







Systematic Review

Ultrasound-Based Deep Learning Models Performance versus Expert Subjective Assessment for Discriminating Adnexal Masses: A Head-to-Head Systematic Review and Meta-Analysis

Mariana Lourenço ^{1,†}, Teresa Arrufat ^{2,†}, Elena Satorres ^{3,†}, Sara Maderuelo ^{3,†}, Blanca Novillo-Del Álamo ³, Stefano Guerriero ⁴, Rodrigo Orozco ⁵ and Juan Luis Alcázar ^{5,6,*}

¹ Department of Obstetrics and Gynecology, Hospital Vila Franca de Xira, 2600-009 Lisboa, Portugal; marianalourenco52@gmail.com

² Department of Obstetrics and Gynecology, Hospital de Castellón, 12004 Castellón, Spain; teresarrufat@gmail.com

³ Department of Obstetrics and Gynecology, Hospital Universitario y Politécnico La Fe, 46026 Valencia, Spain; elenasatorres@gmail.com (E.S.); saramade10@hotmail.com (S.M.); bnalamo@gmail.com (B.N.-D.Á.)

⁴ Department of Obstetrics and Gynecology, Università di Cagliari, Policlinico Universitario Duilio Casula, 09042 Cagliari, Italy; gineca.sguerriero@tiscali.it

⁵ Department of Obstetrics and Gynecology, Hospital QuirónSalud, 29004 Málaga, Spain; rodrigo.orozco@quironsalud.es

⁶ School of Medicine, University of Navarra, 31009 Pamplona, Spain

* Correspondence: jalcazar@unav.es

† These authors contributed equally to this work.



Citation: Lourenço, M.; Arrufat, T.; Satorres, E.; Maderuelo, S.; Novillo-Del Álamo, B.; Guerriero, S.; Orozco, R.; Alcázar, J.L. Ultrasound-Based Deep Learning Models Performance versus Expert Subjective Assessment for Discriminating Adnexal Masses: A Head-to-Head Systematic Review and Meta-Analysis. *Appl. Sci.* **2024**, *14*, 2998. <https://doi.org/10.3390/app14072998>

Academic Editor: Qi-Huang Zheng

Received: 5 March 2024

Revised: 25 March 2024

Accepted: 29 March 2024

Published: 3 April 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: (1) Background: Accurate preoperative diagnosis of ovarian masses is crucial for optimal treatment and postoperative outcomes. Transvaginal ultrasound is the gold standard, but its accuracy depends on operator skill and technology. In the absence of expert imaging, pattern-based approaches have been proposed. The integration of artificial intelligence, specifically deep learning (DL), shows promise in improving diagnostic precision for adnexal masses. Our meta-analysis aims to evaluate DL's performance compared to expert evaluation in diagnosing adnexal masses using ultrasound images. (2) Methods: Studies published between 2000 and 2023 were searched in PubMed, Scopus, Cochrane and Web of Science. The study quality was assessed using QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2). Pooled sensitivity and specificity for both methods were estimated and compared. (3) Results: From 1659 citations, we selected four studies to include in this meta-analysis. The mean prevalence of ovarian cancer was 30.6%. The quality of the studies was good with low risk of bias for index and reference tests, but with high risk of bias for patient selection domain. Pooled sensitivity and specificity were 86.0% and 90.0% for DL and 86.0% and 89.0% for expert accuracy ($p = 0.9883$). (4) Conclusion: We found no significant differences between DL systems and expert evaluations in detecting and differentially diagnosing adnexal masses using ultrasound images.

Keywords: adnexal mass; deep learning; transvaginal ultrasound; ovarian cancer; artificial intelligence

1. Introduction

Adnexal masses, comprising ovarian, tubal, and para-ovarian lesions constitute a common clinical problem and represent a diagnostic challenge in gynecology. Fortunately, most adnexal masses are benign [1]. Benign adnexal masses can be managed conservatively with serial follow-up or surgically with tumor removal, preferentially with minimally invasive surgery by a general gynecologist [2–4]. In addition, fertility preservation is an important issue in patients of reproductive age and with childbearing intentions [5]. On the other hand, ovarian cancer is the second most common gynecological malignancy with an estimated incidence of 11.4 cases/100,000 women-year [6]. However, this cancer is the most

lethal gynecological malignancy, with a 5-year overall survival of 48% [7], and there is no effective screening strategy [8]. Ovarian cancer should be treated in specialized centers, with expert surgeons and adequate resources, where surgical outcomes and patients' survival are clearly improved [9,10].

Therefore, the precise preoperative diagnosis between benign and malignant ovarian masses is essential for selecting appropriate treatment strategies and enhancing patients' postoperative prognoses [11], necessitating the precise preoperative assessments for optimal clinical guidance and therapeutic decision-making [11]. The mainstay for diagnosing adnexal masses relies on imaging techniques, with transvaginal ultrasound being regarded as a first-line imaging tool for assessing adnexal tumors for its non-invasiveness, cost-effectiveness, and ability to provide real-time insights into pelvic pathology and because no other imaging technique has been proven to be superior in terms of diagnostic performance [11,12]. However, the accuracy of ultrasound imaging interpretation is contingent upon both the proficiency of the operator and the available technology [13,14]. In fact, a randomized trial demonstrated how the operator's experience has a measurable impact on the correct management of adnexal masses [15].

In this context, the importance of accurate preoperative diagnosis through transvaginal ultrasound cannot be overstated, as it significantly influences the subsequent clinical management of patients with adnexal masses [16,17]. Because this technique depends on the operator's experience and due to their limited access to expert imaging assessments, to complement their evaluation, numerous pattern-based methodologies and logistical models leveraging image characteristics have been proposed to improve the precision of diagnoses for adnexal masses. For years many attempts have been made to develop scoring systems and classifications that would allow less-experienced examiners achieve good results [18–22].

However, as technological landscapes continue to evolve, the integration of artificial intelligence (AI) has emerged as a revolutionary paradigm in medical diagnostics [23,24]. Artificial intelligence is described as the ability of a computer program to perform processes associated with human intelligence, that can learn and interact [25]. Machine learning is a branch of AI, defined by the ability to learn from data without being explicitly programmed. Machine learning can be understood as a statistical method that gradually improves as it is exposed to more data, by extracting patterns from the data [25].

Deep learning (DL), a subset of machine learning, that entails the creation of convolutional neural networks that emulate the human brain's capacity to learn and make decisions. Deep learning involves learning from vast amounts of data and performs especially well in pattern recognition within data; therefore, it can be particularly helpful in medical imaging. With the exponential growth of medical data, this tool has shown promising results in enhancing the accuracy and efficiency of adnexal mass characterization. In the last decade, several studies have been reported using ultrasound-based ML approaches for the differential diagnosis of adnexal masses [26–44].

However, only some of these used DL models and even fewer compared DL with the diagnostic performance of a human examiner. To the best of our knowledge, no meta-analysis has been reported comparing the diagnostic performance of deep learning and human expert examiner diagnoses. The aim of our meta-analysis and systematic review is to evaluate the performance of DL compared to expert evaluation in the diagnosis of adnexal masses using ultrasound images.

2. Materials and Methods

This meta-analysis was performed according to the PRISMA recommendations (<http://www.prisma-statement.org/>, accessed on 2 November 2023) [45]. This study was registered in PROSPERO (CRD42024502144). Given the nature and design of this study, ethics committee approval was not required.

2.1. Search Strategy

A search was conducted across four electronic databases (PubMed, Scopus, Cochrane, and Web of Science) from January 2000 to October 2023 to identify studies that could meet the eligibility criteria. To prevent potential oversight of relevant studies, we refrained from applying methodological filters in the database searches.

For the research, we utilized the search terms “adnexal mass or ovarian cancer” and “machine learning” in the search field.

2.2. Selection Criteria

Once we had compiled the preliminary list of potential studies, all titles were reviewed to eliminate any duplications. In the next step, we filtered the titles and subsequently examined abstracts to identify and exclude irrelevant articles. This included those not directly related to the topic under review, as well as non-observational studies, such as reviews, case reports, and letters to the editor. To ensure we did not exclude any valid study to our research, all the abstracts were evaluated and validated by five authors. At the end, records were again filtered with a complete reading of the full text of the studies that remained after exclusion.

This meta-analysis had the following inclusion criteria:

- Primary diagnostic prospective or retrospective studies evaluating deep learning models for the diagnosis of adnexal masses and comparing them to expert evaluation.
- Collected Data allows the construction of a 2×2 table to estimate true positive, true negative, false positive, and false negative cases for any of the index tests assessed.
- The reference test was considered histological confirmation.

The exclusion criteria were the following:

- Studies not related to the topic.
- Studies evaluating AI or ML other than DL.
- Secondary studies (other meta-analysis or systematic reviews) or those where insufficient data were provided.
- Studies where other imaging studies such as MRI or CT were considered.
- Articles that did not compare deep learning results with those obtained by expert examiners.
- Studies where histological confirmation was not the gold standard or was not available.
- Studies combining evaluation of biomarkers to ultrasonography.

We used the snowball strategy to identify potentially interesting papers by reading the reference lists of the papers selected for full text reading.

2.3. Data Extraction and Management

Five authors independently retrieved the following data from each study: first author, year of publication, country, study design, number of centers participating, patients' inclusion criteria, patients' exclusion criteria, patients' age, number of patients, number of patients with benign and malignant diagnoses, number of images, suspicions of benignity or malignancy, index test used, number of examiners, number of images employed for training, validation and testing, number of sonographers and whether they were experts or not, the reference standard used, the diagnostic accuracy results, and the time elapsed from the index test to the reference standard test.

Disagreements arising during the process of study selection and data extraction were solved by consensus among these five authors.

2.4. Qualitative Synthesis

The quality assessment of the studies included in the meta-analysis was conducted using the tool provided by the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) [46]. The QUADAS-2 format includes four domains: patient selection, index test, reference standard, and flow and timing. For each domain, the risk of bias and concerns

about applicability (not applying to the domain of flow and timing) were analyzed and rated as low, high, or unclear risk. Five authors independently evaluated the methodological quality. Disagreements were solved by discussion between these authors.

When evaluating risk of bias in patient selection, those with a retrospective design were considered high risk. For the flow-and-timing domain, we considered a high risk of bias when time from the index test assessment to the reference standard was over three months. The fact that sonographers were not blinded to ultrasound findings was considered to pose a high risk of bias. Surgery and anatomopathological study were defined as the reference standards.

2.5. Quantitative Synthesis

All statistical analysis has been performed exclusively with external validation data, excluding internal validation in those studies that reported this information. When there were multiple prediction models, we selected the one that offered the best diagnostic performance in terms of sensitivity and specificity. Likewise, if the primary study included more than one external validation of the same model (for example external validation in two or more centers), we grouped the data. When there were multiple groups of experts, we took the grouped data as presented in the papers.

Pooled specificity and sensitivity were calculated using randomized effects and the comparison between the performance of the model and the experts was done using the bivariate method. Forest plots of the sensitivity and specificity of all studies were plotted. Heterogeneity for sensitivity and specificity was assessed using the I^2 index [47]. If heterogeneity was found, meta-regression analysis was performed using as covariates sample size and malignancy prevalence.

Summary receiver operating characteristic (sROC) curves were plotted to illustrate the relationship between sensitivity and specificity, and the area under the curve (AUC) was calculated. Publication bias would also be assessed. Statistical analysis was performed using STATA version 12.0 for Windows (Stata Corporation, College Station, TX, USA). A p -value < 0.05 was considered statistically significant.

3. Results

3.1. Search Results

The electronic search provided 1659 citations. We excluded 809 duplicate records, and 33 secondary reports. After that, 817 citations remained. After reading titles and abstracts, 795 citations were excluded (papers not related to the topic ($n = 530$), assessing biomarkers ($n = 33$) and other diagnostic methods used ($n = 232$)). Twenty-two papers remained for full-text reading. Eighteen papers were further excluded (no deep learning ($n = 3$), no reference standard ($n = 2$) and no comparison with expert sonographer ($n = 13$)). Therefore, only four studies were ultimately included in this meta-analysis for qualitative and quantitative synthesis [40,42–44]. A flowchart summarizing the literature search is shown in Figure 1.

3.2. Characteristics of Studies

Data of the four studies included are summarized in Table 1. These papers were published between January 2021 and October 2022 and reported the data of 2439 women with ages ranging from 11 to 85 years old. The mean prevalence of ovarian cancer was 30.6% (19.0–48.7%). All the studies were retrospective and in all studies the ultrasound was performed by a selection of proficient experts and the reference standard was the histology obtained after surgery.

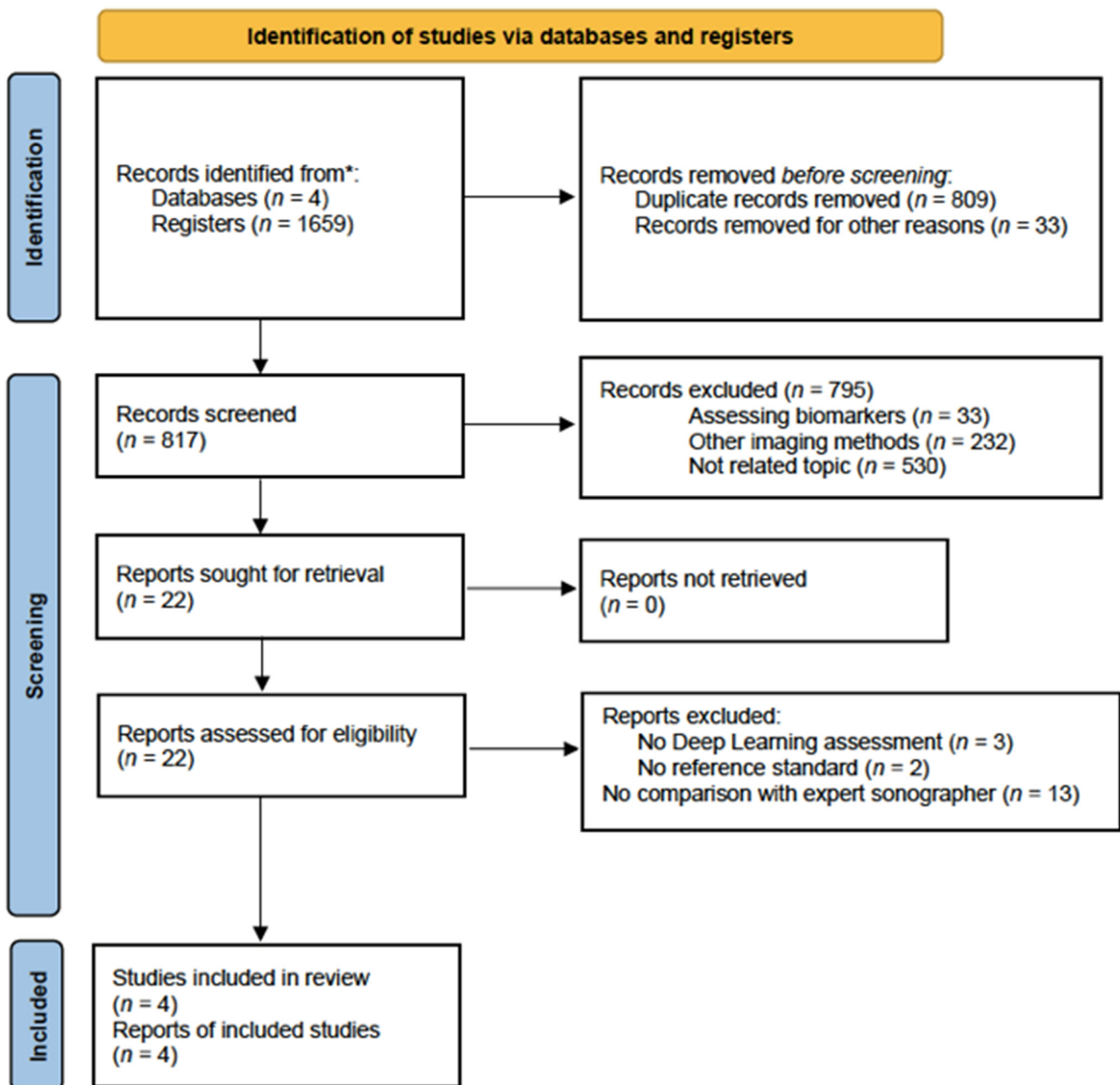


Figure 1. Flowchart showing the study selection process. * From January 2000 to October 2023.

Altogether, there was a total of 2541 adnexal masses, where 1797 were benign, 71 borderline and 673 were malignant masses. Nevertheless, one article did not specify such data for any subgroup [40], and only one specified the number of borderline tumors [43]. For statistical analysis, borderline and malignant tumors were grouped together.

Concerning Christiansen et al. [40] and Chen et al. [42], two DL models were evaluated. In order to calculate pooled values, we chose the model with better sensitivity and specificity results.

In Christiansen et al.'s study [40], the total number of cases was 634 (325 benign cases, 55 borderline, 254 malignant) with a total of 3077 images. In the training dataset there were 508 cases (314 benign and 194 malignant cases). In the validation dataset there were 100 cases (60 benign and 40 malignant cases). Whereas, in the test set there were 150 cases (75 benign and 75 malignant cases).

Table 1. Characteristics of the studies included in this meta-analysis.

Author	Year	Country	Study Design	N Patients	Benign	Malignant and Borderline	N° Centers	Age (Years) *	Reference Standard	Deep Learning Architecture	Time until Surgery (Days) **
Gao [44]	2022	China	Retrospective	1224	991	233	10	External dataset1: 43 (32–52) External dataset2: 38 (27–48)	Histology	DCNN	NR
Christiansen [40]	2021	Sweden	Retrospective	634	325	309	2	NR	Histology	DNN: (VGG16, ResNet50 and MobileNet)	120
Chen [42]	2022	China	Retrospective	422	304	118	1	46.8 (18–85)	Histology	CNN (ResNet-18)	30
Li [43]	2022	China	Retrospective	261	177	84	3	External 1: 48.87 (15, 77) External 2: 43.26 (19, 50)	Histology	CNN (ResNet-18)	NR

* Expressed as median with range in parentheses. ** Maximum time elapsed.

In Chen et al.'s study [42], the total number of cases was 422 (304 benign cases and 118 malignant ones) with a total of 2113 images. In the training dataset there was 296 cases (213 benign and 83 malignant cases) with a total of 1493 images. In the validation dataset there were 41 cases (30 benign and 11 malignant cases) with a total of 189 images. In the test dataset there was a total of 85 cases (61 benign, 24 malignant) with a total 431 images.

In Li et al.'s article [43], the total number of cases for internal validation was 460 with a total of 1217 images. There was a total of 377 images of 134 benign cases, a total of 219 images of 56 malignant cases and 57 images of 15 borderline cases. Regarding external validation, which was formed by two different groups, there was in group 1 a total of 198 cases, with a total of 490 images, 134 images of 43 patients with cancer, 27 images of 8 borderline cases and 151 images of 51 benign masses. In group 2, the total number was 264 (a total of 761 images, 97 images of 25 patients with cancer, 30 images of 8 borderline cases and 401 images of 126 benign masses).

In Gao et al.'s study [44], the total number of cases for internal validation was 868 (3031 images for 266 cases of malignant adnexal masses, and 5385 images of 602 cases of benign adnexal masses). Regarding external validation, two groups were formed with different patients; in external validation 1, the number of cases was 335 (486 images of 67 patients with cancer, and 933 images of 268 benign masses). In external validation 2, the total number was 889 cases (1253 images in the malignant group with 166 cases, and 5257 images from 723 benign cases). In our meta-analysis, we decided to combine the data from both external validation groups to simplify the statistical analysis, where the performances of the model and the experts were compared.

This information is summarized in Table 2.

3.3. Qualitative Synthesis

The risk of biased evaluation and concerns regarding the applicability of the selected studies is shown in Table 3.

All studies were deemed high-risk in the "patient selection" domain due to their retrospective design. Concerning the domain "index test", all studies were considered to have a low risk of bias because the diagnostic test methods were clearly explained.

Table 2. Distribution of histology and the number of images used in the training, internal validation, test and external validation sets for all the studies included in this meta-analysis.

Author		Training SET	Internal Validation Set	Test Set	External Validation Set 1	External Validation Set 2
Christensen [40]	Histology	314 B/194 M	60 B/40 M	75 B/75 M		
	Images per histology	NR	NR	NR	ND	ND
Chen [42]	Histology	213 B/83 M	30 B/11 M	61 B/24 M	ND	ND
	Images per histology	NR	NR	NR		
Li [43]	Histology	882 B/217 M	389 B/71 M	ND	147 B/51 M	231 B/33 M
	Images per histology	2227 B/2270 M	941 B/276 M	ND	329 B/161 M	634 B/127 M
Gao [44]	Histology	101,777 B/3755 M	602 B/266 M	ND	268 B/67 M	723 B/166 M
	Images per histology	541,442 B/344,488 M	5385 B/3031 M	ND	933 B/486 M	5257 B/1253 M

NR: not reported. ND: not done. B: benign. M: malignant.

Table 3. Risk of bias of the studies included in this meta-analysis.

Author	Year	Patient Selection	Index Test	Reference Standard	Flow and Timing
Christiansen [40]	2021	High risk	Low risk	Low risk	High risk
Chen [42]	2022	High risk	Low risk	Low risk	Low risk
Li [43]	2022	High risk	Low risk	Low risk	Unclear
Gao [44]	2022	High risk	Low risk	Low risk	Unclear

The standard reference employed was reliable and widely accepted for determining the true status of the condition (histological confirmation), allowing for accurate comparisons with the diagnostic test results; therefore, all the studies were considered low-risk for the domain “reference standard”.

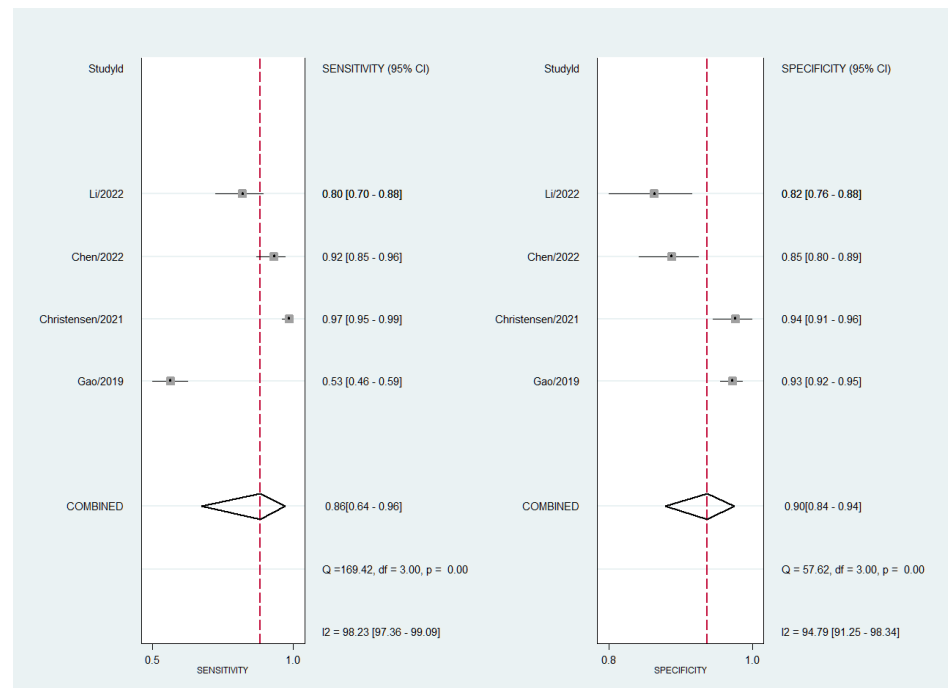
Regarding the domain “flow and timing”, two articles were considered to have an unclear risk of bias because the time elapsed between the ultrasound and laparoscopy with histological confirmation was not described [43,44]. In the remaining two studies, it was considered low risk in one, as the time lapse was only one month [42], and high in the other, as more than three months occurred between the two interventions (four months, precisely) [40]

3.4. Quantitative Synthesis

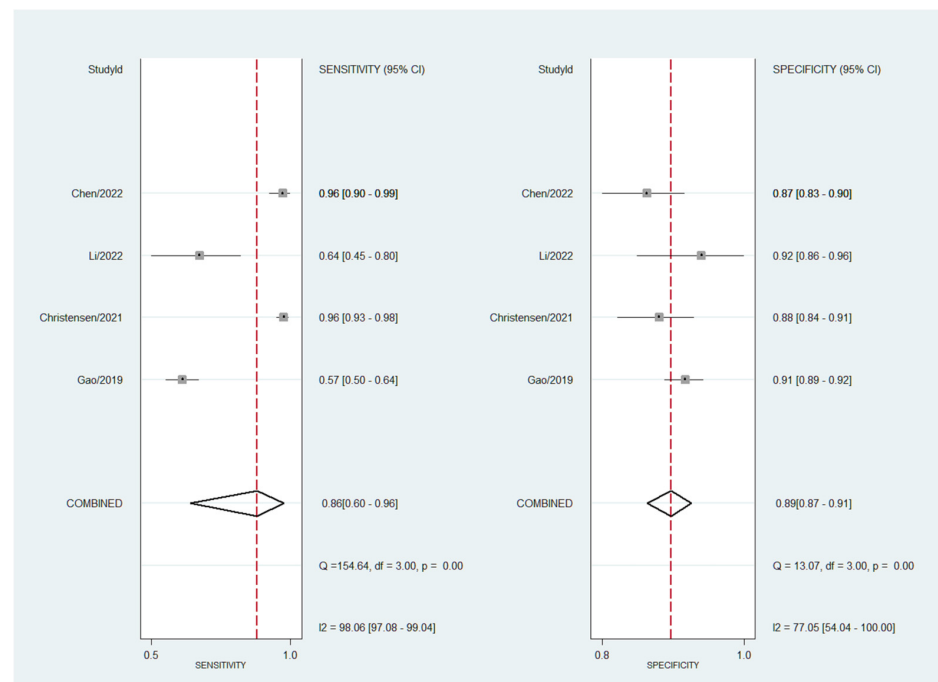
Considering the DL models’ performance for the diagnosis of ovarian adnexal masses, the pooled sensitivity was 86.0% (95% CI, 64.0–96.0%). Whereas, the estimated pooled specificity was 90.0% (95% CI, 84.0–93.0%). Considering experts’ performance, the pooled sensitivity was 86.0% (95% CI 60.0–96.0%). The estimated pooled specificity was 89.0% (95% CI, 87.0–91.0%). When comparing the diagnostic performance of the two approaches, we did not find any statistical difference ($p = 0.9883$).

The heterogeneity observed—meaning the presence of variation in the true effect sizes underlying the different studies—was high for all calculations (Figure 2A,B). The observed

heterogeneity in sensitivity is explained by a difference in the prevalence of the included studies, which was confirmed by meta-regression ($p = 0.0001$).



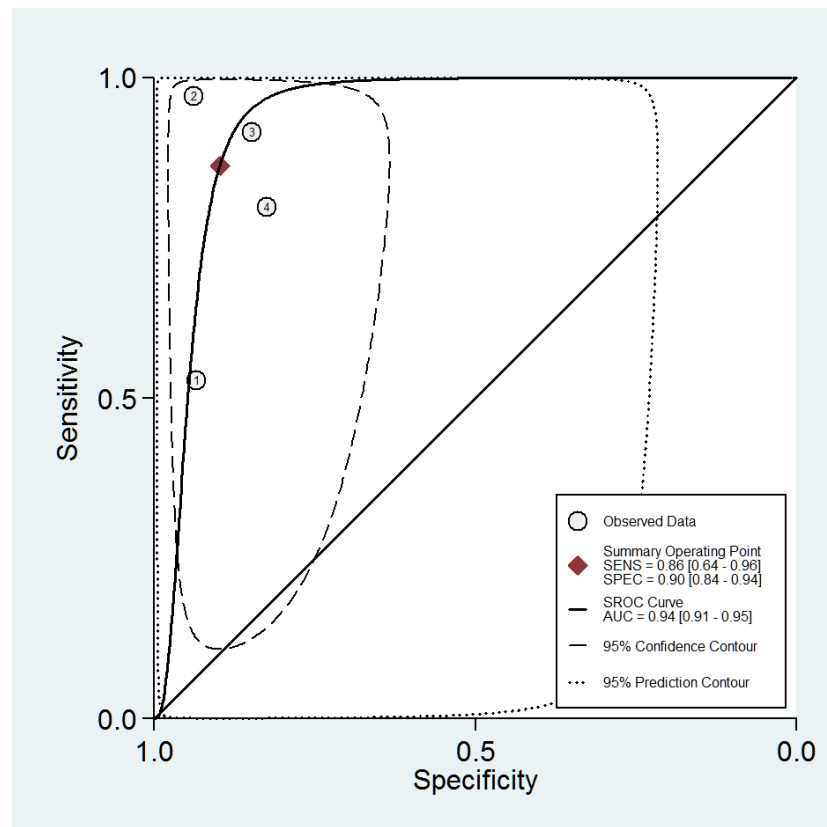
(A)



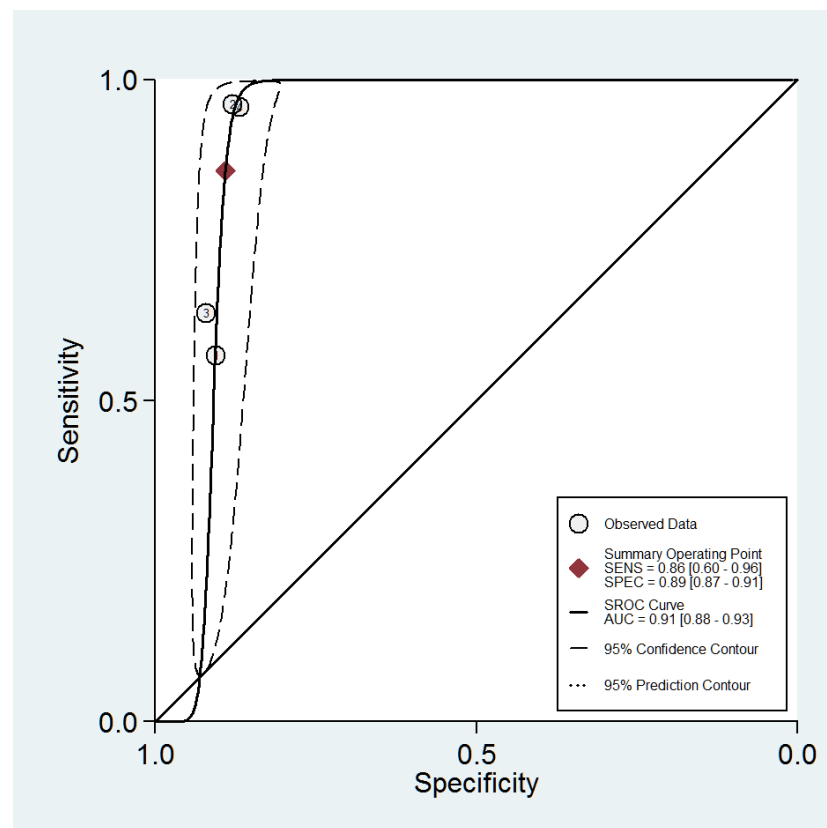
(B)

Figure 2. (A) Forest plot for sensitivity and specificity for DL models. (B) Forest plot for sensitivity and specificity for expert assessment [40,42–44].

The sHROC curves for the diagnostic performances of the DL models and the experts to diagnose ovarian masses are shown in Figure 3A and Figure 3B, respectively. The area under the curve regarding the DL models is 0.94 (95% CI 0.91–0.95). For expert evaluation, the area under the curve is 0.91 (95% CI 0.88–0.93).



(A)



(B)

Figure 3. (A) Summary ROC curves for the DL models. (B) Summary ROC curves for the expert examiners.

Fagan’s nomograms for DL and expert evaluation are shown in Figure 4A and Figure 4B, respectively. When interpreting these figures, we can observe that if a DL model classifies the mass as malignant, the a priori risk (mean prevalence) increases from 31% to 79%. By contrast, when a DL model indicates that the mass is benign, the a priori risk drops to 6% (Figure 4A). Similarly, when an expert examiner classifies a mass as malignant, the a posteriori risk increases up to 78%. Whereas, when an expert examiner classifies the mass as benign, the a posteriori risk decreases to 7%.

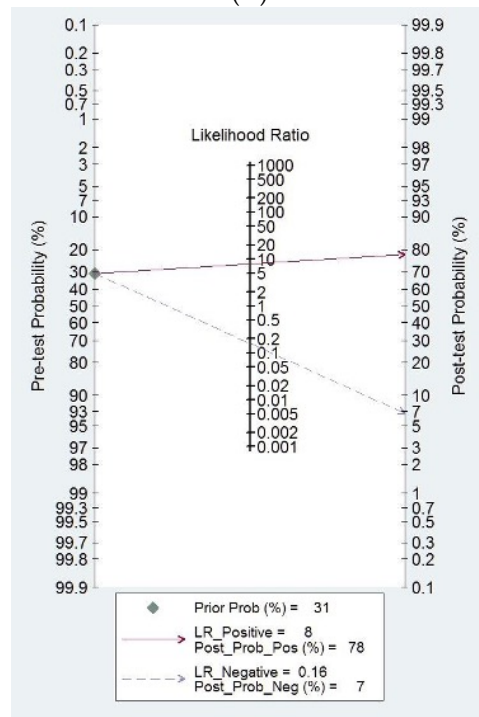
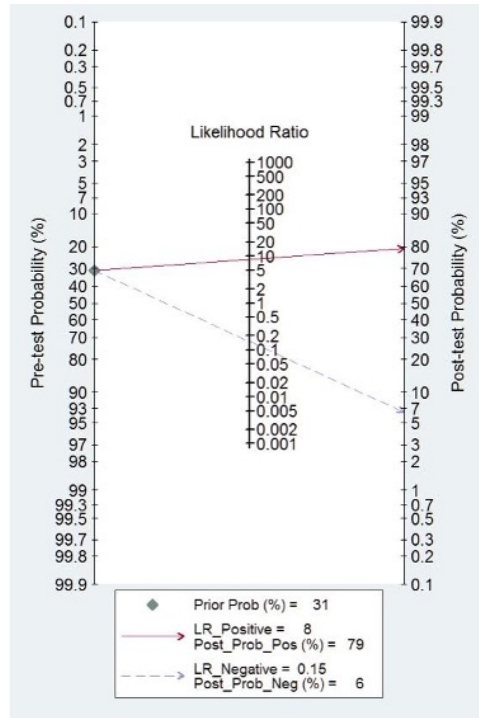


Figure 4. (A) Fagan nomograms for DL models. (B) Fagan nomograms for expert examiners.

Publication bias was not assessed as we only included four studies.

4. Discussion

4.1. Summary of Findings

In our analysis, after a meticulous research and selection process, focusing on a singular imaging assessment method (transvaginal ultrasound) and a specific form of artificial intelligence (DL), we identified only four studies that aligned with our research objectives.

We did find that DL models do not perform better than expert evaluation for discriminating benign from malignant adnexal masses. However significant heterogeneity across studies was observed. This heterogeneity was explained by a different prevalence of malignancy in the studies included, ranging from 19% to 48%. This different prevalence could be explained by the different inclusion and exclusion criteria used in the studies.

The quality of studies was considered as good, related to the index and reference tests, and the flow and timing domain. However, we noticed that only retrospective studies have been published. This means that the methods used as diagnostic tests (DL architectures and criteria for examiner impression diagnosis) as well as the reference test (histopathology) were deemed as having low risk of bias. However, we did consider that the retrospective design in all studies poses a high risk of bias, since this kind of study design may have an inherent risk of bias as the authors cannot control for all potential confounding variables regarding patient selection.

4.2. Interpretation of Findings in Clinical Context

Currently, ultrasound is considered the first-choice imaging technique for discriminating between benign and malignant adnexal masses [48]. Expert examiner subjective impression has been shown as the best diagnostic approach [49]. However, this approach has inherent limitations such as the need for adequate training and experience [13,14]. In fact, a randomized trial demonstrated that examiner experience had a measurable clinical impact in the management of women diagnosed as having an adnexal mass [15].

For this reason, scoring systems and logistic models, such as the Risk of Malignancy index, IOTA simple rules, Risk Assessment based on Simple rules, LR1 model, LR2 model, ADNEX model, GI-RADS and O-RADS classification systems were developed in an attempt to overcome this limitation [21,22,50–54]. No doubt, these approaches have produced a significant advance in the ultrasound discrimination of adnexal masses, allowing less-experienced examiners achieve good results in terms of diagnostic performance [55–63]. However, the expert examiner is still the most cost-effective approach [49].

Artificial intelligence is becoming a relevant issue in medical practice, particularly for diagnosis. In fact, the use of AI models in the differential diagnosis of adnexal masses for detecting ovarian cancer is increasing, as recently described [64,65]. AI models in medical imaging may be based on data and/or image processing [25]. In deep learning models, the input and output are connected by multiple layers of hidden connections, also known as convolutional neural networks [25]. Deep learning involves learning from vast amounts of data and performs especially well in pattern recognition within data; therefore, it can be particularly helpful in medical imaging. The use of these models is attractive because it avoids problems that derive from the interpretation of an image by a human observer, such as lack of experience, availability, or subjectivity in interpretation, making the diagnosis more generalizable and reproducible.

The use of artificial neural networks for discriminating benign from malignant tumors is not a new topic [66–69]. Nevertheless, the performance of these early prediction models was limited. The main reason for this limited performance is that these models were built with few data.

However, in the era of big data and with the progressive integration of DL in medical imaging involving the automatic assessment of digital images by the software, there is an ongoing active assessment of models to ascertain their diagnostic accuracy in the context of adnexal masses. DL represents the latest evolution in the field of ML, characterized by the

utilization of multiple layers of neural networks that contribute increased computational power through heightened neuronal complexity. This advancement, however, comes with a caveat, as deep learning methods are inherently more susceptible to overfitting. Consequently, these methods often necessitate larger datasets to ensure optimal performance. However, it is necessary to determine whether these DL models actually perform better than human beings, particularly expert examiners.

The present study compares the performance of DL systems using ultrasound images for detecting and differentially diagnosing adnexal masses against expert evaluations. As stated above, our analysis revealed that there were no significant differences between the two methods of assessment. The conclusion to be drawn may be influenced by the study design and the limited sample size; however, it is evident that these DL models can be applied to ultrasound imaging in situations where expert opinions are unavailable or inaccessible.

Furthermore, although it was not our objective, it is interesting to note that the pooled diagnostic performance of DL (sensitivity 86%, specificity 90%) does not seem to be better than other simpler logistic models, such as the ADNEX model or the IOTA-LR2 model. In fact, regarding the IOTA-LR2 model, two meta-analyses, including data from 1612 and 1554 women, reported a sensitivity of 92–93% and a specificity of 83–84% for this model [49,55]. Regarding the IOTA ADNEX model, two recent meta-analyses reporting data from more than 17,000 women showed that this model has a sensitivity of 93–94% and a specificity of 75–78% [57,58].

The considerable future potential of DL in various medical diagnostic areas is clear; nevertheless, optimizing the way we conceptualize these models, gather data, and analyze them is essential. We propose future studies and analyses that take these considerations into account.

4.3. Strengths and Limitations

The main strength of our meta-analysis is that, to the best of our knowledge, this is the first reported comparison that is specifically between the performance of DL models and that of expert examiners.

Certainly, there are other recent meta-analyses addressing the performance of artificial intelligence approaches to the differential diagnosis of adnexal masses [24,70,71]. However, our results cannot be compared with these meta-analyses.

Xu et al. reported a meta-analysis including 19 studies using ultrasound-based artificial intelligence models [70]. They included four that used DL (all four of these are also included in our study) and fifteen more using other models. They estimated pooled sensitivity and specificity mixing data from all 19 studies, not discriminating among different approaches such as machine learning, deep learning or artificial neural networks, and did not compare them to expert evaluation. We think that mixing all these data can be misleading, since different approaches may perform differently.

Grigore et al. reported a meta-analysis assessing the performance of logistic models and “artificial intelligence” models [24]. They included 18 studies, but none addressing the role of DL. Therefore, the results of this meta-analysis cannot be compared to ours.

Finally, Ma et al. reported a meta-analysis including 11 studies, three of them using ultrasound-based approaches [71]. Two of them used DL, but did not compare this to expert examination and it was not included in our meta-analysis [36]. Again, the results of this meta-analysis cannot be compared to ours.

As limitations, the inclusion of a relatively small dataset, consisting only of four studies and mostly of a few hundred data points, imposes limitations on the potential advantages afforded by DL models and has to be considered when evaluating the obtained results. On the other hand, the heterogeneity observed among the included studies was high, and with regards to sensitivity it can be explained by a difference in the prevalence, as stated above.

Furthermore, all included investigations were retrospective, using patients’ data derived from medical records. This study design has an inherent risk of bias. This observation

manifests the need to promote the development of prospective AI research in forthcoming endeavors to enhance study validity. We should be cautious regarding the interpretation of our results. In fact, another potential limitation to be considered is that all four studies included in our meta-analysis used different DL models. Christiansen et al. used three different deep neural networks (VGG-16, Resnet 50 and MobileNet) [40]. They reported the diagnostic performance of each model and the combination of all three. We used the latter for our calculations. Chen et al. used two different deep neural networks (Resnet 18 and Resnet 50) [42]. The diagnostic performance was slightly better for the model Resnet 18 and we used these data for our calculations. Li et al. used the LKResnet 18 model, which is a combination of Resnet 18 and LKNet deep neural networks [43]. And finally, Gao et al. used a single deep convolutional neural network (DenseNet-121) [44]. All these models are, certainly, convolutional neural network architectures, but there are differences among them. For example, ResNet18 consists of 18 layers, whereas Resnet50 is composed of 50 layers. On the other hand, LKResnet and DenseNet-121 are combinations of two convolutional neural network architectures. If we look at Figure 2A, we can observe that not all models perform similarly, so some questions arise. Can we use distinct DLs for the same purpose? Are these data generalizable? Can we use these models in different populations? In fact, there is some evidence that these architectures do not perform similarly when compared [72–74].

Another issue is the selection of the images and the region of interest (ROI) within the image for training the DL model. Ultimately, an operator, or operators, should decide which image should be used. Can this affect the performance of the model? Certainly, segmentation is a significant advance when using DL models, as the computer selects the ROI. However, it remains to be shown whether the models perform equally well on images acquired by other expert centers, less-experienced examiners or by examiners not using high-end equipment [10].

Finally, the inclusive approach of incorporating articles written in diverse languages was intended to provide comprehensive insights into ongoing research; however, no non-English articles were identified that aligned with the scope of our investigation.

5. Conclusions

The current study assesses the effectiveness of DL systems in utilizing ultrasound images for the detection and differential diagnosis of adnexal masses when compared to expert evaluations. Our analysis indicates that there were no significant differences between the two assessment methods. However, further research on this topic is needed, as the number of studies is low. In particular, prospective studies that address the actual value of DL models in clinical practice is needed.

Author Contributions: Conceptualization, J.L.A. and S.G.; methodology, J.L.A. and S.G.; software, J.L.A.; validation, J.L.A., S.G. and R.O.; formal analysis, J.L.A., M.L., T.A., E.S., S.M. and B.N.-D.Á.; investigation, J.L.A., M.L., T.A., E.S., S.M. and B.N.-D.Á.; resources, J.L.A.; data curation, J.L.A., M.L., T.A., E.S., S.M. and B.N.-D.Á.; writing—original draft preparation, M.L., T.A., E.S. and S.M.; writing—review and editing, J.L.A., S.G. and R.O.; project administration, J.L.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the study design.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Grimes, D.A.; Hughes, J.M. Use of multiphasic oral contraceptives and hospitalizations of women with functional ovarian cysts in the United States. *Obstet. Gynecol.* **1989**, *73*, 1037–1039. [\[CrossRef\]](#) [\[PubMed\]](#)
- Alcázar, J.L.; Olartecochea, B.; Guerriero, S.; Jurado, M. Expectant management of adnexal masses in selected premenopausal women: A prospective observational study. *Ultrasound Obstet. Gynecol.* **2013**, *41*, 582–588. [\[CrossRef\]](#) [\[PubMed\]](#)
- Froyman, W.; Landolfo, C.; De Cock, B.; Wynants, L.; Sladkevicius, P.; Testa, A.C.; Van Holsbeke, C.; Domali, E.; Fruscio, R.; Epstein, E. Risk of complications in patients with conservatively managed ovarian tumours (IOTA5): A 2-year interim analysis of a multicentre, prospective, cohort study. *Lancet Oncol.* **2019**, *20*, 448–458. [\[CrossRef\]](#)
- Glanc, P.; Benacerraf, B.; Bourne, T.; Brown, D.; Coleman, B.G.; Crum, C.; Dodge, J.; Levine, D.; Pavlik, E.; Ueland, F.R. First International Consensus Report on Adnexal Masses: Management Recommendations. *J. Ultrasound Med.* **2017**, *36*, 849–863. [\[CrossRef\]](#)
- Stein, E.B.; Hansen, J.M.; Maturen, K.E. Fertility-Sparing Approaches in Gynecologic Oncology: Role of Imaging in Treatment Planning. *Radiol. Clin. N. Am.* **2020**, *58*, 401–412. [\[CrossRef\]](#) [\[PubMed\]](#)
- Webb, P.M.; Jordan, S.J. Epidemiology of epithelial ovarian cancer. *Best. Pract. Res. Clin. Obstet. Gynaecol.* **2016**, *S1521-S6934*, 30091–30098. [\[CrossRef\]](#)
- Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2020. *CA Cancer J. Clin.* **2020**, *70*, 7–30. [\[CrossRef\]](#)
- Jacobs, I.J.; Menon, U.; Ryan, A.; Gentry-Maharaj, A.; Burnell, M.; Kalsi, J.K.; Amso, N.N. Ovarian cancer screening and mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): A randomized controlled trial. *Lancet* **2016**, *387*, 945–956. [\[CrossRef\]](#)
- Kim, B.R.; Kim, H.; Joo, S.G.; Jang, E.J.; Jo, J.; Lee, H.; Ryu, H.G. Effect of Hospital Case-Volume on Mortality after Ovarian Cancer Surgery: A Population-Based Retrospective Cohort Study. *Gynecol. Obstet. Investig.* **2022**, *87*, 364–372. [\[CrossRef\]](#)
- Nasioudis, D.; Kahn, R.; Chapman-Davis, E.; Frey, M.K.; Caputo, T.A.; Witkin, S.S.; Holcomb, K. Impact of hospital surgical volume on complete gross resection (CGR) rates following primary debulking surgery for advanced stage epithelial ovarian carcinoma. *Gynecol. Oncol.* **2019**, *154*, 401–404. [\[CrossRef\]](#)
- American College of Obstetricians and Gynecologists' Committee on Practice Bulletins—Gynecology. Practice Bulletin No. 174: Evaluation and Management of Adnexal Masses. *Obstet. Gynecol.* **2016**, *128*, e210–e226. [\[CrossRef\]](#) [\[PubMed\]](#)
- Salvador, S.; Scott, S.; Glanc, P.; Eiriksson, L.; Jang, J.H.; Sebastianelli, A.; Dean, E. Guideline No. 403: Initial Investigation and Management of Adnexal Masses. *J. Obstet. Gynaecol. Can.* **2020**, *42*, 1021–1029.e3. [\[CrossRef\]](#) [\[PubMed\]](#)
- Yazbek, J.; Ameye, L.; Testa, A.C.; Valentin, L.; Timmerman, D.; Holland, T.K.; Van Holsbeke, C.; Jurkovic, D. Confidence of expert ultrasound operators in making a diagnosis of adnexal tumor: Effect on diagnostic accuracy and interobserver agreement. *Ultrasound Obstet. Gynecol.* **2010**, *35*, 89–93. [\[CrossRef\]](#) [\[PubMed\]](#)
- Van Holsbeke, C.; Daemen, A.; Yazbek, J.; Holland, T.K.; Bourne, T.; Mesens, T.; Lannoo, L.; Boes, A.S.; Joos, A.; Van De Vijver, A. Ultrasound experience substantially impacts on diagnostic performance and confidence when adnexal masses are classified using pattern recognition. *Gynecol. Obstet. Investig.* **2010**, *69*, 160–168. [\[CrossRef\]](#) [\[PubMed\]](#)
- Yazbek, J.; Raju, S.K.; Ben-Nagi, J.; Holland, T.K.; Hillaby, K.; Jurkovic, D. Effect of quality of gynaecological ultrasonography on management of patients with suspected ovarian cancer: A randomised controlled trial. *Lancet Oncol.* **2008**, *9*, 124–131. [\[CrossRef\]](#) [\[PubMed\]](#)
- Alcázar, J.L.; Pascual, M.A.; Graupera, B.; Aubá, M.; Errasti, T.; Olartecochea, B.; Ruiz-Zambrana, A.; Hereter, L.; Ajossa, S.; Guerriero, S. External validation of IOTA simple descriptors and simple rules for classifying adnexal masses. *Ultrasound Obstet. Gynecol.* **2016**, *48*, 397–402. [\[CrossRef\]](#) [\[PubMed\]](#)
- Coccia, M.E.; Rizzello, F.; Romanelli, C.; Capezzuoli, T. Adnexal masses: What is the role of ultrasonographic imaging? *Arch. Gynecol. Obstet.* **2014**, *290*, 843–854. [\[CrossRef\]](#) [\[PubMed\]](#)
- Sassone, A.M.; Timor-Tritsch, I.E.; Artner, A.; Westhoff, C.; Warren, W.B. Transvaginal sonographic characterization of ovarian disease: Evaluation of a new scoring system to predict ovarian malignancy. *Obstet. Gynecol.* **1991**, *78*, 70–76.
- Alcázar, J.L.; Mercé, L.T.; Laparte, C.; Jurado, M.; López-García, G. A new scoring system to differentiate benign from malignant adnexal masses. *Am. J. Obstet. Gynecol.* **2003**, *188*, 685–692. [\[CrossRef\]](#)
- Amor, F.; Alcázar, J.L.; Vaccaro, H.; León, M.; Iturra, A. GI-RADS reporting system for ultrasound evaluation of adnexal masses in clinical practice: A prospective multicenter study. *Ultrasound Obstet. Gynecol.* **2011**, *38*, 450–455. [\[CrossRef\]](#)
- Timmerman, D.; Testa, A.C.; Bourne, T.; Ameye, L.; Jurkovic, D.; Van Holsbeke, C.; Paladini, D.; Van Calster, B.; Vergote, I.; Van Huffel, S. Simple ultrasound-based rules for the diagnosis of ovarian cancer. *Ultrasound Obstet. Gynecol.* **2008**, *31*, 681–690. [\[CrossRef\]](#) [\[PubMed\]](#)
- Andreotti, R.F.; Timmerman, D.; Strachowski, L.M.; Froyman, W.; Benacerraf, B.R.; Bennett, G.L.; Bourne, T.; Brown, D.L.; Coleman, B.G.; Frates, M.C.; et al. O-RADS US Risk Stratification and Management System: A Consensus Guideline from the ACR Ovarian-Adnexal Reporting and Data System Committee. *Radiology* **2020**, *294*, 168–185. [\[CrossRef\]](#)
- Shrestha, P.; Poudyal, B.; Yadollahi, S.; Wright, D.; Gregory, A.; Warner, J.; Korfiatis, P.; Green, I.; Rassier, S.; Mariani, A.; 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*, 596–605. [\[CrossRef\]](#)

24. Grigore, M.; Popovici, R.M.; 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*, 469–475. [[CrossRef](#)] [[PubMed](#)]
25. Drukker, L.; Noble, J.A.; Papageorgiou, A.T. Introduction to artificial intelligence in ultrasound imaging in obstetrics and gynecology. *Ultrasound Obstet. Gynecol.* **2020**, *56*, 498–505. [[CrossRef](#)] [[PubMed](#)]
26. Acharya, U.R.; Sree, S.V.; Krishnan, M.M.; Saba, L.; Molinari, F.; Guerriero, S.; Suri, J.S. Ovarian tumor characterization using 3D ultrasound. *Technol. Cancer Res. Treat.* **2012**, *11*, 543–552. [[CrossRef](#)]
27. Alqasemi, U.; Kumavor, P.; Aguirre, A.; Zhu, Q. Recognition algorithm for assisting ovarian cancer diagnosis from coregistered ultrasound and photoacoustic images: Ex vivo study. *J. Biomed. Opt.* **2012**, *17*, 126003. [[CrossRef](#)]
28. Acharya, U.R.; Sree, V.S.; Saba, L.; Molinari, F.; Guerriero, S.; Suri, J.S. Ovarian tumor characterization and classification: A class of GyneScan™ systems. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* **2012**, *2012*, 4446–4449.
29. Acharya, U.R.; Sree, S.V.; Saba, L.; Molinari, F.; Guerriero, S.; Suri, J.S. Ovarian tumor characterization and classification using ultrasound—A new online paradigm. *J. Digit. Imaging* **2013**, *26*, 544–553. [[CrossRef](#)]
30. Acharya, U.R.; Sree, S.V.; Kulshreshtha, S.; Molinari, F.; Koh, J.E.; Saba, L.; Suri, J.S. GyneScan: An improved online paradigm for screening of ovarian cancer via tissue characterization. *Technol. Cancer Res. Treat.* **2014**, *13*, 529–539. [[CrossRef](#)]
31. Acharya, U.R.; Mookiah, M.R.; Sree, S.V.; Yanti, R.; Martis, R.J.; Saba, L.; Molinari, F.; Guerriero, S.; Suri, J.S. Evolutionary algorithm-based classifier parameter tuning for automatic ovarian cancer tissue characterization and classification. *Ultraschall Med.* **2014**, *35*, 237–245. [[PubMed](#)]
32. Pathak, H.; Kulkarni, V. Identification of ovarian mass through ultrasound images using machine learning techniques. In Proceedings of the 2015 IEEE International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), Kolkata, India, 20–22 November 2015; pp. 137–140.
33. Aramendía-Vidaurreta, V.; Cabeza, R.; Villanueva, A.; Navallas, J.; Alcazar, J.L. Ultrasound image discrimination between benign and malignant adnexal masses based on a neural network approach. *Ultrasound Med. Biol.* **2016**, *42*, 742–752. [[CrossRef](#)]
34. Martínez-Más, J.; Bueno-Crespo, A.; Khazendar, S.; Remezal-Solano, M.; Martínez-Cendán, J.P.; Jassim, S.; Du, H.; Al Assam, H.; Bourne, T. Evaluation of machine learning methods with Fourier transform features for classifying ovarian tumors based on ultrasound images. *PLoS ONE* **2019**, *14*, e0219388. [[CrossRef](#)]
35. Akazawa, M.; Hashimoto, K. Artificial intelligence in ovarian cancer diagnosis. *Anticancer Res.* **2020**, *40*, 4795–4800. [[CrossRef](#)] [[PubMed](#)]
36. Wang, H.; Liu, C.; Zhao, Z.; Zhang, C.; Wang, X.; Li, H.; Wu, H.; Liu, X.; Li, C.; Qi, L.; et al. Application of deep convolutional neural networks for discriminating benign, borderline, and malignant serous ovarian tumors from ultrasound images. *Front. Oncol.* **2021**, *11*, 770683. [[CrossRef](#)]
37. Chiappa, V.; Bogani, G.; Interlenghi, M.; Salvatore, C.; Bertolina, F.; Sarpietro, G.; Signorelli, M.; Castiglioni, I. The Adoption of radiomics and machine learning improves the diagnostic processes of women with ovarian masses (the AROMA pilot study). *J. Ultrasound.* **2021**, *24*, 429–437. [[CrossRef](#)]
38. Ștefan, P.A.; Lupean, R.A.; Mișu, C.M.; Lebovici, A.; Oancea, M.D.; Hițu, L.; Duma, D.; Csutak, C. Ultrasonography in the diagnosis of adnexal lesions: The role of texture analysis. *Diagnostics* **2021**, *11*, 812. [[CrossRef](#)] [[PubMed](#)]
39. Al-Karawi, D.; Al-Assam, H.; Du, H.; Sayasneh, A.; Landolfo, C.; Timmerman, D.; Bourne, T.; Jassim, S. 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**, *43*, 124–138. [[CrossRef](#)]
40. Christiansen, F.; Epstein, E.L.; 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 Obstet. Gynecol.* **2021**, *57*, 155–163. [[CrossRef](#)]
41. Guo, X.; Zhao, G. Establishment and verification of logistic regression model for qualitative diagnosis of ovarian cancer based on MRI and ultrasound signs. *Comput. Math. Methods Med.* **2022**, *2022*, 7531371. [[CrossRef](#)]
42. Chen, H.; Yang, B.W.; Qian, L.; Meng, Y.S.; Bai, X.H.; Hong, X.W.; He, X.; Jiang, M.J.; Yuan, F.; Du, Q.W. Deep learning prediction of ovarian malignancy at US compared with O-RADS and expert assessment. *Radiology.* **2022**, *304*, 106–113. [[CrossRef](#)] [[PubMed](#)]
43. Li, J.; Chen, Y.; Zhang, M.; Zhang, P.; He, K.; Yan, F.; Li, J.; Xu, H.; Burkhoff, D.; Luo, Y.; et al. A Deep Learning Model System for Diagnosis and Management of Adnexal Masses. *Cancers* **2022**, *14*, 5291. [[CrossRef](#)] [[PubMed](#)]
44. Gao, Y.; Zeng, S.; Xu, X.; Li, H.; Yao, S.; Song, K.; Li, X.; Chen, L.; Tang, J.; Xing, H.; 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*, e179–e187. [[CrossRef](#)]
45. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *29*, n71. [[CrossRef](#)]
46. Whiting, P.F.; Rutjes, A.W.; Westwood, M.E.; Mallett, S.; Deeks, J.J.; Reitsma, J.B.; Leeflang, M.M.; Sterne, J.A.; Bossuyt, P.M.; QUADAS-2 Group. QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Ann. Intern. Med.* **2011**, *18*, 529–536.
47. Higgins, J.P.; Thompson, S.G.; Deeks, J.J.; Altman, D.G. Measuring inconsistency in meta-analyses. *BMJ* **2003**, *327*, 557–560. [[CrossRef](#)] [[PubMed](#)]

48. Timmerman, D.; Planchamp, F.; Bourne, T.; Landolfo, C.; du Bois, A.; Chiva, L.; Cibula, D.; Concin, N.; Fischerova, D.; Froyman, W. ESGO/ISUOG/ IOTA/ESGE Consensus Statement on pre-operative diagnosis of ovarian tumors. *Int. J. Gynecol. Cancer* **2021**, *31*, 961–982. [[CrossRef](#)]
49. Meys, E.M.; Kaijser, J.; Kruitwagen, R.F.; Slangen, B.F.; Van Calster, B.; Aertgeerts, B.; Verbakel, J.Y.; Timmerman, D.; Van Gorp, T. Subjective assessment versus ultrasound models to diagnose ovarian cancer: A systematic review and meta-analysis. *Eur. J. Cancer* **2016**, *58*, 17–29. [[CrossRef](#)]
50. Jacobs, I.; Oram, D.; Fairbanks, J.; Turner, J.; Frost, C.; Grudzinskas, J.G. A risk of malignancy index incorporating CA 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer. *Br. J. Obstet. Gynaecol.* **1990**, *97*, 922–929. [[CrossRef](#)]
51. Chacón, E.; Dasí, J.; Caballero, C.; Alcázar, J.L. Risk of Ovarian Malignancy Algorithm versus Risk Malignancy Index-I for Preoperative Assessment of Adnexal Masses: A Systematic Review and Meta-Analysis. *Gynecol. Obstet. Investig.* **2019**, *84*, 591–598. [[CrossRef](#)]
52. Timmerman, D.; Testa, A.C.; Bourne, T.; Ferrazzi, E.; Ameye, L.; Konstantinovic, M.L.; Van Calster, B.; Collins, W.P.; Vergote, I.; Van Huffel, S.; et al. Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: A multicenter study by the International Ovarian Tumor Analysis Group. *J. Clin. Oncol.* **2005**, *23*, 8794–8801. [[CrossRef](#)]
53. Van Calster, B.; Van Hoorde, K.; Valentin, L.; Testa, A.C.; Fischerova, D.; Van Holsbeke, C.; Savelli, L.; Franchi, D.; Epstein, E.; Kaijser, J. 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. [[CrossRef](#)] [[PubMed](#)]
54. Amor, F.; Vaccaro, H.; Alcázar, J.L.; León, M.; Craig, J.M.; Martinez, J. Gynecologic imaging reporting and data system: A new proposal for classifying adnexal masses on the basis of sonographic findings. *J. Ultrasound Med.* **2009**, *28*, 285–291. [[CrossRef](#)] [[PubMed](#)]
55. Kaijser, J.; Sayasneh, A.; Van Hoorde, K.; Ghaem-Maghami, S.; Bourne, T.; Timmerman, D.; Van Calster, B. Presurgical diagnosis of adnexal tumours using mathematical models and scoring systems: A systematic review and meta-analysis. *Hum. Reprod. Update* **2014**, *20*, 449–462. [[CrossRef](#)] [[PubMed](#)]
56. Ilundain, A.; Salas, A.; Chacón, E.; Manzour, N.; Alcazar, J.L. IOTA Simple Rules for the differential diagnosis of ovarian adnexal masses: Systematic review and meta-analysis. *Prog. Obstet. Gynecol.* **2018**, *61*, 390–400.
57. Yue, X.; Zhong, L.; Wang, Y.; Zhang, C.; Chen, X.; Wang, S.; Hu, J.; Hu, J.; Wang, C.; Liu, X. Value of Assessment of Different Neoplasias in the Adnexa in the Differential Diagnosis of Malignant Ovarian Tumor and Benign Ovarian Tumor: A Meta-analysis. *Ultrasound Med. Biol.* **2022**, *48*, 730–742. [[CrossRef](#)] [[PubMed](#)]
58. Barreñada, L.; Ledger, A.; Dhiman, P.; Collins, G.; Wynants, L.; Verbakel, J.Y.; Timmerman, D.; Valentin, L.; Van Calster, B. ADNEX risk prediction model for diagnosis of ovarian cancer: Systematic review and meta-analysis of external validation studies. *BMJ Med.* **2024**, *3*, e000817. [[CrossRef](#)] [[PubMed](#)]
59. Guo, W.; Zou, X.; Xu, H.; Zhang, T.; Zhao, Y.; Gao, L.; Duan, W.; Ma, X.; Zhang, L. The diagnostic performance of the Gynecologic Imaging Reporting and Data System (GI-RADS) in adnexal masses. *Ann. Transl. Med.* **2021**, *9*, 398. [[CrossRef](#)] [[PubMed](#)]
60. Alcázar, J.L.; Rodriguez-Guzman, L.; Vara, J.; Amor, F.; Diaz, L.; Vaccaro, H. Gynecologic Imaging and Reporting Data System for classifying adnexal masses. *Minerva Obstet. Gynecol.* **2023**, *75*, 69–79. [[CrossRef](#)] [[PubMed](#)]
61. Vara, J.; Manzour, N.; Chacón, E.; López-Picazo, A.; Linares, M.; Pascual, M.Á.; Guerriero, S.; Alcázar, J.L. Ovarian Adnexal Reporting Data System (O-RADS) for Classifying Adnexal Masses: A Systematic Review and Meta-Analysis. *Cancers* **2022**, *14*, 3151. [[CrossRef](#)]
62. Lee, S.; Lee, J.E.; Hwang, J.A.; Shin, H. O-RADS US: A Systematic Review and Meta-Analysis of Category-specific Malignancy Rates. *Radiology* **2023**, *308*, e223269. [[CrossRef](#)]
63. Zhang, Q.; Dai, X.; Li, W. Systematic Review and Meta-Analysis of O-RADS Ultrasound and O-RADS MRI for Risk Assessment of Ovarian and Adnexal Lesions. *AJR Am. J. Roentgenol.* **2023**, *221*, 21–33. [[CrossRef](#)] [[PubMed](#)]
64. Dhombres, F.; Bonnard, J.; Bailly, K.; Maurice, P.; Papageorghiou, A.T.; Jouannic, J.M. Contributions of Artificial Intelligence Reported in Obstetrics and Gynecology Journals: Systematic Review. *J. Med. Internet Res.* **2022**, *24*, e35465. [[CrossRef](#)] [[PubMed](#)]
65. Jost, E.; Kosian, P.; Jimenez Cruz, J.; Albarqouni, S.; Gembruch, U.; Strizek, B.; Recker, F. Evolving the Era of 5D Ultrasound? A Systematic Literature Review on the Applications for Artificial Intelligence Ultrasound Imaging in Obstetrics and Gynecology. *J. Clin. Med.* **2023**, *12*, 6833. [[CrossRef](#)]
66. Tailor, A.; Jurkovic, D.; Bourne, T.H.; Collins, W.P.; Campbell, S. Sonographic prediction of malignancy in adnexal masses using an artificial neural network. *Br. J. Obstet. Gynaecol.* **1999**, *106*, 21–30. [[CrossRef](#)]
67. Timmerman, D.; Verrelst, H.; Bourne, T.H.; De Moor, B.; Collins, W.P.; Vergote, I.; Vandewalle, J. Artificial neural network models for the preoperative discrimination between malignant and benign adnexal masses. *Ultrasound Obstet. Gynecol.* **1999**, *13*, 17–25. [[CrossRef](#)]
68. Biagiotti, R.; Desii, C.; Vanzi, E.; Gacci, G. Predicting ovarian malignancy: Application of artificial neural networks to transvaginal and color Doppler flow US. *Radiology* **1999**, *210*, 399–403. [[CrossRef](#)]
69. Szpurek, D.; Moszynski, R.; Smolen, A.; Sajdak, S. Artificial neural network computer prediction of ovarian malignancy in women with adnexal masses. *Int. J. Gynecol. Obstet.* **2005**, *89*, 108–113. [[CrossRef](#)] [[PubMed](#)]

70. Xu, H.L.; Gong, T.T.; Liu, F.H.; Chen, H.Y.; Xiao, Q.; Hou, Y.; Huang, Y.; Sun, H.Z.; Shi, Y.; Gao, S.; et al. Artificial intelligence performance in image-based ovarian cancer identification: A systematic review and meta-analysis. *E Clin. Med.* **2022**, *17*, 101662.
71. Ma, L.; Huang, L.; Chen, Y.; Zhang, L.; Nie, D.; He, W.; Qi, X. AI diagnostic performance based on multiple imaging modalities for ovarian tumor: A systematic review and meta-analysis. *Front. Oncol.* **2023**, *13*, 1133491. [[CrossRef](#)]
72. Ismael, A.M.; Şengür, A. Deep learning approaches for COVID-19 detection based on chest X-ray images. *Expert. Syst. Appl.* **2021**, *164*, 114054. [[CrossRef](#)]
73. Li, C.; Zhang, H.; Chen, J.; Shao, S.; Li, X.; Yao, M.; Zheng, Y.; Wu, R.; Shi, J. Deep learning radiomics of ultrasonography for differentiating sclerosing adenosis from breast cancer. *Clin. Hemorheol. Microcirc.* **2023**, *84*, 153–163. [[CrossRef](#)] [[PubMed](#)]
74. Sethy, P.K.; Behera, S.K.; Anitha, K.; Pandey, C.; Khan, M.R. Computer aid screening of COVID-19 using X-ray and CT scan images: An inner comparison. *J. X-ray Sci. Technol.* **2021**, *29*, 197–210. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.