JHWCR JOURNAL OF HEALTH, WELLNESS AND COMMUNITY RESEARCH 3007-0570

Vol. 4 Issue 6, 2026
Open Access | DOI: 10.61919/mty8kz97

Review Article

Diagnostic Accuracy of Artificial Intelligence Models for Ovarian Tumor Classification on Ultrasound: A Systematic Review

Zujaja Naveed Iqbal[1], Sania Naveed[2], Shumaila Parveen[3], Rabia Khattak[4], Rimsha Latif[5], Nawal Fatima[6]

[1] MBBS, ISRA University, Hyderabad, Pakistan.
[2] Resident Radiology, INMOL Cancer Hospital, Lahore, Pakistan.
[3] Postgraduate Gynecology, Jinnah Postgraduate Medical Center (JPMC), Karachi, Pakistan.
[4] Resident Radiologist, Combined Military Hospital, Peshawar, Pakistan.
[5] MBBS, Niazi Medical and Dental College, Sargodha, Pakistan.
[6] MBBS, King Edward Medical University, Lahore, Pakistan.

✉ **Correspondence:** Rabia Khattak, khattakrabia538@gmail.com

Received: 02 March 2026 **Revised:** 10 March 2026 **Accepted:** 15 March 2026 **Published:** 30 March 2026

HOW TO CITE

Zujaja Naveed Iqbal et al. 2026. Diagnostic Accuracy of Artificial Intelligence Models for Ovarian Tumor Classification on Ultrasound: A Systematic Review. Journal of Health, Wellness and Community Research. 4, 6 (Mar. 2026), 1–8. DOI:https://doi.org/10.61919/mty8kz97

ARTICLE INFORMATION

Author Contributions	Concept: ZNI; Design: ZNI, SN; Data Collection: SP, RL, NF; Analysis: SN, RK; Drafting: ZNI, SN, SP, RK, RL, NF.
Ethical Approval	The study was approved by research ethical board of ISRA University, Hyderabad, Pakistan
Informed Consent	Written informed consent was obtained from all participants
Conflict of Interest	The authors declare no conflict of interest.
Funding	No external funding.
Data Availability	Available from the corresponding author on reasonable request.
AI Declaration	No generative AI was used for conceptualization, analysis, or content generation. AI use was limited to language refinement, and the authors remain fully responsible for the accuracy, originality, validity, and integrity of the manuscript's content.

ABSTRACT

Background: Artificial intelligence (AI) may support ultrasound-based characterization of ovarian and adnexal masses, but differences in diagnostic targets, validation methods, and reporting complicate interpretation of performance. Objective: To systematically evaluate the diagnostic accuracy of ultrasound-based AI models for distinguishing benign from malignant ovarian and adnexal masses and examine factors affecting clinical applicability. Methods: PubMed/MEDLINE, Embase, Scopus, and Web of Science were searched for publications from January 2022 through January 2026, with a search update in March 2026. Eligibility specified adult women, ultrasound-based AI classification, clinical comparators, and histopathological verification. Two reviewers independently screened records and assessed full texts. QUADAS-2 appraisal was described, and findings were synthesized narratively without statistical pooling. Results: The search identified 720 records, with 425 remaining after deduplication. Six reports were listed, covering deep learning pipelines, convolutional neural networks, hybrid echogenic component analysis, and multiclass classification. However, one February 2026 publication exceeded the stated eligibility window, and the eligibility of a cyst-subtype classification study remained unresolved. Participant totals, study-level diagnostic estimates, confidence intervals, and QUADAS-2 judgments were not presented. Consequently, diagnostic performance magnitude, precision, comparative superiority, and overall evidence reliability could not be established. Reported workflow and safety benefits lacked extracted outcome data. Conclusion: The assembled literature illustrates several AI approaches to ovarian ultrasound classification, but the present synthesis does not establish their diagnostic accuracy or readiness for routine implementation. Eligibility reconciliation, complete quantitative extraction, and transparent risk-of-bias reporting are necessary before clinical recommendations can be supported. Prospective independent validation should evaluate defined diagnostic tasks and patient-level consequences. Keywords: Artificial Intelligence; Machine Learning; Deep Learning; Ovarian Neoplasms; Ultrasonography; Sensitivity and Specificity; Systematic Review.

INTRODUCTION

Accurate preoperative characterization of ovarian and adnexal masses is essential for distinguishing patients who require specialist evaluation from those who may be managed conservatively. Ultrasound provides morphological information for this assessment, but interpretation can be challenging when benign and malignant lesions share sonographic features. Artificial intelligence (AI) has therefore been investigated as an adjunct to ultrasound interpretation, with the aim of supporting more consistent classification of ovarian masses. Recent studies have evaluated deep learning pipelines and convolutional neural networks for this purpose, reflecting growing interest in computational approaches to ultrasound-based diagnostic assessment (1,2). Their clinical relevance, however, depends on whether reported performance remains reliable beyond the datasets used for model development.

The emerging literature encompasses several related but distinct diagnostic applications. Li et al. investigated a deep learning model system for the diagnosis and management of adnexal masses, extending the focus beyond image classification alone (3). Wu et al. evaluated an AI-based model designed to distinguish benign, borderline, and malignant adnexal masses, highlighting the importance of defining the target condition when interpreting diagnostic performance (4). Other approaches include the hybrid echogenic component-based framework described by Yoeli-Bik et al. and the EfficientOvaNet model developed by Parekh and Desai for multiclass ovarian cyst classification (5,6). These developments illustrate the diversity of computational methods and classification tasks within the field. Nevertheless, performance in cyst subtype categorization cannot be assumed to represent accuracy in distinguishing benign from malignant tumors, and multiclass results require careful interpretation before they can inform a binary clinical decision.

This diversity creates a need for critical synthesis focused on the comparability and applicability of diagnostic evidence. Model architecture alone does not establish clinical usefulness; interpretation also requires attention to patient selection, the reference standard, classification thresholds, and the independence of validation data. Similarly, performance measured using individual images may not directly represent performance at the patient level, particularly when several images originate from the same lesion or patient. Distinguishing internal testing from external validation is therefore central to assessing generalizability. Comparisons with sonographers or established ultrasound scoring systems also require scrutiny of whether the index test and comparator were evaluated in the same patients under comparable conditions. Without these distinctions, apparently favorable results may obscure differences in study design rather than demonstrate consistent diagnostic superiority.

The recent publication of multicenter diagnostic pipelines, hybrid models, and approaches addressing borderline tumors provides a timely rationale for examining how these developments contribute to the specific task of malignancy classification (1,4,5). A systematic review is appropriate because explicit eligibility criteria, transparent study selection, and structured risk-of-bias assessment can distinguish directly relevant diagnostic evidence from adjacent classification research. Its contribution should be to evaluate the strength and clinical applicability of reported performance, rather than to catalogue architectures or assume that diagnostic accuracy establishes improved patient outcomes. Particular attention is needed to the evidence supporting external validity and to the methodological limitations that may affect confidence in individual estimates.

Accordingly, this systematic review aims to evaluate the diagnostic accuracy of AI models applied to ultrasound in adult women undergoing assessment for ovarian or adnexal masses, using histopathological diagnosis as the reference standard. It examines sensitivity, specificity, and area under the receiver operating characteristic curve, alongside comparisons with human interpretation or established ultrasound scoring systems. It also assesses how to study design, validation strategy, treatment of borderline tumors, and risk of bias influence interpretation of the findings. The review asks: how accurately do ultrasound-based AI models distinguish benign from malignant ovarian or adnexal masses, and how applicable is the available evidence to preoperative diagnostic assessment?

MATERIAL AND METHODS

This systematic review examined the diagnostic accuracy of artificial intelligence (AI) models applied to ultrasound images for the classification of ovarian and adnexal masses. The review used a diagnostic test accuracy framework to distinguish the index test, clinical comparator, target condition, and reference standard. Reporting was structured around PRISMA 2020, and the methodological approach was informed by the Cochrane Handbook for

Systematic Reviews of Diagnostic Test Accuracy. The evidence was synthesized narratively to examine reported diagnostic performance alongside differences in model architecture, clinical application, and validation setting.

Eligible studies were primary empirical investigations involving adult women undergoing evaluation for suspected ovarian or adnexal masses. The index tests comprised AI, machine learning, or deep learning models applied to ultrasound imaging for tumor classification. Eligible comparators included human sonographer interpretation, expert subjective assessment, or established ultrasound scoring systems, including the International Ovarian Tumor Analysis rules. Histopathological diagnosis was the required reference standard. Studies were required to provide diagnostic classification data, including true-positive, false-positive, true-negative, and false-negative results, sufficient to determine sensitivity, specificity, and predictive values. The primary diagnostic question concerned differentiation between benign and malignant masses.

Studies were excluded if they evaluated magnetic resonance imaging or computed tomography without an ultrasound-based AI component, used non-AI computer-aided methods, or involved preclinical or non-human investigations. Case reports, review articles, and conference abstracts without sufficient diagnostic data were also excluded. Reviews retrieved during searching were considered sources for identifying potentially eligible primary studies rather than units of analysis.

Electronic searches were conducted in PubMed/MEDLINE, Embase, Scopus, and Web of Science. The stated publication eligibility window extended from January 2022 through January 2026, and the final search update was conducted in March 2026. Search concepts covered ovarian or adnexal masses, AI-based analytical methods, ultrasound imaging, and diagnostic classification. Controlled vocabulary and free-text terms were combined using Boolean operators. The PubMed search expression was: ("Ovary"[Mesh] OR "Ovarian Neoplasms"[Mesh] OR "ovarian tumor"[tiab] OR "ovarian mass"[tiab] OR "adnexal mass"[tiab] OR "ovarian cyst"[tiab] OR "ovarian cancer"[tiab]) AND ("Artificial Intelligence"[Mesh] OR "Machine Learning"[Mesh] OR "Deep Learning"[Mesh] OR "Neural Networks, Computer"[Mesh] OR "artificial intelligence"[tiab] OR "machine learning"[tiab] OR "deep learning"[tiab] OR "convolutional neural network"[tiab] OR "CNN"[tiab] OR "computer-aided diagnosis"[tiab] OR "CAD"[tiab] OR "radiomics"[tiab]) AND ("Ultrasonography"[Mesh] OR "ultrasound"[tiab] OR "sonography"[tiab] OR "transvaginal ultrasound"[tiab] OR "ultrasonographic"[tiab]) AND ("Sensitivity and Specificity"[Mesh] OR "diagnostic accuracy"[tiab] OR "classification"[tiab] OR "benign"[tiab] OR "malignant"[tiab]). Supplementary searching involved examining the reference lists of retrieved systematic reviews, narrative reviews, and eligible primary studies.

Retrieved records were imported into reference management software and deduplicated using automated and manual procedures. Study selection proceeded in two stages. Two reviewers independently screened titles and abstracts against the eligibility criteria. Potentially eligible reports were then retrieved for independent full-text assessment by the same review process. Disagreements at either stage were resolved through discussion, with adjudication by a third senior reviewer when consensus could not be reached.

Study information was organized in a descriptive extraction table covering author and publication year, model architecture, target population and setting, principal classification task, reported findings, and validation characteristics. The table provided the basis for comparing the clinical purposes and technical approaches represented in the included literature. Diagnostic classification, validation findings, and computational characteristics were considered separately when interpreting the studies.

Two reviewers independently assessed methodological quality using the Quality Assessment of Diagnostic Accuracy Studies-2 tool. Risk of bias was evaluated across patient selection, the index test, the reference standard, and flow and timing. Patient selection assessment considered consecutive or random recruitment, inappropriate exclusions, and the clinical spectrum represented. Index-test assessment considered whether model evaluation was conducted without knowledge of the reference-standard findings and whether diagnostic thresholds were prespecified. Reference-standard assessment examined the use of histopathology and whether its interpretation was independent of the AI output. Flow and timing assessment considered the interval between ultrasound and reference-standard assessment, consistency of verification, and inclusion of participants in the analysis. Each domain was classified as having low, high, or unclear risk of bias. Applicability concerns were assessed for patient selection, the index test, and the reference standard. Disagreements were resolved by re-examining the source report and seeking third-reviewer adjudication when necessary.

The synthesis was descriptive and organized around diagnostic application, model architecture, reported performance, and validation characteristics. Findings were discussed in relation to the populations and classification tasks evaluated, with attention to differences between binary malignancy discrimination and multiclass categorization. The narrative also considered the extent to which reported validation supported interpretation beyond model development settings. No pooled diagnostic estimates were generated.

RESULTS

The literature search identified 720 records, of which 425 remained after deduplication. These counts imply removal of 295 duplicate records, assuming that no other exclusions occurred before screening. Six studies were listed in the final narrative synthesis. The numbers excluded during title and abstract screening, assessed at full text, and excluded after full-text review were not supplied, preventing complete reconciliation of the selection pathway (Table 1).

The six reports comprised one publication from 2022, two from 2024, two from 2025, and one from 2026. They examined several related applications of AI to ovarian and adnexal ultrasound assessment (Table 2). Dai et al. evaluated a retrospective multicenter diagnostic pipeline, and Giourga et al. investigated convolutional neural networks for ovarian mass classification (1,2). Li et al. described a broader diagnostic and management model system (3). Wu et al. addressed benign, borderline, and malignant masses as separate categories, while Yoeli-Bik et al. investigated hybrid echogenic component analysis and Parekh and Desai evaluated multiclass cyst categorization (4–6). Study-level patient, lesion, and image counts were absent from the supplied manuscript; therefore, the total population represented by the review could not be established.

Table 1. Study identification and selection

Selection stage	Number	Detail
Records identified	720	Combined search yield reported in the manuscript
Unique records after deduplication	425	Records available for screening
Duplicate records removed	295	Derived from 720 − 425, assuming no other prescreening removals
Records excluded during title and abstract screening	Not reported	Screening-stage counts require confirmation
Full-text reports assessed	Not reported	Full-text assessment was described
Full-text reports excluded	Not reported	Exclusion counts and reasons were not supplied
Studies listed in the narrative synthesis	6	Final eligibility requires reconciliation
Studies contributing to statistical pooling	0	No meta-analysis was presented

The available information does not permit a complete PRISMA flow diagram. Records, reports, and studies should remain distinct when the selection counts are finalized.

Table 2. Characteristics of studies listed in the review

Study	Design and setting described in the manuscript	Population or diagnostic task	Model or analytical approach	Sample size	Validation information
Dai et al., 2024 (1)	Retrospective multicenter study	Patients evaluated for ovarian masses; benign-versus-malignant classification	Deep learning diagnostic pipeline	Not reported	Evaluation across hospital sites described
Giourga et al., 2024 (2)	Ultrasound image classification study	Patients with adnexal masses; binary classification	Convolutional neural networks applied to two-dimensional ultrasound images	Not reported	Validation cohort structure not specified
Li et al., 2022 (3)	Diagnostic-model development and evaluation	Patients with adnexal masses; diagnosis and management support	Deep learning model system	Not reported	Internal and external validation described
Wu et al., 2026 (4)	Multicenter model development and validation	Classification of benign, borderline, and malignant adnexal masses	AI-based multiclass diagnostic model	Not reported	Multicenter validation described
Yoeli-Bik et al., 2025 (5)	Ultrasound-based diagnostic-model evaluation	Classification of adnexal masses using echogenic features	Hybrid AI approach based on echogenic components	Not reported	Validation mentioned without cohort allocation details
Parekh and Desai, 2025 (6)	Ultrasound image classification study	Multiclass ovarian cyst categorization	EfficientOvaNet	Not reported	Computational performance described; clinical validation details not supplied

"Not reported" refers to the supplied review manuscript, not necessarily the primary publication. Patient, mass, and image counts must be extracted separately. The February 2026 publication date of reference 4 conflicts with the stated January 2026 cutoff. Reference 6 requires confirmation of eligibility for the primary diagnostic question.

Table 3. Risk-of-bias and applicability reporting

QUADAS-2 domain	Assessment described in Methods	Study-level judgments available	Synthesis supported
Patient selection	Recruitment method, exclusions, and clinical spectrum	No	Distribution of selection-bias judgments cannot be determined
Index test	Independence from reference-standard findings and prespecified thresholds	No	Index-test bias cannot be summarized
Reference standard	Histopathological verification and independent interpretation	No	Reference-standard bias cannot be summarized
Flow and timing	Examination-to-verification interval and completeness of verification	No	Flow and timing bias cannot be summarized
Applicability	Patient selection, index test, and reference standard	No	Applicability concerns cannot be compared across studies

Absent judgments must not be converted automatically into "unclear risk." That classification requires assessment of the primary report.

Table 4. Descriptive synthesis of model applications and reported findings

Synthesis theme	Relevant studies	Findings described in the submitted manuscript	Evidential qualification
Binary malignancy classification	Dai et al.; Giourga et al.; Yoeli-Bik et al. (1,2,5)	Favorable classification performance using deep learning, convolutional networks, or hybrid echogenic analysis	Numerical performance and uncertainty were not presented
Diagnosis and management support	Li et al. (3)	A comprehensive deep learning model system was described	Effects on management decisions or patient outcomes were not quantified
Borderline-tumor classification	Wu et al. (4)	Separate classification of benign, borderline, and malignant lesions	Class-specific performance and the binary grouping rule were not reported
Cyst subtype categorization	Parekh and Desai (6)	A lightweight architecture for multiclass categorization was described	Relevance to histopathology-confirmed malignancy discrimination remains unresolved
Validation across settings	Dai et al.; Li et al.; Wu et al. (1,3,4)	Multicenter or external evaluation was described	Cohort independence, size, and setting-specific estimates were not documented
Computational efficiency	Parekh and Desai (6)	Computational efficiency and rapid classification were described	Inference time, hardware requirements, and comparative numerical results were not supplied
Clinical workflow, safety, and avoided procedures	Broad claims across the manuscript	Benefits were asserted	No defined outcomes, event counts, or comparative measurements support these claims

Independent QUADAS-2 appraisal was described in the Methods, but study-level risk-of-bias and applicability judgments were not presented. Consequently, the numbers of studies at low, high, or unclear risk could not be reported for any domain. The available synthesis also did not establish whether the mandatory histopathological reference standard and clinical comparator requirements were satisfied by every listed study (Table 4).

The descriptive synthesis identified variation in model architecture and classification purpose. Approaches included deep learning pipelines, convolutional neural networks, a comprehensive diagnostic system, and hybrid echogenic component analysis (1–5). Efficient Ova Net addressed multiclass cyst categorization, representing a different classification task from the review's primary benign-versus-malignant question (6). Three-class assessment incorporating borderline tumors also required separate interpretation because the review did not specify whether borderline lesions were included in the malignant category for binary comparisons (4).

Although the submitted summaries described diagnostic performance favorably, none of the six study entries contained numerical sensitivity, specificity, AUC, accuracy, or confidence intervals (Table 3). Diagnostic classification counts, thresholds, and comparative statistical results were likewise absent. The magnitude and precision of performance, consistency across studies, and superiority over human interpretation or clinical scoring systems therefore could not be determined from the assembled results. No pooled estimates, heterogeneity statistics, subgroup findings, or sensitivity analyses were available.

Multicenter or external evaluation was described for three reports, but validation cohort sizes, independence, and setting-specific estimates were not documented (1,3,4). Computational efficiency was highlighted for EfficientOvaNet without corresponding timing or resource measurements (6). Claims concerning improved workflow, reduced operator variability, safety, and avoidance of invasive procedures were not supported by extracted outcome data (Table 5). The available evidence therefore permits a descriptive account of the reported approaches, but a quantitative diagnostic synthesis requires completion of the primary-study extraction and appraisal.

DISCUSSION

This systematic review examined the diagnostic accuracy of artificial intelligence models applied to ultrasound for differentiating benign from malignant ovarian and adnexal masses. The six reports presented in the synthesis described several computational approaches, including deep learning pipelines, convolutional neural networks, and hybrid echogenic component analysis. However, the available synthesis did not provide study-level diagnostic estimates, confidence intervals, or risk-of-bias judgments sufficient to establish the magnitude, precision, or consistency of performance. The principal finding is therefore that ultrasound-based AI has been investigated across several relevant diagnostic applications, while the evidence assembled in this review remains insufficient to establish comparative superiority or readiness for routine clinical implementation.

The included literature illustrates differences in the intended role of AI within adnexal mass assessment. Dai et al. and Giourga et al. investigated ultrasound-based classification approaches, whereas Li et al. described a broader system addressing diagnosis and management (1–3). These applications share an interest in extracting diagnostic information from ultrasound but do not necessarily address equivalent clinical decisions. A model intended to classify an individual image, for example, requires different evaluation from a system intended to guide referral or management at the patient level. Consequently, evidence of classification performance should not be interpreted as evidence of improved management unless the associated clinical decisions and outcomes have been evaluated directly.

Differences in target conditions further complicate interpretation. Wu et al. addressed benign, borderline, and malignant adnexal masses as distinct categories, while Yoeli-Bik et al. investigated a hybrid echogenic component-based diagnostic approach and Parekh and Desai focused on multiclass cyst categorization (4–6). These studies demonstrate the breadth of the research landscape, but their outcomes cannot be treated as interchangeable. In particular, successful differentiation among cyst subtypes does not independently establish sensitivity for malignancy. Similarly, the classification of borderline tumors can materially affect the interpretation of binary diagnostic results. A clinically coherent synthesis therefore requires explicit definitions of the target condition and consistent treatment of borderline lesions before performance estimates are compared.

The submitted study summaries generally characterized model performance favorably, but the absence of numerical comparisons prevents assessment of whether findings agree in magnitude or reveal clinically important contradictions. Apparent agreement in the direction of conclusions may conceal differences in diagnostic thresholds, disease prevalence, patient selection, and validation procedures. Without comparable estimates and uncertainty measures, it is also not possible to determine whether one architecture offers a meaningful advantage over another. The supplied reference set contains no earlier systematic reviews against which the present findings can be evaluated directly; accordingly, an advance over previous pooled estimates or conclusions cannot be claimed. The contribution of this review is currently limited to organizing the reported approaches and identifying the methodological distinctions necessary for a more informative diagnostic synthesis.

Validation design is central to the clinical interpretation of this evidence. Multicenter or external evaluation was described for several models, but the synthesis did not document the independence, composition, or size of the relevant cohorts (1,3,4). Multicenter recruitment alone does not establish external validation if data from participating institutions contribute to both model development and testing. Similarly, image-level separation may provide an optimistic assessment when images from the same patient appear in different datasets. Evaluation of generalizability therefore requires transparent reporting of patient-level partitioning, institutional separation, model tuning, and threshold selection. These considerations are particularly relevant when assessing performance across different ultrasound devices, acquisition practices, and clinical settings.

The clinical implications must remain proportionate to the evidence presented. Diagnostic discrimination is only one component of a useful decision-support system. Clinical adoption also depends on the consequences of false-negative and false-positive classifications, the intended position of AI within the diagnostic pathway, and whether assistance improves clinician performance under routine conditions. The present synthesis does not establish reductions in unnecessary procedures, diagnostic delays, or operator variability, nor does it demonstrate improved patient outcomes. It therefore cannot support replacing established clinical assessment or making policy recommendations for widespread deployment. Its more defensible implication is that AI evaluation should be tied to a clearly defined clinical task and a suitable comparator rather than to model performance in isolation.

Several limitations arise specifically from this review's methods and reporting. The restriction to publications from January 2022 onward narrows the evidence base, while inclusion of a February 2026 report remains inconsistent with the stated January 2026 endpoint unless publication eligibility was formally extended. The incomplete selection pathway and absence of full-text exclusion reasons limit assessment of how consistently the eligibility criteria were applied. Eligibility also requires reassessment where the reported task is multiclass cyst categorization rather than histopathology-verified malignancy discrimination. Furthermore, the absence of participant totals, diagnostic estimates, and study-level QUADAS-2 findings prevents evaluation of precision, clinical applicability, and the influence of bias. Potential overlap between cohorts was not resolved, and the narrative synthesis cannot determine whether favorable descriptions reflect consistent findings or differences in selective reporting. These limitations constrain interpretation of the assembled evidence rather than demonstrating that the underlying models are ineffective.

Future research should prioritize prospective evaluation of fixed models and prespecified thresholds in independent clinical populations. Studies should separate participants across development and evaluation datasets, report recruitment and reference-standard procedures transparently, and provide patient-level diagnostic counts with confidence intervals. Borderline tumors should be handled according to an explicit clinical rationale, with separate results where appropriate. Comparative evaluations should assess clinicians with and without AI assistance in the same patients and measure referral decisions, additional investigations, missed malignancies, and unnecessary interventions. A subsequent evidence synthesis should first reconcile the search window and study eligibility, complete the diagnostic extraction and risk-of-bias assessment, and then determine whether sufficiently comparable data support a diagnostic meta-analysis.

CONCLUSION

Ultrasound-based artificial intelligence has been investigated using several approaches to ovarian and adnexal mass classification, but the evidence presented in this review is insufficient to establish the magnitude of diagnostic accuracy, superiority over conventional assessment, or readiness for routine clinical use. Differences in classification tasks and incomplete reporting of validation, diagnostic estimates, and risk of bias constrain the clinical interpretation of the synthesis. The immediate priorities are an auditable, quantitatively complete evidence assessment and prospective independent evaluation of clearly defined AI applications before practice or policy recommendations can be justified.

REFERENCES

1. Dai WL, Wu YN, Ling YT, Zhao J, Zhang S, Gu ZW, et al. Development and validation of a deep learning pipeline to diagnose ovarian masses using ultrasound screening: a retrospective multicenter study. EClinicalMedicine. 2024;78.

2. Giourga M, Petropoulos I, Stavros S, Potiris A, Gerede A, Sapantzoglou I, et al. Enhancing ovarian tumor diagnosis: performance of convolutional neural networks in classifying ovarian masses using ultrasound images. J Clin Med. 2024;13(14):4123.

3. Li J, Chen Y, Zhang M, Zhang P, He K, Yan F, et al. A deep learning model system for diagnosis and management of adnexal masses. Cancers (Basel). 2022;14(21):5291.

4. Wu Y, Dai W, Li X, Zhang S, Gong L, Wang J, et al. Development and validation of an artificial intelligence-based model for diagnosing benign, borderline, and malignant adnexal masses. NPJ Precis Oncol. 2026;10(1):106.

5. Yoeli-Bik R, Whitney HM, Li H, Bilecz A, Abramowicz JS, Lan L, et al. Hybrid artificial intelligence echogenic components-based diagnosis of adnexal masses on ultrasound. Med Phys. 2025;52(7):e17983.

6. Parekh A, Desai M. EfficientOvaNet: efficient deep learning model for multiclass classification of benign ovarian cyst using ultrasound images. Neural Comput Appl. 2025;37(30):25495-514.