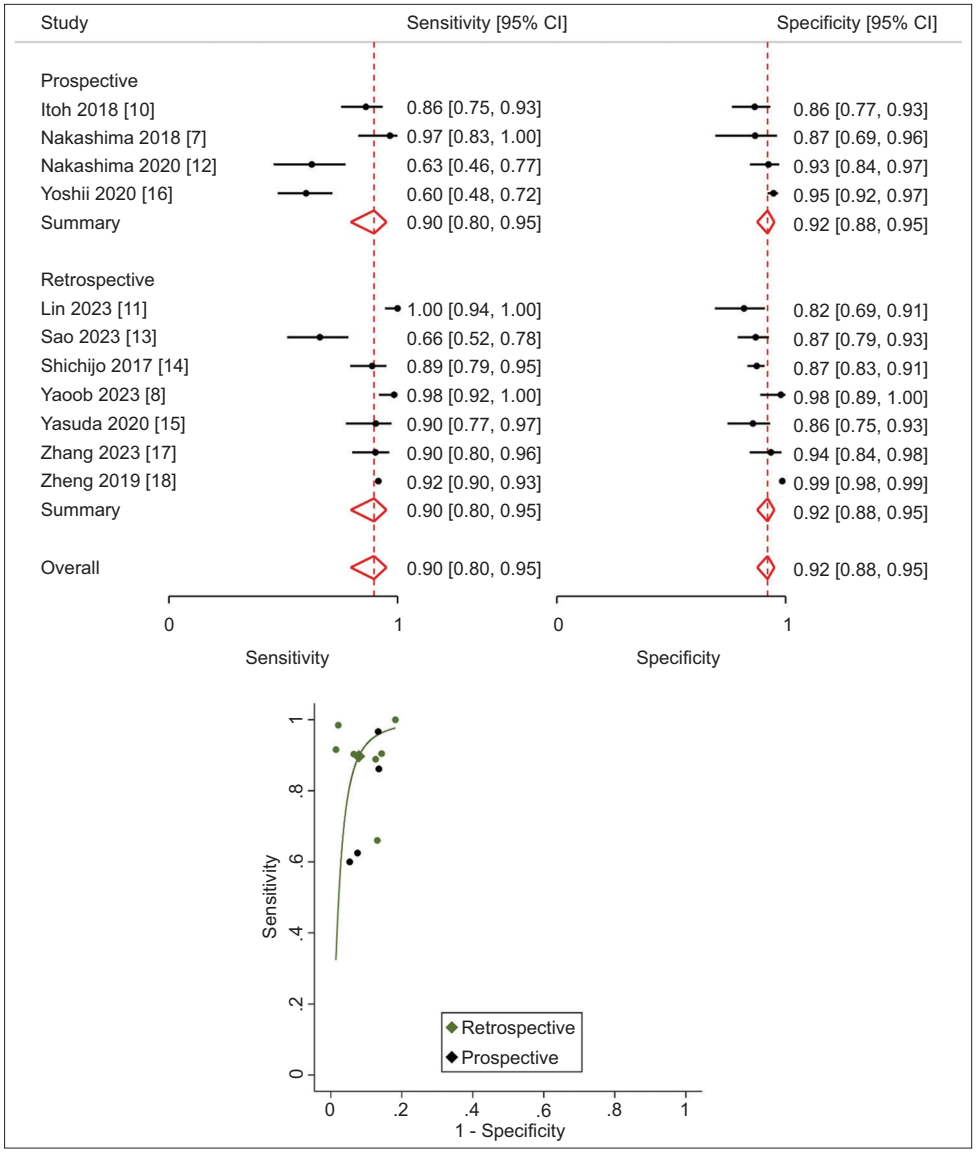
(*Contd...*)

Supplementary Table 1 PRISMA checklist Q. Please cite the Table in your manuscript (*Continued*)

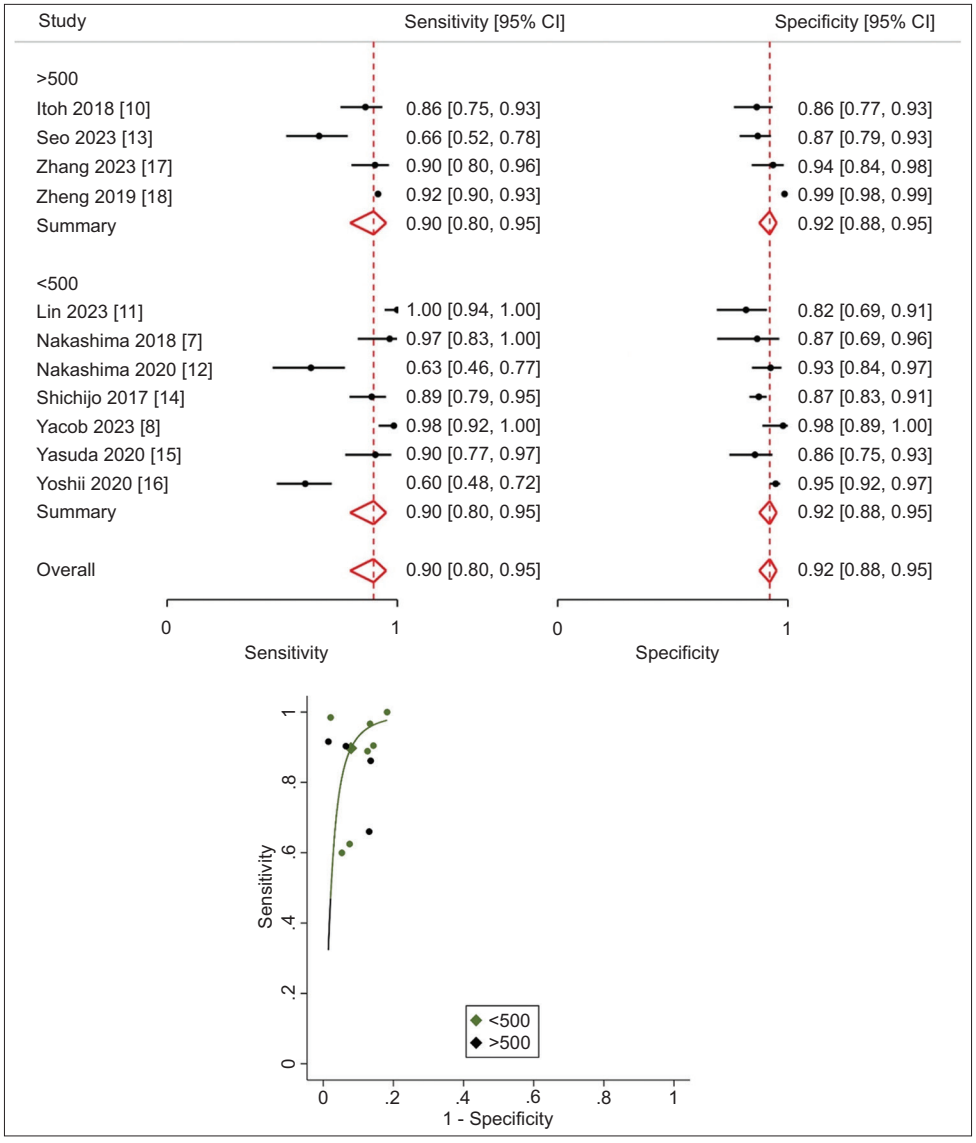
Section and Topic	Item #	Checklist item	Location where item is reported
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	-
		RESULTS	
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	9-10
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	10
Study characteristics	17	Cite each included study and present its characteristics.	24-26
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	10
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	24-26
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	9-15
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	10-15
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	10-15
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	10-15
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	10-15
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	-
		DISCUSSION	
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	16
	23b	Discuss any limitations of the evidence included in the review.	17-18
	23c	Discuss any limitations of the review processes used.	17-18
	23d	Discuss implications of the results for practice, policy, and future research.	17
		OTHER INFORMATION	
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	7
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	7
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	-
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	1
Competing interests	26	Declare any competing interests of review authors.	1
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	1



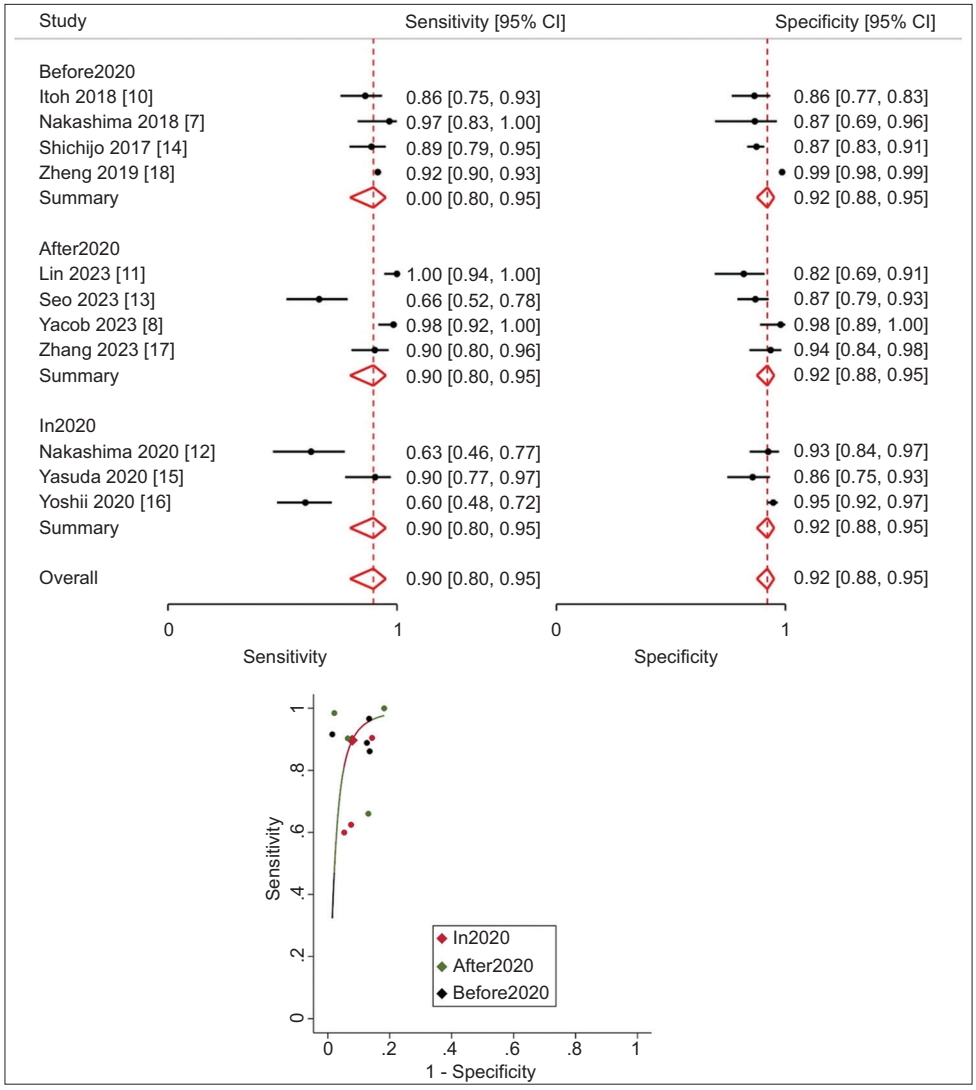
Supplementary Figure 1 Subgroup analysis based on different quality of studies
CI, confidence interval



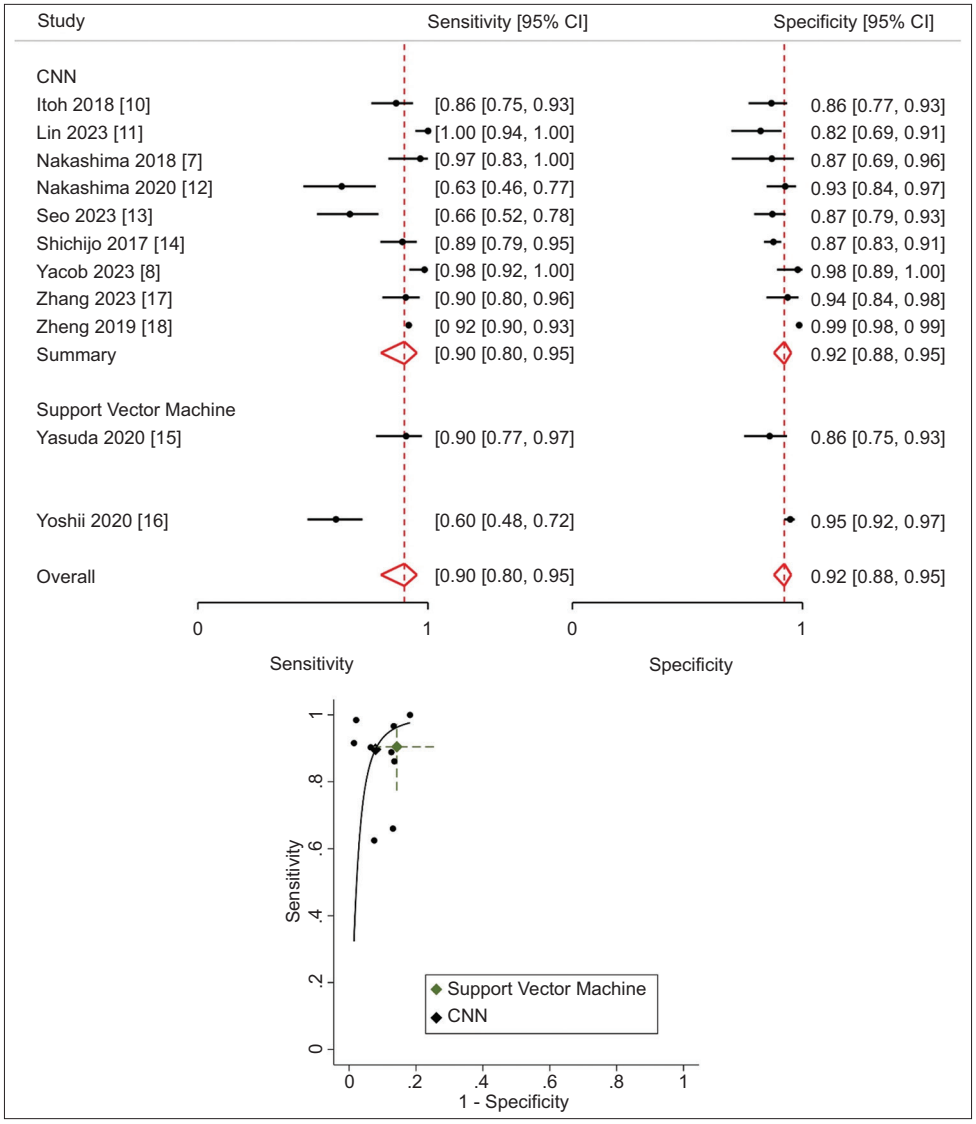
Supplementary Figure 2 Subgroup analysis based on different study formats
CI, confidence interval



Supplementary Figure 3 Subgroup analysis based on number of patients
CI, confidence interval



Supplementary Figure 4 Subgroup analysis based on published years
CI, confidence interval



Supplementary Figure 5 Subgroup analysis based on different artificial intelligence models
CI, confidence interval; CNN, convolutional neural network