Intelligent Oncology 1 (2025) 341–351

Contents lists available at ScienceDirect

Intelligent Oncology

journal homepage: www.keaipublishing.com/en/intonc/

Review

Performance evaluation of artificial intelligence–assisted diagnostic tools for human papillomavirus–related cervical and anal cancers and their precancerous lesions: A systematic review and meta-analysis

Hailin Yang[a,#], Zhuoru Zou[a,#], Zhenghong Li[a], Nyi Nyi Soe[b,c], Jason Ong[b,c,d], Phyu Mon Latt[b,c], Ying Zhang[b,c], Reshi Suthakaran[e], Lei Zhang[a,b,c,f,*]

[a] China-Australia Joint Research Center for Infectious Diseases, School of Public Health, Xi'an Jiaotong University Health Science Center, Xi'an Shaanxi 710061, China
[b] Artificial Intelligence and Modelling in Epidemiology Program, Melbourne Sexual Health Centre, Alfred Health, Melbourne Victoria 3053, Australia
[c] School of Translational Medicine, Faculty of Medicine, Nursing and Health Sciences, Monash University, Melbourne Victoria 3800, Australia
[d] Department of Clinical Research, London School of Hygiene and Tropical Medicine, London WC1E 7HT, United Kingdom
[e] Department of Colorectal Surgery, Western Health, Melbourne Victoria 3021, Australia
[f] Phase I Clinical Trial Research Ward, The Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an Shaanxi 710004, China

ARTICLE INFO

Keywords:
Cervical cancer
Anal cancer
Artificial intelligence
Computer-aided diagnosis
Meta-analysis

ABSTRACT

Detection of high-grade squamous intraepithelial lesions (HSILs) is key for the prevention of human papillomavirus (HPV)–related cancers. In this study, we aimed to identify and consolidate the existing evidence on the diagnostic performances of artificial intelligence (AI)–assisted tools for HPV-related HSILs. We followed the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy and the Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy guidelines and systematically searched PubMed, Embase, Web of Science, IEEE, and the Cochrane Library for eligible studies published between 2000 and January 2024. Studies on AI-assisted colposcopy used to facilitate the diagnosis of cervical or anal HSILs and cancers were included. The revised Quality Assessment of Diagnostic Studies 2 and AI checklists were used to assess the risk of bias. With random-effect models, meta-analyses were performed on performance indicators, including accuracy, sensitivity, specificity, and area under the receiver-operating characteristic curves (AUC). Subgroup and meta-regression analyses were performed to identify the sources of heterogeneity. Twenty-five studies, comprising 21 studies focused on cervical cancer and 4 on anal cancer, were included in the meta-analysis. When differentiating from individuals with <cervical intraepithelial neoplasia grade 2 (CIN2)/low-grade squamous intraepithelial lesion (LSIL)− (lesion grade <CIN2 or ≤LSIL), the AI-assisted diagnostic tools identified individuals with CIN2+/HSIL+ (lesion grade ≥CIN2 or ≥HSIL) at a pooled accuracy of 0.84 (95% confidence interval [CI]: 0.81–0.88), sensitivity of 0.87 (0.81–0.93), specificity of 0.85 (0.81–0.89), and AUC of 0.94 (0.91–0.96). For anal cancer, the AI-assisted diagnostic tools differentiated individuals with anal intraepithelial neoplasia grade 2 (AIN2)+/HSIL+ (≥AIN2 or ≥HSIL) from those with <AIN2/LSIL− (lesion grade <AIN2 or ≤LSIL) at a pooled accuracy of 0.90 (95% CI: 0.88–0.93), sensitivity of 0.94 (0.87–0.99), specificity of 0.88 (0.71–0.98), and AUC of 0.97 (0.95–0.98). All meta-analysis results demonstrated high heterogeneity (I^2 > 90%), but the meta-regression analysis did not reveal any significant contributing sources of heterogeneity. Based on this systematic review and meta-analysis of existing literature, we concluded that AI-assisted diagnostic tools have demonstrated promising predictive performance in diagnosing HPV-related cervical and anal precancerous lesions and cancers.

* Corresponding author.
E-mail address: lei.zhang1@monash.edu (L. Zhang).
# These authors contributed equally to this work.

https://doi.org/10.1016/j.intonc.2025.10.002
Received 24 August 2025; Received in revised form 28 September 2025; Accepted 8 October 2025
Available online 17 October 2025
2950-2616/© 2025 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Cervical cancer is the fourth most common cancer in terms of both incidence and mortality in women, with an estimated 660,000 new cases and 350,000 deaths worldwide in 2022 alone.[1] Human papillomavirus (HPV) infection is the key cause of cervical cancer.[2,3] Low-risk HPV subtypes usually do not lead to severe clinical outcomes, but high-risk HPV subtypes are carcinogenic.[4] High-risk HPVs cause most of the anogenital cancers (vulvar, vaginal, penile, and anal cancers) and a subset of head and neck cancers (i.e., tongue, tonsil, and oropharyngeal cancers).[5] A recent report has indicated that the incidence rate of anal squamous cell carcinoma has been increasing in both men and women in developed countries, such as Australia, Canada, the United States, and many European countries.[6] This trend is believed to be mainly driven by the widespread prevalence of HPV infection.[7,8] The risk of anal cancer also varies among populations, with the highest risk reported in people living with HIV, followed by men who have sex with men and transgender women.[9] Early detection and treatment of high-grade squamous intraepithelial lesions (HSILs) are crucial for the effective prevention of both anal and cervical cancers.[10]

Conventional cervical cancer screening techniques, including HPV DNA testing, cervical cytology, and visual inspection with acetic acid (VIA), are widely used, but they have several limitations.[11–13] These approaches typically require access to laboratory infrastructure, extended processing times, and the involvement of experienced clinicians and trained pathologists. Screening for anal cancer, which has pathophysiological parallels to cervical cancer, faces similar challenges. In a meta-analysis of anal cytology, the pooled sensitivities for low-grade squamous intraepithelial lesions (LSILs) and HSILs were only 0.61 (95% confidence interval [CI]: 0.42–0.77) and 0.27 (0.10–0.52), respectively.[14] Consequently, there is an increasing consensus that anal cytology alone may be insufficient and should be supplemented with direct visualization techniques, such as high-resolution anoscopy (HRA), to strengthen diagnostic accuracy. HRA and colposcopy enable real-time image acquisition and thus could facilitate the visual identification of HSILs.[15] High-resolution imaging has emerged as a promising standalone screening tool for HSILs and cancers owing to its superior sensitivity compared with conventional methods.[16]

The International Anal Neoplasia Society consensus guidelines underscore the importance of prioritizing high-risk individuals for immediate HRA referrals, especially in resource-limited settings, and advocate for the expansion of HRA screening infrastructure.[17] Consistent with these guidelines, the Acquired Immunodeficiency Syndrome Institute of the New York State Department of Health recommends HRA as the standard of care for individuals with abnormal anal cytological findings.[18]

Screening for HPV-related cancers involves the analysis of extensive imaging data, a process well suited to artificial intelligence (AI)–based approaches that enable automated image recognition, feature extraction, classification, and data processing. In recent years, AI-based diagnostic applications have gained momentum, demonstrating substantial potential in clinical practice.[19] The development of AI has significantly streamlined existing screening and diagnostic workflows,[20,21] facilitating real-time identification of HSILs and enabling point-of-care diagnosis. This reduction in diagnostic turnaround time supports earlier treatment initiation. By contrast, conventional manual analysis in HPV-related cancer screening remains prohibitively expensive, time-consuming, labor-intensive, and susceptible to human error.[22] These limitations are particularly acute in resource-constrained settings, where health-care systems often face disproportionate disease burdens and lack sufficient diagnostic capacity.[23] The integration of AI in diagnostic modalities provides the potential to reduce time demands, lessen the dependence on specialist expertise, and reduce subjective biases. In particular, deep learning has attracted increasing attention owing to its exceptional performance in computer vision tasks, including facial recognition and image classification.[24] Its capabilities extend across various domains of cancer research and clinical practice, especially in the automatic and accurate detection of malignancies from medical imaging. This technology holds considerable promise in alleviating the workload of pathologists and radiologists by automating routine, repetitive tasks.[25]

The application of AI in the early screening and diagnosis of HPV-related cancers provides a promising strategy to enhance diagnostic precision, with the potential to transform clinical pathways, particularly in resource-limited settings.[26–28] In recent years, AI has been increasingly used in cervical cancer diagnosis, not only in colposcopy but also in cytology systems for the automated analysis of cervical smears[29] and magnetic resonance imaging to improve preoperative staging.[30] In the context of anal cancer screening, HRA has also facilitated the rapid and accurate identification of HPV-associated anal lesions. Concurrently, innovations in high-resolution microendoscopy have enabled real-time imaging of squamous epithelial nuclei, supporting in vivo diagnosis and allowing for more targeted biopsies.[31,32] These technological advances have further accelerated the integration of AI into cervical and anal cancer screening protocols.

To date, numerous studies have investigated the application of AI in colposcopic and anoscopic image analyses.[33,34] In most cases, AI models were trained on microscopic visual images, using histopathological findings as the reference standard. In this study, we aimed to systematically synthesize the current evidence on the diagnostic performance of AI in detecting HPV-related HSILs and malignancies in the cervix and anus.

2. Methods

2.1. Protocol and registration

This study followed the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy and the Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy guidelines.[35] It was registered with PROSPERO (CRD42024527465).

2.2. Literature search strategy

A comprehensive literature search for studies published between 2000 and January 2024 was conducted in PubMed, Embase, Web of Science, IEEE, and the Cochrane Library. The search strategy incorporated four key concepts combined using logical operators: (i) the diseases of interest (cervical and anal cancers), (ii) the application of an AI-based algorithm (including automated detection, machine learning, deep learning, or AI), (iii) items related to diagnosis and screening (including diagnosis, screening, classification, discrimination, colposcopy, colposcopist, and colposcopic), and (iv) the performance metrics of the algorithms (including sensitivity, specificity, accuracy, area under the receiver operating characteristic curve [AUC], and calibration). The full search strategies and results are reported in the supplementary material (Table S1).

2.3. Studies selection

To ensure comparability, studies were included on the basis of the following inclusion criteria: Studies that evaluated an AI-based algorithm designed to differentiate between categories, such as cervical intraepithelial neoplasia grade (CIN) ≥ 2 (CIN2+, positive) and normal or benign conditions (negative), and HSILs (positive) and LSILs (negative) in the anus. Histopathological results were required as the gold standard for assessing algorithm accuracy for precancerous and cancerous lesions. For negative cases (normal or benign conditions), histopathology, normal cytology, and normal colposcopy were considered acceptable. This approach accounts for the fact that while histopathology remains the most reliable diagnostic method, some screening procedures perform biopsies only after suspicious findings are obtained

from cytology or colposcopy. We did not include studies that did not explicitly state the reference diagnostic method, those not in English, and those published outside the January 2000 to January 2024 timeframe. Review articles, patents, conference abstracts, and preprints were also not included.

In this study, the term *colposcopic images* refers to those captured during colposcopic examinations prior to any operative procedures. For instance, after the application of saline, the transformation zones and regions of interest were examined after the sequential use of 3%–5% acetic acid and Lugol's iodine solution.

2.4. Data collection and extraction

The literature search was conducted by a single reviewer (Y. H. L) who gathered data from PubMed, Embase, Web of Science, IEEE, and the Cochrane Library into Endnote 20.0, removing duplicates. Two reviewers (Y. H. L and L. Z. H) independently assessed all publications in a two-stage process: screening the titles and abstracts, and then reviewing the full texts. Disagreements were resolved by consulting a third reviewer (Z. Z. R).

For all the included studies, the following data were extracted: publication details (including title, first author, and year of publication), AI algorithms used, reference gold standard, number of patients included, dataset size and partitioning (training, validation, and testing), and the performance of the algorithms. The most dominant result pertaining to algorithm performance was the confusion matrix, which included true positives, false positives, true negatives, and false negatives. The results of some included studies did not directly give the confusion matrix value, and we derived part of the data retrospectively from sample size, accuracy, sensitivity, and specificity. Other primary outcomes included accuracy, sensitivity, specificity, Youden index, F1 score, AUC, and Kappa value.

To ensure clarity and consistency, the following symbols were used to denote the severity thresholds of diseases: <[disease grade], lesions with a severity less than the specified grade (e.g., <CIN2 signifies lesions less severe than CIN2); [disease grade]−, lesions with a severity less than or equal to the specified grade (e.g., LSIL− represents LSIL or less severe lesions); and [disease grade]+, lesions with a severity greater than or equal to the specified grade (e.g., CIN2+ represents CIN2 or more severe lesions). In the lesion grade, "/" represents "or," for example, <CIN2/LSIL− means that the lesion grade is lower than CIN2 or lower than/equal to LSIL).

2.5. Quality assessment

The qualities of the included studies were evaluated using the Quality Assessment of Diagnostic Studies-AI (QUADAS-AI)[36] checklists. Two reviewers (Y. H. L and L. Z. H) independently evaluated all the included studies using the QUADAS-AI. Disagreements were resolved by consulting a third evaluator (Z. Z. R). This tool represents an extension and revision of the QUADAS-2[37] and QUADAS-C[38] guidelines, and uses participant selection, index test, reference standard, and workflow to evaluate the risk of bias in studies assessing the accuracy of AI diagnostic experiments. The risk of bias and applicability concerns were evaluated with reference standards, index tests, and patient selection. Studies that used biopsies as controls were considered low-risk in the reference standard dimension (see details in Table S1).

2.6. Data analysis

Accuracy, sensitivity, specificity, and AUC estimates for each test are presented as forest plots and transformed using Freeman-Tukey double arcsine transformation[39–41] (except for AUC). Subgroup analyses were performed for studies with varying diagnostic thresholds. Statistical heterogeneity was assessed using the I^2 statistic, with values > 50% indicating substantial heterogeneity. Subgroup and meta-regression analyses were performed to identify potential sources of heterogeneity. When the number of subgroup studies was less than 3, the I^2 and P values were considered invalid.[42] We presented this graphically with a summary receiver-operating characteristic, including the summary point, AUC, 95% CI, and prediction contours. The Deeks funnel plot was used to detect potential publication bias, with P values >0.1 indicating no publication bias.

Meta-regression analyses were conducted using relevant covariates, including diagnostic threshold, the availability of data sources (studies with image data publicly or not publicly acccessible), the adoption of multimodal learning, the use of cross-validation procedures, the application of convolutional neural networks (CNNs) or their extensions, the implementation of transfer learning, and whether histological biopsy—considered the current gold standard for clinical diagnosis—was used as the reference standard. All were compared with scenarios where certain technology was not used, such as "using transfer learning" compared with "not using transfer learning" as a control. The relative diagnostic odds ratio was used to compare the effect of two different subgroups on diagnostic power, and the P value was used to test whether the relative diagnostic odds ratio was statistically significant. All meta-analyses and additional analyses were performed using STATA 15 (StataCorp LP, College Station, TX, USA), and Meta-disc 1.4. A significance level of 0.05 was applied to all test results.

3. Results

3.1. Study characteristics

A total of 938 records were retrieved from multiple databases. After the screening process and application of the inclusion and exclusion criteria, 25 studies were included in the meta-analysis to determine the diagnostic accuracy of AI-based tools (Figure 1). Of these studies, 21 focused on cervical cancer and 4 on anal cancer. In the studies that evaluated more than one algorithm, only the algorithm that yielded the highest diagnostic accuracy was included in the analysis. As studies that reported results under different diagnostic thresholds and results derived from images obtained using different methods were treated as separate data points, 38 data points were included in the final analysis. Of these data points, 34 were for cervical cancer and 4 were for anal cancer (Figure 1). Of the image data informing these studies, 89% (34/38) were collected from a clinical setting, while the rest were from public sources. The study data were collected from accumulated data in hospital departments, such as the gynecology and pathology departments. Most studies used advanced machine learning techniques, including CNNs or their extended frameworks in 84% (32/38) of the studies, cross-validation methods in 74% (28/38), and transfer learning in 82% (31/38; Table 1).

3.2. Quality assessment and publication bias

The risk of bias was assessed using the QUADAS-2 and QUADAS-AI checklists. Overall, 15 of the 25 studies had a high risk of bias in the patient selection, largely due to the recruitment of participants from clinical or institutional settings, which may limit generalizability. For the index test domain, 23 studies were considered high-risk, primarily due to the lack of external validation in the remaining studies. We found a moderate risk of bias in 1 study and a high risk of bias in 7 studies. However, most studies (22/25, 88%) were rated as having low applicability concerns for patient selection. All the studies were deemed low-risk with respect to the index test, and 72% (18/25) were rated as low-risk concerning the applicability of the reference standard (Details in Figure 2 and Table S2). The Deeks' funnel plot analysis showed no evidence of significant publication bias (P = 0.119; Supplementary material).

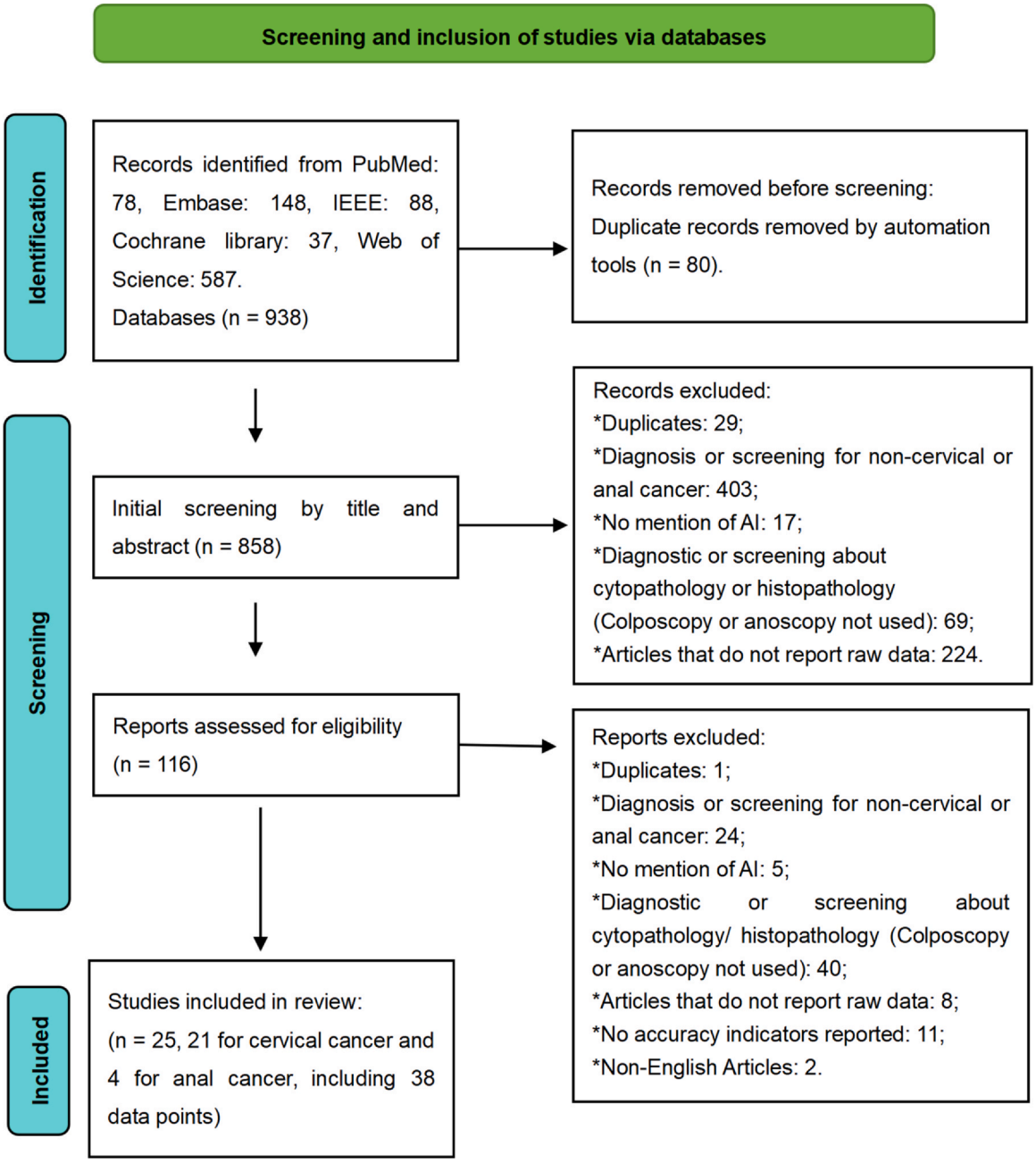


Figure 1. Flow diagram of publication screening.

3.3. Meta-analysis of key performance indicators

Overall, the AI-assisted diagnostic tools for cervical and anal cancers demonstrated high diagnostic accuracy (Table 2 and Figure 3. For cervical cancer, when distinguishing individuals with CIN1+ or LSIL+ from those with normal cytology, the AI-assisted tools achieved a pooled sensitivity of 0.91 (95% CI: 0.87–0.95), specificity of 0.89 (0.78–0.97), and AUC of 0.97 (0.95–0.98), with an overall accuracy of 0.90 (0.83–0.95). When identifying cases with CIN2+ or HSIL+ relative to <CIN2 or LSIL−, their pooled sensitivity was 0.87 (0.81–0.93), specificity was 0.85 (0.81–0.89), and AUC was 0.94 (0.91–0.96), corresponding to an accuracy of 0.84 (0.81–0.88).

Performance in anal cancer diagnosis was similarly high. Compared with individuals with < anal intraepithelial neoplasia grade 2 (AIN2) and LSIL, those with AIN2+ and HSIL were identified with a pooled sensitivity of 0.94 (0.87–0.99), specificity of 0.88 (0.71–0.98), and AUC of 0.97 (0.95–0.98), aligning with a pooled accuracy of 0.90 (0.88–0.93). Among the various diagnostic thresholds examined, the threshold that identified CIN3+ cases exhibited the highest pooled sensitivity of 0.96 (0.92–0.99) but was also accompanied by the lowest specificity of 0.65 (0.62–0.69). The only comparison groups with AUC values < 0.95 were <CIN2 vs. CIN2+ and < LSIL vs. HSIL+. On the other hand, the comparison between the VIA+ and VIA− groups yielded the highest AUC value of 0.98 (0.96–0.99).

3.4. Subgroup analyses of key performance indicators

For cervical cancer, studies based on public datasets consistently demonstrated superior diagnostic metrics, with a pooled sensitivity, specificity, and accuracy of 0.95 (95% CI: 0.90–0.99, 0.86–0.99, and 0.87–0.99, respectively), and an AUC of 0.96 (0.94–0.97). By contrast, studies that used non-public datasets showed comparatively lower performance, with a sensitivity of 0.89 (95% CI: 0.86–0.92), specificity of 0.85 (0.80–0.90), AUC of 0.95 (0.93–0.97), and accuracy of 0.88

Table 1
Summary of marker-based motion detection publications.

Author	Year of publication	Sample size (number of images)	Effective sample size	Study location	AI-powered method (Intervention AI methodology)	Standard diagnostic method	Classification of region of interest	Classification threshold	Type of image	TP	FP	TN	FN	Accuracy	Sensitivity (recall)	Specificity	PPV	NPV	AUC
Adweb et al.[43]	2021	(4800)	3168	NA	DCNN (ReLU-ResNet)	histopathology	automated	Normal vs. CIN1 +	Colposcopy	1565	30	1538	35	0.98	0.91	0.96	0.98	0.98	0.93
Elakkiya et al.[44]	2022	(3105)	621	NA	FSOD-GAN (combines Faster R-CNN and Generative Adversarial Network)	NA	automated	Normal vs. CIN1 +	Colposcopy	392	7	220	2	0.98	0.99	0.97	0.98	0.99	NA
Nurmaini et al.[45]	2023	(250)	110	Indonesia	CNNs (CervicoXNet)	NA	automated	Normal vs. CIN1 +	Colposcopy	59	1	49	1	0.98	0.98	0.98	0.98	0.98	NA
Deo et al.[46]	2023	110 (392)	40	India	Entropic features and CNN	biopsy	NA	Normal vs. CIN1 +	Colposcopy	20	0	17	3	0.93	0.97	1	1	0.85	NA
Fan et al.[47]	2022	7106 (7108)	891	NA	Colposcopic multimodal fusion CNN	biopsy	automated	Normal vs. LSIL +	Colposcopy	561	29	265	36	0.93	0.94	0.9	0.95	0.88	NA
Chandran et al.[48]	2021	(11,266)	1884	NA	DCNN (VGG − 19)	NA	artificial	Normal vs. CIN1 +	Colposcopy	559	49	1230	46	0.92	0.92	0.96	0.92	0.95	NA
Fan et al.	2022	7106 (7106)	891	NA	Colposcopic multimodal fusion CNN	biopsy	automated	Normal vs. LSIL +	Colposcopy	557	35	264	35	0.92	0.94	0.88	0.94	0.88	NA
Liu et al.[33]	2021	7530 (15,278)	1506	China	CNN (ResNet)	NA	artificial	Normal vs. LSIL +	Colposcopy	687	47	647	125	0.89	0.93	0.85	0.84	0.94	0.953
Yuan et al.[49]	2020	(22,396)	2233	China	Mask R-CNN	pathology	automated	Normal vs. LSIL +	Colposcopy	1022	856	180	175	0.88	0.85	0.89	0.2	0.96	NA
Yuan et al.	2020	22,330	3533	China	Mask R-CNN	pathology	automated	Normal vs. LSIL +	Colposcopy	624	297	2499	113	0.88	0.85	0.89	0.21	0.95	NA
Xue et al.[50]	2020	19,435 (101,268)	3887	China	CNN	histopathology	automated	Normal vs. LSIL +	Colposcopy	2451	54	1151	231	0.86	0.72	0.94	0.98	0.83	NA
Yuan et al.	2020	22,330	3575	China	Mask R-CNN	pathology	automated	Normal vs. LSIL +	Colposcopy	667	311	2477	120	0.84	0.85	0.82	0.85	0.83	0.93
Liu et al.	2021	7530 (15,276)	1506	China	CNN (ResNet)	NA	artificial	Normal vs. LSIL +	Colposcopy	579	247	447	233	0.68	0.64	0.71	0.68	0.71	0.69
Bountris et al.[51]	2014	740	100	Greece	CDSS (probabilistic neural network and multilayer perceptron network)	biopsy	NA	<CIN2 vs. CIN2 +	Cytology	23	4	71	2	0.94	0.89	0.97	0.89	0.97	NA
Kim et al.[52]	2023	(886)	886	Republic of Korea	NA	biopsy	NA	<CIN2 vs. CIN2 +	Colposcopy	692	14	164	16	0.97	0.98	0.92	0.98	0.91	0.95
Fan et al.	2022	7106 (7109)	891	NA	Colposcopic multimodal fusion CNN	biopsy	automated	LSIL- vs. HSIL +	Colposcopy	537	8	289	57	0.93	0.9	0.97	0.99	0.84	NA
Fan et al.	2022	7106 (7107)	891	NA	Colposcopic multimodal fusion CNN	biopsy	automated	LSIL- vs. HSIL +	Colposcopy	533	9	288	61	0.92	0.9	0.97	0.98	0.83	NA
Hu et al.[53]	2023	5188 (13,484)	8917	Zambia	Applicable, but not explicitly described in the article.	biopsy	NA	<CIN2 vs. CIN2 +	Colposcopy	212	1480	7209	16	0.86	0.85	0.86	0.88	0.84	0.91
Pal et al.[54]	2021	(9327)	8259	NA	BH-NasNet − 1-NN	NA	NA	<CIN2 vs. CIN2 +	Colposcopy	71	1213	6961	14	0.85	0.84	0.85	0.06	0.99	NA
Hu et al.[55]	2019	9406 (9406)	203	Costa Rica	Faster R-CNN	biopsy	automated	<CIN2 vs. CIN2 +	Colposcopy	92	13	82	16	0.83	0.93	0.93	0.13	0.99	0.91

(*continued on next page*)

Table 1 (*continued*)

Author	Year of publication	Sample size (number of images)	Effective sample size	Study location	AI-powered method (Intervention AI methodology)	Standard diagnostic method	Classification of region of interest	Classification threshold	Type of image	TP	FP	TN	FN	Accuracy	Sensitivity (recall)	Specificity	PPV	NPV	AUC
Fu et al.[56]	2022	2160	1452	China	DCNN (VGG−16)	biopsy	NA	LSIL- vs. HSIL+	Colposcopy	77	233	1131	11	0.83	0.86	0.83	0.26	0.99	0.92
Liu et al.	2021	7530 (15,281)	1506	China	CNN (ResNet)	NA	artificial	LSIL- vs. HSIL+	Colposcopy	349	216	866	75	0.81	0.82	0.8	0.62	0.92	0.9
Xue et al.	2020	19,435 (101,267)	3887	China	CNN	histopathology	automated	LSIL- vs. HSIL+	Colposcopy	1074	145	2248	420	0.81	0.66	0.9	0.88	0.84	NA
Park et al.[57]	2008	29	29	NA	KNN	histopathology	automated	LSIL- vs. HSIL+	Colposcopy	15	1	7	6	0.76	0.82	0.73	0.94	0.54	NA
Hunt et al.[58]	2022	1486	314	NA	multi-task CNN	histopathology	automated	<CIN2 vs. CIN2+	Colposcopy	87	86	137	4	0.71	0.96	0.61	0.5	0.97	NA
Hunt et al.	2022	1486	314	NA	multi-task CNN	histopathology	automated	<CIN2 vs. CIN2+	HRME	84	91	132	7	0.69	0.93	0.59	0.48	0.95	0.86
Liu et al.	2021	7530 (15,279)	1506	China	CNN (ResNet)	NA	artificial	LSIL- vs. HSIL+	Colposcopy	307	373	709	117	0.68	0.72	0.66	0.45	0.86	0.7
Hunt et al.	2022	1486	314	NA	multi-task CNN	histopathology	automated	<CIN3 vs. CIN3+	Colposcopy	69	105	137	3	0.66	0.96	0.57	0.4	0.98	NA
Hunt et al.	2022	1486	314	NA	multi-task CNN	histopathology	automated	<CIN3 vs. CIN3+	HRME	69	107	135	3	0.65	0.96	0.56	0.39	0.98	0.85
Yue et al.[59]	2020	679 (4753)	679	China	CNN	histopathology	automated	Normal, CIN1, CIN2/3, cervical cancer	Colposcopy	255	7	404	13	0.96	0.95	0.98	0.97	0.97	0.96
Harsono et al.[60]	2022	199 (398)	65	Indonesia	Gradient Boosting Classifier	Oncologist assessed images	automated	VIA- vs. VIA+	Cervicogram	8	2	53	2	0.94	0.8	0.96	0.8	0.96	0.85
Nurmaini et al.	2023	(779)	179	Indonesia	CNNs (CervicoXNet)	NA	automated	VIA- vs. VIA+	Cervicogram	121	1	57	0	0.98	1	0.98	0.88	1	NA
Bae et al.[61]	2020	20 (240)	120	Republic of Korea	KNN	histopathology	automated	VIA- vs. VIA+	Colposcopy	23	14	74	9	0.78	0.75	0.8	0.62	0.89	0.81
Ye et al.[62]	2021	(4119)	824	Republic of Korea	Extreme Gradient Boost	Negative:visible lesions or symptoms; Positive:clearly containing lesions (naked eye)	automated	VIA- vs. VIA+	Cervicogram	381	27	370	46	0.74	0.78	0.71	0.93	0.89	0.82
Saraiva et al.[27]	2022	71 (5026)	503	France	CNN	HRA-guided biopsies	automated	LSIL vs. HSIL*	HRA	146	29	312	16	0.91	0.9	0.92	0.82	0.96	0.97
Brenes et al.[31]	2023	49 (104)	99	USA	MTN	histopathology	NA	<AIN2 vs. AIN2+	HRME	36	24	36	3	0.88	0.91	0.87	0.65	0.97	0.96
Brenes et al.[32]	2023	67 (99)	104	NA	MTN	biopsy, HRA	automated	<AIN2 vs. AIN2+	HRME	20	11	71	2	0.73	0.92	0.6	0.6	0.92	0.84
Saraiva et al.[28]	2024	(27,770)	8532	France	CNN	HRA-guided biopsies	automated	LSIL vs. HSIL*	HRA	4007	36	4384	105	0.96	0.96	0.94	0.99	0.89	0.99

Abbreviations: 1. TP: true positive 2. FP: false positive 3. TN: true negative 4. FN: false negative 5. PPV: positive predictive value 6. NPV: negative predictive value 7. AUC: area under the receiver operating characteristic curve 8. NA: not available 9. CNN: convolutional neural network 10. R-CNN: regions with CNN features 11. KNN: K-nearest neighbor 12. HRA: high-resolution anoscopy 13. HRME: high-resolution microendoscopy 14. MTN: multitask deep learning network 15. BH: batch-hard loss minimization.
*Note: There are only two classifications, with no more severe or less severe lesion levels.

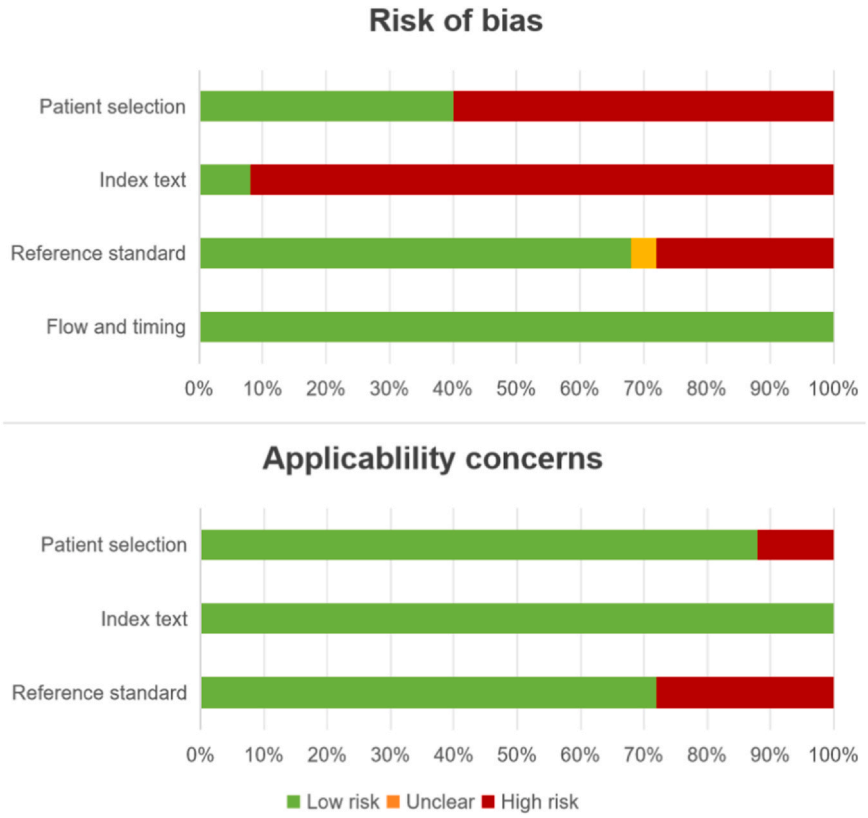


Figure 2. Quality assessment of included articles using the QUADAS-AI criteria.

(0.82–0.89). For anal cancer, among the included covariates, only the model differed; the other covariates were consistent. The CNN-based model showed a high accuracy of 0.98 and a sensitivity and specificity close to 1 (0.97 and 0.99, respectively).

3.5. Heterogeneity and meta-regression analysis

Substantial heterogeneity was observed across all the subgroup analyses, as evidenced by the I^2 values exceeding 50%. Meta-regression analysis was performed to identify potential sources of heterogeneity, including diagnostic thresholds, data source type, CNN usage, cross-validation, transfer learning, and reference standards. However, none of these covariates reached statistical significance ($P > 0.05$), and no single source of heterogeneity could be definitively identified.

4. Discussion

This study evaluated the diagnostic performances of AI-assisted tools in the detection of cervical and anal cancers and their precancerous lesions. Our meta-analysis of the included 25 studies confirmed that AI-assisted tools have high diagnostic performance for cervical cancer, with an overall sensitivity of 0.90 (95% CI: 0.87–0.93), specificity of 0.87 (0.82–0.91), AUC of 0.96 (0.94–0.97), and accuracy of 0.87 (0.84–0.90). AI-assisted diagnostic tools for anal cancer also exhibited comparable diagnostic performance, with a pooled sensitivity of 0.94 (0.87–0.99), specificity of 0.88 (0.71–0.98), and AUC of 0.97 (0.95–0.98) in distinguishing AIN2+/HSIL+ cases from <AIN2/LSIL− cases. These metrics collectively reflect an overall accuracy of 0.90 (95% CI: 0.88–0.93). Compared with earlier studies that used portable diagnostic devices for cervical cancer in low- and middle-income countries, which reported a sensitivity of 0.79 and specificity of 0.83,[63] the present study found a markedly improved diagnostic performance. This likely reflects advancements in AI model architectures and training methods, and improvements in device accuracies.

AI-assisted colposcopy and anoscopy have demonstrated notable precision in identifying cervical and anal cancers and their precursor lesions, underlining their value for clinical deployment. The timeliness and accuracy of cancer detection are critical determinants of treatment outcomes, and AI contributes substantially to these goals. AI enables scalable diagnostic support, with performance metrics comparable with expert clinicians, while allowing for automation and reproducibility.[64] In our subgroup and meta-regression analyses, key model characteristics, including diagnostic threshold, data source, use of CNNs, cross-validation, transfer learning, and reference standard, were examined for their potential contributions to heterogeneity. Most studies (17/25) used CNNs or related architectures, underscoring their central role in image-based classification tasks in oncology.[65] Datasets derived from public sources yielded superior performance, with both sensitivity and specificity reaching 0.96,[66] which may reflect the exclusion of low-quality or mislabeled images from public datasets. By contrast, studies with image data not publicly accessible are often narrower in scope and more susceptible to bias.[67] Despite the substantial heterogeneity observed, the meta-regression analysis did not show any statistically significant moderators, likely indicating the influence of unmeasured confounders.

Several factors may explain this residual heterogeneity. First, variations in image quality and acquisition protocols are fundamental drivers. Data from multiple centers and equipment types result in differences in resolution, contrast, and illumination that directly affect model performance. Second, the subjectivity of reference standards introduces unavoidable labeling noise. Interpretation of medical images, particularly borderline cases, carries inherent interobserver variability, which means that the gold standard for training and evaluation may itself be a source of variation.[68,69] Third, latent heterogeneity in model architectures extends well beyond the binary distinction of whether CNNs were used. The included studies implemented diverse CNN variants with differing receptive fields and inductive biases. They showed that variations in hyperparameters, such as learning rates, optimizers, data augmentation strategies, and regularization

Table 2
Pooled effects and heterogeneity in subgroup analysis.

	No. of data points	Accuracy	Sensitivity	Specificity	AUC of SROC	Relative diagnostic odds rate	*P* value
Cervical cancer:	34	0.87 (0.84—0.90)	0.90 (0.87—0.93)	0.87 (0.82—0.91)	0.96 (0.94—0.97)		
Threshold						0.60 (0.17—2.14)[a]	0.4135
Normal vs. CIN1 + /LSIL +	13	0.90 (0.83—0.95)	0.91 (0.87—0.95)	0.89 (0.78—0.97)	0.97 (0.95—0.98)		
(<CIN2 vs. CIN2 +) or (LSIL- vs. HSIL +)	14	0.84 (0.81—0.88)	0.87 (0.81—0.93)	0.85 (0.81—0.89)	0.94 (0.91—0.96)		
<CIN3 vs. CIN3 +	2	0.65 (0.62—0.69)	0.96 (0.92—0.99)	0.56 (0.52—0.61)	/		
Normal, CIN1, CIN2/3, cervical cancer	1	0.97 (0.96—0.98)	0.95 (0.92—0.97)	0.98 (0.97—0.99)	/		
VIA- vs. VIA +	4	0.93 (0.84—0.98)	0.90 (0.73—0.97)	0.95 (0.86—0.98)	0.98 (0.96—0.99)		
Data sources						5.03 (0.54—46.88)	0.1490
Public	4	0.95 (0.87—0.99)	0.95 (0.90—0.99)	0.95 (0.86—0.99)	0.99 (0.98—1.00)		
Proprietary (ref)	30	0.88 (0.82—0.89)	0.89 (0.86—0.92)	0.85 (0.80—0.90)	0.95 (0.93—0.97)		
Model						0.52 (0.08—3.37)	0.4757
CNN and extensions	28	0.86 (0.83—0.89)	0.90 (0.87—0.93)	0.85 (0.80—0.90)	0.96 (0.93—0.97)		
Others (ref)	6	0.93 (0.88—0.97)	0.92 (0.86—0.97)	0.94 (0.88—0.99)	0.97 (0.96—0.98)		
Cross-validation						1.50 (0.30—7.45)	0.7807
Yes	10	0.92 (0.90—0.94)	0.91 (0.88—0.94)	0.95 (0.91—0.97)	0.97 (0.95–0.98)		
No (ref)	24	0.86 (0.81—0.89)	0.90 (0.86—0.94)	0.83 (0.77—0.88)	0.96 (0.94—0.97)		
Transfer learning						5.30 (0.64—43.86)	0.1168
Yes	7	0.90 (0.87—0.93)	0.90 (0.87—0.93)	0.91 (0.87—0.94)	0.96 (0.94—0.97)		
No (ref)	27	0.86 (0.82—0.91)	0.90 (0.86—0.93)	0.86 (0.79—0.92)	0.96 (0.94—0.97)		
Multi-modal learning						0.23 (0.05—1.10)	0.0637
Yes	13	0.84 (0.78—0.90)	0.86 (0.82—0.90)	0.83 (0.72—0.92)	0.92 (0.89—0.94)		
No (ref)	25	0.89 (0.85—0.92)	0.93 (0.89—0.95)	0.89 (0.84—0.93)	0.97 (0.96—0.98)		
Reference standard						0.64 (0.12—3.35)	0.5819
Biopsy	23	0.86 (0.81—0.90)	0.90 (0.87—0.94)	0.85 (0.78—0.91)	0.96 (0.93—0.97)		
Others (ref)	11	0.90 (0.84—0.94)	0.89 (0.82—0.95)	0.90 (0.83—0.95)	0.97 (0.95—0.98)		
Anal cancer:							
Threshold						/	/
(<AIN2 vs. AIN2 +) or (LSIL vs. HSIL)	4	0.90 (0.88—0.93)	0.94 (0.87—0.99)	0.88 (0.71—0.98)	0.97 (0.95—0.98)		
Model						20.44 (0—187065)	0.2906
CNN	2	0.98 (0.98—0.98)	0.97 (0.97—0.98)	0.99 (0.99—0.99)	/		
MTN (ref)	2	0.81 (0.75—0.86)	0.92 (0.83—0.98)	0.77 (0.69—0.83)	/		

Abbreviations: 1. CIN: cervical intraepithelial neoplasia 2. LSIL: low-grade squamous intraepithelial lesion 3. HSIL: high-grade squamous intraepithelial lesion 4. VIA: visual inspection with acetic acid 5. AIN: anal intraepithelial neoplasia 6. CNN: convolutional neural network 7. MTN: multitask deep learning network 8. AUC: area under the receiver operating characteristic curve 9. SROC: summary receiver operating characteristic.
All I^2 were more than 50% and all *P*-values of I^2 were less than 0.01 (Details in Table S3).
Ref denotes the reference group during meta-regression. [a]: The (< CIN2 vs. CIN2 +) or (LSIL- vs. HSIL +) group was used to compare with the other threshold groups.

techniques, are substantial sources of heterogeneity that meta-regression, constrained by limited reporting details, could not adequately capture.

Therefore, the high heterogeneity observed in our study is not an analytical flaw but a true reflection of the reproducibility and generalization challenges currently faced in the field of medical AI. Future research should prioritize adopting stricter reporting standards (e.g., adhering to the Checklist for Artificial Intelligence in Medical Imaging[70] or the Standards for Reporting Diagnostic Accuracy-AI guidelines[71]), advocating for code and model sharing, and ultimately advancing meta-analyses based on individual participant data. These will enable the direct examination of the underlying technical and clinical covariates. However, achieving robustness and generalizability is only the first technical hurdle in the long pathway toward clinical adoption. Beyond these methodological challenges, the translation of AI research into routine clinical practice faces even broader, multifaceted barriers that encompass issues of data equity, algorithmic fairness, and legal accountability.[72] Despite the promising advances in AI applications, the translation of research findings into clinical practice continues to encounter multifaceted challenges. These include data quality assurance, clinical validation, algorithmic bias, and resulting health disparities, as well as unresolved legal and regulatory complexities. A fundamental limitation in developing clinically robust AI systems is the scarcity of high-quality, representative datasets.[73] Many existing studies rely on hospital-based cohorts of high-risk or prescreened participants, which may increase diagnostic sensitivity but compromise generalizability to broader patient populations. Furthermore, while HRA is considered the gold standard for anal cancer screening, it is not routinely taught in medical training.[74,75] However, studies have demonstrated high acceptance of HRA among Nigerian men who have sex with men undergoing voluntary screening,[76] highlighting future opportunities for wider implementation.

Algorithmic transparency is another critical concern.[77] Machine learning systems can inadvertently amplify societal biases embedded in historical medical data, particularly undermining diagnostic accuracy in minority populations.[78] The absence of a comprehensive legal framework governing medical AI further hinders clinical adoption. This challenge is compounded by persistent ambiguity around responsibility when AI-related errors occur.[79] Establishing robust accountability protocols involving algorithm developers, clinicians, and health-care institutions remains a pressing need. However, allocating liability is complex, as responsibility may rest with developers, whose systems have “black box” features; the institutions that deploy them; or the clinicians making the final decisions.[80] Current malpractice frameworks are ill-equipped to manage such multi-party liability. Moreover, regulators face the ongoing difficulty of evaluating AI systems that evolve dynamically with new data inputs.[81] These uncertainties discourage health-care institutions from investing in AI adoption.

Finally, algorithmic bias requires a more comprehensive approach. Biases can permeate the entire development pipeline, from systematic exclusion of low-risk groups in training datasets to optimization

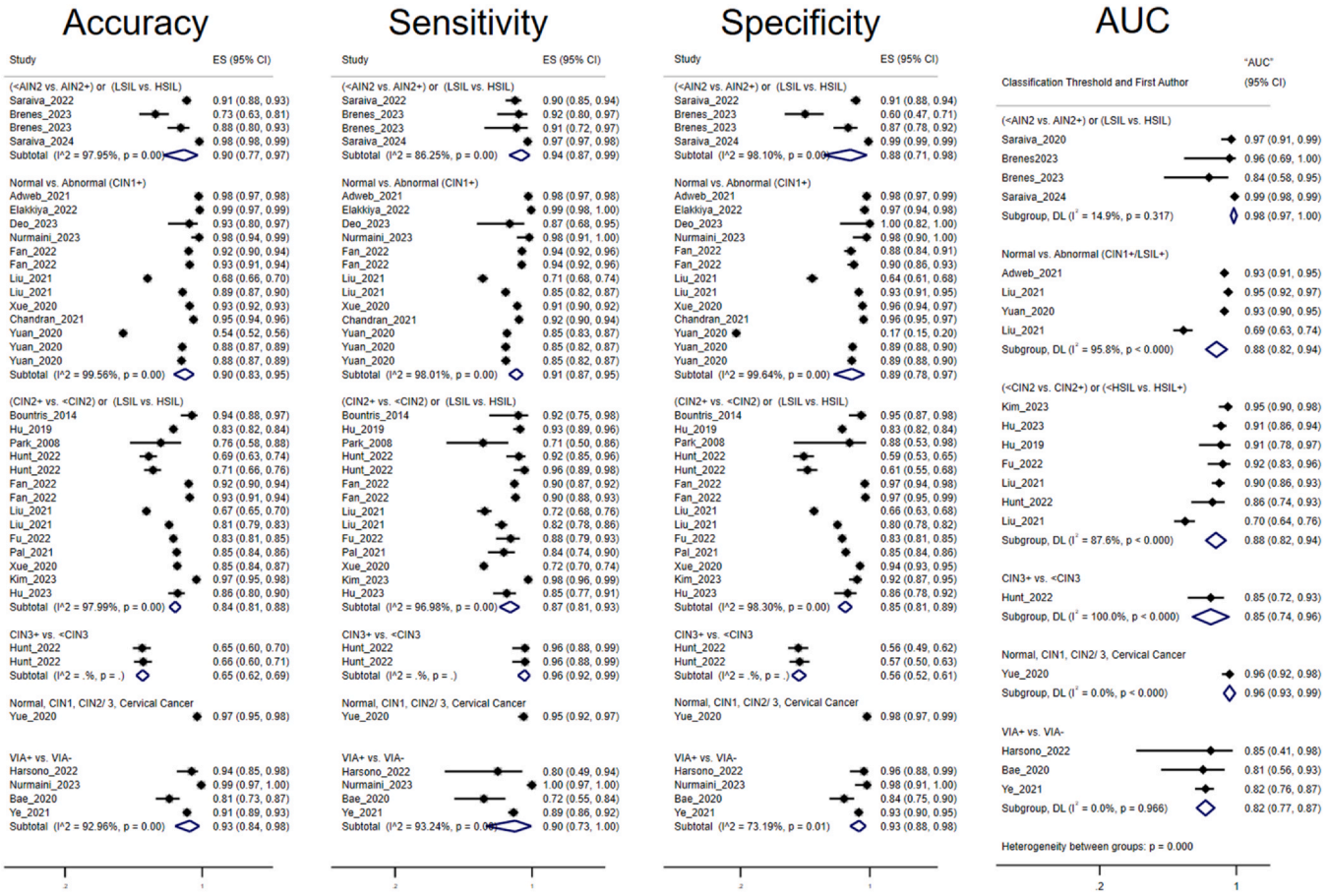


Figure 3. Forest plot of accuracy, sensitivity, specificity and AUC for different diagnostic thresholds. Abbreviation: AUC: area under the receiver operating characteristic curve.

objectives that prioritize overall accuracy at the expense of minority performance to clinicians' uncritical reliance on AI outputs in practice.[82] Such mechanisms risk not only perpetuating but also amplifying existing health-care disparities. Therefore, future research must move beyond aggregate performance reporting. It should mandate subgroup analyses, implement routine performance auditing across diverse demographic groups, and prioritize the design of equitable, transparent, and accessible AI tools from the outset.

Several limitations of the current AI-based approaches warrant attention. First, model training is highly sensitive to data quality, which could not be independently verified in this review. The presence of partial verification bias, especially in studies where negative test results were not confirmed by histology, likely led to an overestimation of specificity. Second, the lack of external validation across most included studies introduced a high risk of overfitting, particularly in small or homogeneous datasets, limiting our ability to draw definitive conclusions about the generalizability of AI performance to truly independent populations. This highlights a critical gap in the current literature and represents a priority direction for future research. Third, the interpretability of AI models remains limited. Many deep learning models operate as "black boxes," lacking transparency in how predictions are generated. This can reduce trust in AI outputs among clinicians and patients, especially in contexts demanding diagnostic accountability. Incorporating systematic uncertainty quantification and explainability measures is essential to support clinical adoption. Fourth, the limited number of studies on anal cancer compared with those on cervical cancer, combined with the frequent reliance on aggregated statistics instead of raw confusion matrices, hindered detailed analyses. The high heterogeneity observed across the studies further complicated the synthesis. Although we included all available covariates in the meta-regression analysis, we found no significant source of heterogeneity, which could be due to unreported or unmeasurable factors. Only two studies implemented external validation protocols, and heavy reliance on internal test sets may artificially boost performance estimates.

5. Conclusions

This systematic review and meta-analysis examined the diagnostic performances of AI-assisted tools in the detection of cervical and anal cancers and their precancerous lesions. Although relatively few studies have specifically addressed anal cancer, AI-assisted tools have demonstrated high accuracy across both cancer types. Future studies should investigate whether AI-assisted screening translates into improved clinical outcomes, such as reduced missed diagnoses, fewer unnecessary interventions, and lower recurrence rates, and assess its health economic impact, especially in low- and middle-income countries.

CRediT authorship contribution statement

Hailin Yang: Conceptualization, Methodology, Validation, Writing – original draft, Writing – review and editing. **Zhuoru Zou:** Conceptualization, Methodology, Validation, Writing – review and editing. **Zhenghong Li:** Conceptualization. **Nyi Nyi Soe:** Writing – review and editing. **Jason Ong:** Writing – review and editing. **Phyu Mon Latt:** Writing – review and editing. **Ying Zhang:** Writing – review and editing. **Reshi Suthakaran:** Writing – review and editing. **Lei Zhang:** Conceptualization, Methodology, Supervision, Writing – review and editing.

Declaration of Competing Interest

Lei Zhang is an editorial board member for Intelligent Oncology and was not involved in the editorial review or the decision to publish this article. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.intonc.2025.10.002.

References

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2024;74(3):229–263.
2. Brianti P, De Flammineis E, Mercuri SR. Review of HPV-related diseases and cancers. *New Microbiol.* 2017;40(2):80–85.
3. Walboomers JM, Jacobs MV, Manos MM, et al. Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *J Pathol.* 1999;189(1):12–19.
4. Ferrall L, Lin KY, Roden RBS, Hung CF, Wu TC. Cervical cancer immunotherapy: facts and hopes. *Clin Cancer Res.* 2021;27(18):4953–4973.
5. Serrano B, Brotons M, Bosch FX, Bruni L. Epidemiology and burden of HPV-related disease. *Best Pract Res Clin Obstet Gynaecol.* 2018;47:14–26.
6. Islami F, Ferlay J, Lortet-Tieulent J, Bray F, Jemal A. International trends in anal cancer incidence rates. *Int J Epidemiol.* 2017;46(3):924–938.
7. Deshmukh AA, Suk R, Shiels MS, et al. Recent trends in squamous cell carcinoma of the anus incidence and mortality in the United States, 2001-2015. *J Natl Cancer Inst.* 2020;112(8):829–838.
8. Kang Y-J, Smith M, Canfell K. Anal cancer in high-income countries: Increasing burden of disease. *PLoS One.* 2018;13(10):e0205105.
9. Clifford GM, Georges D, Shiels MS, et al. A meta-analysis of anal cancer incidence by risk group: toward a unified anal cancer risk scale. *Int J Cancer.* 2021;148(1):38–47.
10. Palefsky JM, Lee JY, Jay N, et al. Treatment of anal high-grade squamous intraepithelial lesions to prevent anal cancer. *N Engl J Med.* 2022;386(24):2273–2282.
11. Koliopoulos G, Nyaga VN, Santesso N, et al. Cytology versus HPV testing for cervical cancer screening in the general population. *Cochrane Database Syst Rev.* 2017;8(8):CD008587.
12. Preventive Services Task Force US, Curry SJ, Krist AH, et al. Screening for cervical cancer: US preventive services task force recommendation statement. *JAMA.* 2018;320(7):674–686.
13. Adsul P, Manjunath N, Srinivas V, Arun A, Madhivanan P. Implementing community-based cervical cancer screening programs using visual inspection with acetic acid in India: A systematic review. *Cancer Epidemiol.* 2017;49:161–174.
14. Clarke MA, Wentzensen N. Strategies for screening and early detection of anal cancers: A narrative and systematic review and meta-analysis of cytology, HPV testing, and other biomarkers. *Cancer Cytopathol.* 2018;126(7):447–460.
15. Palefsky JM. Practising high-resolution anoscopy. *Sex Health.* 2012;9(6):580–586.
16. Lam JM, Hoch JS, Tinmouth J, Sano M, Raboud J, Salit IE. Cost-effectiveness of screening for anal precancers in HIV-positive men. *AIDS.* 2011;25(5):635–642.
17. Stier EA, Clarke MA, Deshmukh AA, et al. International anal neoplasia society's consensus guidelines for anal cancer screening. *Int J Cancer.* 2024;154(10):1694–1702.
18. Institute NYSD of HA. HIV Clinical Resource. anal dysplasia and cancer. New York, NY: New York State Department of Health; 2007;20007.
19. Alexander A, Jiang A, Ferreira C, Zurkiya D. An intelligent future for medical imaging: a market outlook on artificial intelligence for medical imaging. *J Am Coll Radiol.* 2020;17(1 Pt B):165–170.
20. Hou X, Shen G, Zhou L, Li Y, Wang T, Ma X. Artificial intelligence in cervical cancer screening and diagnosis. *Front Oncol.* 2022;12:851367.
21. Wang CW, Liou YA, Lin YJ, et al. Artificial intelligence-assisted fast screening cervical high grade squamous intraepithelial lesion and squamous cell carcinoma diagnosis and treatment planning. *Sci Rep.* 2021;11(1):16244.
22. Boon SS, Luk HY, Xiao C, Chen Z, Chan PKS. Review of the standard and advanced screening, staging systems and treatment modalities for cervical cancer. *Cancers (Basel).* 2022;14(12):2913.
23. Robinson N, Stoffel C, Haider S. Global women's health is more than maternal health: a review of gynecology care needs in low-resource settings. *Obstet Gynecol Surv.* 2015;70(3):211–222.
24. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521(7553):436–444.
25. Liu PR, Lu L, Zhang JY, Huo TT, Liu SX, Ye ZW. Application of artificial intelligence in medicine: an overview. *Curr Med Sci.* 2021;41(6):1105–1115.
26. Aswathy R, Sumathi S. The evolving landscape of cervical cancer: breakthroughs in screening and therapy through integrating biotechnology and artificial intelligence. *Mol Biotechnol.* 2025;67(3):925–941.
27. Saraiva MM, Spindler L, Fathallah N, et al. Artificial intelligence and high-resolution anoscopy: automatic identification of anal squamous cell carcinoma precursors using a convolutional neural network. *Tech Coloproctol.* 2022;26(11):893–900.
28. Saraiva MM, Spindler L, Fathallah N, et al. Deep learning in high-resolution anoscopy: assessing the impact of staining and therapeutic manipulation on automated detection of anal cancer precursors. *Clin Transl Gastroenterol.* 2024;15(4):e00681.
29. Du H, Dai W, Zhou Q, et al. AI-assisted system improves the work efficiency of cytologists via excluding cytology-negative slides and accelerating the slide interpretation. *Front Oncol.* 2023;13:1290112.
30. Shrestha P, Poudyal B, Yadollahi S, et al. A systematic review on the use of artificial intelligence in gynecologic imaging - Background, state of the art, and future directions. *Gynecol Oncol.* 2022;166(3):596–605.
31. Brenes D, Kortum A, Carns J, et al. Automated in vivo high-resolution imaging to detect human papillomavirus-associated anal precancer in persons living with HIV. *Clin Transl Gastroenterol.* 2023;14(2):e00558.
32. Brenes D, Kortum A, Coole J, et al. Deployment and assessment of a deep learning model for real-time detection of anal precancer with high frame rate high-resolution microendoscopy. *Sci Rep.* 2023;13(1):22267.
33. Liu L, Wang Y, Liu X, et al. Computer-aided diagnostic system based on deep learning for classifying colposcopy images. *Ann Transl Med.* 2021;9(13):1045.
34. Brandão M, Mendes F, Martins M, et al. Revolutionizing women's health: a comprehensive review of artificial intelligence advancements in gynecology. *J Clin Med.* 2024;13(4):1061.
35. McInnes MDF, Moher D, Thombs BD, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. *JAMA.* 2018;319(4):388–396.
36. Sounderajah V, Ashrafian H, Rose S, et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies QUADAS-AI. *Nat Med.* 2021;27(10):1663–1665.
37. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med.* 2011;155(8):529–536.
38. Yang B, Mallett S, Takwoingi Y, et al. QUADAS-C: a tool for assessing risk of bias in comparative diagnostic accuracy studies. *Ann Intern Med.* 2021;174(11):1592–1599.
39. Chen Y, Chen D, Wang Y, Han Y. Using Freeman-Tukey double arcsine transformation in meta-analysis of single proportions. *Aesthetic Plast Surg.* 2023;47(Suppl 1):83–84.
40. Freeman MF, Tukey JW. Tukey Transformations related to the angular and the square root. *The Annals of Mathematical Statistics.* Published online, 1950;607–611.
41. Schwarzer G, Chemaitelly H, Abu-Raddad LJ, Rücker G. Seriously misleading results using inverse of Freeman-Tukey double arcsine transformation in meta-analysis of single proportions. *Res Synth Methods.* 2019;10(3):476–483.
42. Nyaga VN, Arbyn M, Aerts M. Metaprop: a Stata command to perform meta-analysis of binomial data. *Arch Public Health.* 2014;72(1):39.
43. Adweb KMA, Cavus N, Sekeroglu B. Cervical cancer diagnosis using very deep networks over different activation functions. *IEEE Access.* 2021;9:46612–46625.
44. Elakkiya R, Subramaniyaswamy V, Vijayakumar V, Mahanti A. Cervical cancer diagnostics healthcare system using hybrid object detection adversarial networks. *IEEE J Biomed Health Inform.* 2022;26(4):1464–1471.
45. Nurmaini S, Rachmatullah MN, Agustiansyah P, et al. CervicoXNet: an automated cervicogram interpretation network. *Med Biol Eng Comput.* 2023;61(9):2405–2416.
46. Deo BS, Sah AN, Shukla S, et al. Cervical pre-cancer classification using entropic features and CNN: in vivo validation with a handheld fluorescence probe. *J Biophotonics.* 2024;17(3):e202300363.
47. Fan Y, Ma H, Fu Y, Liang X, Yu H, Liu Y. Colposcopic multimodal fusion for the classification of cervical lesions. *Phys Med Biol.* 2022;67(13): 10.1088/1361-6560/ac73d4.
48. Chandran V, Sumithra MG, Karthick A, et al. Diagnosis of cervical cancer based on ensemble deep learning network using colposcopy images. *Biomed Res Int.* 2021;2021:5584004.
49. Yuan C, Yao Y, Cheng B, et al. The application of deep learning based diagnostic system to cervical squamous intraepithelial lesions recognition in colposcopy images. *Sci Rep.* 2020;10(1):11639.
50. Xue P, Tang C, Li Q, et al. Development and validation of an artificial intelligence system for grading colposcopic impressions and guiding biopsies. *BMC Med.* 2020;18(1):406.
51. Bountris P, Haritou M, Pouliakis A, et al. An intelligent clinical decision support system for patient-specific predictions to improve cervical intraepithelial neoplasia detection. *Biomed Res Int.* 2014;2014:341483.
52. Kim S, An H, Cho HW, et al. Pivotal clinical study to evaluate the efficacy and safety of assistive artificial intelligence-based software for cervical cancer diagnosis. *J Clin Med.* 2023;12(12):4024.
53. Hu L, Mwanahamuntu MH, Sahasrabuddhe VV, et al. Internal validation of automated visual evaluation (AVE) on smartphone images for cervical cancer screening in a prospective study in Zambia. Preprint. *medRxiv*. 2024;2023.07.19.23292888. Published 2024 Feb 12.
54. Pal A, Xue Z, Befano B, et al. Deep metric learning for cervical image classification. *IEEE Access.* 2021;9:53266–53275.
55. Hu L, Bell D, Antani S, et al. An observational study of deep learning and automated evaluation of cervical images for cancer screening. *J Natl Cancer Inst.* 2019;111(9):923–932.
56. Fu L, Xia W, Shi W, et al. Deep learning based cervical screening by the cross-modal integration of colposcopy, cytology, and HPV test. *Int J Med Inform.* 2022;159:104675.
57. Park SY, Follen M, Milbourne A, et al. Automated image analysis of digital colposcopy for the detection of cervical neoplasia. *J Biomed Opt.* 2008;13(1):014029.
58. Hunt B, Fregnani JHTG, Brenes D, et al. Cervical lesion assessment using real-time microendoscopy image analysis in Brazil: the CLARA study. *Int J Cancer.* 2021;149(2):431–441.
59. Yue Z, Ding S, Zhao W, et al. Automatic CIN grades prediction of sequential cervigram image using LSTM with multistate CNN features. *IEEE J Biomed Health Inform.* 2020;24(3):844–854.

60. Harsono AB, Susiarno H, Suardi D, et al. Cervical pre-cancerous lesion detection: development of smartphone-based VIA application using artificial intelligence. *BMC Res Notes.* 2022;15(1):356.
61. Bae JK, Roh HJ, You JS, et al. Quantitative screening of cervical cancers for low-resource settings: pilot study of smartphone-based endoscopic visual inspection after acetic acid using machine learning techniques. *JMIR Mhealth Uhealth.* 2020;8(3):e16467.
62. Park YR, Seol YJ, Kim YJ, Kim KG. Comparison on the deep learning performance of a field of view variable color images of uterine cervix. In: *International Forum on Medical Imaging in Asia 2021*. Vol 11792. SPIE; 2021:99-102.
63. Taghavi K, Rohner E, Basu P, Low N, Rutjes A, Bohlius J. Screening test accuracy of portable devices that can be used to perform colposcopy for detecting CIN2+ in low- and middle-income countries: a systematic review and meta-analysis. *BMC Womens Health.* 2020;20(1):253.
64. Bhinder B, Gilvary C, Madhukar NS, Elemento O. Artificial intelligence in cancer research and precision medicine. *Cancer Discov.* 2021;11(4):900–915.
65. Chen ZH, Lin L, Wu CF, Li CF, Xu RH, Sun Y. Artificial intelligence for assisting cancer diagnosis and treatment in the era of precision medicine. *Cancer Commun (Lond).* 2021;41(11):1100–1115.
66. Jeronimo J, Long LR, Neve L, Michael B, Antani S, Schiffman M. Digital tools for collecting data from cervigrams for research and training in colposcopy. *J Low Genit Tract Dis.* 2006;10(1):16–25.
67. Varoquaux G, Cheplygina V. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ Digit Med.* 2022;5(1):48.
68. Fraser CG, Harris EK. Generation and application of data on biological variation in clinical chemistry. *Crit Rev Clin Lab Sci.* 1989;27(5):409–437.
69. Jungo A, Meier R, Ermis E, et al. On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2018:682–690.
70. Tejani AS, Klontzas ME, Gatti AA, et al. Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. *Radiol Artif Intell.* 2024;6(4):e240300.
71. Sounderajah V, Guni A, Liu X, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nat Med.* 2025;31(10):3283–3289.
72. Zhao C, López BS. Translating artificial intelligence algorithms into clinical practice: addressing implementation barriers in healthcare. *Front Emerg Comput Sci Inf Technol.* 2025;2(07):1–6.
73. Reddy S, Shaikh S. The long road ahead: navigating obstacles and building bridges for clinical integration of artificial intelligence technologies. *J Med Artif Intell.* Published online. 2024.
74. Hillman RJ, Cuming T, Darragh T, et al. 2016 IANS international guidelines for practice standards in the detection of anal cancer precursors. *J Low Genit Tract Dis.* 2016;20(4):283–291.
75. Ghebre R, Berry-Lawhorn JM, D'Souza G. State of the science: screening, surveillance, and epidemiology of hpv-related malignancies. *Am Soc Clin Oncol Educ Book.* 2021;41:1–12.
76. Nowak RG, Nnaji CH, Dauda W, et al. Satisfaction with high-resolution anoscopy for anal cancer screening among men who have sex with men: a cross-sectional survey in Abuja, Nigeria. *BMC Cancer.* 2020;20(1):98.
77. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* 2019;17(1):195.
78. Crawford K, Calo R. There is a blind spot in AI research. *Nature.* 2016;538(7625):311–313.
79. Price 2nd WN, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. *JAMA.* 2019;322(18):1765–1766.
80. Chen A, Wang C, Zhang X. Reflection on the equitable attribution of responsibility for artificial intelligence-assisted diagnosis and treatment decisions. *Intell Med.* 2023;3(02):139–143.
81. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med.* 2023;6(1):120.
82. Yang SR, Chien JT, Lee CY. Advancements in clinical evaluation and regulatory frameworks for AI-driven software as a medical device (SaMD). *IEEE Open J Eng Med Biol.* Published online. 2024.