

A Systematic Review and Descriptive Quantitative Synthesis of Artificial Intelligence in Cervical Cancer Screening and Diagnosis

H Aswini

Vellore Institute of Technology University

T Benil

benil.t@vit.ac.in

Vellore Institute of Technology University

Systematic Review

Keywords: Artificial Intelligence, Cervical cancer, Cervical screening, Deep learning, Diagnostic accuracy, Systematic review.

Posted Date: October 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10982718/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Artificial intelligence (AI) has been frequently utilised in cervical cancer screening, diagnosis, lesion assessment and clinical decision support, but the evidence remains methodologically heterogeneous. This systematic review examined the application of AI in cytology, colposcopy, visual examination, histopathology, and multimodal clinical data. Also reviewed PubMed, Scopus, Web of Science and IEEE Xplore to identify 4094 records. After the exclusion of 2053 duplicates and applying pre-defined eligibility criteria, 898 studies were included. Directly extractable diagnostic-performance measures were available from 60 studies. Accuracy was reported in 40 studies, sensitivity in 19, specificity in 16 and AUC in 12 studies. A total of 13 studies reported both sensitivity and specificity. There was considerable heterogeneity in populations, datasets, reference standards, diagnostic tasks, AI methods, validation procedures and performance reporting. Complete TP, FP, TN and FN data were not available across the eligible evidence base and pooled diagnostic accuracy estimation and summary receiver-operating-characteristic analysis were not possible. Therefore, the quantitative results were summarized descriptively. Evidence indicates broad application of AI approaches but also limitations in terms of comparability, external validation, reporting completeness and clinical generalizability. Future research should focus on standardized reporting, multicenter datasets, independent external validation and prospective clinical evaluation.

Introduction

Cervical cancer remains a major public health Challenge in worldwide, although there are advances in progress in prevention, screening, early diagnosis and treatment. The primary cause is persistent infection with high-risk human papillomavirus (HPV). Early detection and treatment of precancerous lesions can greatly reduce the burden of disease. Methods of screening and diagnosis currently used are HPV testing, cervical cytology, visual inspection, colposcopy and histopathology¹⁻⁶. Cytology remains an important screening tool and colposcopy and histopathology are used to further evaluate and diagnose suspicious lesions^{2,5,7-10}. However, traditional evaluation may be influenced by the need for skilled personnel, variability in interpretation, image or specimen quality, and differences in diagnostic resources and clinical practice^{2-7,9}.

Artificial intelligence (AI), especially machine learning and deep learning, has been more and more researched for cervical cell classification, lesion detection and segmentation, screening, diagnostic classification and clinical prediction^{2,7,11-15}. Methodological development has progressed from traditional machine learning and image analysis to convolutional neural networks, ensemble methods, transformers, attention-based and explainable models, and multimodal systems^{11,16-19}. Applications now include cytology and whole-slide imaging^{8,12,20,21}, colposcopy and digital visual assessment^{2,9,13,14,22,23}, histopathology^{5,6,10,24} and multimodal HPV-related triage^{18,25-27}.

Despite this expansion, the evidence remains clinically and methodologically heterogeneous. Studies differ in the populations, geographic settings, datasets, imaging modalities, diagnostic tasks, reference

standards, model architectures, validation strategies, and reported performance^{23,28–30}. Thus, accuracy, sensitivity, specificity and AUC are not directly comparable without consideration of the diagnostic question, target population, reference standard, threshold and evaluation design. Synthesis of diagnostic test accuracy also requires sufficiently complete and comparable study-level information, thus incomplete contingency data can preclude reliable quantitative pooling^{31,32}. AI applications related to screening and diagnosis of cervical cancer have been evaluated in previous systematic and scoping reviews, including diagnostic accuracy, cytology based on deep learning, and other applications of AI^{33–37}. With ongoing development in cytology, colposcopy, visual assessment, HPV-based triage, histopathology, multimodal systems and clinical prediction, an updated assessment of the evidence base is required.

Therefore, this systematic review was conducted to evaluate the use of AI in screening and diagnosis of cervical cancer based on the evidence obtained from PubMed, Scopus, Web of Science, and IEEE Xplore. Objectives included 1) characterisation of clinical, imaging, data set, diagnostic, and AI-related features, 2) summary of reported diagnostic-performance measures, 3) assessment of feasibility of formal quantitative synthesis, 4) evaluation of methodological quality and risk of bias where appropriate, and 5) identification of priorities for comparability, generalisability, reproducibility, and clinical implementation.

Methods

Search strategy and study selection

This systematic review was performed according to the PRISMA 2020 statement³⁸ and the PRISMA-DTA guidelines^{31,32}, where applicable. Methods A search was performed in databases such as PubMed, Scopus, Web of Science and IEEE Xplore from their inception to 13 August 2026 for studies on artificial intelligence (AI) in cervical cancer screening and diagnosis. The keywords for the search are cervical cancer, HPV, cervical cytology, pap smear, liquid-based cytology, colposcopy, visual inspection, histopathology, artificial intelligence, machine learning, deep learning, computer vision, classification, segmentation, and computer-aided diagnosis.

Eligible studies were original research investigating AI or computational methods for cervical cancer or precancerous-lesion screening, detection, classification, segmentation, diagnosis, triage, or risk assessment. The exclusion criteria consisted reviews, meta-analyses, editorials, letters, commentaries, book chapters, dissertations, inadequate conference abstracts, duplicated publications, and studies outside the defined parameters. Records were deduplicated and screened by title and abstract and full-text review was conducted against pre-defined eligibility criteria. The study selection process and reasons for exclusions are summarised in Fig. 1.

Data extraction and evidence classification

Data were extracted using a standardized framework that included publication characteristics, study population, dataset, imaging modality, clinical task, AI methodology, preprocessing, augmentation, training, and validation strategy, reference standard, comparator, diagnostic-performance measures, explainability, uncertainty, multimodal inputs, class imbalance, limitations, and suitability for quantitative synthesis. Studies were classified by clinical function, diagnostic modality, AI approach, dataset characteristics, validation strategy. Imaging studies were classified as cytology, liquid based cytology, colposcopy, visual assessment, or histopathology and HPV-related and other clinical information were recorded separately. Approaches taken by AI were conventional machine learning, deep learning, CNNs, transformers, ensemble methods and multimodal frameworks. Validation strategies were classified as internal hold-out, cross-validation, independent test-set or external/multicenter validation.

Quantitative synthesis and risk of bias

Studies reporting at least one directly extractable accuracy, sensitivity, specificity, or AUC measure were considered eligible for quantitative description. Conventional diagnostic test-accuracy pooling was considered only when sufficient information was available to establish comparable diagnostic endpoints and, where required, TP, FP, TN, and FN contingency data^{31,32}. Missing diagnostic data was neither reconstructed nor imputed and performance measures were retained as reported for descriptive synthesis. The risk of bias of patient selection, index test, reference standard and flow and timing was assessed in those studies with sufficient methodological details to allow a reliable assessment using QUADAS-2. Evidence synthesis Qualitative analysis of all included studies and systematic quantitative description of the relevant performance data.

Results

Study selection and characteristics

The systematic review detected 4,094 articles originating from PubMed (970), Scopus (1,586), Web of Science (1,410) and IEEE Xplore (128). After removing 2,053 duplicates, 2,041 records were screened, and 1,292 articles were selected for full-text review, of which 898 studies met the pre-determined eligibility criteria. The studies investigated cervical cytology, Papanicolaou smear, liquid-based cytology, colposcopy, visual inspection, histopathology, HPV-related assessment and multimodal clinical information. Figure 2 shows a substantial increase in the number of publications from 2015 to 2026, and Table 1 presents the distributions of publication periods and study designs.

Table 1
Characteristics of the 898 studies included in the systematic review

Characteristic	Category	Number (n)	Percentage (%)
Publication period	2015–2019	64	7.1
	2020–2021	123	13.7
	2022–2023	232	25.8
	2024–2025	337	37.5
	2026	142	15.8
Study design	Primary AI/ML diagnostic or screening study	485	54.0
	Primary diagnostic image study	41	4.6
	Diagnostic image study	39	4.3
	Diagnostic image-analysis study	21	2.3
	Clinical diagnostic validation	15	1.7
	Diagnostic imaging study	13	1.4
	Primary diagnostic imaging study	11	1.2
	Diagnostic cytology image study	11	1.2
	Diagnostic pathology image study	7	0.8
	Clinical prediction study	7	0.8
	Other study-design descriptions	248	27.6
Total	Included evidence base	898	100.0

Among the 413 studies with an explicitly populated diagnostic-task field, classification/grading was the most frequently reported task, followed by detection and screening (Fig. 3). Among the 411 studies with an explicitly extracted primary AI methodology, CNN-based/deep-vision approaches were the most frequently identified methodological family, followed by hybrid/ensemble/multimodal and other deep-learning approaches (Fig. 4).

Diagnostic-performance evidence and synthesis feasibility

Among the 898 included studies, 60 provided at least one directly extractable core diagnostic-performance measure. Accuracy was reported in 40 studies (66.7%), sensitivity in 19 (31.7%), specificity in 16 (26.7%), and AUC in 12 (20.0%); 13 studies (21.7%) reported both sensitivity and specificity (Table 2). These categories were not mutually exclusive. For visual clarity, Fig. 5 presents selected

performance values from 19 of these 60 studies; the remaining eligible studies were retained in the quantitative evidence base and Table 2 but were not displayed individually in the figure.

Table 2
Quantitative diagnostic-performance evidence among the eligible studies.

Quantitative-synthesis characteristic	Number of studies (n)	Percentage (%)
Studies with ≥ 1 directly reported core performance measure	60	100.0
Accuracy reported	40	66.7
Sensitivity reported	19	31.7
Specificity reported	16	26.7
AUC reported	12	20.0
Both sensitivity and specificity reported	13	21.7
Complete TP/FP/TN/FN contingency data available	0	0.0
Conventional pooled diagnostic test-accuracy meta-analysis feasible	0	0.0

Substantial heterogeneity was observed across endpoints, reference standards, populations, datasets, evaluation methods, thresholds, and reporting formats. Complete TP, FP, TN, and FN contingency data were unavailable for the 60 quantitatively eligible studies. Consequently, pooled sensitivity, specificity, AUC, summary receiver-operating-characteristic analysis, and diagnostic odds ratios were not calculated, and missing information was neither reconstructed nor imputed. Directly reported performance measures were therefore retained as a descriptive synthesis (Fig. 5).

Risk of bias Assessment

A QUADAS-2 assessment was professionally conducted for 15 studies with sufficient methodological information. Three investigations (20.0%) were considered low risk, 8 (53.3%) uncertain risk and 4 (26.7%) high risk. Figure 6: Domain specific evaluations for patient selection, primary test, benchmark standard and workflow and timing of procedure.

Discussion

Evidence landscape and methodological heterogeneity

This systematic review extracted 898 studies addressing AI applications in cervical cytology, colposcopy, visual assessment, histopathology, assessment related to HPV, multimodal systems and clinical prediction^{2,4,5,7,8,15,18}. The methodological landscape comprised CNN and other deep learning

approaches, including a transformer, hybrid, ensemble, multimodal, and conventional machine learning methods^{11,16–19,24,39,40}. This diversity is due to rapid methodological development and substantial heterogeneity in populations, datasets, diagnostic tasks, reference standards, model development, and validation, which limits the ability to directly compare reported performance^{23,28–32}. Only 60 of the 898 studies provided directly extractable core diagnostic-performance measures. Accuracy was the most frequently reported metric, followed by sensitivity, specificity, and AUC. These measures are not directly interchangeable because accuracy is influenced by class distribution and disease prevalence, while sensitivity and specificity depend on the operating threshold and AUC summarizes discrimination across thresholds without establishing clinical utility at a specific operating point^{31,32}.

Quantitative synthesis, methodological quality and clinical translation

The available performance evidence did not support conventional pooled diagnostic test-accuracy meta-analysis. None of the 60 quantitatively eligible studies provided complete TP, FP, TN, and FN information, while substantial heterogeneity existed in populations, datasets, reference standards, thresholds, and validation procedures^{31,32}. Accordingly, missing information was neither reconstructed nor imputed, and pooled sensitivity, specificity, AUC, summary ROC analyses, and diagnostic odds ratios were not calculated. This conservative approach avoided generating summary estimates from fundamentally non-comparable evidence.

Formal QUADAS-2 assessment was possible for 15 studies, of which 3 were classified as low risk, 8 as unclear risk, and 4 as high risk. The prevalence of unclear or high-risk judgements further emphasizes the need for transparent reporting of patient selection, index tests, reference standards, study flow, dataset construction, preprocessing, model development, validation, and subgroup performance^{29,31,32,38}. Independent external, multicentre, and prospective validation is particularly important because performance obtained from curated datasets or internal train-test splits may not generalize across institutions, populations, imaging conditions, and clinical workflows^{23,28–30}. Explainability should also complement predictive performance when AI is intended to support clinical decisions^{19,29}.

Strengths, limitations and future directions

A major strength of this review is its broad coverage of PubMed, Scopus, Web of Science, and IEEE Xplore and its separation of overall systematic-review eligibility from quantitative-synthesis eligibility. This enabled characterization of the wider evidence base while avoiding inappropriate pooling of heterogeneous diagnostic measures. However, substantial variation in populations, datasets, diagnostic tasks, reference standards, performance reporting, and validation limited comparability and clinical generalizability; incomplete diagnostic reporting also prevented reconstruction of contingency tables and formal pooled estimates^{28–30}.

Future research should prioritize multicentre datasets with patient-level inclusion criteria, standardized reference standards, transparent reporting, independent external validation, and prospective evaluation^{23,28–30}. Complete reporting of diagnostic-performance measures and TP/FP/TN/FN data would improve the feasibility of future diagnostic meta-analyses^{31,32}, while multimodal integration of imaging, HPV, and clinical information and development of interpretable AI may support more robust clinical decision-making^{18,19,25–27}.

Conclusion

This systematic review identified 898 studies investigating AI applications in cervical cancer screening and diagnosis, including cytology, colposcopy, visual assessment, histopathology, HPV-related evaluation, multimodal systems, and clinical prediction. Despite the availability of 60 studies with directly extractable diagnostic-performance indicators, considerable heterogeneity and lack of contingency data precluded robust traditional quantitative synthesis. The evidence demonstrates broad application of AI, but persistent limitations in comparability, reporting completeness, external validation and clinical generalizability. Future studies should aim at standardized reporting, comprehensive diagnostic performance indicators, multi-centre data, independent external validation, prospective evaluation, and clinically meaningful AI models to facilitate the reliable integration into cervical cancer screening and diagnostic techniques.

Declarations

Competing Interests

The authors declare no competing interests.

Ethics Statement

This study is a meta-analysis, and it used data obtained from the published literature without the recruitment of human subjects or the collection of new human samples, so it not required ethical approval and informed consent.

Additional Information

Correspondence and requests for materials should be addressed to Benil T.

Funding

This study received no specific funding from public, commercial or not-for-profit organisations.

Author Contribution

H. Aswini conceptualised and designed the research, developed the methodology, performed the literature review, chose the studies and handled data, carried out the formal analysis and visualisation, and wrote the manuscript. Benil T. conceptualised the study, supervised the methodology, validated the results, and critically reviewed and edited the manuscript. The final version of the manuscript was reviewed and approved by both authors.

Acknowledgement

The authors acknowledge the Vellore Institute of Technology (VIT), Chennai, India, for providing the research facilities and academic support necessary for the conduct of this study. The authors would like to thank the researchers whose published studies formed the basis for this systematic review and quantitative analysis.

Data Availability

The datasets generated and analysed in this study are available from the corresponding author on reasonable request. The study-level bibliographic dataset, screening dataset, evidence-mapping dataset and quantitative diagnostic-performance dataset are available to support the findings of this study. The datasets supporting the quantitative analyses and evidence-mapping results can be made available in machine-readable formats on reasonable request.

References

1. Organization, W. H. *Global Strategy to Accelerate the Elimination of Cervical Cancer as a Public Health Problem*. (2020).
2. Kim, J., Park, C. M., Kim, S. Y. & Cho, A. Convolutional neural network-based classification of cervical intraepithelial neoplasias using colposcopic image segmentation for acetowhite epithelium. *Sci. Rep.* **12**, 17228 (2022).
3. Harsono, A. B. Cervical pre-cancerous lesion detection: development of smartphone-based VIA application using artificial intelligence. *BMC Res. Notes* **15**, 356 (2022).
4. Parham, G. P. Validation in Zambia of a cervical screening strategy including HPV genotyping and artificial intelligence (AI)-based automated visual evaluation. *Infect. Agent. Cancer* **18**, 61 (2023).
5. Li, C. Cervical histopathology image classification using multilayer hidden conditional random fields and weakly supervised learning. *IEEE Access* **7**, 90378–90397 (2019).
6. Pal, A. Deep multiple-instance learning for abnormal cell detection in cervical histopathology images. *Comput. Biol. Med.* **138**, 104890 (2021).

7. Bao, H. Artificial intelligence-assisted cytology for detection of cervical intraepithelial neoplasia or invasive cancer: A multicenter, clinical-based, observational study. *Gynecol. Oncol.* **159**, 171–178 (2020).
8. Wang, J. Artificial intelligence enables precision diagnosis of cervical cytology grades and cervical cancer. *Nat. Commun.* **15**, 4369 (2024).
9. Tamang, P., Gupta, M. & Thatal, A. Digital colposcopy image analysis techniques requirements and their role in clinical diagnosis: a systematic review. *Expert Rev. Med. Devices* **21**, 955–969 (2024).
10. Liu, Q. A histopathologic image analysis for the classification of endocervical adenocarcinoma Silva patterns depend on weakly supervised deep learning. *Am. J. Pathol.* **194**, 735–746 (2024).
11. Tan, X. Automatic model for cervical cancer screening based on convolutional neural network: a retrospective, multicohort, multicenter study. *Cancer Cell Int.* **21**, 35 (2021).
12. Kanavati, F. A deep learning model for cervical cancer screening on liquid-based cytology specimens in whole slide images. *Cancers (Basel)*. **14**, 1159 (2022).
13. Cho, B.-J. Classification of cervical neoplasms on colposcopic photography using deep learning. *Sci. Rep.* **10**, 13652 (2020).
14. Ma, J.-H. Computer-aided diagnosis of cervical dysplasia using colposcopic images. *Front. Oncol.* **12**, 905623 (2022).
15. Hariprasad, R., Navamani, T. M., Rote, T. R. & Chauhan, I. Design and development of an efficient risk prediction model for cervical cancer. *IEEE Access* **11**, 74290–74300 (2023).
16. Khan, A., Han, S., Ilyas, N., Lee, Y. M. & Lee, B. CervixFormer: A multi-scale swin transformer-based cervical pap-smear WSI classification framework. *Comput. Methods Programs Biomed.* **240**, 107718 (2023).
17. Alsalatie, M., Alquran, H., Mustafa, W. A., Yacob, Y. M. & Alayed, A. A. Analysis of cytology Pap smear images based on ensemble deep learning approach. *Diagnostics* **12**, 2756 (2022).
18. Fu, L. Deep learning based cervical screening by the cross-modal integration of colposcopy, cytology, and HPV test. *Int. J. Med. Inform.* **159**, 104675 (2022).
19. Khare, S., Booth, B. B., Blanes-Vidal, V., Nadimi, E. & Petersen, L. K. An explainable attention model for cervical precancer risk classification using colposcopic images. *Comput. Methods Programs Biomed.* **271**, 108976 (2025).
20. Cheng, S. Robust whole slide image analysis for cervical cancer screening using deep learning. *Nat. Commun.* **12**, 5639 (2021).
21. Tang, H.-P. Cervical cytology screening facilitated by an artificial intelligence microscope: A preliminary study. *Cancer Cytopathol.* **129**, 693–700 (2021).
22. Skerrett, E. Multicontrast Pocket Colposcopy Cervical Cancer Diagnostic Algorithm for Referral Populations. *BME Front.* **2022**, 9823184 (2022).
23. Wu, T., Wang, Y., Cui, X., Xue, P. & Qiao, Y. AI-based identification method for cervical transformation zone within digital colposcopy: Development and multicenter validation study. *JMIR Cancer* **11**,

- e69672 (2025).
24. Xue, D. An application of transfer learning and ensemble learning techniques for cervical histopathology image classification. *IEEE Access* **8**, 104603–104618 (2020).
 25. Xue, P. Assessing artificial intelligence enabled liquid-based cytology for triaging HPV-positive women: a population-based cross-sectional study. *Acta Obstet. Gynecol. Scand.* **102**, 1026–1033 (2023).
 26. Wittenborn, J. CIN2 + detection in high-risk HPV patients with no or minor cervical cytologic abnormalities: a clinical approach validated by machine learning. *Arch. Gynecol. Obstet.* **307**, 881–890 (2023).
 27. Stegmüller, T. Self-supervised learning-based cervical cytology for the triage of HPV-positive women in resource-limited settings and low-data regime. *Comput. Biol. Med.* **169**, 107809 (2024).
 28. Dhar, S. A prospective study to assess the utility of artificial intelligence-driven cervical screening using Smart Scope versus conventional method. *J. Med. Heal. Sci.* **17**, 160–166 (2026).
 29. Ahmed, S. R. Reproducible and clinically translatable deep neural networks for cervical screening. *Sci. Rep.* **13**, 21772 (2023).
 30. Elvatun, S. Cross-population evaluation of cervical cancer risk prediction algorithms. *Int. J. Med. Inform.* **181**, 105297 (2024).
 31. McInnes, M. D. F. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. *JAMA* **319**, 388–396 (2018).
 32. Salameh, J. P. Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. *BMJ* **370**, m2632 (2020).
 33. Liu, L. Performance of artificial intelligence for diagnosing cervical intraepithelial neoplasia and cervical cancer: A systematic review and meta-analysis. *eClinicalMedicine* **80**, 102992 (2025).
 34. Allahqoli, L. Diagnosis of cervical cancer and pre-cancerous lesions by artificial intelligence: A systematic review. *Diagnostics* **12**, 2771 (2022).
 35. Jiang, P. A systematic review of deep learning-based cervical cytology screening: From cell identification to whole slide image analysis. *Artif. Intell. Rev.* **56**, 2687–2758 (2023).
 36. Vargas-Cardona, H. D. Artificial intelligence for cervical cancer screening: Scoping review, 2009–2022. *Int. J. Gynaecol. Obstet.* **165**, 566–578 (2024).
 37. Akazawa, M. & Hashimoto, K. Artificial intelligence in gynecologic cancers: Current status and future challenges – A systematic review. *Artif. Intell. Med.* **120**, 102164 (2021).
 38. Page, M. J. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **372**, n71 (2021).
 39. Cai, D. HiCervix: An extensive hierarchical dataset and benchmark for cervical cytology classification. *IEEE Trans. Med. Imaging* **43**, 4344–4355 (2024).
 40. Wu, N., Jia, D., Zhang, C. & Li, Z. Cervical cell extraction network based on optimized YOLO. *Math. Biosci. Eng.* **20**, 2364–2381 (2023).

Figures

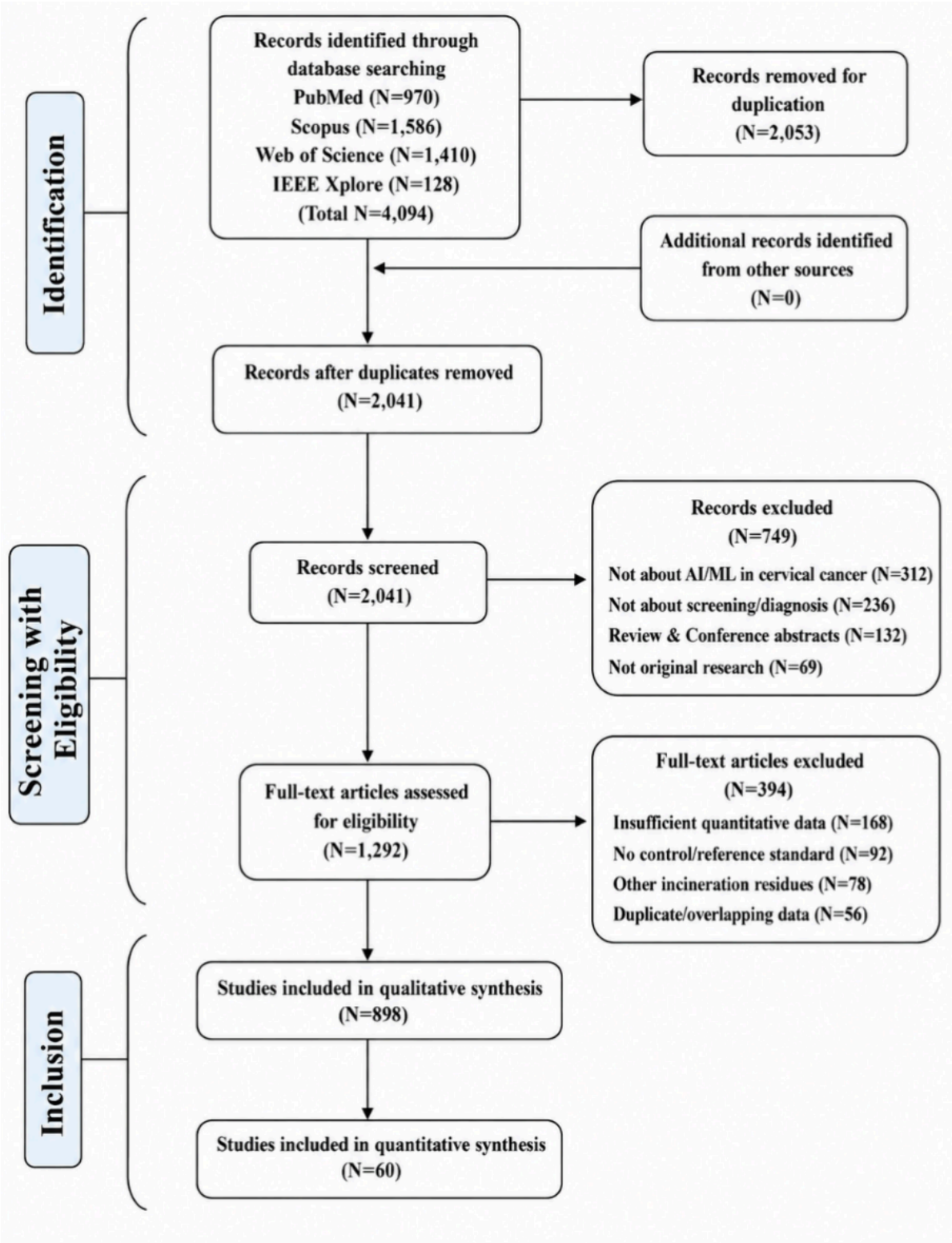


Figure 1

PRISMA 2020 flow diagram of study identification and selection.

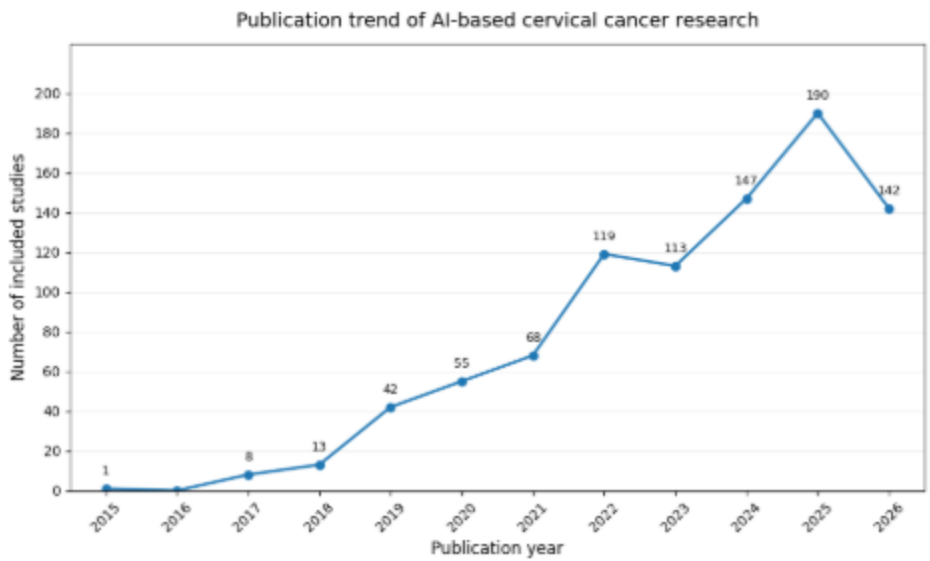


Figure 2

Publication trend of AI-based cervical cancer research from 2015 to 2026.

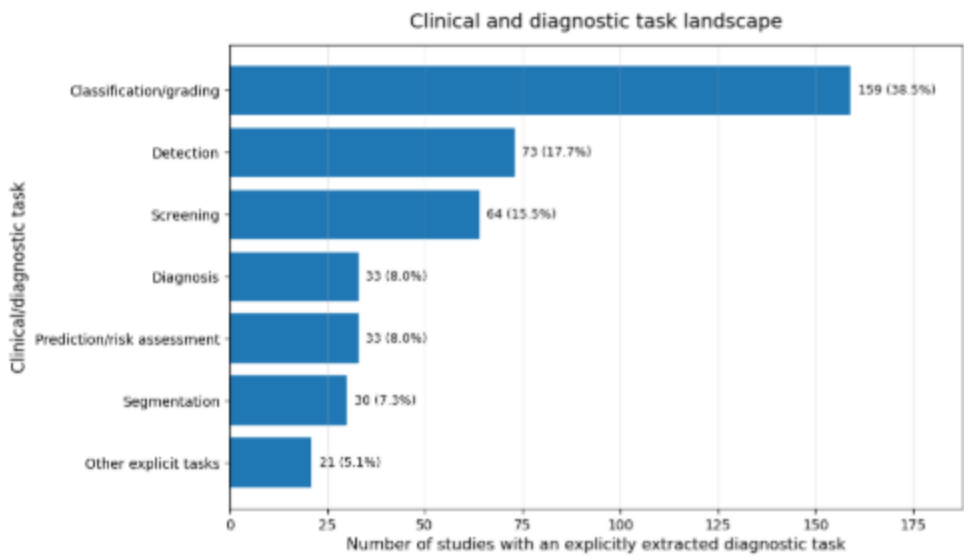


Figure 3

Clinical and diagnostic task landscape of the included studies.

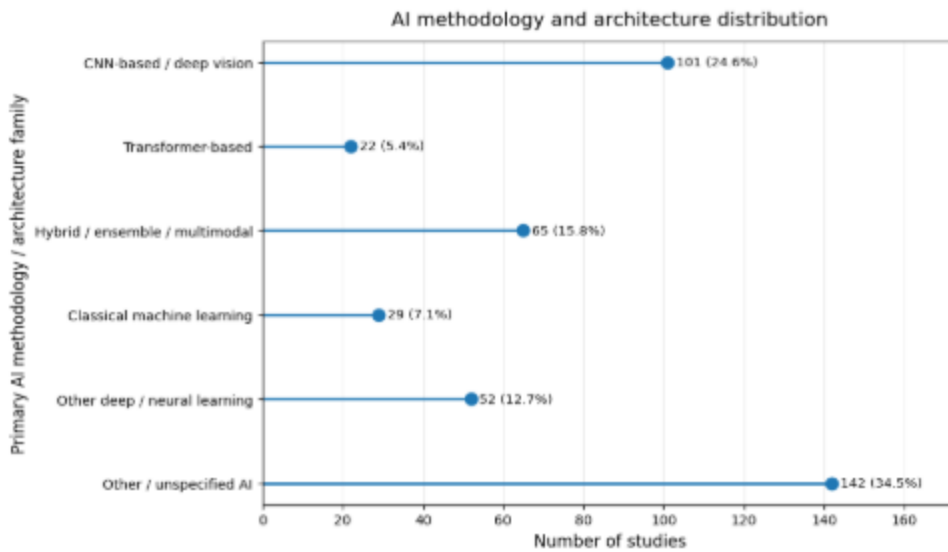


Figure 4

AI methodology and architecture distribution.

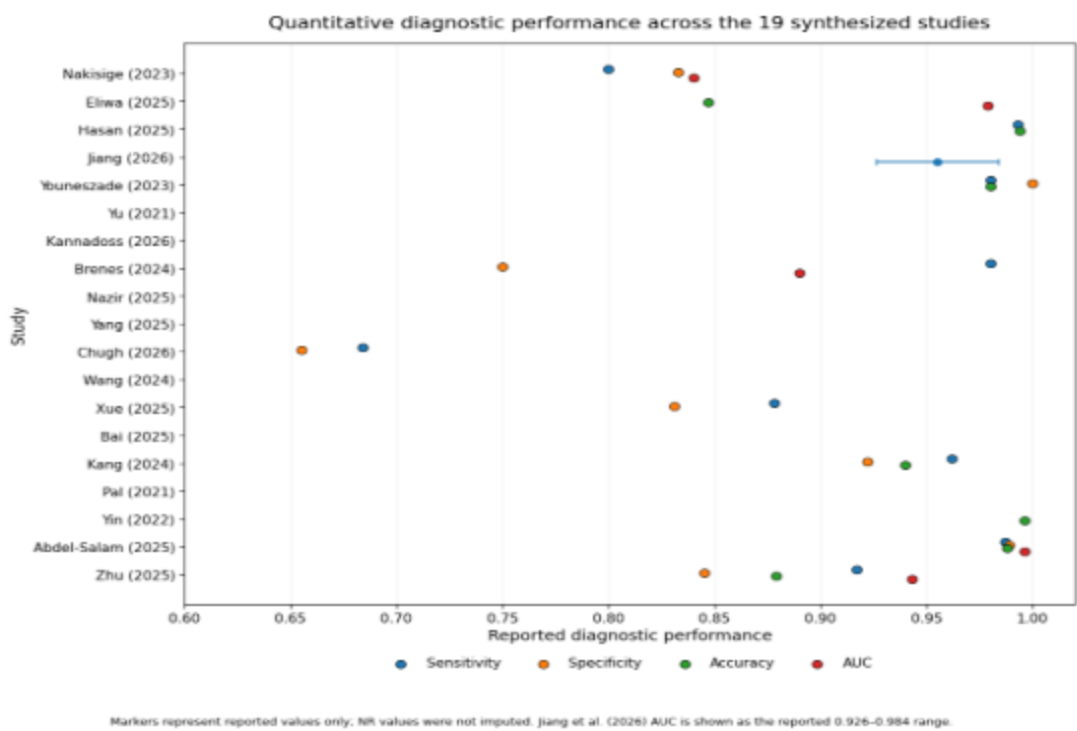


Figure 5

Selected reported diagnostic-performance values from 19 studies within the quantitatively eligible evidence base.

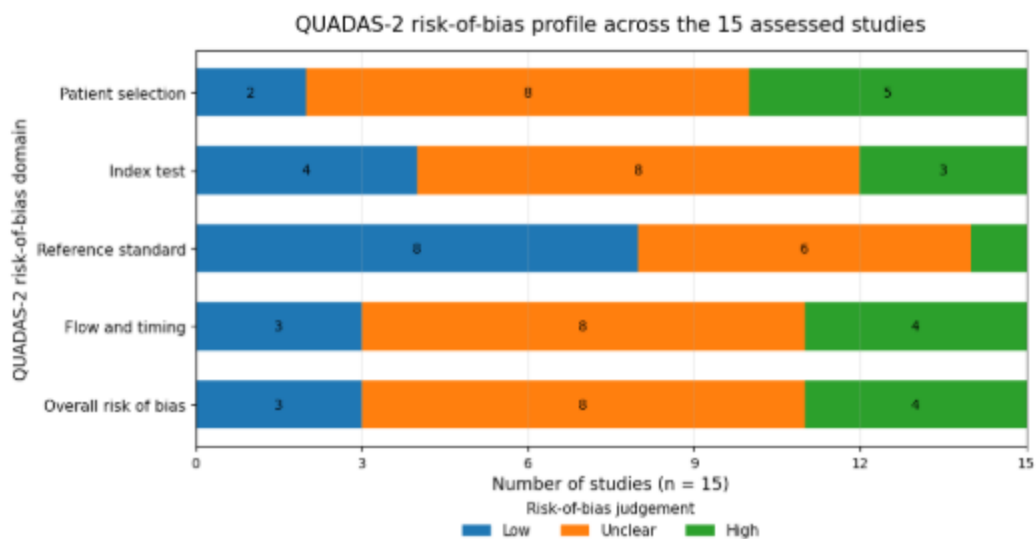


Figure 6

QUADAS-2 risk-of-bias profile across the 15 formally assessed studies.