Gynecologic and Obstetric Investigation

Gynecol Obstet Invest , DOI: 10.1159/000545850
Received: April 10, 2024
Accepted: April 2, 2025
Published online: May 8, 2025

# Artificial intelligence applied to ultrasound diagnosis of pelvic gynecological tumors: a systematic review and meta-analysis

Geysels A, Garofalo G, Timmerman S, Barreñada L, De Moor B, Timmerman D, Froyman W, Van Calster B

ISSN: 0378-7346 (Print), eISSN: 1423-002X (Online)
https://www.karger.com/GOI
Gynecologic and Obstetric Investigation

Disclaimer:
Accepted, unedited article not yet assigned to an issue. The statements, opinions and data contained in this publication are solely those of the individual authors and contributors and not of the publisher and the editor(s). The publisher and the editor(s) disclaim responsibility for any injury to persons or property resulting from any ideas, methods, instructions or products referred to the content.

Copyright:
This article is licensed under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC) (http://www.karger.com/Services/OpenAccessLicense). Usage and distribution for commercial purposes requires written permission.

© 2025 The Author(s). Published by S. Karger AG, Basel

Accepted Manuscript

Downloaded from http://karger.com/goi/article-pdf/doi/10.1159/000545850/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025

**Artificial intelligence applied to ultrasound diagnosis of pelvic gynecological tumors: a systematic review and meta-analysis**

Axel Geysels[1,#], Giulia Garofalo[2,3,4,#], Stefan Timmerman[2,3], Lasai Barreñada[2], Bart De Moor[1], Dirk Timmerman[2,3], Wouter Froyman[2,3,**], and Ben Van Calster[2,5,6,**]

[1]Department of Electrical Engineering, STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, Leuven, Belgium
[2]Department of Development and Regeneration, KU Leuven, Leuven, Belgium
[3]Department of Obstetrics and Gynaecology, University Hospitals Leuven, Leuven, Belgium
[4]Department of Gynaecology and Obstetrics, Université Libre de Bruxelles (ULB) Hôpital Universitaire de Bruxelles (H.U.B.), Hôpital Erasme, Brussels, Belgium
[5]Department of Biomedical Data Sciences, Leiden University Medical Centre, Leiden, Netherlands
[6]Leuven Unit for Health Technology Assessment Research (LUHTAR), KU Leuven, Leuven, Belgium

[#]Joint first authors
[**]Joint last authors
Corresponding author: Prof. Ben Van Calster, Department of Development and Regeneration, KU Leuven, Herestraat 49, 3000 Leuven, Belgium. E-mail address: ben.vancalster@kuleuven.be

**Running head:** AI for US diagnosis of gynecological tumors
**Keywords:** Gynecological oncology; Artificial intelligence; Ovarian cancer; Endometrial cancer; Myometrial cancer; Ultrasound; Meta-analysis

Accepted Manuscript

Downloaded from http://karger.com/goi/article-pdf/doi/10.1159/000545850/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025

**Abstract**

Objective: To perform a systematic review on artificial intelligence (AI) studies focused on identifying and differentiating pelvic gynecological tumors on ultrasound scans.

Methods: Studies developing or validating AI models for diagnosing gynecological pelvic tumors on ultrasound scans were eligible for inclusion. We systematically searched PubMed, Embase, Web of Science, and Cochrane Central from their database inception until April 30th, 2024. To assess the quality of the included studies, we adapted the QUADAS-2 risk of bias tool to address the unique challenges of AI in medical imaging. Using multi-level random effects models, we performed a meta-analysis to generate summary estimates of the area under the receiver operating characteristic curve (AUC), sensitivity, and specificity. To provide a reference point of current diagnostic support tools for ultrasound examiners, we descriptively compared the pooled performance to that of the well-recognized ADNEX model on external validation. Subgroup analyses were performed to explore sources of heterogeneity.

Results: From 9151 records retrieved, 44 studies were eligible: 40 on ovarian, three on endometrial, and one on myometrial pathology. Overall, 95% were at high risk of bias – primarily due to inappropriate study inclusion criteria, the absence of a patient-level split of training and testing image sets, and no calibration assessment. For ovarian tumors, the summary AUC for AI models distinguishing benign from malignant tumors was 0.89 (95% CI: 0.85-0.92). In lower-risk studies (at least three low-risk domains), the summary AUC dropped to 0.87 (0.83-0.90), with deep learning models outperforming radiomics-based machine learning approaches in this subset. Only five studies included an external validation, and six evaluated calibration performance. In a recent systematic review of external validation studies, the ADNEX model had a pooled AUC of 0.93 (0.91-0.94) in studies at low risk of bias. Studies on endometrial and myometrial pathologies were reported individually.

Conclusion: Although AI models show promising discriminative performances for diagnosing gynecological tumors on ultrasound, most studies have methodological shortcomings that result in a high risk of bias. In addition, the ADNEX model appears to outperform most AI approaches for ovarian tumors. Future research should emphasize robust study designs – ideally large, multicenter, and prospective cohorts that mirror real-world populations – along with external validation, proper calibration, and standardized reporting.

Registration: This study was pre-registered with Open Science Framework (OSF): https://doi.org/10.17605/osf.io/bhkst.

Downloaded from http://karger.com/goi/article-pdf/doi/10.1159/000545850/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025

## Introduction

Pelvic gynecological cancers pose a substantial global health challenge. Ensuring early and accurate detection, as well as correct patient referral to specialized centers, is crucial for improving outcomes [1,2]. Additionally, it is important to avoid unnecessary surgeries in patients with benign tumors, thereby limiting surgical risks and preserving fertility in younger patients. Ultrasound imaging is widely accepted as the first-line modality for evaluating these tumors, because of its strong diagnostic capabilities, good safety profile, and cost-effectiveness. However, interpreting ultrasound images remains challenging due to its inherent subjectivity and reliance on the operator's expertise [3]. To aid in clinical decision-making, several ultrasound-based risk prediction models have been developed, with the extensively validated Assessment of Different NEoplasias in the adneXa (ADNEX) model for ovarian tumors diagnosis being the most prominent [4,5]. However, these models still depend on manual assessment of ultrasound variables, thereby requiring adequately trained examiners, and remain susceptible to interobserver variability [6].

In recent years, artificial intelligence (AI)-based imaging techniques have shown promise in various medical domains, claiming performances comparable with expert clinicians on diagnostic tasks [7,8]. Applied directly to ultrasound scans, AI has the potential to streamline image interpretation, reduce operator-dependence, and improve diagnostic performance. However, despite growing research interest, routine clinical integration of AI into gynecological ultrasound applications remains limited, evidenced by the relatively few AI-based devices that received approval from the United States Food and Drug Administration (FDA) [9].

Previous reviews [10–19] have examined AI models for diagnosing gynecological tumors, yet many focused exclusively on ovarian tumors, lacked a risk of bias assessment, or included multiple imaging modalities rather than only ultrasound. Only three reviews performed a meta-analysis, with most treating AI models merely as diagnostic tests at a single threshold, rather than fully assessing their discriminative performance across multiple thresholds, characteristic of true risk prediction models.

Therefore, the objectives of this systematic review were to: (i) describe studies applying AI to ultrasound for diagnosing pelvic gynecological tumors (qualitative synthesis), (ii) perform a risk of bias assessment, (iii) conduct a meta-analysis (quantitative synthesis) examining both diagnostic test performance (based on single threshold) and full discriminative performance (across all thresholds), and (iv) benchmark these results against existing state-of-the-art diagnostic support tools.

Downloaded from http://karger.com/goi/article-pdf/doi/10.1159/000545850/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025

## Methods

Protocol and registration

We registered this study prospectively on Open Science Framework (OSF; registration: https://doi.org/10.17605/osf.io/bhkst, protocol: https://osf.io/6ps3f) and reported it in accordance with the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines (Note S1). While our initial intention was to perform a scoping review to broadly map the emerging literature on AI in gynecologic ultrasound, the relatively small number of identified articles prompted us to refine our approach into a systematic review with meta-analysis, thereby enabling a more detailed quantitative assessment of the included studies [20].

Eligibility criteria

We included studies in our systematic review evaluating diagnostic AI models using ultrasound scans to aid in the disease detection and differentiation of pelvic gynecological tumors – tumors originating from the female reproductive system, such as those in the ovaries, fallopian tubes, myometrium, endometrium, or cervix. Exclusion criteria were: (i) studies with primary objectives other than diagnosing benign versus malignant tumors, (ii) studies implementing AI models based only on descriptive ultrasound variables (e.g., measurement, morphological scores) collected by an ultrasound examiner rather than the imaging data itself, (iii) studies using imaging modalities other than ultrasound, and (iv) studies assessing non-pelvic gynecological tumors.

Information sources and search strategy

The search strategy spanned four databases: PubMed, Embase, Web of Science and Cochrane Central, from database inception up to April 30$^{th}$, 2024. Three main search "branches" were developed: gynecological pathology (a), ultrasound modality (b) and artificial intelligence (c). For the first branch, five different concepts were assigned: ovarian tumors (a1); tubal tumors (a2); myometrial tumors (a3); endometrial tumors (a4); cervical tumors (a5). The final equation used was "(*a1* OR *a2* OR *a3* OR *a4* OR *a5*) AND *b* AND *c*" with customized search strings for each database detailed in Note S2. Backward snowballing (i.e., backward reference checking) was performed on systematic reviews of AI applied to the diagnosis of pelvic gynecological tumors to identify any articles that may have been missed during the initial search [21]. There were no language restrictions, no filters were applied concerning article type (e.g., conference papers, abstract, etc.), and non-indexed articles were also considered.

Study selection

The studies identified from the different databases were imported in the Rayyan web application for automatic de-duplication [22]. Two reviewers (GG and ST) performed title and abstract screening independently and blinded according to our eligibility criteria, after which relevant full-text articles were retrieved. If only an abstract was available or if the full-text could not be retrieved, we contacted the authors to request the full-text (Table S1). Any disagreement was resolved through unblinded consensus before proceeding.

Data extraction

Data extraction was performed independently by two reviewers (AG and GG), each collecting a predetermined set of relevant variables, shown in Figure S1. Any discrepancies were resolved by a third reviewer (ST). Using a standardized extraction sheet, we collected two categories of variables. The first, referred to as "study variables", included aspects such as study methodology, reference standards, and data curation processes. The second, termed "statistical variables", included



Downloaded from http://karger.com/goi/article-pdf/doi/10.1159/000545850/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025

Accepted Manuscript

samples sizes (patient and image), modelling techniques, interpretability methods, validation techniques, and reported performances.
The reported performance metrics and sample size data – indicating either the number of images or patients, depending on the study – were used to derive the true positives, true negatives, false positives, and false negatives for reconstructing the 2×2 tables (assuming a specific threshold). For studies with multiclass outputs, we combined the classes into binary (benign versus malignant) categorization. When insufficient data were available for reconstruction, the authors were contacted for additional information (Table S1).

Risk of bias assessment

To assess the risk of bias, we used an adapted version of the QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies [23]) tool. At the time of the assessment, no dedicated AI-focused frameworks were available (QUADAS-AI [24] and PROBAST-AI [25] have been announced but not yet published). Therefore, we integrated signaling questions from PROBAST [26] (Prediction Model Risk of Bias Assessment) and introduced additional AI-for-medical-imaging items into the QUADAS-2 tool, resulting into two distinct index test domains: model development and model evaluation (Table S2-3). Studies were grouped by anatomical tumor type, and we provided a summary of the risk of bias for each domain. Following QUADAS-2 guidelines, any study rated as "high risk" in one or more domains was at overall "risk of bias" [23]. To allow for a more fine-grained analysis, studies with at least three domains rated as low risk were classified as "less high risk", whereas those with fewer than three low-risk domains were classified as "very high risk".
Studies that did not meet either of the two "fatal items" in the "index test: model development" domain – (i) no train-test split or (ii) no patient-wise train-test split when multiple images per patient were used – were deemed to have major methodological flaws, as this likely produces biased performance estimates and compromises the study's validity. Consequently, such studies were excluded from the meta-analysis following guidance from the Cochrane Handbook for Systematic Reviews of Interventions [27]. Three reviewers (GG, AG, ST) independently performed the quality assessment, with conflicts resolved by consensus.

Statistical analysis

We conducted meta-analyses of the reported model performances, including only studies without major methodological flaws and with sufficient data to reconstruct the 2×2 tables (Table S4). To ensure a fair comparison, models requiring clinical inputs were excluded, although manual ROI delineation for radiomics studies was permitted. In cases when multiple internal validation strategies were applied to the same model, we selected the most robust approach (e.g., resampling techniques such as bootstrapping or cross-validation) [28]. For ensemble models that combine multiple models to produce one single output prediction, we reported only the ensemble's overall performance, provided no clinical inputs were used in the ensemble.
We performed a meta-analysis of diagnostic test accuracy (sensitivity and specificity) using a bivariate, multi-level random effects model and performed a separate multi-level meta-analysis of overall predictive performance (AUC) of those studies that reported the AUC (the AUC meta-analysis was not pre-specified in the protocol but introduced post-hoc to limit optimism by relying on test accuracy results based on statistically optimized thresholds). Forest plots were produced to visualize the results. As pre-specified in the protocol, summary receiver operating characteristic (SROC) curves were generated as well.
The multi-level approach accounts for the hierarchical nature of the data, as some studies presented multiple models and certain authors conducted multiple studies [29]. We chose to cluster by author group rather than by study, since multiple studies from the same group



Downloaded from http://karger.com/gol/article-pdf/doi/10.1159/000545850/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025

consistently used at least part of the same development dataset. To further prevent any single author group for disproportionally influencing the meta-analysis by reporting numerous models, we imposed a cutoff and included only the 10 models within each group that had the highest reported diagnostic odds ratio (DOR), while still ensuring at least one model per study was represented. Heterogeneity was assessed at two levels: within-cluster (variation across multiple models or studies from the same author) and between-cluster (variation across different authors). We quantified heterogeneity using $I^2$ (a relative measure of variability) and $\tau^2$ (an absolute measure of variability) [30]. To determine whether a multi-level model provided a superior fit, we employed a likelihood ratio test. We descriptively compared our AUC summary estimates to the performance of the ADNEX model with and without the CA-125 predictor based on a recent systematic review of external validation studies deemed at low risk of bias [5].

To explore sources of heterogeneity, we conducted subgroup analyses stratified by: (i) risk of bias (categorized as "less high" versus "very high"); (ii) the modelling approach (machine learning, deep learning, or rule-based) – both across all studies and within the "less high risk" subgroup; (iii) publication year (split at the median year of publication); and (iv) patient sample size (split at the median number of patients).

Publication bias was evaluated using a funnel plot and Egger's regression test, with both DOR and AUC serving as performance metrics. A p-value less than 0.05 was considered statistically significant. All analyses were performed with R software (v.4.4.1) using the "metafor" package [31].

Downloaded from http://karger.com/gol/article-pdf/doi/10.1159/000545850/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025

## Results

Study selection

Figure 1 shows the PRISMA flowchart of the screening process and the number of records that were excluded per step. Ultimately, 44 studies were eligible for inclusion. We subdivided the included studies based on anatomical tumor type, identifying 40 articles concerning adnexal masses, three related to endometrial pathology and one on myometrial pathology. Given the limited number of studies for the latter two pathologies, each will be described individually.

### *Ovarian tumors (40)*

Study characteristics

The characteristics of the 40 studies on ovarian tumors [32–71] can be found in Table 1. Of these, 26 studies (65%) [38,40,42–55,57,58,61,63,64,66–70] were published in the last four years (Figure S2) and all but one [62] were development studies. In terms of study design, 31 studies (78%) [38,40–61,63,64,66–71]) used retrospective cohorts, two studies (5%) [53,62] used prospective cohorts, 19 (48%) [39,41,43–48,50–53,56,58,63,64,66,69,70]) were single-center, and nine (23%) [39,40,42,49,54,55,59,60,62]) were multicenter. Inclusion and exclusion criteria varied considerably (e.g., different reported maximal intervals between ultrasound examination and surgery). Most studies (n=33, 83%) [38–61,63–71] used 2D ultrasound scans for model development; in six (15%) [32–37] 2D images were extracted from 3D volumes, and in one exclusively 3D scans were used [62]. Imaging systems ranged from single-vendor (13 (33%) [32–36,39,45,48,52,56,62,66,69]) to multi-vendor (16 (40%) studies [40–43,46,47,49–51,53,54,58,63,64,70,71]). Details on data curation were scarce: only 15 studies (38%) [32–37,42,45,46,50,54,55,60,63,69] described image selection procedure, and 12 (30%) [40,46,48,51,54,56,58–60,62,63,66] described data anonymization. Most studies (n=34, 85%) [32–36,38,39,41–53,55–60,62–64,66,68–71] used histology as the exclusive primary reference standard; one study [37] used radiologic assessment, and two (5%) [40,54] combined histology with ultrasound follow-up over time (these latter cases being presumed benign).

The statistical variables are summarized in Table 2, with 18 studies (45%) [32–39,48,52,56,57,59,60,64,65,68,69] exclusively presenting machine learning models using radiomics features, 15 (38%) [41,42,44–46,49,50,54,55,58,61,63,66,67,70] using deep learning models, four (10%) [40,43,47,51] implementing both machine learning and deep learning models, and two (6%) [53,71] employing rule-based systems that operate on a predefined set of rules (details in Table S5).

Manual delineation of the region-of-interest (ROI) was required in 24 studies (60%) [32,35–39,43,44,47,48,50,52,54,56,59,60,62,64,65,68–72], but only nine (23%) [43,44,47,48,50,53,62,64,70] described in detail the delineation procedure and delineator experience. By contrast, seven studies (18%) [40,49,55,57,58,61,66] used fully automatic delineation, while ten studies (25%) [33,41–43,46,47,51,59,63,68] had at least one model that did not require ROI delineation.

Validation techniques varied considerably. Fourteen studies (35%) [32–37,39,52,56,58–60,68,70] relied exclusively on cross-validation, whereas 20 studies (50%) [40–51,54,55,61,63–67] used a single training-test split; two studies (5%) [38,57] did both. Three development studies (9%) [53,69,71] applied no validation strategy and assessed performance only on their training set, and one study [62] was a dedicated validation study, assessing existing models' performance in its entire cohort. Full model discrimination with AUC was reported in 21 studies (53%) [37,39–45,47,48,50–52,54,55,58,60,63,64,69,70].



Downloaded from http://karger.com/goi/article-pdf/doi/10.1159/000545850/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025

Eleven studies (28%) [42,45,46,48–51,54,56,58,70] compared AI model performance to a clinician's subjective assessment or established risk stratification tools such as the Ovarian-Adnexal Reporting & Data System (O-RADS). Six studies (15%) [43,47,48,51,53,62] included clinical-measured features alongside images as inputs for their models. One study [54] investigated whether less experienced clinicians (3-5 years) could improve their diagnostic accuracy by having access to all model predictions, and another study [42] tested if combining model predictions and expert assessment (7-23 years) in uncertain cases could yield better outcomes. To improve model interpretability, eight studies (20%) [43,44,46,47,51,58,63,70] produced saliency maps (i.e., heatmaps highlighting input regions most relevant to predictions), whereas two studies (5%) [40,51] explicitly evaluated feature importance in their machine learning models. Only five development studies (15%) [40,42,49,51,53] conducted external validation. Among the 21 studies (53%) [32–36,38,40,42,44–46,49–51,54,55,59,60,63,64,66] using multiple images per patient, seven (33%) [40,42,49–51,54,64] performed patient-level evaluations; all other studies applied image-level analyses. Calibration was assessed in just six studies [40,42,47–49,54] (15%).

Risk of bias assessment

Overall, 95% of the studies (38 out of 40) [32–71] were at "high risk of bias" (Figure S3). Eight (20%) [35,37,42,48,56,58,69,70] were deemed high risk for patient selection, owing either due to non-consecutive sampling [35,37] or to inappropriate inclusion/exclusion criteria [42,48,56,58,69,70]. For model development, 21 studies (53%) [32–36,38,44–46,53–55,57,59–61,63,65,67,69,71] were high risk. Three (8%) [53,69,71] only reported training-set performance, and ten (25%) [32–36,44–46,55,63] used multiple images per patient without a patient-wise train-test split, both considered "fatal items" (Table S2). Other issues included enforcing a 50–50 benign-to-malignant ratio during cross-validation [38,57,59,60] or in a fixed test set [54].

Model evaluation was deemed high risk in 35 studies (88%) [32–71], primarily because they did not assess calibration or define the risk thresholds used to evaluate the model. Histology served as the reference standard in 34 studies (85%) [32–36,38,39,41–53,55–60,62–64,66,68–71] (low risk of bias), while flow and timing risks were unclear in 24 studies (60%) [32–36,39,40,42,44,46,48,49,56–58,61–63,65–68,70,71] due to unreported intervals between ultrasound scan and diagnosis. Two studies (5%) [40,54] used both histology and conservative follow-up as reference standards; although this can introduce verification bias, it also avoids selection bias by not limiting the analysis to surgical data [23]. Consequently, no increased risk of bias was assigned for these two studies. Overall, applicability concerns were limited.

Performance summary

We included 25 ovarian cancer studies (63%) [37–40,42,43,47–52,54,56–60,62,64–68,70] in the meta-analysis, contributing a total of 76 AI models. Other studies were excluded because they failed to meet the "fatal items" of the risk of bias assessment [32–36,44–46,53,55,63,69,71] or lacked insufficient data to reconstruct a 2×2 table [41,61]. Among the included studies, only 15 (38%) [37,39,40,42,43,47,48,50–52,54,58,60,64,70] reported an AUC value, comprising 50 models. A multi-level modelling approach provided a significantly better fit compared to the standard two-level model for both diagnostic test accuracy ($\chi_1^2$=17, p=4e-5) and AUC ($\chi_1^2$=16, p=7e-5). Overall, the summary sensitivity was 0.86 (95% CI: 0.82-0.89) and summary specificity 0.84 (0.80-0.88) (Figure S4). Among the studies reporting AUC, the pooled AUC was 0.89 (0.85-0.92). Both meta-analyses showed heterogeneity, mainly driven by between-cluster variation across different authors.



Downloaded from http://karger.com/gol/article-pdf/doi/10.1159/000545850/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025

Although 95% of the included studies were at high risk of bias, 14 studies (35%) [38,40,42,43,47–52,54,59,60,64] were classified at "less high risk" by having at least three low-risk domains (Table S2). In this subgroup, the summary sensitivity and specificity were at 0.82 (0.76-0.86) and 0.81 (0.75-0.85), respectively (Figure 2). Among the 11 studies (28%) [40,42,43,47,48,50,51,53,54,60,64] that reported an AUC in this subgroup, the pooled AUC also decreased to 0.87 (0.83-0.90) (Figure 3). Within these "less high risk" studies, deep learning models generally matched or slightly outperformed machine learning models. In a recent systematic review of external validation studies at low risk of bias, ADNEX with CA-125 had a pooled AUC of 0.93 (0.91-0.94) and ADNEX without CA-125 had a pooled AUC of 0.91 (0.90-0.93). Additional subgroup analyses can be found in Figures S5-S17.

Of the six studies (15%) [40,42,47–49,54] that performed calibration analyses, three [47,49,54] showed good subjective alignment to the 45-degree line, whereas the remaining studies demonstrated overestimation [40] or underestimation [42,48] of the predicted probabilities. Funnel plots (Figures S18-19) showed asymmetry (Egger's test: p=5e-10 for the diagnostic test model and p=3e-6 for the AUC model). These results suggest possible issues with performance overestimations due to small-study effects, high risk of bias, and publication bias.

***Endometrial tumors (3)***

Study Characteristics

We identified three studies investigating endometrial pathologies: Michail et al. [73] and Wang et al. [74] focused on binary classification (benign or malignant), while Fang et al. [75] addressed a three-class problem (polyps, hyperplasia, and cancer). Study details and statistical variables are provided in Tables 1 and 2, with inclusion and exclusion criteria listed in Table S6. None of the studies provided data curation details.

Risk of bias assessment

All three studies were rated as high risk of bias. For patient selection, Fang et al. [75] had inappropriate inclusion criteria, and Wang et al. [74] used a case-control design. Regarding model development, Michail et al. [73] and Wang et al. [74] did not perform a train-test split, reporting only training-set performance and thereby not complying with the "fatal" item criteria (Table S2). Although Wang et al. [74] provided threshold-specific metrics (accuracy, sensitivity, specificity, and kappa index), the exact threshold applied was not specified. None of the studies assessed calibration performance. Histology served as the reference standard in Michail et al [73] and Wang et al. [74], while Fang et al. [75] did not specify a reference standard. No study reported the time interval between scan date and reference standard. Overall, applicability concerns were limited.

Performance summary

Due to the non-compliance with the "fatal" risk of bias items, the results of Michail et al. [73] and Wang et al. [74] should be interpreted with caution. Michail et al. [73] evaluated logistic regression models applied on different ROIs, reporting AUC values of 0.86-0.90, which increased to 0.91-0.96 after applying a contrast-enhancing filter that facilitated manual ROI segmentation. Wang et al. [74] noted a sensitivity of 0.89 and specificity of 0.94, which rose to 0.91 and 0.98, respectively, in the experimental group. Fang et al. [75] achieved an aggregated AUC of 0.87 (averaged across all three classes and five cross-validation folds). Sensitivity and specificity from Fang et al.'s [75] models are shown in Figure S20.

***Myometrial tumors (1)***



Downloaded from http://karger.com/gol/article-pdf/doi/10.1159/000545850/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025

Study Characteristics

Chiappa et al. [72] conducted the only included study on myometrial tumors. This single-center, retrospective pilot study aimed to predict malignancy risk in uterine myometrial lesions. Tables 1 and 2 provide the study and statistical variables, respectively, while the complete inclusion criteria are available in Table S6 (no exclusion criteria were reported).

Risk of bias assessment

The study was considered high risk of bias in the index test – model evaluation domain. Specifically, a "top three" Support Vector Machines (SVM) ensemble was selected post-hoc based on validation results of the full SVM ensemble, and no calibration assessment was performed. By contrast, all other assessed domains were deemed low risk of bias, and no applicability concerns were identified.

Performance summary

Chiappa et al. [72] evaluated two SVM ensembles. The first, comprising ten SVMs, reported an AUC of 0.83 (standard deviation 0.04). The second, a "top three" SVM ensemble, achieved an AUC of 0.86 (standard deviation 0.03) (Figure S21). Sensitivity and specificity are shown in Figure S22.

Downloaded from http://karger.com/goi/article-pdf/doi/10.1159/000545850/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025

## Discussion

Main findings

Most studies have focused on ovarian tumors, with relatively few investigating other pathologies such as endometrial or myometrial tumors.

Study methodology and quality varied substantially. Most studies were rated as high risk of bias (via the adapted QUADAS-2 assessment), primarily due to inappropriate inclusion/exclusion criteria, lack of patient-level train-test splits when multiple images per patient were used, and the absence of calibration assessments.

Overall, AI models demonstrated high accuracy in differentiating benign from malignant tumors. For ovarian tumors, we compared the pooled estimates for these AI models (summary AUC: 0.89, 95% CI: 0.85-0.92) to those of the extensively validated ADNEX diagnostic risk model (AUC: 0.93, 95% CI: 0.91-0.94), which relies on descriptive ultrasound variables, using results from a recent systematic review of external validation studies [5]. AI models underperformed relative to ADNEX – particularly in studies with less high risk of bias – a result that aligns with a 2019 systematic review, though that review predates many recent deep learning studies [76]. Within the less high-risk subgroup, deep learning approaches slightly outperformed radiomics-based machine learning. Funnel plots and Egger's regression tests suggested publication bias, and only a limited number of studies performed external validation or calibration assessments.

Our findings of high discriminatory performance but poor study quality of the AI models is consistent with previous reviews [10–19]. By adapting QUADAS-2 for medical imaging-based AI, excluding methodologically flawed studies, accounting for clustering effects among models from the same author, and performing both diagnostic test and AUC meta-analyses, we aimed to minimize bias and strengthen confidence in our overall findings.

Strengths and limitations

A key strength of our study is the dual evaluation of AI models, both as diagnostic tests and as complete predictive risk models. We also present a descriptive comparison with ADNEX – a leading diagnostic risk model for ovarian tumor diagnosis. However, because the AI models and ADNEX were not evaluated on the same dataset, we did not undertake a formal head-to-head statistical comparison. In addition, we performed an in-depth data extraction and applied an adapted QUADAS-2 risk of bias assessment designed to address the methodological challenges of medical imaging-based AI.

There are, however, some limitations. Some authors of this review are members of the steering committee of the International Ovarian Tumor Analysis (IOTA) consortium and contributed to some of the included studies. To mitigate potential conflicts of interest, we followed a predefined protocol and involved independent researchers for both data extraction and quality assessment. Further constraints of our findings (as opposed to the study itself) include the lack of a standardized bias assessment tool for AI prediction models, forcing us to adapt the QUADAS-2 tool. Additionally, few studies addressed pathologies other than ovarian tumors, did external validations, or calibration assessments, preventing us from conducting separate meta-analyses for these categories.

Study implications

Our systematic review found that most AI models for gynecologic ultrasound focus on ovarian pathology, while other gynecologic tumors (such as endometrial, cervical, or myometrial) remain understudied and should be explored more in the future. We recommend following guidelines such as TRIPOD+AI [77] to ensure transparent reporting, including detailed descriptions of data sources,



Downloaded from http://karger.com/gol/article-pdf/doi/10.1159/000545850/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025

model development steps, anonymization practices, and performance metrics. In particular, model discrimination is best evaluated using AUC, reported in about half of the studies, as it reflects performance across all thresholds, whereas traditional diagnostic test metrics (e.g., sensitivity and specificity) are less relevant for risk models unless a specific "actionable" threshold is known in clinical practice [78]. Calibration, reported in only six studies [40,42,47–49,54], is also an important yet under-reported measure of model performance, which evaluates how closely a model's estimated risks of malignancy agree with the actual outcomes in clinical practice.

High-quality study design is crucial for building representative datasets. Ideally, this involves large, multicenter, and prospective cohorts that mirror real-world populations. After a model is trained and internally validated, external validation on an independent dataset (performed in only five studies [40,42,49,51,53]) is essential to confirm generalizability. Clinical impact can be further established by comparing or combining AI with clinical experts or established risk models under realistic conditions. To improve trust and encourage adoption, studies should include explainability techniques that clarify how a model arrives at its predictions. For example, saliency maps (implemented in nine studies [43,44,46,47,51,58,63,70]) or feature-importance analyses (used in only two studies [40,51]) can enhance interpretability. However, each approach has its limitations, and more research is needed to better understand these "black box" models [79,80]. Finally, exploring other ultrasound modalities – such as 3D imaging or video – could provide a more comprehensive representation of the tumor's morphology and vascularization, potentially improving diagnostic accuracy [81,82].

Conclusion

Although AI models demonstrate promising accuracy in diagnosing pelvic gynecologic tumors on ultrasound, most studies suffer from high risk of bias. In ovarian tumors, the well-established ADNEX model still appears to outperform most AI-based approaches (both deep learning and radiomics-based machine learning). Future studies should emphasize robust study designs, external validation, performance evaluation across all thresholds, proper calibration, and standardized reporting – essential steps toward achieving successful clinical integration.

**Acknowledgements**

The authors wish to thank Thomas Vandendriessche, Chayenne Van Meel, Norin Hamouda and Krizia Tuand, the biomedical reference librarians of the KU Leuven Libraries – 2Bergen – learning Centre Désiré Collen (Leuven, Belgium), for their help in conducting the systematic literature search.

**Statement of ethics**

A statement of ethics is not applicable because this study is based exclusively on published literature.

**Conflict of interest statement**

BVC, DT and WF serve as members of the steering committee of the International Ovarian Tumor Analysis (IOTA) consortium, with BVC and DT being involved in the development of the ADNEX model. All other authors have no conflicts of interests.

**Funding sources**

This work is part of the IOTA7 study which is supported by the FWO (projects G049312N), IWT (project IWT-TBM 070706), KU Leuven Internal Funds (BOF 2015: C24/15/037) and UZ Leuven Internal Funds (KOOR). BDM is supported by KU Leuven Internal Funds (projects iBOF/23/064,



Downloaded from http://karger.com/goi/article-pdf/doi/10.1159/000545850/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025

C3/20/117, C3I/21/00316), Industrial Research Fund and several Leuven Research and Development bilateral industrial projects, Flemish Government Agencies: FWO: SBO project S005319N, TBM Project T001919N, and FWO PhD Grant 11K5623N, EWI: the Flanders AI Research Program, VLAIO: CSBO (HBC.2021.0076) and Baekeland PhD (HBC.20192204), EU: European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant agreement No 885682). DT is a Senior Clinical Investigator of the FWO. WF is funded by KU Leuven Internal Funds (C3I/21/00316) and has a postdoctoral mandate of the Foundation against Cancer (2023-031). BVC is funded by the Research Foundation – Flanders (FWO: grant G097322N), Internal Funds KU Leuven (grant C24M/20/064), and University Hospitals Leuven (grant COPREDICT).

**Author contributions**

AG, GG and ST contributed to the design of the study methodology. GG and ST contributed to the search strategy and initial article screening. AG, GG and ST contributed to the data extraction and risk of bias evaluation. AG, GG, LB, and BVC contributed to data analysis and interpretation. AG and GG contributed to the writing of the manuscript while ST, LB, DT, WF, BVC, and BDM contributed to the critical revision of the manuscript. All authors approved the final version of the manuscript. The corresponding author attests that all listed authors meet authorship criteria and that no other meeting the criteria have been omitted.

**Data availability statement**

All data generated or analyzed during this study are included in this article and its supplementary material files. Further enquiries can be directed to the corresponding author.

Downloaded from http://karger.com/goi/article-pdf/doi/10.1159/000545850/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025

**References**


1 Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, Rosso S, Coebergh JWW, Comber H, et al. Cancer incidence and mortality patterns in Europe: Estimates for 40 countries in 2012. Eur J Cancer. 2013;49(6):1374–403.

2 Heintz A, Odicino F, Maissonneuve P, Quinn M, Benedet J, Creasman W, et al. Carcinoma of the Ovary. International Journal of Gynecology & Obstetrics. 2006;95(S1):S161–92.

3 Van Holsbeke C, Daemen A, Yazbek J, Holland TK, Bourne T, Mesens T, et al. Ultrasound experience substantially impacts on diagnostic performance and confidence when adnexal masses are classified using pattern recognition. Gynecol Obstet Invest. 2010;69(3):160–8.

4 Van Calster B, Van Hoorde K, Valentin L, Testa AC, Fischerova D, Van Holsbeke C, et al. Evaluating the risk of ovarian cancer before surgery using the ADNEX model to differentiate between benign, borderline, early and advanced stage invasive, and secondary metastatic tumours: Prospective multicentre diagnostic study. BMJ. 2014;349:g5920.

5 Barreñada L, Ledger A, Dhiman P, Collins G, Wynants L, Verbakel JY, et al. ADNEX risk prediction model for diagnosis of ovarian cancer: systematic review and meta-analysis of external validation studies. BMJ Medicine. 2024;3(1):e000817.

6 Sladkevicius P, Valentin L. Interobserver agreement in describing the ultrasound appearance of adnexal masses and in calculating the risk of malignancy using logistic regression models. Clinical Cancer Research. 2015;21(3):594–601.

7 Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88.

8 Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022;28(1):31–8.

9 Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med. 2020;3(1):118.

10 Grigore M, Popovici R, Gafitanu D, Himiniuc L, Murarasu M, Micu R. Logistic models and artificial intelligence in the sonographic assessment of adnexal masses-a systematic review of the literature. Med Ultrason. 2020;22(4):469–75.

11 Akazawa M, Hashimoto K. Artificial intelligence in gynecologic cancers: Current status and future challenges – A systematic review. Artif Intell Med. 2021;120:102164.

12 Shrestha P, Poudyal B, Yadollahi S, E. Wright D, V. Gregory A, D. Warner J, 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.

13 Xu H-L, Gong T-T, Liu F-H, Chen H-Y, Xiao Q, Hou Y, et al. Artificial intelligence performance in image-based ovarian cancer identification: A systematic review and meta-analysis. EClinicalMedicine. 2022;53:101662.

14 Ma L, Huang L, Chen Y, Zhang L, Nie D, He W, et al. AI diagnostic performance based on multiple imaging modalities for ovarian tumor: A systematic review and meta-analysis. Front Oncol. 2023;13. DOI: 10.3389/fonc.2023.1133491

15 Adusumilli P, Ravikumar N, Hall G, Swift S, Orsi N, Scarsbrook A. Radiomics in the evaluation of ovarian masses — a systematic review. Insights Imaging. 2023;14(1):165.

16 Koch AH, Jeelof LS, Muntinga CLP, Gootzen TA, van de Kruis NMA, Nederend J, et al. Analysis of computer-aided diagnostics in the preoperative diagnosis of ovarian cancer: a systematic review. Insights Imaging. 2023;14(1):34.

17 Sadeghi MH, Sina S, Omidi H, Farshchitabrizi AH, Alavi M. Deep learning in ovarian cancer diagnosis: a comprehensive review of various imaging modalities. Pol J Radiol. 2024;89:30–48.

18 Mitchell S, Nikolopoulos M, El-Zarka A, Al-Karawi D, Al-Zaidi S, Ghai A, et al. Artificial Intelligence in Ultrasound Diagnoses of Ovarian Cancer: A Systematic Review and Meta-Analysis. Cancers (Basel). 2024;16(2). DOI: 10.3390/cancers16020422




Downloaded from http://karger.com/goi/article-pdf/doi/10.1159/000545850/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025

19 Moro F, Ciancia M, Zace D, Vagni M, Tran HE, Giudice MT, et al. Role of artificial intelligence applied to ultrasound in gynecology oncology: A systematic review. Int J Cancer. 2024;155(10):1832–45.
20 Peters MDJ, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth. 2020;18(10):2119–26.
21 Wohlin C. Guidelines for snowballing in systematic literature studies and a replication in software engineering. Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering (EASE '14). Association for Computing Machinery; 2014; pp 1–10.
22 Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev. 2016;5(1):210.
23 Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. Ann Intern Med. 2011;155(8):529–36.
24 Sounderajah V, Ashrafian H, Rose S, Shah NH, Ghassemi M, Golub R, et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. Nat Med. 2021;27(10):1663–5.
25 Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open. 2021;11(7):e048008.
26 Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51–8.
27 Boutron I, Page MJ, Higgins JP, Altman DG, Lundh A, Hróbjartsson A, et al. Chapter 7: Considering bias and conflicts of interest among the included studies. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al., editors. Cochrane Handbook for Systematic Reviews of Interventions. 2019; pp 177–204.
28 Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. PLoS Med. 2014;11(10):e1001744.
29 Konstantopoulos S. Fixed effects and variance components estimation in three-level meta-analysis. Res Synth Methods. 2011;2(1):61–76.
30 Van den Noortgate W, López-López JA, Marín-Martínez F, Sánchez-Meca J. Three-level meta-analysis of dependent effect sizes. Behav Res Methods. 2013;45(2):576–94.
31 Viechtbauer W. Conducting Meta-Analyses in R with the metafor Package. J Stat Softw. 2010;36(3):1–48.
32 Acharya UR, Sree S. V, Saba L, Molinari F, Guerriero S, Suri JS. Ovarian tumor characterization and classification: A class of GyneScan™ systems. 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE; 2012; pp 4446–9.
33 Acharya UR, Sree SV, Saba L, Molinari F, Guerriero S, Suri JS. Ovarian tumor characterization and classification using ultrasound - A new online paradigm. J Digit Imaging. 2013;26(3):544–53.
34 Acharya UR, Sree SV, Krishnan MMR, Saba L, Molinari F, Guerriero S, et al. Ovarian tumor characterization using 3D ultrasound. Technol Cancer Res Treat. 2012;11(6):543–52.
35 Acharya UR, Sree SV, Kulshreshtha S, Molinari F, Koh JEW, Saba L, et al. GyneScan: An improved online paradigm for screening of ovarian cancer via tissue characterization. Technol Cancer Res Treat. 2014;13(6):529–39.
36 Acharya UR, Mookiah MRK, Vinitha Sree S, Yanti R, Martis RJ, Saba L, et al. Evolutionary algorithm-based classifier parameter tuning for automatic ovarian cancer tissue characterization and classification. Ultraschall in der Medizin. 2014;35(3):237–45.

Downloaded from http://karger.com/goi/article-pdf/doi/10.1159/000545850/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025

37 Acharya UR, Akter A, Chowriappa P, Dua S, Raghavendra U, Koh JEW, et al. Use of Nonlinear Features for Automated Characterization of Suspicious Ovarian Tumors Using Ultrasound Images in Fuzzy Forest Framework. International Journal of Fuzzy Systems. 2018;20(4):1385–402.
38 Al-karawi D, Al-Assam H, Du H, Sayasneh A, Landolfo C, Timmerman D, et al. An Evaluation of the Effectiveness of Image-based Texture Features Extracted from Static B-mode Ultrasound Images in Distinguishing between Benign and Malignant Ovarian Masses. Ultrason Imaging. 2021 May;43(3):124–38.
39 Aramendía-Vidaurreta V, Cabeza R, Villanueva A, Navallas J, Alcázar JL. Ultrasound Image Discrimination between Benign and Malignant Adnexal Masses Based on a Neural Network Approach. Ultrasound Med Biol. 2016;42(3):742–52.
40 Barcroft JF, Linton-Reid K, Landolfo C, Al-Memar M, Parker N, Kyriacou C, et al. Machine learning and radiomics for segmentation and classification of adnexal masses on ultrasound. NPJ Precis Oncol. 2024;8(1):41.
41 Wu C, Wang Y, Wang F. Deep learning for ovarian tumor classification with ultrasound images. Advances in Multimedia Information Processing – PCM 2018. Springer; 2018; pp 395–406.
42 Gao Y, Zeng S, Xu X, Li H, Yao S, Song K, et al. Deep learning-enabled pelvic ultrasound images for accurate diagnosis of ovarian cancer in China: a retrospective, multicentre, diagnostic study. Lancet Digit Health. 2022;4(3):e179–87.
43 Du Y, Guo W, Xiao Y, Chen H, Yao J, Wu J. Ultrasound-based deep learning radiomics model for differentiating benign, borderline, and malignant ovarian tumours: a multi-class classification exploratory study. BMC Med Imaging. 2024;24(1):89.
44 Wu M, Cui G, Lv S, Chen L, Tian Z, Yang M, et al. Deep convolutional neural networks for multiple histologic types of ovarian tumors classification in ultrasound images. Front Oncol. 2023;13. DOI: 10.3389/fonc.2023.1154200
45 Xi M, Zheng R, Wang M, Shi X, Chen C, Qian J, et al. Ultrasonographic diagnosis of ovarian tumors through the deep convolutional neural network. Ginekol Pol. 2024;95(3):181–9.
46 Hsu ST, Su YJ, Hung CH, Chen MJ, Lu CH, Kuo CE. Automatic ovarian tumors recognition system based on ensemble convolutional neural network with ultrasound imaging. BMC Med Inform Decis Mak. 2022;22(1):298.
47 Du Y, Xiao Y, Guo W, Yao J, Lan T, Li S, et al. Development and validation of an ultrasound-based deep learning radiomics nomogram for predicting the malignant risk of ovarian tumours. Biomed Eng Online. 2024;23(1):41.
48 Qi L, Chen D, Li C, Li J, Wang J, Zhang C, et al. Diagnosis of Ovarian Neoplasms Using Nomogram in Combination With Ultrasound Image-Based Radiomics Signature and Clinical Factors. Front Genet. 2021;12. DOI: 10.3389/fgene.2021.753948
49 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). DOI: 10.3390/cancers14215291
50 Chen H, Yang BW, Qian L, Meng YS, Bai XH, Hong XW, et al. Deep Learning Prediction of Ovarian Malignancy at US Compared with O-RADS and Expert Assessment. Radiology. 2022;304(1):106–13.
51 Xiang H, Xiao Y, Li F, Li C, Liu L, Deng T, et al. Development and validation of an interpretable model integrating multimodal information for improving ovarian cancer diagnosis. Nat Commun. 2024;15(1):2681.
52 Chiappa V, Bogani G, Interlenghi M, Salvatore C, Bertolina F, Sarpietro G, et al. The Adoption of Radiomics and machine learning improves the diagnostic processes of women with Ovarian MAsses (the AROMA pilot study). J Ultrasound. 2021;24(4):429–37.
53 Chiappa V, Interlenghi M, Bogani G, Salvatore C, Bertolina F, Sarpietro G, et al. A decision support system based on radiomics and machine learning to predict the risk of malignancy of ovarian masses from transvaginal ultrasonography and serum CA-125. Eur Radiol Exp. 2021;5(1):28.

Downloaded from http://karger.com/goi/article-pdf/doi/10.1159/000545850/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025

54 Christiansen F, Epstein EL, Smedberg E, Åkerlund M, Smith K, Epstein E. Ultrasound image analysis using deep neural networks for discriminating between benign and malignant ovarian tumors: comparison with expert subjective assessment. Ultrasound in Obstetrics & Gynecology. 2021;57(1):155–63.
55 Deeparani M, Kalamani M. Gynecological Healthcare: Unveiling Pelvic Masses Classification through Evolutionary Gravitational Neocognitron Neural Network Optimized with Nomadic People Optimizer. Diagnostics. 2023;13(19):3131.
56 Faschingbauer F, Beckmann MW, Weyert Goecke T, Renner S, Häberle L, Benz M, et al. Automatic texture-based analysis in ultrasound imaging of ovarian masses. Ultraschall in der Medizin. 2013;34(2):145–50.
57 Hussein IJ, Burhanuddin MA, Mohammed MA, Benameur N, Maashi MS, Maashi MS. Fully-automatic identification of gynaecological abnormality using a new adaptive frequency filter and histogram of oriented gradients (HOG). Expert Syst. 2022 Mar;39(3):e12789.
58 Jung Y, Kim T, Han MR, Kim S, Kim G, Lee S, et al. Ovarian tumor diagnosis using deep convolutional neural networks and a denoising convolutional autoencoder. Sci Rep. 2022;12(1):17024.
59 Khazendar S, Sayasneh A, Al-Assam H, Du H. Automated characterisation of ultrasound images of ovarian tumours: the diagnostic accuracy of a support vector machine and image processing with a local binary pattern operator. Facts Views Vis Obgyn. 2015;7(1):7–15.
60 Martínez-Más J, Bueno-Crespo A, Khazendar S, Remezal-Solano M, Martínez-Cendán JP, Jassim S, et al. Evaluation of machine learning methods with Fourier Transform features for classifying ovarian tumors based on ultrasound images. PLoS One. 2019 Jul;14(7):e0219388.
61 Kongara SR, Prakasha S, Brindha A, Pathak SK, Miya J, Taqui SN, et al. Performance evaluation of optimized convolutional neural network mechanism in the detection and classification of ovarian cancer. Multimed Tools Appl. 2024;83(28):71311–34.
62 Lucidarme O, Akakpo JP, Granberg S, Sideri M, Levavi H, Schneider A, et al. A new computer-aided diagnostic tool for non-invasive characterisation of malignant ovarian masses: Results of a multicentre validation study. Eur Radiol. 2010;20(8):1822–30.
63 Miao K, Zhao N, Lv Q, He X, Xu M, Dong X, et al. Prediction of benign and malignant ovarian tumors using Resnet34 on ultrasound images. Journal of Obstetrics and Gynaecology Research. 2023;49(12):2910–7.
64 Moro F, Vagni M, Tran HE, Bernardini F, Mascilini F, Ciccarone F, et al. Radiomics analysis of ultrasound images to discriminate between benign and malignant adnexal masses with solid ultrasound morphology. Ultrasound in Obstetrics & Gynecology. 2024 DOI: 10.1002/uog.27680
65 Pathak H, Kulkarni V. Identification of ovarian mass through ultrasound images using machine learning techniques. 2015 IEEE International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN). IEEE; 2015; pp 137–40.
66 Pham T-L, Le V-H. Ovarian Tumors Detection and Classification from Ultrasound Images Based on YOLOv8. Journal of Advances in Information Technology. 2024;15(2):264–75.
67 Raja P, Suresh P. Variety of ovarian cysts detection and classification using 2D Convolutional Neural Network. Multimed Tools Appl. 2024;83(16):49473–91.
68 Sheela S, Sumathi M. An Evaluation of Effectiveness of a Texture Feature Based Computerized Diagnostic Model in Classifying the Ovarian Cyst as Benign and Malignant from Static 2D B-Mode Ultrasound Images. Curr Med Imaging. 2023;19(3):292–305.
69 Ștefan PA, Lupean RA, Mihu CM, Lebovici A, Oancea MD, Hîțu L, et al. Ultrasonography in the diagnosis of adnexal lesions: The role of texture analysis. Diagnostics. 2021;11(5). DOI: 10.3390/diagnostics11050812

Downloaded from http://karger.com/goi/article-pdf/doi/10.1159/000545850/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025

70 Wang H, Liu C, Zhao Z, Zhang C, Wang X, Li H, et al. Application of Deep Convolutional Neural Networks for Discriminating Benign, Borderline, and Malignant Serous Ovarian Tumors From Ultrasound Images. Front Oncol. 2021;11. DOI: 10.3389/fonc.2021.770683
71 Zimmer Y, Tepper R, Akselrod S. An automatic approach for morphological analysis and malignancy evaluation of ovarian masses using B-scans. Ultrasound Med Biol. 2003;29(11):1561–70.
72 Chiappa V, Interlenghi M, Salvatore C, Bertolina F, Bogani G, Ditto A, et al. Using rADioMIcs and machine learning with ultrasonography for the differential diagnosis of myometRiAL tumors (the ADMIRAL pilot study). Radiomics and differential diagnosis of myometrial tumors. Gynecol Oncol. 2021;161(3):838–44.
73 Michail G, Karahaliou A, Skiadopoulos S, Kalogeropoulou C, Terzis G, Boniatis I, et al. Texture analysis of perimenopausal and post-menopausal endometrial tissue in grayscale transvaginal ultrasonography. British Journal of Radiology. 2007;80(956):609–16.
74 Wang X, Zhang R. Clinical Value Analysis of Combined Vaginal Ultrasound, Magnetic Resonance Dispersion Weighted Imaging, and Multilayer Spiral CT in the Diagnosis of Endometrial Cancer Using Deep VGG-16 AdaBoost Hybrid Classifier. J Oncol. 2022;2022:7677004.
75 Fang Y, Wei Y, Liu X, Qin L, Gao Y, Yu Z, et al. A self-supervised classification model for endometrial diseases. J Cancer Res Clin Oncol. 2023;149(20):17855–63.
76 Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12–22.
77 Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.
78 Riley RD, Archer L, Snell KIE, Ensor J, Dhiman P, Martin GP, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ. 2024;384:e074820.
79 Saporta A, Gui X, Agrawal A, Pareek A, Truong SQH, Nguyen CDT, et al. Benchmarking saliency methods for chest X-ray interpretation. Nat Mach Intell. 2022;4(10):867–78.
80 Huang X, Marques-Silva J. On the failings of Shapley values for explainability. International Journal of Approximate Reasoning. 2024;171:109112.
81 Sladkevicius P, Jokubkiene L, Timmerman D, Fischerova D, Van Holsbeke C, Franchi D, et al. Vessel morphology depicted by three-dimensional power Doppler ultrasound as second-stage test in adnexal tumors that are difficult to classify: prospective diagnostic accuracy study. Ultrasound in Obstetrics & Gynecology. 2021;57(2):324–34.
82 Xydias EM, Kalantzi S, Tsakos E, Ntanika A, Beis N, Prior M, et al. Comparison of 3D ultrasound, 2D ultrasound and 3D Doppler in the diagnosis of endometrial carcinoma in patients with uterine bleeding: A systematic review and meta-analysis. European Journal of Obstetrics and Gynecology and Reproductive Biology. 2022;277:42–52.

Downloaded from http://karger.com/goi/article-pdf/doi/10.1159/000545850/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025

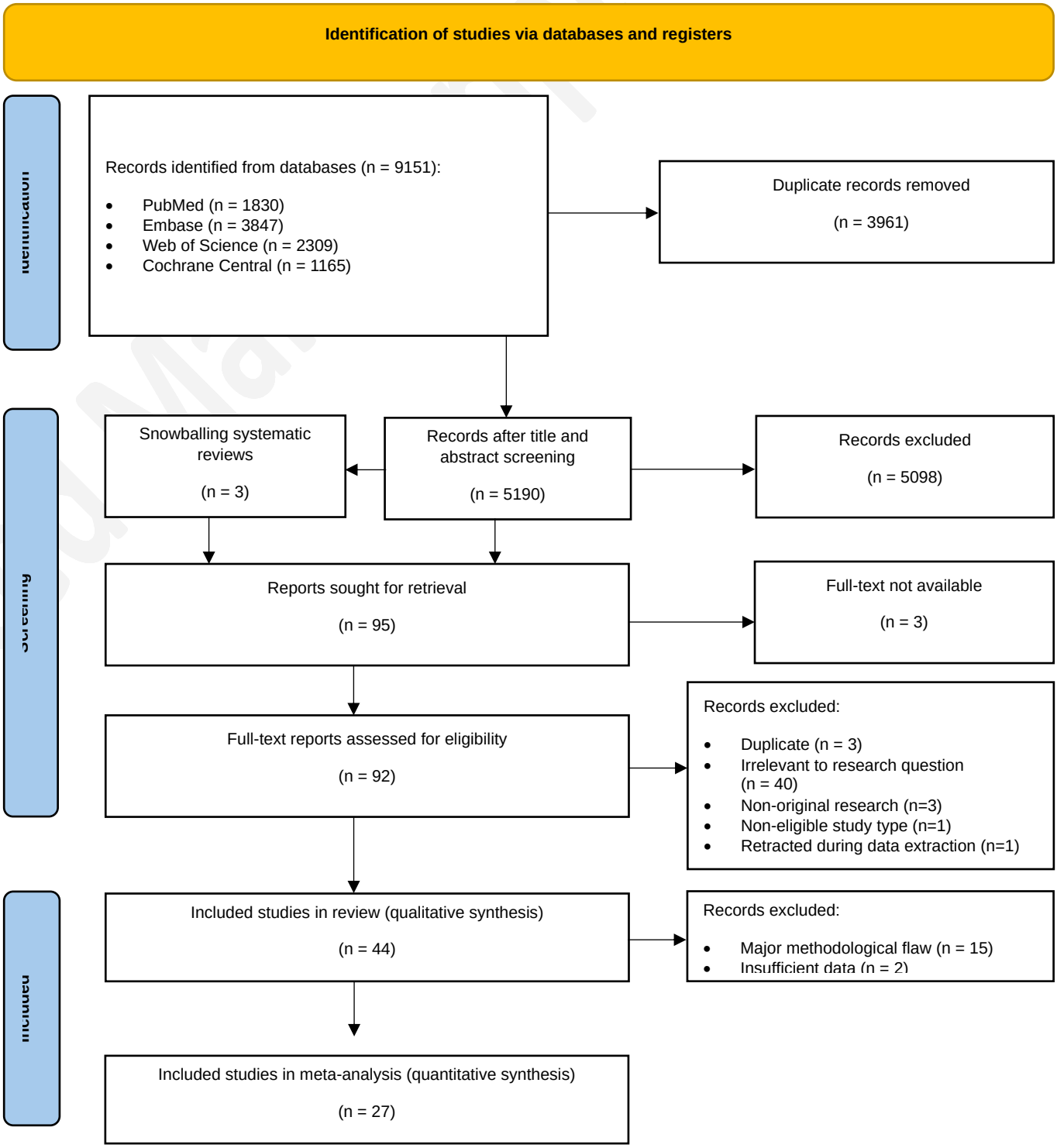
**Figure legends**

**Figure 1** PRISMA Flowchart of the study selection.

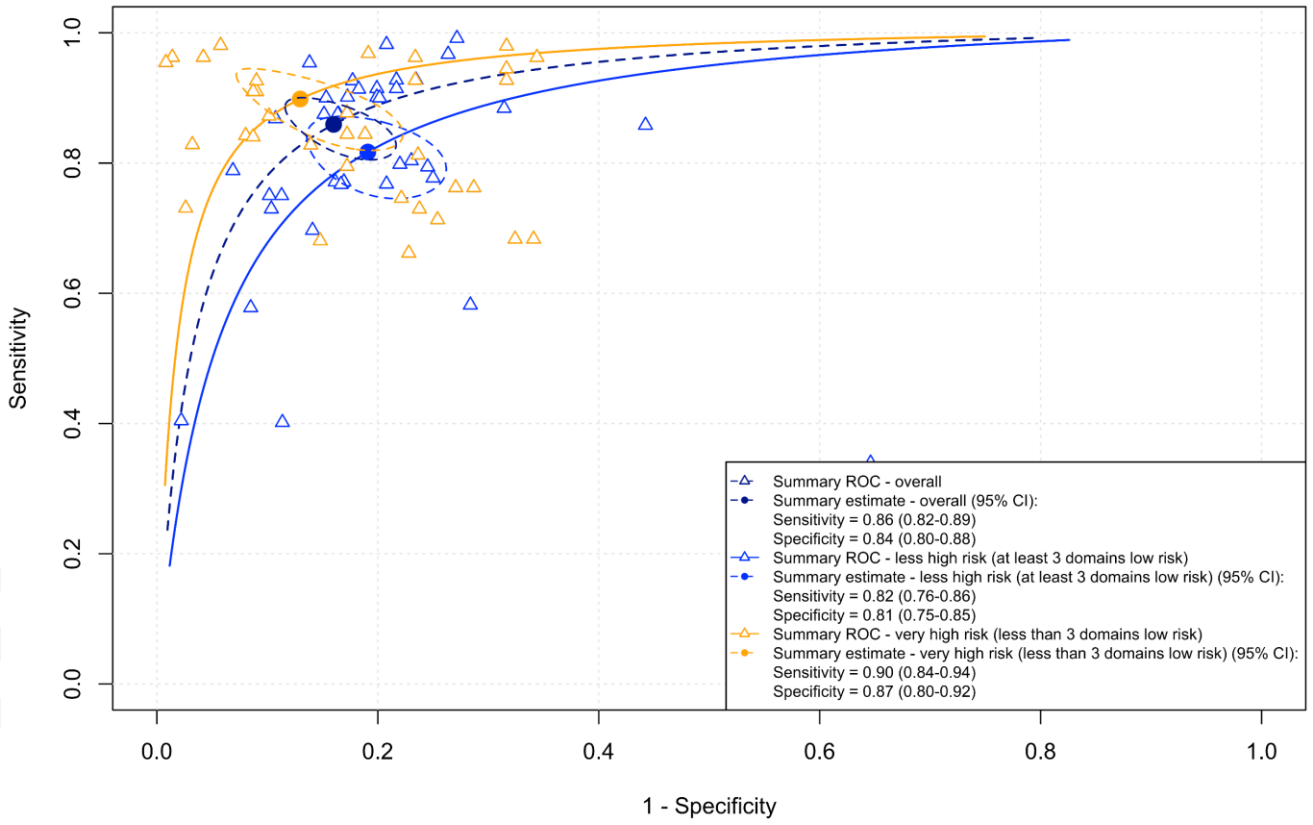
**Figure 2** Summary ROC graph with 95% CI stratified by the risk of bias based on the adapted QUADAS-2 assessment, spanning 40 models from 14 ovarian cancer studies with "less high risk" (at least 3 low risk domains) and 36 models from 11 ovarian cancer studies with "very high risk" (less than 3 low risk domains). *Abbreviations*: ADNEX: Assessment of Different NEoplasias in the adneXa; CA-125: Cancer Antigen-125; CI: Confidence Interval; ROC: Receiver Operating Characteristic Curve

**Figure 3** Forest plot of reported AUC with 95% CI stratified by the risk of bias based on the adapted QUADAS-2 assessment, spanning 34 models from 11 ovarian cancer studies with "less high risk" (at least 3 low risk domains) and 16 models from 4 ovarian cancer studies with "very high risk" (less than 3 low risk domains). A descriptive comparison with ADNEX on external validation – both with and without CA-125 – is also shown. *Abbreviations*: ADNEX: Assessment of Different NEoplasias in the adneXa; AUC: Area Under the Receiver Operating Characteristic Curve; CA-125: Cancer Antigen-125; CI: Confidence Interval; ROC: Receiver Operating Characteristic Curve

Downloaded from http://karger.com/gol/article-pdf/doi/10.1159/000545850/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025

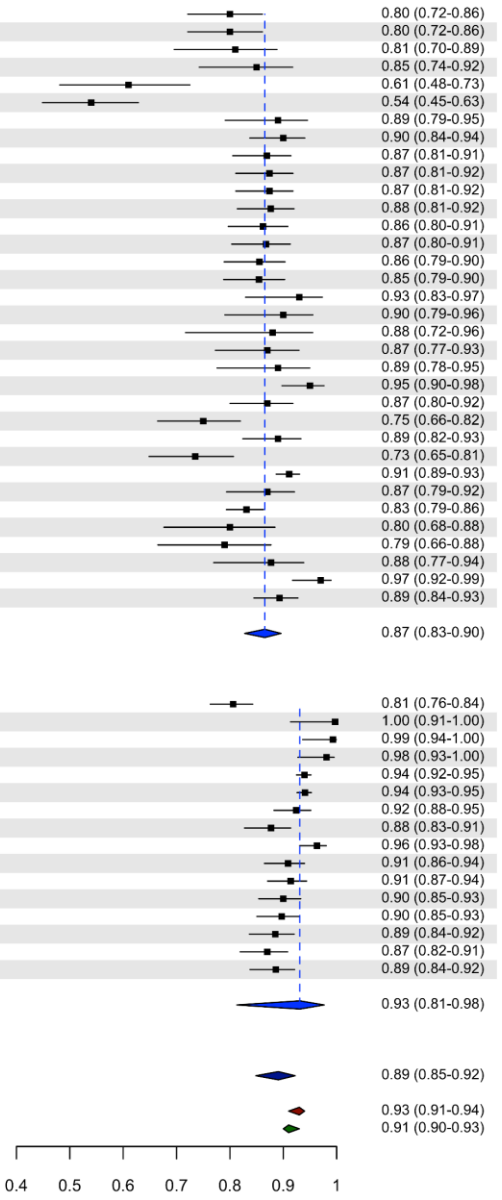



Downloaded from http://karger.com/gol/article-pdf/doi/10.1159/000545850/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025




Downloaded from http://karger.com/gol/article-pdf/doi/10.1159/000545850/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025

| Model | Author group (cluster) | TP | FP | FN | TN | AUC (95% CI) |
|---|---|---|---|---|---|---|
| **Less high risk (at least 3 domains low risk)** | | | | | | |
| Barcroft et al. (2024) - Model 1 | Barcroft et al. | 45 | 20 | 13 | 106 | 0.80 (0.72-0.86) |
| Barcroft et al. (2024) - Model 2 | Barcroft et al. | 45 | 21 | 13 | 105 | 0.80 (0.72-0.86) |
| Barcroft et al. (2024) - Model 3 | Barcroft et al. | 22 | 20 | 5 | 68 | 0.81 (0.70-0.89) |
| Barcroft et al. (2024) - Model 4 | Barcroft et al. | 21 | 18 | 6 | 70 | 0.85 (0.74-0.92) |
| Barcroft et al. (2024) - Model 5 | Barcroft et al. | 9 | 57 | 18 | 31 | 0.61 (0.48-0.73) |
| Barcroft et al. (2024) - Model 6 | Barcroft et al. | 17 | 91 | 41 | 35 | 0.54 (0.45-0.63) |
| Barcroft et al. (2024) - Model 7 | Barcroft et al. | 27 | 18 | 0 | 70 | 0.89 (0.79-0.95) |
| Barcroft et al. (2024) - Model 8 | Barcroft et al. | 58 | 34 | 0 | 92 | 0.90 (0.84-0.94) |
| Martínez-Más et al. (2019) - Model 1 | Buckingham-Imperial-KU Leuven | 70 | 24 | 5 | 88 | 0.87 (0.81-0.91) |
| Martínez-Más et al. (2019) - Model 2 | Buckingham-Imperial-KU Leuven | 68 | 19 | 7 | 93 | 0.87 (0.81-0.92) |
| Martínez-Más et al. (2019) - Model 3 | Buckingham-Imperial-KU Leuven | 69 | 22 | 6 | 90 | 0.87 (0.81-0.92) |
| Martínez-Más et al. (2019) - Model 4 | Buckingham-Imperial-KU Leuven | 70 | 26 | 5 | 86 | 0.88 (0.81-0.92) |
| Martínez-Más et al. (2019) - Model 5 | Buckingham-Imperial-KU Leuven | 69 | 24 | 6 | 88 | 0.86 (0.80-0.91) |
| Martínez-Más et al. (2019) - Model 6 | Buckingham-Imperial-KU Leuven | 68 | 22 | 7 | 90 | 0.87 (0.80-0.91) |
| Martínez-Más et al. (2019) - Model 7 | Buckingham-Imperial-KU Leuven | 68 | 22 | 7 | 90 | 0.86 (0.79-0.90) |
| Martínez-Más et al. (2019) - Model 8 | Buckingham-Imperial-KU Leuven | 66 | 18 | 9 | 94 | 0.85 (0.79-0.90) |
| Chen et al. (2022) - Model 1 | Chen et al. | 22 | 9 | 2 | 52 | 0.93 (0.83-0.97) |
| Chen et al. (2022) - Model 2 | Chen et al. | 22 | 12 | 2 | 49 | 0.90 (0.79-0.96) |
| Chiappa et al. (2021A) - Model 1 | Chiappa et al. | 10 | 5 | 3 | 48 | 0.88 (0.72-0.96) |
| Chiappa et al. (2021A) - Model 2 | Chiappa et al. | 44 | 6 | 13 | 32 | 0.87 (0.77-0.93) |
| Chiappa et al. (2021A) - Model 3 | Chiappa et al. | 45 | 5 | 11 | 19 | 0.89 (0.78-0.95) |
| Christiansen et al. (2021) - Model 1 | Christiansen et al. | 72 | 10 | 3 | 65 | 0.95 (0.90-0.98) |
| Du et al. (2024A) - Model 1 | Du et al. | 42 | 15 | 18 | 94 | 0.87 (0.80-0.92) |
| Du et al. (2024A) - Model 2 | Du et al. | 24 | 12 | 36 | 97 | 0.75 (0.66-0.82) |
| Du et al. (2024B) - Model 1 | Du et al. | 44 | 11 | 16 | 99 | 0.89 (0.82-0.93) |
| Du et al. (2024B) - Model 2 | Du et al. | 35 | 31 | 25 | 79 | 0.73 (0.65-0.81) |
| Gao et al. (2022) - Model 1 | Gao et al. | 210 | 41 | 56 | 561 | 0.91 (0.89-0.93) |
| Gao et al. (2022) - Model 2 | Gao et al. | 27 | 1 | 40 | 66 | 0.87 (0.79-0.92) |
| Gao et al. (2022) - Model 3 | Gao et al. | 96 | 61 | 70 | 662 | 0.83 (0.79-0.86) |
| Moro et al. (2024) - Model 1 | Moro et al. | 57 | 6 | 16 | 19 | 0.80 (0.68-0.88) |
| Moro et al. (2024) - Model 2 | Moro et al. | 63 | 11 | 10 | 14 | 0.79 (0.66-0.88) |
| Qi et al. (2021) - Model 1 | Tianjin | 40 | 3 | 13 | 27 | 0.88 (0.77-0.94) |
| Xiang et al. (2024) - Model 1 | Xiang et al. | 73 | 14 | 2 | 40 | 0.97 (0.92-0.99) |
| Xiang et al. (2024) - Model 2 | Xiang et al. | 72 | 96 | 9 | 210 | 0.89 (0.84-0.93) |
| **RE model - subgroup** ($I^2_{total}$ = 81.9%, $I^2_{between}$ = 32.0%, $\tau^2_{between}$ = 0.113, $I^2_{within}$ = 49.9%, $\tau^2_{within}$ = 0.176) | | | | | | 0.87 (0.83-0.90) |
| **Very high risk (less than 3 domains low risk)** | | | | | | |
| Acharya et al. (2018) - Model 1 | Acharya et al. | 188 | 56 | 43 | 182 | 0.81 (0.76-0.84) |
| Aramendia-Vidaurreta et al. (2016) - Model 1 | Aramendia-Vidaurreta et al. | 38 | 1 | 1 | 105 | 1.00 (0.91-1.00) |
| Aramendia-Vidaurreta et al. (2016) - Model 2 | Aramendia-Vidaurreta et al. | 38 | 4 | 1 | 102 | 0.99 (0.94-1.00) |
| Aramendia-Vidaurreta et al. (2016) - Model 3 | Aramendia-Vidaurreta et al. | 38 | 4 | 1 | 102 | 0.98 (0.93-1.00) |
| Jung et al. (2022) - Model 1 | Jung et al. | 451 | 66 | 84 | 760 | 0.94 (0.92-0.95) |
| Jung et al. (2022) - Model 2 | Jung et al. | 467 | 85 | 68 | 756 | 0.94 (0.93-0.95) |
| Wang H. et al. (2021) - Model 1 | Tianjin | 142 | 3 | 29 | 105 | 0.92 (0.88-0.95) |
| Wang H. et al. (2021) - Model 2 | Tianjin | 168 | 34 | 3 | 74 | 0.88 (0.83-0.91) |
| Wang H. et al. (2021) - Model 3 | Tianjin | 156 | 9 | 15 | 99 | 0.96 (0.93-0.98) |
| Wang H. et al. (2021) - Model 4 | Tianjin | 165 | 25 | 6 | 83 | 0.91 (0.86-0.94) |
| Wang H. et al. (2021) - Model 5 | Tianjin | 144 | 9 | 27 | 99 | 0.91 (0.87-0.94) |
| Wang H. et al. (2021) - Model 6 | Tianjin | 165 | 37 | 6 | 71 | 0.90 (0.85-0.93) |
| Wang H. et al. (2021) - Model 7 | Tianjin | 159 | 25 | 12 | 83 | 0.90 (0.85-0.93) |
| Wang H. et al. (2021) - Model 8 | Tianjin | 159 | 25 | 12 | 83 | 0.89 (0.84-0.92) |
| Wang H. et al. (2021) - Model 9 | Tianjin | 162 | 34 | 9 | 74 | 0.87 (0.82-0.91) |
| Wang H. et al. (2021) - Model 10 | Tianjin | 159 | 34 | 12 | 74 | 0.89 (0.84-0.92) |
| **RE model - subgroup** ($I^2_{total}$ = 96.8%, $I^2_{between}$ = 96.8%, $\tau^2_{between}$ = 1.260, $I^2_{within}$ = 0.0%, $\tau^2_{within}$ = 0.000) | | | | | | 0.93 (0.81-0.98) |
| Test for subgroup differences: $Q_M$=144.96, df=1, p < 0.01 | | | | | | |
| **RE model - overall** ($I^2_{total}$ = 90.3%, $I^2_{between}$ = 67.2%, $\tau^2_{between}$ = 0.373, $I^2_{within}$ = 23.0%, $\tau^2_{within}$ = 0.128) | | | | | | 0.89 (0.85-0.92) |
| ADNEX with CA-125 | | | | | | 0.93 (0.91-0.94) |
| ADNEX without CA-125 | | | | | | 0.91 (0.90-0.93) |




Downloaded from http://karger.com/gol/article-pdf/doi/10.1159/000545850/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025

**Table 1** Study variables extracted from the included studies, categorized by anatomical tumor type.

| Author (year) | Objective | Patients | Retrospective/ Prospective | Reference standard | Modality | Mono-/ Multicenter | Single-/ Multi-vendor scanning | Experience level scanner(s) | Anonymization |
|---|---|---|---|---|---|---|---|---|---|
| **Ovarian cancer** | | | | | | | | | |
| Acharya et al. (2012)[32] | Diagnosis (benign vs malignant) | 20 | NR | Histology | 2D from 3D | NR | Single-vendor (GE) | NR | NR |
| Acharya et al. (2012)[34] | Diagnosis (benign vs malignant) | 20 | NR | Histology | 2D from 3D | NR | Single-vendor (GE) | NR | NR |
| Acharya et al. (2013)[33] | Diagnosis (benign vs malignant) | 20 | NR | Histology | 2D from 3D | NR | Single-vendor (GE) | NR | NR |
| Acharya et al. (2014)[36] | Diagnosis (benign vs malignant) | 20 | NR | Histology | 2D from 3D | NR | Single-vendor (GE) | NR | NR |
| Acharya et al. (2014)[35] | Diagnosis (benign vs malignant) | 20 | NR | Histology | 2D from 3D | NR | Single-vendor (GE) | NR | NR |
| Acharya et al. (2018)[37] | Diagnosis (suspicious vs non-suspicious) | 469 | NR | SA | 2D from 3D | NR | NR | NR | NR |
| Al-Karawi et al. (2021)[38] | Diagnosis (benign vs malignant) | 232 | Retrospective | Histology | 2D | Multicenter (IOTA) | NR | Experienced radiologists | NR |
| Aramendia-Vidaurreta et al. (2016)[39] | Diagnosis (benign vs malignant) | 145 | NR | Histology | 2D | Monocenter | Single-vendor (GE) | Examiner with >12 years of experience | NR |
| Barcroft et al. (2024)[40] | Diagnosis (benign vs malignant) | 577 | Retrospective | Histology, SA | 2D | Multicenter (one center for development, other center for external validation) | Multi-vendor (GE, Samsung, Esaote) | Examiners with varying experience (under expert supervision) | Yes |
| Chen et al. (2022)[50] | Diagnosis (benign vs malignant) | 422 | Retrospective | Histology | 2D | Monocenter | Multi-vendor (GE, Philips) | Expert with 14 of ultrasound experience | NR |
| Chiappa et al. (2021)[52] | Diagnosis (benign vs malignant for three groups of ovarian masses) | 241 | Retrospective | Histology | 2D | Monocenter | Single-vendor (GE) | NR | NR |
| Chiappa et al. (2021)[53] | Diagnosis (benign vs malignant) | 274 | 1st cohort: Retrospective 2nd cohort: Prospective | Histology | 2D | Monocenter | Multi-vendor (retrospective part: GE, Prospective part: Samsung) | Gynecologist with >10 years experience | NR |

Accepted Manuscript

Downloaded from http://karger.com/gol/article-pdf/doi/10.1159/000545850/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025

**Table 1** (continued)

| Author (year) | Objective | Patients | Retrospective/ Prospective | Reference standard | Modality | Mono-/ Multicenter | Single-/ Multi-vendor scanning | Experience level scanner(s) | Anonymization |
|---|---|---|---|---|---|---|---|---|---|
| Christiansen et al. (2021)[54] | Diagnosis (benign vs malignant) | 758 | Retrospective | Histology, SA | 2D | Multicenter | Multi-vendor (GE, Samsung) | Each exam performed by one of six examiners with substantial experience (7-23 years) | Yes |
| Deeparani et al. (2023)[55] | Diagnosis (benign vs borderline vs malignant) | 634 | Retrospective | Histology | 2D | Multicenter | NR | Each exam performed by one of six examiners with substantial experience (7-23 years) | NR |
| Du et al. (2024)[43] | Diagnosis (benign vs borderline vs malignant) | 849 | Retrospective | Histology | 2D | Monocenter | Multi-vendor (GE, Mindray) | NR | NR |
| Du et al. (2024)[47] | Diagnosis (benign vs malignant) | 849 | Retrospective | Histology | 2D | Monocenter | Multi-vendor (GE, Mindray) | NR | NR |
| Faschingbauer et al. (2013)[56] | Diagnosis (2 benign classes vs malignant) | 105 | Retrospective | Histology | 2D | Monocenter | Single-vendor (GE) | Expert ultrasonographer (level III EFSUMB) | Yes |
| Gao et al. (2022)[42] | Diagnosis (benign vs malignant) | 106400 | Retrospective | Histology | 2D | Multicenter | Multi-vendor (Toshiba, Siemens, Phillips, GE) | Examiners with >3 years experience | NR |
| Hsu et al. (2022)[46] | Diagnosis (benign vs malignant) | 587 | Retrospective | Histology | 2D | Monocenter | Multi-vendor (GE, Mindray) | NR | Yes |
| Hussein et al. (2022)[57] | Diagnosis | NR | Retrospective | NR | 2D | NR | NR | NR | NR |
| Jung et al. (2022)[58] | Diagnosis (normal vs 3 benign classes vs malignant) | 1154 | Retrospective | Histology | 2D | Monocenter | Multi-vendor (GE, Samsung) | Expert ultrasonographer | Yes |
| Khazendar et al. (2015)[59] | Diagnosis (benign vs malignant) | 177 | Retrospective | Histology | 2D | Multicenter (IOTA) | NR | NR | Yes |
| Kongara et al. (2024)[61] | Diagnosis (normal vs benign vs malignant) | NR | Retrospective | NR | 2D | NR | NR | NR | NR |
| Li et al. (2022)[49] | Diagnosis (5 benign classes vs borderline vs malignant) | 1559 | Retrospective | Histology | 2D | Multicenter (one center for development, other two centers for external validation) | Multi-vendor (GE, Mindray, Philips, Samsung, Hitachi, Siemens) | NR | NR |
| Lucidarme et al. (2010)[62] | Diagnosis (benign vs malignant) | 264 | Prospective | Histology | 3D | Multicenter | Single-vendor (GE) | Five radiologists with more than 10 years experience | Yes |

**Table 1** (continued)

| Author (year) | Objective | Patients | Retrospective/ Prospective | Reference standard | Modality | Mono-/ Multicenter | Single-/ Multi-vendor scanning | Experience level scanner(s) | Anonymization |
|---|---|---|---|---|---|---|---|---|---|
| Martinez-Mas et al. (2019)[60] | Diagnosis (benign vs malignant) | 187 | Retrospective | Histology | 2D | Multicenter (IOTA) | NR | Experienced radiologist | Yes |
| Miao et al. (2023)[63] | Diagnosis (benign vs malignant) | 1350 | Retrospective | Histology | 2D | Monocenter | Multi-vendor (GE, Samsung, Esaote) | Clinicians with at least 5 years of experience in ultrasound diagnosis | Yes |
| Moro et al. (2024)[64] | Diagnosis (benign vs malignant) | 326 | Retrospective | Histology | 2D | Monocenter | Multi-vendor (GE, Samsung) | Experienced ultrasound examiners (EFSUMB level II-III) | NR |
| Pathak et al. (2015)[65] | Diagnosis (benign vs malignant) | NR | NR | NR | 2D | NR | NR | NR | NR |
| Pham et al. (2024)[66] | Detection & diagnosis (tumor vs no tumor & 6 benign vs 2 malignant classes) | 247 | Retrospective | Histology | 2D | Monocenter | Single-vendor (Mindra) | NR | Yes |
| Qi et al. (2021)[48] | Diagnosis (benign vs borderline vs malignant) | 265 | Retrospective | Histology | 2D | Monocenter | Single-vendor (Phillips) | NR | Yes |



Downloaded from http://karger.com/gol/article-pdf/doi/10.1159/000545850/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025

Accepted Manuscript

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Raja et al. (2024)[67] | Diagnosis (normal vs 6 benign vs malignant) | NR | Retrospective | NR | 2D | NR | NR | NR | NR |
| Sheela et al. (2023)[68] | Diagnosis (benign vs malignant) | 216 | Retrospective | Histology | 2D | NR | NR | NR | NR |
| Stefan et al. (2021)[69] | Diagnosis (benign vs malignant) | 120 | Retrospective | Histology | 2D | Monocenter | Single-vendor (Toshiba) | Gynecologists with at least 7 years of experience | NR |
| Wang H. et al. (2021)[70] | Diagnosis (benign vs borderline vs malignant) | 265 | Retrospective | Histology | 2D | Monocenter | Multi-vendor (Phillips, GE, Samsung) | Ultrasound specialist | NR |
| Wu C. et al. (2018)[41] | Diagnosis (benign vs borderline vs malignant) | NR | Retrospective | Histology | 2D | Monocenter | Multi-vendor (GE, Philips) | NR | NR |
| Wu M. et al. (2023)[44] | Diagnosis (6 benign vs 1 malignant class) | 328 | Retrospective | Histology | 2D | Monocenter | NR | NR | NR |
| Xi et al. (2024)[45] | Diagnosis (benign vs malignant) | 405 | Retrospective | Histology | 2D | Monocenter | Single-vendor (GE) | NR | NR |

**Table 1** (continued)

| Author (year) | Objective | Patients | Retrospective/ Prospective | Reference standard | Modality | Mono-/ Multicenter | Single-/ Multi-vendor scanning | Experience level scanner(s) | Anonymization |
|---|---|---|---|---|---|---|---|---|---|
| Xiang et al. (2024)[51] | Diagnosis (benign vs malignant) | 724 | Retrospective | Histology | 2D | Monocenter | Multi-vendor (GE, Aplio) | NR | Yes |
| Zimmer et al. (2003)[71] | Diagnosis (benign vs malignant) | NR | Retrospective | Histology | 2D | NR | Multi-vendor (Acuson, Aloka) | NR | NR |
| **Endometrial cancer** | | | | | | | | | |
| Fang et al. (2023)[75] | Diagnosis (2 benign classes vs malignant) | 734 | Retrospective | NR | 2D | Multicenter | Multi-vendor (GE, Samsung, Philips) | NR | Yes |
| Michail et al. (2007)[73] | Diagnosis (benign vs malignant) | 65 | NR | Histology | 2D | Monocenter | Single-vendor (Phillips) | One experienced radiologist | NR |
| Wang X. et al. (2022)[74] | Diagnosis (normal vs abnormal) | 100 | NR | Histology | 2D | Monocenter | NR | NR | NR |
| **Myometrial cancer** | | | | | | | | | |
| Chiappa et al. (2021)[72] | Diagnosis (benign vs malignant) | 70 | Retrospective | Histology | 2D | Monocenter | Multi-vendor (GE, Samsung) | Two examiners with >10 years experience | NR |

*Abbreviations*: EFSUMB: European Federation for Ultrasound in Medicine and Biology; GE: General Electric; IOTA: International Ovarian Tumor Analysis; NR: Not Reported; SA: Subjective Assessment

Accepted Manuscript

Downloaded from http://karger.com/gol/article-pdf/doi/10.1159/000545850/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025

**Table 2** Statistical variables extracted from the included studies, categorized by anatomical tumor type.

| Author (year) | Patients | Images | Images/ Patient | ROI delineation | Modelling technique | Type of internal validation | Performance metric(s) | Patient- or Image-wise evaluation | AI with/vs clinician | Auxiliary interpretability techniques | External validation | Calibration assessment |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Ovarian cancer** | | | | | | | | | | | | |
| Acharya et al. (2012)[32] | 20 | 2000 | 100 | NR | ML (with radiomics) | 10-Fold CV | ACC, SENS, SPEC, PPV | Image-wise | No | No | No | No |
| Acharya et al. (2012)[34] | 20 | 2000 | 100 | Manual | ML (with radiomics) | 10-Fold CV | ACC, SENS, SPEC, PPV | Image-wise | No | No | No | No |
| Acharya et al. (2013)[33] | 20 | 2000 | 100 | None | ML (with radiomics) | 10-Fold CV | ACC, SENS, SPEC, PPV | Image-wise | No | No | No | No |
| Acharya et al. (2014)[36] | 20 | 2600 | 130 | Manual | ML (with radiomics) | 10-Fold CV | ACC, SENS, SPEC, PPV | Image-wise | No | No | No | No |
| Acharya et al. (2014)[35] | 20 | 2600 | 130 | Manual | ML (with radiomics) | 10-Fold CV | ACC, SENS, SPEC, PPV | Image-wise | No | No | No | No |
| Acharya et al. (2018)[37] | 469 | 469 | 1.00 | Manual | ML (with radiomics) | 10-Fold CV | ACC, SENS, SPEC, AUC¶ | Same† | No | No | No | No |
| Al-Karawi et al. (2021)[38] | 232 | 242 | 1.04 | Manual | ML (with radiomics) | L1O-CV, Train-Test (50%-50%) | ACC, SENS, SPEC | Image-wise | No | No | No | No |
| Aramendia-Vidaurreta et al. (2016)[39] | 145 | 145 | 1.00 | Manual | ML (with radiomics) | 10-Fold CV | ACC, SENS, SPEC, PPV, AUC | Same† | No | No | No | No |
| Barcroft et al. (2024)[40] | 577 | 1444 | 2.50 | Automatic | ML (with radiomics), DL | Train-Test (80%-20%) | PREC, REC*, SPEC, F1, AUC | Patient-wise | No | Yes (Feature Importance) | Yes (Development Study) | Yes |
| Chen et al. (2022)[50] | 422 | 2113 | 5.01 | Manual | DL | Train-Validation-Test (70%-10%-20%) | SENS, SPEC, PPV, NPV, LR+, LR-, AUC | Patient-wise | Yes (vs O-RADS, SA) | No | No | No |
| Chiappa et al. (2021)[52] | 241 | 241 | 1.00 | Manual | ML (with radiomics) | (Nested) 10-Fold CV | ACC, SENS, SPEC, AUC | Same† | No | No | No | No |
| Chiappa et al. (2021)[53] | 274 | 274 | 1.00 | Manual | Rule-Based DSS | No split (Development Study) | ACC, SENS, SPEC, PPV, NPV | Same† | Yes (with clinical feature) | No | Yes (Development Study) | No |

Accepted Manuscript

Downloaded from http://karger.com/gol/article-pdf/doi/10.1159/000545850.pdf/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025

**Table 2** (continued)

| Author (year) | Patients | Images | Images/ Patient | ROI delineation | Modelling technique | Type of internal validation | Performance metric(s) | Patient- or Image-wise evaluation | AI with/vs clinician | Auxiliary interpretability techniques | External validation | Calibration assessment |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Christiansen et al. (2021)[54] | 758 | 3077 | 4.06 | Manual | DL | Train-Validation-Test (67%-13.2%-19.8%) | ACC, SENS, SPEC, AUC | Patient-wise | Yes (with SA for uncertain model predictions, vs: SA) | No | No | Yes |
| Deeparani et al. (2023)[55] | 634 | 39942 (after DA) | 63 (after DA) | Automatic | DL (with radiomics) | Train-Test (60%-40%) | ACC, SENS, SPEC, PREC, F1, AUC | Image-wise | No | No | No | No |
| Du et al. (2024)[43] | 849 | 849 | 1.00 | Manual, None | DL, ML (with radiomics) | Train-Test (80%-20%) | ACC, PREC, REC*, F1, AUC | Same† | Yes (with clinical feature) | Yes (Saliency Maps) | No | No |
| Du et al. (2024)[47] | 849 | 849 | 1.00 | Manual, None | DL, ML (with radiomics) | Train-Test (80%-20%) | ACC, PREC, REC*, F1, AUC | Same† | Yes (with clinical feature) | Yes (Saliency Maps) | No | Yes |
| Faschingbauer et al. (2013)[56] | 105 | 105 | 1.00 | Manual | ML (with radiomics) | L1O-CV | SENS, SPEC, YOUD | Same† | Yes (vs SA) | No | No | No |
| Gao et al. (2022)[42] | 106400 | 584346 (Training+ Internal Validation) | 5.49 | None | DL | Train-Validation-Test (89.7%-8.97%-1.31%) | ACC, SENS, SPEC, PPV, NPV, F1, AUC | Patient-wise | Yes (with SA for all model predictions, vs: SA) | No | Yes (Development Study) | Yes |
| Hsu et al. (2022)[46] | 587 | 1896 | 3.23 | None | DL | Train-Test (70%-30%) | ACC, SENS, SPEC | Image-wise | Yes (vs SA) | Yes (Saliency Maps) | No | No |
| Hussein et al. (2022)[57] | NR | 125 (+250 breast images) | NR | Automatic | ML (with radiomics) | (Nested) 5-Fold CV, Train-Test (45.9%-55.1%) | ACC, SENS, SPEC | Image-wise | No | No | No | No |
| Jung et al. (2022)[58] | 1154 | 1613 | 1.40 (1.00)‡ | Automatic | DL | 5-Fold CV | ACC, SENS, SPEC, PPV, NPV, AUC | Same‡ | Yes (vs SA) | Yes (Saliency Maps) | No | No |
| Khazendar et al. (2015)[59] | 177 | 187 | 1.06 | Manual, None | ML (with radiomics) | L1O-CV | ACC, SENS, SPEC | Image-wise | No | No | No | No |
| Kongara et al. (2024)[61] | NR | 1820 | NR | Automatic | DL (with radiomics) | Train-Test (83.5%-16.5%) | ACC, PREC, REC*, F1 | Image-wise | No | No | No | No |

**Table 2** (continued)

| Author (year) | Patients | Images | Images/ Patient | ROI delineation | Modelling technique | Type of internal validation | Performance metric(s) | Patient- or Image-wise evaluation | AI with/vs clinician | Auxiliary interpretability techniques | External validation | Calibration assessment |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Li et al. (2022)[49] | 1559 | 5714 | 3.67 | Automatic | DL | Train-Test (70.5%-29.5%) | ACC, SENS, SPEC, PPV, NPV, F1 | Patient-wise | Yes (vs SA) | No | Yes (Development Study) | Yes |
| Lucidarme et al. (2010)[62] | 264 | 375 | 1.42 (1.00)§ | Manual | NR | No split (Validation Study) | SENS, SPEC | Same§ | Yes (with clinical feature) | No | Yes (Validation Study) | No |
| Martinez-Mas et al. (2019)[60] | 177 | 187 | 1.06 | Manual | ML (with radiomics) | L1O-CV | ACC, SENS, SPEC, AUC | Image-wise | No | No | No | No |
| Miao et al. (2023)[63] | 1350 | 2340 | 1.73 | None | DL | Train-Test (80%-20%) | SENS, SPEC, PPV, NPV, LR+, LR-, AUC | Image-wise | No | Yes (Saliency Maps) | No | No |
| Moro et al. (2024)[64] | 326 | 775 | 2.38 | Manual | ML (with radiomics) | Train-Test (70%-30%) | ACC, SENS, SPEC, LR+, LR-, AUC | Patient-wise | No | No | No | No |
| Pathak et al. (2015)[65] | NR | 120 | NR | Manual | ML (with radiomics) | Train-Test (58.3%-41.7%) | ACC, SENS, SPEC | Image-wise | No | No | No | No |



Accepted Manuscript

Downloaded from http://karger.com/gol/article-pdf/doi/10.1159/000545850/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pham et al. (2024)[66] | 247 | 1469 | 5.98 | Automatic | DL | Train-Test (68.1%-31.9%) | PREC, REC*, mAP50¶ | Image-Wise | No | No | No | No |
| Qi et al. (2021)[48] | 265 | 279 | 1.05 (1.00)§ | Manual | ML (with radiomics) | Train-Test (70%-30%) | ACC, SENS, SPEC, AUC | Same§ | Yes (with clinical feature, vs SA) | No | No | Yes |
| Raja et al. (2024)[67] | NR | 240 | NR | NR | DL | Train-Test (66.7%-33.3%) | ACC, PREC, REC*, F1 | Image-wise | No | No | No | No |
| Sheela et al. (2023)[68] | NR | 250 | NR | Manual, None | ML (with radiomics) | L1O-CV | ACC, SENS, SPEC | Image-wise | No | No | No | No |
| Stefan et al. (2021)[69] | 120 | 123 | 1.03 (1.00)§ | Manual | ML (with radiomics) | No split (Development Study) | ACC, SENS, SPEC, AUC¶ | Same§ | No | No | No | No |
| Wang H. et al. (2021)[70] | 265 | 279 | 1.05 (1.00)§ | Manual | DL | 3-Fold CV | ACC, SENS, SPEC, F1, AUC | Same§ | Yes (vs SA) | Yes (Saliency Maps) | No | No |
| Wu C. et al. (2018)[41] | NR | 988 | NR | None | DL | Train-Validation | ACC, AUC | Image-wise | No | No | No | No |

**Table 2** (continued)

| **Author (year)** | **Patients** | **Images** | **Images/ Patient** | **ROI delineation** | **Modelling technique** | **Type of internal validation** | **Performance metric(s)** | **Patient- or Image-wise evaluation** | **AI with/vs clinician** | **Auxiliary interpretability techniques** | **External validation** | **Calibration assessment** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Wu M. et al. (2023)[44] | 328 | 1142 | 3.48 | Manual | DL | Train-Validation (70%-10%-20%) | ACC, SENS, SPEC, PPV, NPV, F1, AUC | Image-wise | No | Yes (Saliency Maps) | No | No |
| Xi et al. (2024)[45] | 405 | 1103 | 2.72 | NR | DL | Train-Test (70%-30%) | ACC, SENS, SPEC, PPV, NPV, AUC | Image-wise | Yes (vs SA) | No | No | No |
| Xiang et al. (2024)[51] | 724 | 3972 | 5.49 | None | DL, (ML without radiomics) | Train-Validation-Test (70%-10%-20%) | ACC, SENS, SPEC, PPV, NPV, AUC | Patient-Wise | Yes (with clinical feature, O-RADS, vs O-RADS) | Yes (Saliency Maps and Feature Importance) | Yes (Development Study) | No |
| Zimmer et al. (2003)[71] | NR | 28 | NR | Manual | Rule-Based DSS | No split (Development Study) | SENS, SPEC, PPV, NPV | Image-wise | No | No | No | No |
| **Endometrial cancer** | | | | | | | | | | | | |
| Fang et al. (2023)[75] | 734 | 1875 | 2.55 | None | DL | 5-Fold CV | ACC, PREC, REC*, F1, AUC | Patient-wise | No | Yes (Saliency Maps) | No | No |
| Michail et al. (2007)[73] | 65 | 65 | 1.00 | Manual | ML (with radiomics) | No split | SENS, SPEC, AUC | Same† | No | No | No | No |
| Wang X. et al. (2022)[74] | 100 | 100 | 1.00 | Automatic | DL | Train-Test | ACC, SENS, SPEC, Kappa | Same† | No | No | No | No |
| **Myometrial cancer** | | | | | | | | | | | | |
| Chiappa et al. (2021)[72] | 70 | 70 | 1.00 | Manual | ML (with radiomics) | (Nested) 10-Fold CV | ACC, SENS, SPEC, AUC, PPV, NPV | Same† | No | No | No | No |

*Recall is equivalent to sensitivity. †Sample size of the number images corresponds to the number of patients. ‡Sample size of the number images corresponds to the number of ovaries. §Sample size of the number images corresponds to the number of ovarian tumors. ¶This metric was only evaluated for some implemented models. *Abbreviations*: AUC: Area under the Receiver Operating Characteristic curve; ACC: Accuracy; CV: Cross-Validation; DA: Data Augmentation; DSS: Decision Support System; DL: Deep Learning; L1O-CV: Leave-one-Out Cross-Validation; LR+: Positive Likelihood Ratio; LR-: Negative Likelihood Ratio; ML: Machine Learning; mAP50: Mean Average Precision at 50% Intersection over Union (IoU); NPV: Negative Predictive Value; NR: Not Reported; PREC: Precision; PPV: Positive Predictive Value; REC: Recall; RoB: Risk of Bias; ROI: Region of Interest; SA: Subjective Assessment; SENS: Sensitivity; SPEC: Specificity; YOU: Youden index

Accepted Manuscript

Downloaded from http://karger.com/gor/article-pdf/doi/10.1159/000545850/4373044/000545850.pdf by KU Leuven Libraries user on 20 May 2025