

CRITICAL REVIEW

Open Access



# Analysis of computer-aided diagnostics in the preoperative diagnosis of ovarian cancer: a systematic review

Anna H. Koch<sup>1\*</sup> , Lara S. Jeelof<sup>1</sup>, Caroline L. P. Muntinga<sup>1</sup>, T. A. Gootzen<sup>1</sup>, Nienke M. A. van de Kruis<sup>1</sup>, Joost Nederend<sup>2</sup>, Tim Boers<sup>3</sup>, Fons van der Sommen<sup>3</sup> and Jurgen M. J. Piek<sup>1</sup>

## Abstract

**Objectives** Different noninvasive imaging methods to predict the chance of malignancy of ovarian tumors are available. However, their predictive value is limited due to subjectivity of the reviewer. Therefore, more objective prediction models are needed. Computer-aided diagnostics (CAD) could be such a model, since it lacks bias that comes with currently used models. In this study, we evaluated the available data on CAD in predicting the chance of malignancy of ovarian tumors.

**Methods** We searched for all published studies investigating diagnostic accuracy of CAD based on ultrasound, CT and MRI in pre-surgical patients with an ovarian tumor compared to reference standards.

**Results** In thirty-one included studies, extracted features from three different imaging techniques were used in different mathematical models. All studies assessed CAD based on machine learning on ultrasound, CT scan and MRI scan images. Per imaging method, subsequently ultrasound, CT and MRI, sensitivities ranged from 40.3 to 100%; 84.6–100% and 66.7–100% and specificities ranged from 76.3–100%; 69–100% and 77.8–100%. Results could not be pooled, due to broad heterogeneity. Although the majority of studies report high performances, they are at considerable risk of overfitting due to the absence of an independent test set.

**Conclusion** Based on this literature review, different CAD for ultrasound, CT scans and MRI scans seem promising to aid physicians in assessing ovarian tumors through their objective and potentially cost-effective character. However, performance should be evaluated per imaging technique. Prospective and larger datasets with external validation are desired to make their results generalizable.

## Key Points

- Computer-aided diagnostics has potential to predict the nature of ovarian tumors.
- Literature shows heterogeneous sensitivity and specificity of machine learning on ultrasound images, CT-scan images and MRI-scan images.
- More prospective studies on other computer-aided techniques and imaging modalities should be performed with an external validation set.

**Keywords** Diagnosis, Computer-assisted, Machine learning, Ovarian neoplasms

\*Correspondence:

Anna H. Koch  
anna.koch@catharinaziekenhuis.nl

Full list of author information is available at the end of the article

## Introduction

An accurate preoperative diagnosis of an ovarian tumor into either benign, borderline or malignant is important for multiple reasons: (1) for the patients' surgical workup and treatment planning, (2) for the patients' mental well-being and (3) for correct use of diagnostic algorithms [1]. Currently, most women diagnosed with an ovarian tumor are initially evaluated with transvaginal ultrasound and serum CA125. For a more objective approach, different ultrasound-based models, to discriminate between benign, borderline and malignant ovarian tumors, have been constructed over time. One of the first widely used models is the risk of malignancy index (RMI) which combines five ultrasound variables with serum CA125 and postmenopausal status [2]. Other models have been developed by the International Ovarian tumor analysis (IOTA) group, such as the Assessment of Different NEoplasias in the adneXa (ADNEX) model, which combines six ultrasound features together with patients age, serum CA125 and type of center (oncology referral center vs other) [3, 4]. However, for both models the reported sensitivity lies around 98% and 71%, and the specificity around 85% and 62% [5]. In addition, two other classification models were introduced by radiologists and gynecologists: (1) the GI-RADs (Gynecologic Imaging Reporting and Data System) score, for diagnosis of adnexal masses by pelvic ultrasound, and (2) the O-RADS (Ovarian-adnexal reporting and data system) data system, both showing a sensitivity of 92.7% and 93.6% and a specificity of 96.8% and 92.8% [6, 7]. Nevertheless, research has shown that ultrasound features are often misclassified by inexperienced examiners [8].

Nowadays, preoperative computer tomography (CT) and/or magnetic resonance imaging (MRI) is performed to pre-surgically assess the nature of an ovarian tumor and to predict the presence of metastatic disease. MRI has proven to be able to discriminate between benign and malignant ovarian tumors with a sensitivity of 96% and a specificity of 91%. The O-RADS MRI has a sensitivity of 93% and a specificity of 91% for score 5 (malignant) with a comparable reading between senior and junior radiologists [7, 9, 10]. However, for spiral CT scans no diagnostic studies are available. Research conducted with multidetector CT scans shows an accuracy of 90 to 93% in adnexal mass characterization [11].

For clinicians, ideally, when using any test a 100% sensitivity and specificity is desired. For imaging prediction models, this means that no malignant tumors are missed and no benign tumors are classified as malignant to prevent unnecessary surgical procedures on benign ovarian tumors [12, 13]. Hence, diagnostic accuracy with a higher sensitivity at the detriment of the specificity is favorable. The currently used imaging prediction models show high

performance in ovarian tumor classification; nevertheless, they are greatly affected by subjective assessment and users' experience. Therefore, evaluation of more independent strategies to determine the nature of ovarian tumors among these different imaging modalities is needed.

Over the past three decades, several computer-aided diagnostics (CADs) have been developed for accurate ovarian cancer prediction, mainly on ultrasound, all using predefined hand-selected features to build their classifiers [14–16]. Computer-aided diagnostics is used to assist clinical professionals within different medical specialties, such as dermatology, neurology and pathology [17–20]. Furthermore, it can aid radiologists' image interpretations and extract features from medical images, which are not visible for the human eye, giving it a cost-effective potential as well [21]. Still, within the field of gynecologic oncology it is relatively new compared to other medical specialties [22].

In this study, we assess the available literature on CAD in preoperatively predicting the chance of an ovarian malignancy.

## Materials and methods

We searched for all published studies investigating diagnostic accuracy of CAD based on ultrasound, CT and MRI in patients with an ovarian tumor. Search terms used were: 'ovaria', 'ovarian neoplasms', 'ovarian neoplasm', 'ovarian masses', 'ovarian lesion', 'ovarian tumor', 'adnexal', 'adnexal mass', 'ovarian cancer', 'ovarian malignancy', 'ovary', 'classification of ovarian', 'machine learning', 'computer aided', 'Diagnosis Computer-Assisted', 'computer assisted-diagnosis', 'artificial intelligence', 'Neural Networks, Computer', 'convolutional neural network', 'radiomics', 'decision support system', 'decision support technique', 'decision support techniques', 'machine learning classifier', 'machine learning classifiers', 'diagnosis', 'diagnostic accuracy', 'presurgical', 'preoperative', 'preoperative diagnosis', 'preoperative evaluation', 'Tomography, X-ray Computed', 'ct-scan', 'ultrasound', 'echography', 'gynecological ultrasound', 'ultrasonography', 'magnetic resonance imaging', 'nuclear magnetic resonance imaging' and 'MRI'. We used 'title abstract' (tiab) and 'Mesh' added to each search term. The exact search syntax per database is provided in Additional file 1: Appendix 1.

The search was last performed on 9th 2022 by two independent reviewers and a research librarian was consulted for support in this matter.

We searched for papers published in English in Cochrane Central Register of Controlled Trials, MEDLINE, Embase, Scopus and PubMed. Additionally, we searched trial registries for ongoing and registered trials on Clinicaltrials.gov. To identify additional trials,

references of all included studies by the initial search were hand searched to add relevant trials.

All studies that investigated diagnostic accuracy of CAD based on ultrasound, CT and MRI images in patients with an adnexal mass were included. Case reports, summaries, animal studies, meta-analyses, comments, editorials, conference abstracts and other irrelevant article types were excluded.

### Selection of studies

Titles and abstracts retrieved by the search were imported into the reference manager database Covidence [23]. Duplicates were removed and two reviewers independently screened the records. Subsequently, full-text versions of potentially relevant studies were obtained and assessed for eligibility by the same researchers. Studies were qualified if the following criteria were met: (1) accurate disease type, e.g., benign, borderline or malignant ovarian tumors, (2) appropriate clinical setting, for example, no ex vivo studies, (3) description of overfitting techniques and reference standard, (4) use of correct classifier, e.g., none of the features selected to construct the CAD were manually measured, as done by Timmerman et al., Biagiotti et al. or Zimmerman et al. [14–16] and (5) diagnostic accuracy had to be reported, namely sensitivity, specificity or area under the curve (AUC). Disagreements were resolved through discussion until consensus was reached, or by consulting a third member of the review team. The selection process was visualized in a PRISMA flowchart (Fig. 1).

### Data extraction and management

Two reviewers independently extracted the following data from each included study: study design, year of publication, country where the study was conducted, inclusion and exclusion criteria or population description, number of participants, menopausal status, mean CA125 serum levels of included participants, number of images, intervention compared to histology, type of classifier and features used to develop the CAD, duration of follow-up, reference standard and results. When multiple classifiers were described, the best performing one was selected. Supplementary appendices were assessed for additional study details and corresponding authors were contacted by email on study details if necessary. Discrepancies were resolved through discussion and consensus, or by consulting a third member of the review team. Study outcomes were type of classifier, whether an external validation set was used, if CAD was compared to or combined with other models or subjective assessment (SA), sensitivity, specificity, accuracy and AUC, when mentioned in the included study. Other diagnostic accuracy values were also considered. We aimed to perform a

meta-analysis of the CAD methods that used an external validation set, for which *Review Manager* (RevMan) software (v5.4.1) and *Meta-DiSc* software were utilized [24]. Heterogeneity was assessed by using the I<sup>2</sup> statistics, which describes the percentage of variability due directly to heterogeneity, with >50% representing moderate heterogeneity and >75% indicating high heterogeneity and Moses-Littenberg SROC (summary receiver operating curve) plot [25, 26].

### Assessment of risk of bias of included studies

Two independent reviewers assessed the methodological quality of each included study by using the Prediction Model Study Risk of Bias Assessment Tool (PROBAST) together with additional questions from the quality assessment of diagnosis accuracy study (QUADAS-2) tool and the quality in prognostic studies (QUIPS) tool. Discrepancies were resolved through discussion and consensus, or by consulting a third member of the review team. Different risk of bias assessment tools were used because different types of study designs were included. Studies that evaluated multivariable diagnostic or prognostic prediction models were reviewed using PROBAST. PROBAST assesses four key domains: participants, predictors, outcome and analysis. Studies that evaluated diagnostic tests of prognostic factors were reviewed by a few questions from the QUADAS-2 tool and the QUIPS tool [27–30]. Furthermore, seven signaling questions were composed by independent technical members of the study team to assess risk of bias based on the used CAD model, called ‘CAD model risk of bias screening questions.’ These two members were not aware of the content of the articles included. These signaling questions are described in Additional file 1: Appendix 2.

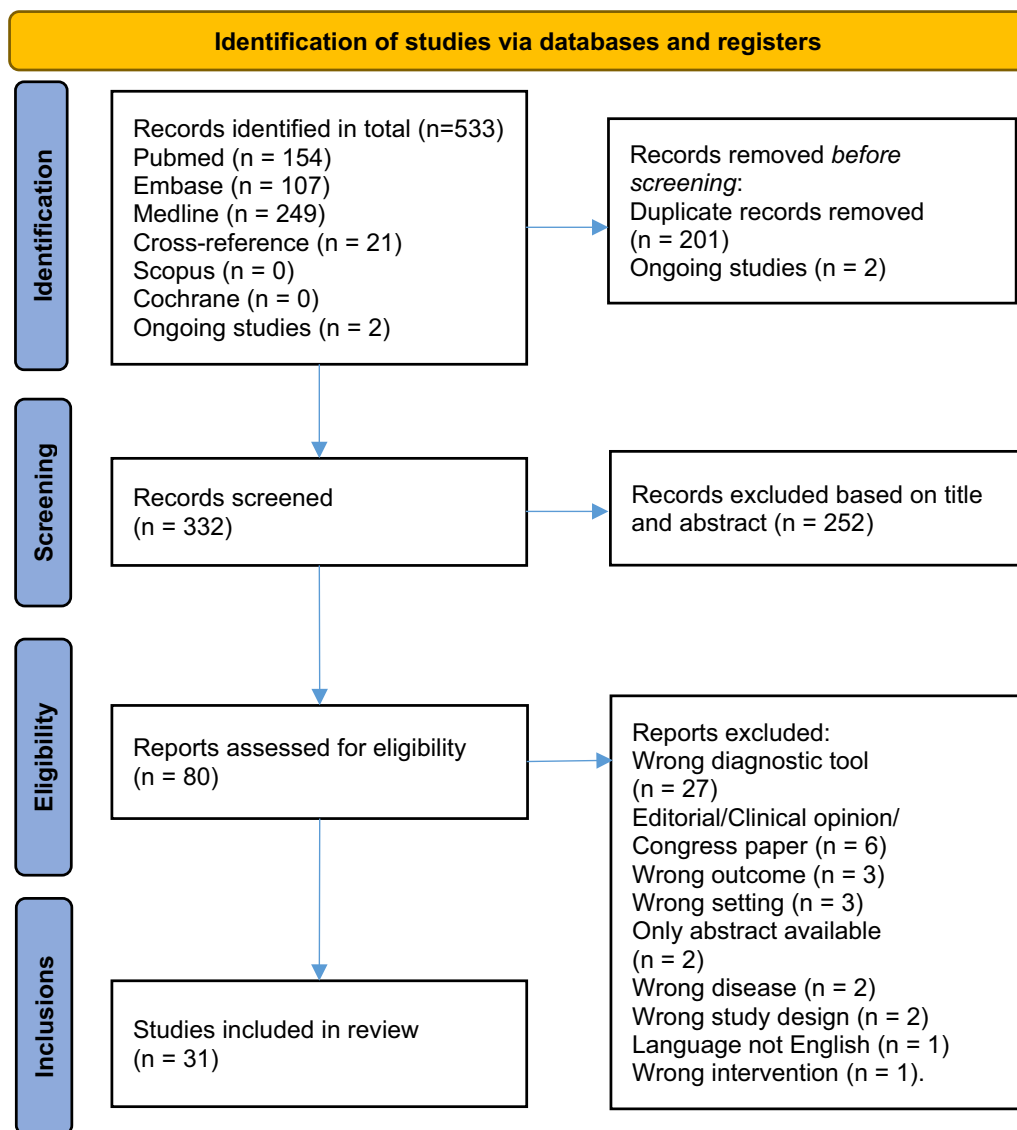
The signaling questions were used to determine whether risk of bias was low, high or unclear.

The extraction of study data, comparisons in data tables and preparation of a ‘Summary of findings’ table were performed before writing the results and conclusions of this review.

The protocol of this systematic review was registered with PROSPERO (Registration number CRD42020189910).

### Results

After the search was performed and cross-reference articles were added, a total of 532 articles were retrieved. Subsequently, duplicates were removed and 331 articles remained for screening on title and abstract. Seventy-one articles were eligible for full-text reading. Two studies on CAD and ovarian tumors were found on ClinicalTrials.gov. Both trials are open for accrual and are using CAD in diagnosing (1) malignant



**Fig. 1** PRISMA flowchart [32]

ovarian tumor with CT (NCT05174377) and (2) endometriosis-related ovarian cancer (NCT05161949). Most articles were excluded because they were not using CAD, not assessing ovarian tumors or because a wrong type of classifier was used. A summary of the selection process is shown in a PRISMA flowchart (Fig. 1) [31, 32].

After screening the title, abstract and full-text thirty-one studies were included in this systematic review.

#### Description of included studies

Thirty-one studies were included in this review, of which twenty-two ultrasound-based studies [33–53, 62], three

CT-based studies [54–56] and six MRI-based studies [57–62]. A detailed overview of the included studies is presented in Additional file 2: Table 1a–c. There were twenty-two retrospective studies of which nineteen are case–control studies and two are cohort studies. Six studies have a prospective case–control design, and one is a cohort study. Women of all ages were included in the studies. Only seven studies used external validation datasets to assess the performance of their classifier: four ultrasound, one CT and two MRI studies [33, 35, 51, 53, 56, 59, 61]. The same dataset was used in ten studies to develop and test different classifiers [43, 45, 49, 50, 52, 54, 56, 59, 61, 62]. In most studies, the region of interest

(ROI) was annotated manually. Two studies did not mention histology as definite diagnosis [41, 42]. Only three studies combined CAD with clinical features. Eleven studies compared the CAD with subjective assessment (SA) of a reviewer or combined the CAD model with SA performance of the reviewer [33, 35–38, 48, 50, 58, 60–62]. Table 1a–c presents the results of each study.

In the included studies, fourteen different machine learning modalities were employed: Seventeen were different types of deep machine learning, and the remaining were conventional machine learning. Fourteen studies used classification [34, 39–47, 50, 51, 53, 63], and remaining studies used segmentation to predict the nature of the ovarian tumor. With classification, a class label (e.g., benign or malignant) is predicted by analyzing its input, which is often numerical data (e.g., images). With segmentation, each pixel in an image is assigned to a predefined category (e.g., malignant or non-malignant), whereby certain image characteristics are shared by pixels with identical labels [64]. The input for the segmentation studies was usually different types of grayscale patterns, e.g., gray-level size zone matrix or wavelet features. The input for the classification studies was global images with or without clinical variables added.

#### Pooling of diagnostic accuracy

A meta-analysis on the seven studies that used an external validation set to test their CAD model was attempted; however, due to heterogeneity, missing diagnostic accuracy rates and unclear data, this could not be executed [33, 35, 51, 53, 56, 59, 61]. An additional sub-analysis of studies using CAD on ultrasound imaging was performed, which showed great heterogeneity as well. The remaining twenty-four studies without an independent validation per imaging modality were not pooled due to heterogeneity at forehand.

#### Risk of bias in studies per imaging modality

A general overview of risk of bias of per imaging modality of the included studies is presented in the ‘Risk of bias’ summary (Table 2a–c).

#### Ultrasound:

##### Participants

Risk of bias based on selection of participants was considered low in ten studies. In five studies, risk of selection bias was unclear, because inclusion of participants was not clearly described. Seven studies were graded with high risk of selection bias because they described neither little information about baseline patient characteristics nor inclusion or exclusion criteria.

#### Predictors

Risk of bias based on predictors was considered low for twenty studies, because predictors were defined and assessed in the same way for all participants and predictor assessments were made before results were known. For two studies, this was unclear due to missing information on this matter.

#### Outcome

Risk of bias based on outcome or its determination was considered low for eighteen studies, because in these studies the outcome was predetermined appropriately. Risk of bias was scored unclear in one study, because there was no clear description of the reference standards used and high in three studies, since reference standards were not described.

#### Analysis

Risk of bias based on analysis was considered low for eight studies, because analysis was properly performed. In three studies, risk of bias based on analysis was unclear, because analysis was not clearly described. Eleven studies described very little of the analysis process, and therefore, these studies were considered containing high risk of bias.

#### CAD model

Risk of bias based on CAD model bias screening questions was considered low in four studies. In nine studies, the risk of bias based on CAD model bias screening questions was assessed as unclear, because it was unclear how overfitting mitigation techniques and cross-validation were used or if the data were reproducible or validated in other centers. Risk of bias based on CAD model bias screening questions was reckoned high in nine studies. This was due to overfitting mitigation techniques which were not used or incorrectly used, the training set was not independent from the test set or did not have enough power, or no cross-validation was used and data were not reproducible or not validated in other settings.

#### CT and MRI

##### Participants

Risk of bias based on selection of participants was considered low in all nine studies, because of transparent description of patient selection.

**Table 1** Results depicted per image modality: ultrasound, CT and MRI

Included studies	Study setting	Patients (n)	Samples (n)	CAD-model	Features (n)	Performance ACC	Performance AUC	Performance Sensitivity	Performance Specificity	Performance Other	CAD model evaluation method	Compared to other models or reviewer(s)*
<i>α: CAD ultrasound (22)</i>												
Gao et al. [33]	Retrospective Case-control	107,624	575,930 images	DCNN	121 layers	(1) 86.9%	(1) 0.870	(1) 40.3%	(1) 91.6%	Brier-score	1 internal validation set	Radiologist alone (3) Radiologist with DCNN (4)
			103,370 benign			(2) 85.3%	(2) 0.831	(2) 57.8%	(2) 98.5%	F1-score	2 external validation set (1 + 2)	
			4254 malignant			(3) 81.1%	(3) N/A	(3) 55.5%	(3) 87.5%	PPV		
Chiappa et al. [34]	Retrospective Case-control	241	241 images	SVM	853	(4) 87.6%	(4) N/A	(4) 82.7%	(4) 88.7%	NPV	Training-Validation Testing Nested-tenfold validation	N/A
			115 benign			80.00%	0.83	78.00%	83.00%	N/A		
			126 malignant									
Chiappa et al. [35]	Retrospective & Prospective Case-control	274	274 images	DSS	306 motley	(1) 87.9%	N/A	(1) 99.2%	(1) 75.9%	PPV	External validation in prospective cohort (n = 35) tenfold cross validation	2 gynecologists with DCNN (1 + 2) on internal & external dataset
			239									
			123 benign			(2) 88.7%		(2) 98.4%	(2) 78.5%	NPV		
		35	116 malignant		4 clinical	(1) 91.4%		(1) 100.0%	(1) 80.0%			
			35			(2) 91.4%		(2) 100.0%	(2) 80.0%			
			15 benign									
			20 malignant									





**Table 1** (continued)

Included studies	Study setting	Patients (n)	Samples (n)	CAD-model	Features (n)	Performance ACC	Performance AUC	Performance Sensitivity	Performance Specificity	Performance Other	CAD model evaluation method	Compared to other models or reviewer(s)*
Wang et al. [38]	Retrospective Case-control	265	279 images 108 benign	VGG	N/A	(1) 91.4%	(1) 0.963	(1) 91.4%	(1) 91.4%	F1-score	Transfer learning	Sonographist (3) task C (benign-borderline-malignant)
			65 borderline 106 malignant	GoogleNet		(2) 75.3%	(2) N/A	(2) 80.0% / 45.5% / 88.9%	(2) 89.7% / 95.8% / 75.4%		threefold-cross validation	task A benign – malignant (1)
				<b>ResNet</b>		(3) 66.7%	(3) N/A	(3) 75.0% / 47.4% / 68.4%	(3) 81.8% / 85.2% / 82.5%			
				MobileNet task A + C (1) + (2)								
Martinez-Mas et al. [39]	Retrospective Case-control	187	384 images 112 benign	SVM	N/A	87.70%	0.874	92.00%	80.00%	N/A	LOO-CV	N/A
			75 malignant	KNN							N = 30	
				LD								
				<b>ELM</b>								
Zhang et al. [40]	Retrospective Case-control	N/A	428 images 357 malignant 71 benign 1400 images 277 malignant 299 benign	<b>Cost-sensitive RF</b>	N/A	99.20%	0.997	99.70%	95.60%	N/A	Transfer learning	N/A
				VGGNet							Training 71.5%	
				GoogleNet							Validation 14.3%	
				FCNN							Testing 14.3%	
				AlexNet							tenfold-cross validation	
Acharya et al. [41]	N/A	469	469	KNN	39	80.60%	0.806	81.40%	76.30%	N/A	tenfold-cross validation	N/A



Table 1 (continued)

Included studies	Study setting	Patients (n)	Samples (n)	CAD-model	Features (n)	Performance ACC	Performance AUC	Performance Sensitivity	Performance Specificity	Performance Other	CAD model evaluation method	Compared to other models or reviewer(s)*		
Aramendia-Vidaurreta et al. [46]	Cohort	145	238 suspicious 281 non-suspicious	RF										
				FF										
				FRNN										
	N/A	145	145 images 106 benign	MLP	40	98.80%	0.997	98.50%	98.90%	PPV	Training 80% Validation 10% Testing 10% tenfold cross validation	N/A		
	Case-control		39 malignant		1 clinical									
	Khazendar et al. [47]	Retrospective Cohort	177	187 images 112 benign	SVM	1	78.00%	N/A	80.00%	77.00%	T-test	Training and testing set	N/A	
LBP on enhanced image														
Acharya et al. *** [44]	Retrospective Case-control	20	75 malignant 10 benign	SVM	11	(1) 100%	N/A	(1) 100%	(1) 100%	N/A	Training and testing set	N/A		
				KNN										
Acharya et al. **** [42]	Prospective cohort	23	20 10 benign	PNN	23	99.81%	N/A	99.92%	99.69%	PPV	Training 90% Testing 10%	N/A		

**Table 1** (continued)

Included studies	Study setting	Patients (n)	Samples (n)	CAD-model	Features (n)	Performance ACC	Performance AUC	Performance Sensitivity	Performance Specificity	Performance Other	CAD model evaluation method	Compared to other models or reviewer(s)*	
Acharya et al. [45]	Prospective Case-control	10	10 malignant								tenfold-cross validation		
			2600 images 1300 benign 1300 malignant										
			20	DT	4	N/A	N/A	94.30%	99.70%	PPV	Training and testing set	N/A	
			10 benign							TP rate	tenfold cross validation		
			10 malignant							FP rate TN rate FN rate			
			2000 images 1000 benign 1000 malignant										
Faschingbauer et al. [48]	Retrospective Case-control	105	105	SVM-ABTA	(1) 16	(1) N/A	N/A	(1) 69%	(1) 86%	Youden-index	Training and testing set	Level III gynaecologists (5)	
			70 benign										
			35 malignant										
				Malignant (1)	(2) 16	(2) N/A		(2) 72%	(2) 81%		onefold cross validation		
				Dermoid cysts (2)	(3) 16	(3) N/A	(3) 82%	(3) 96%					
				Functional cysts (3) Overall (4)	(4) 16	(4) 74.3%	(4) N/A	(4) N/A					
Acharya et al. [43]	Retrospective Cohort	20	20	SVM-RBF	14	(5) 83.75% 99.90%	N/A	100%	99.80%	PPV	Training and testing set	N/A	
			10 benign							TP rate	tenfold cross validation		

**Table 1** (continued)

Included studies	Study setting	Patients (n)	Samples (n)	CAD-model	Features (n)	Performance ACC	Performance AUC	Performance Sensitivity	Performance Specificity	Performance Other	CAD model evaluation method	Compared to other models or reviewer(s)*
			10 malignant							FP rate TN rate FN rate		
			2000 images 1000 benign 1000 malignant									
Vaes et al. [49]	Prospective Case-control	197	291 adnexal masses	OVHS + RMI1	N/A	N/A	N/A	88%	95%	N/A	Training 70%	N/A
			125 benign 166 malignant	OVHS + RMI2 <b>OVHS + RMI3</b>							Testing 30% 100 times a random subsampling process Training 60%	
Vaes et al. [50]	Prospective Case-control	197	197 ultrasound images—365 ovarian tumors	LR (1)	(1) 9	N/A	(1) 0.97	(1) 83%	(1) 98%	N/A	Training 60%	RMI (3)
			77—normal 125—benign	NN (2)	(2) N/A (3) 7		(2) 0.93 (3) 0.80	(2) 80% (3) 69%	(2) 86% (3) 79%		Testing 40% 100 bootstrap resampled data sets with AICC selection	LR2 (4) NN2 (5)
			166—malignant		(4) 6		(4) 0.85	(4) 79%	(4) 70%			
Lucidarme et al. [52]	Prospective Case-control	264	375 ovaries	OVHS	(5) 7 N/A	N/A	(5) 0.87 N/A	(5) > 99% 98%	(5) 10% 88%	PPV	One group	N/A
			107 normal 127 benign 141 malignant							NPV TP rate FP rate		

Table 1 (continued)

Included studies	Study setting	Patients (n)	Samples (n)	CAD-model	Features (n)	Performance ACC	Performance AUC	Performance Sensitivity	Performance Specificity	Performance Other	CAD model evaluation method	Compared to other models or reviewer(s)*
Lu et al. [51]	N/A	425	359 sonographic opinion 104 normal ovaries 119 benign 136 malignant	SVM (1)	(1) 10	(1) 84.38%	(1) 0.918	(1) 85.19%	(1) 83.96%	PPV	Training 62%  Testing 38%	RMI (2)  LR1 (3)
	Case control		291 benign		(2) 7	(2) 76.88&	(2) 0.873	(2) 81.48%	(2) 74.53%	NPV		LR2 (4)
			134 malignant		(3) 12	(3) 80.63%	(3) 0.911	(3) 81.48%	(3) 80.19%		1 internal test set 1 external validation set 30-fold cross-validation	
					(4) 6	(4) 78.75%	(4) 0.916	(4) 81.48%	(4) 77.36%			
Zimmer et al. [53]	Retrospective Case-control	163	163 images	Bayes method	4	82.10%		80%	100%	PPV	Training 85%  External validation 15%	N/A
by CADCT (3) Li et al. [54]	Retrospective Case-control	140	25 transparent cyst  67 turbid cyst 50 significantly solid 21 solid	Radiomics segmentation models	(1) 10	(1) 97.6%	(1) 0.99	(1) 95.7%	(1) 100%	N/V	Training 61%  Testing 29%	N/A
			62 benign 72 malignant		4 clinical	(2) 90.2%	(2) 0.97	(2) 100%	(2) 82.6%			
					(2) 11							

Table 1 (continued)

Included studies	Study setting	Patients (n)	Samples (n)	CAD-model	Features (n)	Performance ACC	Performance AUC	Performance Sensitivity	Performance Specificity	Performance Other	CAD model evaluation method	Compared to other models or reviewer(s)*	
Park et al. [55]	Retrospective Case-control	427	427	RF	5 Clinical 8	N/A	0.88	91%	69%	N/A	tenfold cross validation	N/A	
			348 benign	LR									
			79 malignant										
			160 images	Nomogram (int. val)		14	(1) 89.7%	(1) 0.897	(1) 94.7%	(1) 85.0%	N/A	Training 59%	N/A
Li et al. [56]	Retrospective Case-control	160	134	Nomogram (ext. val)							Testing 24%		
			62 benign		(2) 88.0%	(2) 0.880	(2) 84.6%	(2) 91.7%		External validation 17%			
			72 malignant							tenfold cross validation			
			External dataset N/A										
c: CAD/MRI (6) Liu et al. [57]	Retrospective Case-control	196	196	Radiomics segmentation	(1) 396	(1) 99.0%	(1) 1.0	(1) 100%	(1) 98.0%	PPV	Random Training 50% Testing 50%	N/A	
			91 borderline	models*									
			10 malignant	3D sagit (1) 2D coron (2)	(2) 396	(2) 78.9%	(2) 0.82	(2) 72.9%	(2) 85.1%		NPV		
			104	PK-model	(1) 7	(1) 84.2%	N/A	(1) 66.7%	(1) 100%	N/A	Training 70% Validation 30% 3-class classification task	Radiologists (2)	
Song et al. [58]	Prospective Case-control	82	33 benign		(2) N/A	(2) 68.4%		66.70%	93.80%		50-fold cross-validation	benign	
			18 borderline					70%	77.80%			borderline	

Table 1 (continued)

Included studies	Study setting	Patients (n)	Samples (n)	CAD-model	Features (n)	Performance ACC	Performance AUC	Performance Sensitivity	Performance Specificity	Performance Other	CAD model evaluation method	Compared to other models or reviewer(s)*
Jian et al. ** [59]	Retrospective Case-control	501	53 malignant					(2) 66.7% 66.70%	(2) 92.3% 81.3%			malignant
			501	MICNN	512	76.70%	0.884	70% 74.80%	77.80% 80.80%	F1 score	Training 68%	N/A
			165 border-line 336 malignant	EMP LMP							(centers A-B) External validation set 32% (centers C-H)	
Jian et al. *** [62]	Retrospective Case-control	501	22,977	MICNN MAC-net	512	82.70%	0.878	N/A	N/A	F1 score	Training 76% Validation 23%	N/A
			501									
			165 border-line 336 malignant									
Li et al. [61]	Retrospective Case-control	501	501	MP-ST (1)	(1) 851	(1) N/A	(1) 0.920	(1) N/A	(1) N/A	N/A	Training 50%	Radiologists (3)
			165 border-line	CE-T1WI (2)	(2) 851	(2) N/A	(2) 0.801	(2) N/A	(2) N/A		Internal validation 18%	
			336 malignant		(3) N/A	(3) N/A	(3) 0.797	(3) 80.5%	(3) 78.9%		(centers A-B) External validation 32% (centers C-H)	
Zhang et al. [60]	Retrospective Case-control	280	72 benign	SVM (b-m) (1)	(1) 84	(1) 90.6%	(1) 0.9670	(1) 90.3%	(1) 91.3%	PPV	Randomly	Radiologists (3)

**Table 1** (continued)

Included studies	Study setting	Patients (n)	Samples (n)	CAD-model	Features (n)	Performance ACC	Performance AUC	Performance Sensitivity	Performance Specificity	Performance Other	CAD model evaluation method	Compared to other models or reviewer(s)*
			100 type I EOC	SVM (l-l) (2)	(2) 56	(2) 83.3%	(2) 0.8228	(2) 76.5%	(2) 86.5%	NPV	LOOCV 70%	
			81 type 2 EOC		(3) N/A	(3) 83.5%	(3) N/A	(3) 82.3%	(3) 86.9%	TP rate FP rate TN rate FN rate	Testing 30%	

AUC = Area Under the Curve; PPV = positive predictive value; NPV = negative predictive value; SVM = standard vector machine; DCNN = (deep) Convolutional Neural Network; N/A = not applicable; DSS = decision support system, based on 3 radiomics models VGGNet, ResNet, MobileNet; SA = subjective assessment of an expert (gynaecologist/sonographer); SR = IOTA Simple Rules model; SRL = IOTA simple rules risk model; IDI = integrated discrimination improvement; KNN = k-nearest neighbor; LD = Linear Discriminant; ELM = Extreme Machine Learning (\*\*\*linear-gaussian in this example); LOO-CV = Leave-One-Out Cross Validation procedure; FCNN = Fully Connected Convolutional Neural Network; RF = Random Forest; FRNN = Fuzzy-Rough Nearest Neighbor; FF = fuzzy forest; MLP Multilayer Perceptron Networks; LBP = Local Binary Pattern; PNN = Probabilistic Neural Network; DT = Decision Tree; ABTA = automatic texture based algorithm; RBF = Radial Basis Function; OVHS = Ovarian Histoscanning; RMI = Risk of Malignancy; LR = Logistic Regression; NN = Neural Network; AICC = Akaike information corrected criterion; **Bold = best performing classifier**

N/A = not applicable; LR = logistic regression; RF = random forest; **Bold = best performing classifier**

MICNN = Multiple instance convolutional neural network; EMP = early multiparametric; LMP = late multiparametric; PK model = pharmacokinetic model; MP-solid = multiparametric solid tumor model; CE-T1WI = Contrast-enhanced T1WI model; **Bold = best performing**

\* = radiologist, gynaecologist, sonographer or other(s)

\*\* = Unable to split data set in 70% and 30% training and validation sets, due to limited number malignant tumors, therefore classifier was run twice with different variables

\*\*\* = Acharya et al. [44]—GyneScan: An improved online paradigm for screening of ovarian cancer via tissue characterization

\*\*\*\* = Acharya et al. [42]—Evolutionary algorithm-based classifier parameter tuning for automatic ovarian cancer tissue characterization and classification

\* Different segmentation models were constructed using 3D and 2D MRI in coronal and sagittal plane;

\*\*Jian et al. [59]—MRI-Based Multiple Instance Convolutional Neural Network for Increased Accuracy in the Differentiation of Borderline and Malignant Epithelial Ovarian Tumours;

\*\*\*Jian et al. [64]—Multiple instance convolutional neural network with modality-based attention and contextual multi-instance learning pooling layer for effective differentiation between borderline and malignant epithelial ovarian tumours



**Table 2** 'Risk of bias' summary: review authors' judgements about each risk of bias item for each included study

	Participants	Predictors	Outcome	Analysis	CAD model
<i>'Risk of bias' summary: Per item for each included study—Ultrasound</i>					
Gao et al. [33]	Low	Low	Low	High	Low
Chiappa et al. [34]	Low	Low	Low	Unclear	Unclear
Chiappa et al. [35]	Low	Low	Low	Low	Unclear
Christiansen et al. [36]	Low	Low	Low	Low	Low
Qi et al. [37]	Low	Low	Low	Low	High
Stefan et al. [63]	Unclear	Low	Low	Low	High
Wang et al. [38]	Low	Low	Low	Unclear	Unclear
Martinez-Mas et al. [39]	High	Low	Low	High	High
Zhang et al. [40]	High	Unclear	High	High	Low
Acharya et al. [41]	High	Low	Low	High	High
Aramendia-Vidaurreta et al. [46]	Unclear	Low	Low	Low	Unclear
Khazendar et al. [47]	Unclear	Low	Low	Low	High
Acharya et al. [44]	High	Low	Low	High	High
Acharya et al. [42]	High	Low	High	High	High
Acharya et al. [45]	Unclear	Unclear	Low	High	High
Faschingbauer et al. [48]	Low	Low	Low	High	Unclear
Acharya et al. [43]	High	Low	High	High	Unclear
Vaes et al. [49]	Unclear	Low	Low	High	Unclear
Vaes et al. [50]	Low	Low	Low	Low	Unclear
Lucidarme et al. [52]	Low	Low	Low	Unclear	High
Lu et al. [51]	Low	Low	Low	Low	Low
Zimmer et al. [53]	High	Low	Unclear	High	Unclear
<i>'Risk of bias' summary: Per item for each included study—CT</i>					
Li et al. [54]	Low	Low	Low	Low	Low
Park et al. [55]	Low	Low	High	Low	Low
Li et al. [56]	Low	Low	Low	Low	Low
<i>'Risk of bias' summary: Per item for each included study—MRI</i>					
Liu et al. [57]	Low	Low	Low	Low	Unclear
Song et al. [58]	Low	Low	Low	High	Unclear
Jian et al. [59]	Low	Low	Low	Low	Low
Jian et al. [62]	Low	High	Low	Unclear	Unclear
Li et al. [61]	Low	Low	Low	High	Low
Zhang et al. [60]	Low	Low	Low	Low	Low

**Predictors**

Risk of bias based on predictors was regarded low for eight studies, because the researchers were clear on how predictors were determined and characterized before outcome was known. Only one study reported that the outcome was known at forehand when assessing the predictors.

**Outcome**

Risk of bias based on outcome or its determination was considered low for eight studies, because it was predetermined appropriately. For one study, this was high because outcome differed among participants.

**Analysis**

Risk of bias based on analysis was considered low for six studies, because analysis was accurately carried out. In two studies, risk of bias based on analysis was high, because analysis was not properly performed. In one study, the sub-results were inconclusive and therefore the study was considered containing an unclear risk of bias.

**CAD model**

Risk of bias based on CAD model bias screening questions was assessed low in six studies. In three studies, risk of bias based on CAD model bias screening questions was considered unclear, because the use of overfitting

mitigation techniques was not mentioned or they were not executed correctly, and it was unclear if executed correctly it was unclear if the dataset was reproducible or validated in other settings.

## Discussion

This systematic review shows numerous studies that use CAD to assess the nature of an ovarian tumor. Due to the large heterogeneity, we were not able to pool data. However, highest performance as measured by AUC was seen in both CT- and MRI-based CAD models.

A meta-analysis was endeavored for the seven studies that used an external dataset for validation. However, this could not be executed for multiple reasons. One study, describing a CAD-MRI model for differentiating borderline from malignant ovarian tumors, only mentioned the sensitivity and specificity for radiologists' performance and for the model only the AUC [61]. Another study was unclear about which data were used to calculate the diagnostic performance of their model [56]. Consequently, for both studies it was not possible to calculate diagnostic accuracy rates, such as true positive (TP), true negative (NT) values and to use them in the meta-analysis.

For the five remaining studies, heterogeneity proved to be too large with an I2 of 92.8% and 90.7%. In an additional subgroup analysis of only ultrasound CAD models, this was also apparent with an I2 of 94.3% and 83.5%. These analyses can be found in Additional file 1: Appendix 3. This heterogeneity can be explained by (1) different types of CAD models using either conventional or deep learning techniques, (2) different inclusion and exclusion criteria and (3) type of imaging modality used. Among the twenty-four studies without an independent dataset, pooling of the results was not viable since the data were too diverse. This was illustrated by differences in imaging techniques used, e.g., 2D or 3D ultrasound and CT, or 2D, 3D or pharmacokinetic MRI. Furthermore, different CAD techniques were applied, e.g., conventional and deep learning machine learning models. Moreover, some studies combined clinical features such as patients' age, menopausal status or serum CA125 to support the classifiers. Finally, different outcome measurements per classifier were found, such as benign, malignant and borderline in combination with a different tumor subtype, such as mucinous ovarian tumors.

All studies assessed computer-aided diagnostics based on machine learning. We found that classifying the nature of an ovarian tumor by CAD on ultrasound images results in sensitivities of 40.3% to 100% and specificities of 76.3% to 100%. For CT, sensitivities of 84.6% to 100% and specificities of 69% to 100% were described.

For MRI, sensitivities and specificities ranged between 66.7% and 100% and 77.8% and 100%, respectively. Even though some studies report high performances, they are at risk for overfitting due to the lack of an independent test set. Twenty-three studies lacked an independent test set for evaluating model performance.

With conventional machine learning techniques, features extracted from medical imagery are used to optimize a mathematical model for predicting new, unseen data. A model should be built based on a training set of images and validated in a test set. If the model is too tightly fitted to the training data and does not generalize toward new data, it is called overfitting. Overfitting occurs more often with conventional machine learning, where many parameters are hand-selected instead of being learned from the data, especially when the model is not validated on an independent test set [64].

## Ultrasound

Earlier published studies assessing ultrasound prediction models show reasonable sensitivity (72–77%) and specificity (85–89%) for the RMI [65, 66]. An external validation of the IOTA ADNEX model showed a better performance, with a sensitivity of 98% (95% CI 93–100%), but with low specificity of 62% at a cutoff value for malignancy of 10% (95% CI 55%–68%) [5]. The GI-RADs and the O-RADs perform better with a sensitivity of 92.7% and 93.6% and a specificity of 97.5% and 92.8%, subsequently [6]. However, all these models depend on specific terminology and expertise of their users. Furthermore, interpretation of ultrasound imaging regarding ovarian tumors has shown to be difficult for novel clinicians and for clinicians who do not perform ultrasonography on a regular basis [8, 9]. Based on the amount of studies included in this review assessing the CAD technique for ultrasound, CAD can be a promising tool to aid clinicians in determining the origin of ovarian tumors. Moreover, when comparing CAD models' performances with experienced clinicians or existing models they achieve similar or even better diagnostic accuracy. Nevertheless, this performance comparison was performed in only three studies. Even though overfitting mitigation techniques were applied in twenty-one ultrasound studies, only four studies used external validation. Thus, a high risk of overfitting is present, which could lead to an unreliable performance.

## CT

The diagnostic performance of CT in preoperatively classifying the origin of an ovarian tumor is primary known for multidetector computer tomography (MDCT), with a diagnostic accuracy of 90–93% [11]. Therefore, no fair comparison on CAD for CT can be made. However, the

performance of CAD for CT is indeed promising based on the included studies in this review. The models show a high diagnostic accuracy and low selection bias. Nonetheless, only three studies in total assessed CAD for CT of which only one study utilized an independent validation, thus risking overfitting.

For CAD on CT scans, more research is needed to further evaluate its potential benefits.

## MRI

The diagnostic accuracy for MRI in ovarian tumor classification has a sensitivity and specificity of 96% and 91%, respectively [7, 9]. For the O-RADs MRI score, this is comparable with a sensitivity of 93% and a specificity of 91% and it shows a similar performance among junior and senior radiologists ( $\kappa=0.784$ ; 95% CI, 0.743–0.824) [9, 10]. CAD for MRI as an additional diagnostic method for ovarian tumors has the potential to aid radiologists due to its high diagnostic performance as a single model or when compared to SA of radiologists. However, caution is needed when using MRI-CAD as a supplementary tool. First, due to the absence of international guidelines when to conduct an MRI for ovarian tumors classification a selection bias is being created. Moreover, the performance of the MRI has no further clinical consequences for the patient. However, if radiologists are trained with MRI O-RADs classification model, the usage of MRI can have an additional beneficial effect on ovarian tumor classification, especially when classifying benign and or possibly malignant lesions [67]. However, for the O-RADs MRI familiarity and expertise are essential to use the scoring system [7, 10].

Second, only one out of six studies showed a low overall risk of bias on using MRI CAD [59]. Unfortunately, the authors did not compare their CAD to ovarian tumor characterization by radiologists or to other models, such as the O-RADs model. Hence, one study alone cannot support clinical implementation of MRI CAD. Moreover, although in three studies CAD outperformed the radiologists' performance, no external validation sets were used in these studies and risk of bias was mostly unclear [58, 59, 60, 61]. Furthermore, only two of the six studies used an external validation set [59, 61]. Another study used 3D MRI for their model building, showing good results; however, this is a rather expensive MRI technique [57]. Finally, two studies used the same dataset. Therefore, only limited evidence to support the usage of MRI-CAD additionally is available [59, 63].

Hence, more studies should be undertaken with external validation sets in order to be able to implement these CAD-MRI models in clinical practice.

## Trends among publications

Over the last three decades, different trends among included studies in the CAD field are observed.

An increasing number of publications presented clear inclusion and exclusion criteria for data before using it to construct a CAD model [33–39, 44, 48–50, 52, 54–63]. In addition, more studies used statistical tests to select the most promising features to include into the CAD model and in most articles this was precisely described [34, 37, 41, 43–46, 48–51, 54–58, 60–63]. Furthermore, study cohorts became substantially larger [33]. Finally, clinicians are more involved in the CAD model construction, e.g., for the delineation of the images. Thus, uniformity among studies is improved, making studies more comparable.

Regarding the outcomes, almost all studies used the same outcome measurements, i.e., sensitivity, specificity, accuracy and area under the curve (AUC). More connection with the clinical setting is observed. In particular, the comparison of the CAD model to either assessment of scans by clinicians such as radiologists, sonographers or gynecologists or to commonly used models in ultrasound (RMI or LR1-2) is now included [33, 35–38, 48, 51, 60, 61, 68].

Hence, the difficult technical matter of a CAD model development is made more comprehensible for clinicians.

Finally, more deep learning models have been developed in recent years, showing the potential of this new type of CAD. If these trends continue, it could substantially contribute to patient care.

Previous studies have shown that depending on the imaging technique used the interobserver agreement is low for many features and are prone to contain significant measurement errors when used by inexperienced clinicians. Therefore, more uncertainties in measured features within these imaging techniques can lead to diminished accuracies of a model. It is therefore important to develop new techniques with less inter- and intra-observer variability to reach higher test performances to prevent unnecessary referrals to tertiary centers and unnecessary stress for the patient. Based on this literature review, computer-aided ultrasound, CT and MRI techniques based on different (deep) neural networks and conventional machine learning techniques such as support vector machines are promising. They can either be used as a single entity or combined with SA or with other prediction models. They could potentially offer a noninvasive and cost-effective method in the future. However, this is only shown in eight studies of which five are ultrasound studies and three MRI studies. Of these studies, four used independent validation sets, of which three within ultrasound CAD and one within an MRI CAD. For the remaining studies, lack of a validation cohort might

cause a high risk of overfitting. The CT CAD models seem to perform fairly but they consist of small datasets and are in the absence of a SA and only one study used an external validation set; therefore, risk of overfitting is present.

Furthermore, CAD as a technique within the gynecology–oncology is slowly gaining field in comparison with other oncology specialties. Combining datasets with larger test sets is needed in prospective cohorts [22, 33, 69].

It is likely that deep learning in assessing the nature of an ovarian tumor will reach higher test performances than traditional machine learning. For MRI and CT, the number of studies in this review is limited and needs to be broadened [22].

### Strengths and weaknesses

To the best of our knowledge, this is the most comprehensive review on computer-aided diagnostics for differentiating benign from borderline and malignant ovarian tumors on ultrasound, MRI and CT scans. We have worked by a clearly defined protocol that was first submitted to PROSPERO, to provide transparency in the review process and avoid reporting bias. There was no substantial disagreement in inclusion of articles by the authors, and this can be regarded as a strong point in the review process. A meta-analysis of the studies with an external validation set was attempted. A limitation of this review is the heterogeneity between studies, the lack of independent validation sets and comparison with SA.

### Conclusions

In conclusion, this review shows that CAD certainly has potential as a noninvasive model to preoperatively predict whether an ovarian tumor is benign, borderline or malignant and thus can aid the physician with assessment of ovarian tumors. However, this depends on the type of imaging modality assessed and thus should be evaluated per imaging technique. CAD for CT displays the best performance overall. However, the three studies included are all lacking an external validation. The results of CAD for MRI were similar; however, more studies used external validation to test their CAD. Nevertheless, the risk of bias for the domain ‘CAD model’ for half of the studies was found to be unclear. Furthermore, it is important to take into account that MRI is clinically less relevant for detecting and classifying ovarian tumors. Finally, most research has been done on CAD for ultrasound, of which the results are reasonable in comparison with existing models, but has limited external validation and risks overfitting. Moreover, included studies per image modality show great heterogeneity, and thus, results most likely cannot be generalized to other data.

Studies in which all methods are validated in the same population should be performed in order to prove which techniques demonstrate the best diagnostic performance. Above all, it is important that new CAD techniques are tested and validated with an independent, prospectively collected dataset.

### Future perspectives

In the near future, it is likely that CAD will facilitate diagnostics and will be used as a decision support system by clinicians, depending on the imaging modality the CAD is developed for. The performance of CAD for discriminating the nature of an ovarian tumor on CT and MRI is good, and studies assessing these two imaging techniques show a low risk of bias. Consequently, a majority of research should focus on these two imaging modalities. Particularly, since both MRI and CT are more standardized than ultrasound imaging and therefore more suitable for CAD development. However, it should be taken into account that MRI is less clinically relevant in diagnosing ovarian tumors. In addition, in order to increase accuracy, CAD for CT or MRI could be combined with clinical markers, e.g., menopausal age or liquid biopsies, such as circulating cell free tumor DNA (ct-DNA). Implementation of CAD for ultrasound in clinical practice will presumably take longer due to the dynamic character of this imaging method and the high and unclear risk of bias.

### Abbreviations

ADNEX model	Assessment of different NEoplasias in the adneXa
CA125	Cancer antigen 125
CAD	Computer-aided diagnostics
CT	Computer tomography
ct-DNA	Circulating cell-free tumor DNA
GI-RADS	Gynecologic imaging reporting and data system for diagnosis of adnexal masses (AMs) by pelvic ultrasound (US)
MDCT	Multidetector computer tomography
MRI	Magnetic resonance imaging
O-RADS	Ovarian-adnexal reporting and data system
PROBAST	Prediction Model Study Risk of Bias Assessment Tool
QUADAS-2	Quality Assessment of Diagnosis Accuracy Study
QUIPS	Quality in prognostic studies
RMI	Risk of malignancy index
ROI	Region of interest
SA	Subjective assessment
SROC	Summary receiver operating curve
TP	True positive
TN	True negative

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13244-022-01345-x>.

**Additional file 1.** Appendix 1. Search syntax. Appendix 2. Signaling questions. Appendix 3. Meta-analysis results.

**Additional file 2.** Tables 1a–c.

## Author contributions

JMJP and JN contributed to conceptualization; CLPM, NMK, JL and JMJP were involved in protocol finalization; CLPM and TB contributed to bias screening questions; TB and FvdS were involved in technical contributions; AHK, TAG, NMK, LJ and JMJP contributed to article selection; JMJP was involved in article consensus; AHK, CLPM and LJ contributed to manuscript writing; AHK and TAG were involved in visualization; CLPM, LJ and JMJP contributed to writing—review and editing; JMJP and FvdS were involved in supervision. All authors read and approved the final manuscript.

## Funding

The authors state that this work has not received any funding.

## Availability of data and materials

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

## Author details

<sup>1</sup>Department of Gynaecology and Obstetrics and Catharina Cancer Institute, Catharina Hospital, 5623 EJ Eindhoven, Noord-Brabant, The Netherlands.

<sup>2</sup>Department of Radiology, Catharina Hospital, 5623 EJ Eindhoven, Noord-Brabant, The Netherlands. <sup>3</sup>Department of Electrical Engineering, VCA Group, University of Technology Eindhoven, 5600 MB Eindhoven, Noord-Brabant, The Netherlands.

Received: 17 May 2022 Accepted: 5 December 2022

Published online: 15 February 2023

## References

- Geomini PM, Kruitwagen RF, Bremer GL, Massuger L, Mol BW (2011) Should we centralise care for the patient suspected of having ovarian malignancy? *Gynecol Oncol* 122(1):95–99. <https://doi.org/10.1016/j.ygyno.2011.03.005>
- Jacobs I, Oram D, Fairbanks J, Turner J, Frost C, Grudzinskas JG (1990) A risk of malignancy index incorporating CA 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer. *Br J Obstet Gynaecol* 97:922–929. <https://doi.org/10.1111/j.1471-0528.1990.tb02448.x>
- Van Calster B, Van Hoorde K, Valentin L et al (2014) 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* 349:g5920. <https://doi.org/10.1136/bmj.g5920>
- Van Calster B, Valentin L, Froyman W et al (2020) Validation of models to diagnose ovarian cancer in patients managed surgically or conservatively: multicentre cohort study. *BMJ* 370:m2614. <https://doi.org/10.1136/bmj.m2614>
- Meys EMJ, Jeelof LS, Achten NMJ et al (2017) Estimating risk of malignancy in adnexal masses: external validation of the ADNEX model and comparison with other frequently used ultrasound methods. *Ultrasound Obstet Gynecol* 49(6):784–792. <https://doi.org/10.1002/uog.17225>
- Basha MAA, Metwally MI, Gamil SA et al (2021) Comparison of O-RADS, GI-RADS, and IOTA simple rules regarding malignancy rate, validity, and reliability for diagnosis of adnexal masses. *Eur Radiol* 31(2):674–684. <https://doi.org/10.1007/s00330-020-07143-7>
- Timmerman D, Planchamp F, Bourne T et al (2021) ESGO/ISUOG/IOTA/ESGE Consensus Statement on pre-operative diagnosis of ovarian tumors. *Int J Gynecol Cancer* 31(7):961–982. <https://doi.org/10.1136/ijgc-2021-002565>
- Meys E, Rutten I, Kruitwagen R, et al. Simple Rules, Not So Simple: The Use of International Ovarian Tumor Analysis (IOTA) Terminology and Simple Rules in Inexperienced Hands in a Prospective Multicenter Cohort Study. *Ultraschall Med.* 2017;38(6):633–641. "Simple Rules" - nicht so einfach: Anwendung der "International Ovarian Tumor Analysis" (IOTA)-Terminologie und der "Simple Rules" in unerfahrenen Händen in einer prospektiven multizentrischen Kohortenstudie. doi:<https://doi.org/10.1055/s-0043-113819>
- Shimada K, Matsumoto K, Mimura T et al (2018) Ultrasound-based logistic regression model LR2 versus magnetic resonance imaging for discriminating between benign and malignant adnexal masses: a prospective study. *Int J Clin Oncol* 23(3):514–521. <https://doi.org/10.1007/s10147-017-1222-y>
- Thomassin-Naggara I, Poncelet E, Jalaguier-Coudray A et al (2020) Ovarian-adnexal reporting data system magnetic resonance imaging (O-RADS MRI) score for risk stratification of sonographically indeterminate adnexal masses. *JAMA Netw Open.* 3(1):e1919896. <https://doi.org/10.1001/jamanetworkopen.2019.19896>
- Mukhtar S, Khan SA, Hussain M, Adil SO (2017) Role of multidetector computed tomography in evaluation of ovarian lesions in women clinically suspected of malignancy. *Asian Pac J Cancer Prev* 18(8):2059–2062. <https://doi.org/10.22034/apjcp.2017.18.8.2059>
- Walker SP, The ROC (2019) Curve redefined—optimizing sensitivity (and specificity) to the lived reality of cancer. *N Engl J Med* 380(17):1594–1595. <https://doi.org/10.1056/NEJMp1814951>
- Lange RT, Lippa SM (2017) Sensitivity and specificity should never be interpreted in isolation without consideration of other clinical utility metrics. *Clin Neuropsychol* 31(6–7):1015–1028. <https://doi.org/10.1080/13854046.2017.1335438>
- Biagiotti R, Desii C, Vanzi E, Gacci G (1999) Predicting ovarian malignancy: application of artificial neural networks to transvaginal and color Doppler flow US. *Radiology* 210(2):399–403. <https://doi.org/10.1148/radiology.210.2.r99fe18399>
- Timmerman D, Verrelst H, Bourne TH et al (1999) Artificial neural network models for the preoperative discrimination between malignant and benign adnexal masses. *Ultrasound Obstet Gynecol* 13(1):17–25. <https://doi.org/10.1046/j.1469-0705.1999.13010017.x>
- Zimmer Y, Tepper R, Akselrod S (1999) Computerized quantification of structures within ovarian cysts using ultrasound images. *Ultrasound Med Biol* 25(2):189–200. [https://doi.org/10.1016/s0301-5629\(98\)00150-1](https://doi.org/10.1016/s0301-5629(98)00150-1)
- Chilamkurthy S, Ghosh R, Tanamala S et al (2018) Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *Lancet* 392(10162):2388–2396. [https://doi.org/10.1016/s0140-6736\(18\)31645-3](https://doi.org/10.1016/s0140-6736(18)31645-3)
- Esteve A, Kuprel B, Novoa RA et al (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542(7639):115–118. <https://doi.org/10.1038/nature21056>
- Ehteshami Bejnordi B, Veta M, Johannes van Diest P et al (2017) Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 318(22):2199–2210. <https://doi.org/10.1001/jama.2017.14585>
- Lakhani P, Sundaram B (2017) Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* 284(2):574–582. <https://doi.org/10.1148/radiol.2017162326>
- Lambin P, Leijenaar RTH, Deist TM et al (2017) Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol* 14(12):749–762. <https://doi.org/10.1038/nrclinonc.2017.141>
- Mysona DP, Kapp DS, Rohatgi A et al (2021) Applying artificial intelligence to gynecologic oncology: a review. *Obstet Gynecol Surv* 76(5):292–301. <https://doi.org/10.1097/ogx.0000000000000902>
- Veritas Health Innovation M, Australia. Covidence systematic review software. website. Covidence systematic review software, Veritas Health Innovation. Updated 2022. Accessed 09–05–2022, <https://www.covidence.org/>
- Zamora J, Abaira V, Muriel A, Khan K, Coomarasamy A (2006) Meta-DiSc: a software for meta-analysis of test accuracy data. *BMC Med Res Methodol* 6:31. <https://doi.org/10.1186/1471-2288-6-31>



25. Higgins JP, Thompson SG, Deeks JJ, Altman DG (2003) Measuring inconsistency in meta-analyses. *BMJ* 327(7414):557–560. <https://doi.org/10.1136/bmj.327.7414.557>
26. Moses LE, Shapiro D, Littenberg B (1993) Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med* 12(14):1293–1316. <https://doi.org/10.1002/sim.4780121403>
27. Wolff RF, Moons KGM, Riley RD et al (2019) PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 170(1):51–58. <https://doi.org/10.7326/m18-1376>
28. Moons KGM, Wolff RF, Riley RD et al (2019) PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Ann Intern Med* 170(1):W1–W33. <https://doi.org/10.7326/M18-1377>
29. Whiting PF, Rutjes AW, Westwood ME et al (2011) QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 155(8):529–536. <https://doi.org/10.7326/0003-4819-155-8-201110180-00009>
30. Hayden JA, van der Windt DA, Cartwright JL, Cote P, Bombardier C (2013) Assessing bias in studies of prognostic factors. *Ann Intern Med* 158(4):280–286. <https://doi.org/10.7326/0003-4819-158-4-201302190-00009>
31. Moher D, Liberati A, Tetzlaff J, Altman DG (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 339:b2535. <https://doi.org/10.1136/bmj.b2535>
32. Page MJ, McKenzie JE, Bossuyt PM et al (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372:n71. <https://doi.org/10.1136/bmj.n71>
33. Gao Y, Zeng S, Xu X et al (2022) Deep learning-enabled pelvic ultrasound images for accurate diagnosis of ovarian cancer in China: a retrospective, multicentre, diagnostic study. *Lancet Digit Health* 4(3):e179–e187. [https://doi.org/10.1016/s2589-7500\(21\)00278-8](https://doi.org/10.1016/s2589-7500(21)00278-8)
34. Chiappa V, Bogani G, Interlenghi M et al (2021) The Adoption of Radiomics and machine learning improves the diagnostic processes of women with Ovarian Masses (the AROMA pilot study). *J Ultrasound* 24(4):429–437. <https://doi.org/10.1007/s40477-020-00503-5>
35. Chiappa V, Interlenghi M, Bogani G et al (2021) 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* 5(1):28. <https://doi.org/10.1186/s41747-021-00226-0>
36. Christiansen F, Epstein EL, Smedberg E, Akerlund M, Smith K, Epstein E (2021) Ultrasound image analysis using deep neural networks for discriminating between benign and malignant ovarian tumors: comparison with expert subjective assessment. *Ultrasound Obstet Gynecol* 57(1):155–163. <https://doi.org/10.1002/uog.23530>
37. Qi L, Chen D, Li C et al (2021) Diagnosis of ovarian neoplasms using nomogram in combination with ultrasound image-based radiomics signature and clinical factors. *Front Genet* 12:753948. <https://doi.org/10.3389/fgene.2021.753948>
38. Wang H, Liu C, Zhao Z et al (2021) Application of deep convolutional neural networks for discriminating benign, borderline, and malignant serous ovarian tumors from ultrasound images. *Front Oncol*. <https://doi.org/10.3389/fonc.2021.770683>
39. Martinez-Mas J, Bueno-Crespo A, Khazendar S et al (2019) Evaluation of machine learning methods with Fourier Transform features for classifying ovarian tumors based on ultrasound images. *PLoS One* 14(7):e0219388. <https://doi.org/10.1371/journal.pone.0219388>
40. Zhang L, Huang J, Liu L (2019) Improved deep learning network based in combination with cost-sensitive learning for early detection of ovarian cancer in color ultrasound detecting system. *J Med Syst* 43(8):251. <https://doi.org/10.1007/s10916-019-1356-8>
41. Acharya UR, Akter A, Chowriappa P et al (2018) Use of nonlinear features for automated characterization of suspicious ovarian tumors using ultrasound images in fuzzy forest framework. *Int J Fuzzy Syst* 20(4):1385–1402. <https://doi.org/10.1007/s40815-018-0456-9>
42. Acharya UR, Mookiah MR, Vinitha Sree S et al (2014) Evolutionary algorithm-based classifier parameter tuning for automatic ovarian cancer tissue characterization and classification. *Ultraschall Med* 35(3):237–245. <https://doi.org/10.1055/s-0032-1330336>
43. Acharya UR, Sree SV, Krishnan MM et al (2012) Ovarian tumor characterization using 3D ultrasound. *Technol Cancer Res Treat* 11(6):543–552. <https://doi.org/10.7785/tcrt.2012.500272>
44. Acharya UR, Sree SV, Kulshreshtha S et al (2014) GyneScan: an improved online paradigm for screening of ovarian cancer via tissue characterization. *Technol Cancer Res Treat* 13(6):529–539. <https://doi.org/10.7785/tcrtexpress.2013.600273>
45. Acharya UR, Sree SV, Saba L, Molinari F, Guerriero S, Suri JS (2013) Ovarian tumor characterization and classification using ultrasound—a new online paradigm. *J Digit Imaging* 26(3):544–553. <https://doi.org/10.1007/s10278-012-9553-8>
46. Aramendia-Vidaurreta V, Cabeza R, Villanueva A, Navallas J, Alcazar JL (2016) Ultrasound image discrimination between benign and malignant adnexal masses based on a neural network approach. *Ultrasound Med Biol* 42(3):742–752. <https://doi.org/10.1016/j.ultrasmedbio.2015.11.014>
47. Khazendar S, Sayasneh A, Al-Assam H et al (2015) 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* 7(1):7–15
48. Faschingbauer F, Beckmann MW, Weyert Goecke T et al (2013) Automatic texture-based analysis in ultrasound imaging of ovarian masses. *Ultraschall Med* 34(2):145–150. <https://doi.org/10.1055/s-0031-1299331>
49. Vaes E, Manchanda R, Autier P et al (2012) Differential diagnosis of adnexal masses: sequential use of the risk of malignancy index and HistoScanning, a novel computer-aided diagnostic tool. *Ultrasound Obstet Gynecol* 39(1):91–98. <https://doi.org/10.1002/uog.9079>
50. Vaes E, Manchanda R, Nir R et al (2011) Mathematical models to discriminate between benign and malignant adnexal masses: potential diagnostic improvement using ovarian HistoScanning. *Int J Gynecol Cancer* 21(1):35–43. <https://doi.org/10.1097/IGC.0b013e3182000528>
51. Lu C, Van Gestel T, Suykens JA, Van Huffel S, Vergote I, Timmerman D (2003) Preoperative prediction of malignancy of ovarian tumors using least squares support vector machines. *Artif Intell Med* 28(3):281–306. [https://doi.org/10.1016/s0933-3657\(03\)00051-4](https://doi.org/10.1016/s0933-3657(03)00051-4)
52. Lucidarme O, Akakpo JP, Granberg S et al (2010) A new computer-aided diagnostic tool for non-invasive characterisation of malignant ovarian masses: results of a multicentre validation study. *Eur Radiol* 20(8):1822–1830. <https://doi.org/10.1007/s00330-010-1750-6>
53. Zimmer Y, Tepper R, Akseilrod S (2003) An automatic approach for morphological analysis and malignancy evaluation of ovarian masses using B-scans. *Ultrasound Med Biol* 29(11):1561–1570. <https://doi.org/10.1016/j.ultrasmedbio.2003.08.013>
54. Li S, Liu J, Xiong Y et al (2022) Application values of 2D and 3D radiomics models based on CT plain scan in differentiating benign from malignant ovarian tumors. *Biomed Res Int* 2022:5952296. <https://doi.org/10.1155/2022/5952296>
55. Park H, Qin L, Guerra P, Bay CP, Shinagare AB (2021) Decoding incidental ovarian lesions: use of texture analysis and machine learning for characterization and detection of malignancy. *Abdom Radiol (NY)* 46(6):2376–2383. <https://doi.org/10.1007/s00261-020-02668-3>
56. Li S, Liu J, Xiong Y et al (2021) A radiomics approach for automated diagnosis of ovarian neoplasm malignancy in computed tomography. *Sci Rep* 11(1):8730. <https://doi.org/10.1038/s41598-021-87775-x>
57. Liu X, Wang T, Zhang G et al (2022) Two-dimensional and three-dimensional T2 weighted imaging-based radiomic signatures for the preoperative discrimination of ovarian borderline tumors and malignant tumors. *J Ovarian Res* 15(1):22. <https://doi.org/10.1186/s13048-022-00943-z>
58. Song XL, Ren JL, Zhao D, Wang L, Ren H, Niu J (2021) Radiomics derived from dynamic contrast-enhanced MRI pharmacokinetic protocol features: the value of precision diagnosis ovarian neoplasms. *Eur Radiol* 31(1):368–378. <https://doi.org/10.1007/s00330-020-07112-0>
59. Jian J, Li Y, Xia W et al (2021) MRI-based multiple instance convolutional neural network for increased accuracy in the differentiation of borderline and malignant epithelial ovarian tumors. *J Magn Reson Imaging*. <https://doi.org/10.1002/jmri.28008>
60. Zhang H, Mao Y, Chen X et al (2019) Magnetic resonance imaging radiomics in categorizing ovarian masses and predicting clinical outcome: a preliminary study. *Eur Radiol* 29(7):3358–3371. <https://doi.org/10.1007/s00330-019-06124-9>

61. Li Y, Jian J, Pickhardt PJ et al (2020) MRI-based machine learning for differentiating borderline from malignant epithelial ovarian tumors: a multicenter study. *J Magn Reson Imaging* 52(3):897–904. <https://doi.org/10.1002/jmri.27084>
62. Jian J, Xia W, Zhang R et al (2021) Multiple instance convolutional neural network with modality-based attention and contextual multi-instance learning pooling layer for effective differentiation between borderline and malignant epithelial ovarian tumors. *Artif Intell Med* 121:102194. <https://doi.org/10.1016/j.artmed.2021.102194>
63. Ștefan P-A, Lupean R-A, Mihu CM et al (2021) Ultrasonography in the diagnosis of adnexal lesions: the role of texture analysis. *Diagnostics* 11(5):812
64. van der Sommen F, de Groof J, Struyvenberg M et al (2020) Machine learning in GI endoscopy: practical guidance in how to interpret a novel field. *Gut* 69(11):2035–2045. <https://doi.org/10.1136/gutjnl-2019-320466>
65. Chacon E, Dasi J, Caballero C, Alcazar JL (2019) Risk of ovarian malignancy algorithm versus risk malignancy index-I for preoperative assessment of adnexal masses: a systematic review and meta-analysis. *Gynecol Obstet Invest* 84(6):591–598. <https://doi.org/10.1159/000501681>
66. Mulder EE, Gelderblom ME, Schoot D, Vergeldt TF, Nijssen DL, Piek JM (2020) External validation of risk of malignancy index compared to IOTA simple rules. *Acta Radiol*. 62:673–678. <https://doi.org/10.1177/0284185120933990>
67. Sadowski EA, Maturen KE, Rockall A et al (2021) Ovary: MRI characterisation and O-RADS MRI. *Br J Radiol* 94(1125):20210157. <https://doi.org/10.1259/bjr.20210157>
68. Song H, Bak S, Kim I et al (2021) An application of machine learning that uses the magnetic resonance imaging metric, mean apparent diffusion coefficient, to differentiate between the histological types of ovarian cancer. *J Clin Med* 11(1):229. <https://doi.org/10.3390/jcm11010229>
69. Forstner R (2020) Early detection of ovarian cancer. *Eur Radiol* 30(10):5370–5373. <https://doi.org/10.1007/s00330-020-06937-z>

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:**

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

---

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)