Artificial intelligence performance in image-based ovarian cancer identification: A systematic review and meta-analysis

He-Li Xu,[a,b,c,1] Ting-Ting Gong,[d,1] Fang-Hua Liu,[a,b,c] Hong-Yu Chen,[a,b,c] Qian Xiao,[a] Yang Hou,[e] Ying Huang,[f] Hong-Zan Sun,[e] Yu Shi,[e] Song Gao,[d] Yan Lou,[g] Qing Chang,[a,b,c] Yu-Hong Zhao,[a,b,c] Qing-Lei Gao,[h] and Qi-Jun Wu [a,b,c,d,*]

[a]Department of Clinical Epidemiology, Shengjing Hospital of China Medical University, Shenyang, China
[b]Clinical Research Center, Shengjing Hospital of China Medical University, Shenyang, China
[c]Key Laboratory of Precision Medical Research on Major Chronic Disease, Shengjing Hospital of China Medical University, Shenyang, China
[d]Department of Obstetrics and Gynecology, Shengjing Hospital of China Medical University, Shenyang, China
[e]Department of Radiology, Shengjing Hospital of China Medical University, Shenyang, China
[f]Department of Ultrasound, Shengjing Hospital of China Medical University, Shenyang, China
[g]Department of Intelligent Medicine, China Medical University, China
[h]National Clinical Research Center for Obstetrics and Gynecology, Cancer Biology Research Centre (Key Laboratory of the Ministry of Education) and Department of Gynecology and Obstetrics, Tongji Hospital, Wuhan, China

eClinicalMedicine
2022;53: 101662
Published online xxx
https://doi.org/10.1016/j.eclinm.2022.101662

Summary

Background Accurate identification of ovarian cancer (OC) is of paramount importance in clinical treatment success. Artificial intelligence (AI) is a potentially reliable assistant for the medical imaging recognition. We systematically review articles on the diagnostic performance of AI in OC from medical imaging for the first time.

Methods The Medline, Embase, IEEE, PubMed, Web of Science, and the Cochrane library databases were searched for related studies published until August 1, 2022. Inclusion criteria were studies that developed or used AI algorithms in the diagnosis of OC from medical images. The binary diagnostic accuracy data were extracted to derive the outcomes of interest: sensitivity (SE), specificity (SP), and Area Under the Curve (AUC). The study was registered with the PROSPERO, CRD42022324611.

Findings Thirty-four eligible studies were identified, of which twenty-eight studies were included in the meta-analysis with a pooled SE of 88% (95%CI: 85–90%), SP of 85% (82–88%), and AUC of 0.93 (0.91–0.95). Analysis for different algorithms revealed a pooled SE of 89% (85–92%) and SP of 88% (82–92%) for machine learning; and a pooled SE of 88% (84–91%) and SP of 84% (80–87%) for deep learning. Acceptable diagnostic performance was demonstrated in subgroup analyses stratified by imaging modalities (Ultrasound, Magnetic Resonance Imaging, or Computed Tomography), sample size (≤300 or >300), AI algorithms versus clinicians, year of publication (before or after 2020), geographical distribution (Asia or non Asia), and the different risk of bias levels (≥3 domain low risk or < 3 domain low risk).

Interpretation AI algorithms exhibited favorable performance for the diagnosis of OC through medical imaging. More rigorous reporting standards that address specific challenges of AI research could improve future studies.

Funding This work was supported by the Natural Science Foundation of China (No. 82073647 to Q-JW and No. 82103914 to T-TG), LiaoNing Revitalization Talents Program (No. XLYC1907102 to Q-JW), and 345 Talent Project of Shengjing Hospital of China Medical University (No. M0268 to Q-JW and No. M0952 to T-TG).

Keywords: Artificial intelligence; Medical imaging; Meta-analysis; Ovarian cancer

Abbreviations: AI, Artificial intelligence; OC, Ovarian cancer; SE, Sensitivity; SP, Specificity; AUC, Area Under the Curve; US, Ultrasound; MRI, Magnetic Resonance Imaging; CT, Computed Tomography; ML, Machine learning; DL, Deep learning; XAI, Explainable artificial intelligence

*Corresponding author at: Department of Clinical Epidemiology, Department of Obstetrics and Gynecology, Clinical Research Center, Shengjing Hospital of China Medical University, Address: No. 36, San Hao Street, Shenyang, Liaoning 110004, PR China.
E-mail address: wuqj@sj-hospital.org (Q.-J. Wu).

[1] Joint first authors.

Research in context

Evidence before this study

The accurate preoperative differentiation between benign and malignant masses of the ovary is crucial for determining the appropriate treatment strategies and improving the postoperative quality of life. Imaging is an useful tool in medical science and is invoked in clinical practice to facilitate decision making for the diagnosis, staging, and treatment. The advances of artificial intelligence (AI) might help to bridge the gap between the intense demand for diagnostic from imaging and relatively limited healthcare resources. Up to date, there is a lack of quantitative synthesis to comprehensively summarize the available evidence of the AI-based methods on ovarian cancer (OC) detection. The Medline, Embase, IEEE, Pubmed, Web of Science, and the Cochrane library were systematically searched for studies that developed an AI algorithm for the diagnostic performance of OC from medical imaging, published until August 1, 2022. Only English language articles were considered. We performed a systematic review and meta-analysis of published data on diagnostic performance of AI algorithms and radiomics models for OC detection.

Added value of this study

To our best knowledge, this is the first systematic review and meta-analysis specifically dedicated to AI system performance in the diagnosis of OC. We are strictly in line with the guidelines for diagnostic reviews, and conducted a comprehensive literature search in both medical databases and engineering and technology databases to ensure the rigor of the study. After a careful selection of research on relevant topics, we found that AI algorithms excelled in the identification of OC using medical radiography imaging.

Implications of all the available evidence

AI algorithms exhibited favorable performance for the diagnosis of OC through medical imaging. More rigorous reporting standards that address specific challenges of AI research could improve future studies.

Introduction

Ovarian tumors comprise a remarkably heterogeneous group of benign, borderline, and malignant lesions and exhibit extensive morphological characteristics.[1,2] Among these, ovarian cancer (OC) is the most lethal gynecological malignancy.[3] While malignant ovarian neoplasms may need a more aggressive surgical approach, benign masses can either be safely monitored or undergo simple resection allowing for a fertility- and ovary-sparing approach.[4] Therefore, accurate preoperative differentiation between benign and malignant masses of the ovary is crucial for determining the appropriate treatment strategies and improving the postoperative quality of life.[5]

Imaging is a useful tool in medical science and is invoked in clinical practice to facilitate decision making for the diagnosis, staging, and treatment.[6,7] The ultrasound (US) is commonly used to recognize the presence of an ovarian mass and to determine between benign and malignant lesions.[8] Magnetic resonance imaging (MRI) plays a significant role in characterizing ovarian tumors due to its high soft-tissue resolution, and it is recommended in assessing the need for surgery for an adnexal mass.[9] Computed tomography (CT) may be helpful for judging the gross extent of hematogenous, peritoneal, and lymphatic spread of OC: because of its ability to evaluate the liver, paraaortic region, omentum, and mesentery.[10] While studies have reported the utility of PET CT in diagnosing ovarian tumors, its cost-effectiveness for this purpose remains unproven. Currently, US and MRI are the most commonly used imaging modalities for the diagnosis and characterization of ovarian tumors.[11] Of note, the diagnosis of OC has been conventionally dependent on the subjective assessment of radiologists or gynecologists who use their clinical practice experience to scrutinize imaging features and examine ovarian tumors with high heterogeneity.[12,13] Owing to the intricacy generated by inadequate or absent radiology in resource-poor health regions and the influence of wide disparity in the human rater expertise, making a proper and immediate diagnosis from medical imaging is challenging.[14,15]

The advances of artificial intelligence (AI) might help to bridge the gap between the intense demand for diagnostic from imaging and relatively limited healthcare resources.[16] Meanwhile, as an interesting research hotspot, radiomics is described as a new 'data-driven' approach for extracting large sets of quantitative signatures from radiological images.[17] These data can be subsequently analyzed using conventional biostatistics or AI methods.[18] With sophisticated image processing methods, all medical images are transferred to mineable high-throughput image features, which thereafter can be used to correlate these processed feature signatures with pathology diagnoses or treatment responses.[19] Radiomics models and AI algorithms have shown promising results in integrating medical images for the detection of OC.[20] For example, aramedia-vidaurreta et al.[21] emphasized that a machine learning (ML) algorithm based on US images achieved a diagnostic accuracy of 0.98 in one hundred and forty-five patients. Additionally, a deep learning (DL) model was used to automatically discriminate between benign and malignant ovarian tumor images, with high accuracy of 87.6%.[22] Even though, researchers have still tried different ways, including but not limited to improving image quality, expanding sample sizes, and optimizing algorithms, to raise diagnostic accuracy.[23]

Up to date, there is a lack of quantitative synthesis to comprehensively summarize the available evidence of the AI-based methods on OC detection. Therefore, the purpose of this study is to first perform a systematic review and meta-analysis of published data on the diagnostic performance of AI algorithms and radiomics models for OC detection.

Methods

Protocol registration and study design

The study was registered in the PROSPERO (CRD42022324611). The meta-analysis was conducted following the PRISMA,[24] MOOSE,[25] and CHARMS[26] reporting guidelines.

Search strategy and eligibility criteria

The Medline, Embase, IEEE, Pubmed, Web of Science, and the Cochrane library were systematically searched for studies that developed an AI algorithm for the diagnostic performance of OC from medical imaging, published until August 1, 2022. Only English-language articles were considered. Supplementary Note 1 summarizes the search strategy used in each database.

Eligible studies reported the AI technologies for the diagnosis of OC from medical radiology images with diagnostic outcomes, such as sensitivity (SE), and specificity (SP), or detailed information on 2×2 contingency tables. The following studies were excluded: duplicate publications; reviews; editorials; non-human samples; histopathology images; combining non-image information; no classification task; and no AI model. Two reviewers (H-LX and F-HL) independently screened the titles and abstracts according to these eligibility criteria, and relevant articles for full text were downloaded and reviewed. Disagreement was discussed with a third author (Q-JW) and subsequently resolved via consensus.

Data extraction

Two reviewers (H-LX and H-YC) extracted study characteristics and diagnostic performance independently using a standardized data extraction sheet. Disagreements were resolved by discussion or a third investigator (F-HL) was consulted.

The diagnostic accuracy data including true-positive (TP), false-positive (FP), true-negative (TN), and false-negative (FN) were extracted directly into contingency tables, and were used to calculate SE and SP. If a study provided multiple contingency tables for the same or different AI algorithms, we assumed that they were independent of each other. Supplementary Table 1 summarizes the contingency tables extracted from included studies.

Study quality assessment

All selected studies were assessed for quality with the use of quality assessment of diagnostic accuracy studies-AI (QUADAS-AI) criteria[27] by two independent reviewers (H-LX and T-TG). The details are listed in Supplementary Table 2. This guideline includes four domains (patient selection, index test, reference standard, flow, and timing) in the risk of bias and three domains (patient selection, index test, reference standard) in applicability concerns. This new tool is an AI-specific extension to QUADAS-2[28] and QUADAS-C,[29] providing researchers with a specific framework to evaluate the risk of bias and applicability when conducting reviews that evaluate AI-centered diagnostic test accuracy. Conflicts were discussed with a third collaborator (F-HL).

Meta-analysis

A hierarchical summary receiver-operating characteristic curve (SROC) was fitted to evaluate the accuracy of the AI model. We plotted the combined curve with corresponding 95% confidence region and 95% prediction region around averaged SE, SP, and Area Under the Curve (AUC) estimates in SROC figures. When same or different AI models were tested within the same paper, the proposed model with the best accuracy was used for further meta-analysis. Heterogeneity was assessed using the I^2 statistic. Subgroup and regression analyses were performed to explore potential sources of heterogeneity. The random effects model was conducted because of the assumed differences between studies. The risk of publication bias was evaluated using funnel plot and regression test.

Seven sub-analysis were performed: (1) according to sample size (≤300 or >300); (2) according to AI algorithms (ML or DL); (3) according to imaging modalities (CT, US, or MRI); (4) according to the pooled performance using the same dataset (AI algorithms or human clinicians); (5) according to the year of publication (before or after 2020); (6) according to the geographical distribution (Asia or non Asia); (7) according to different risk of bias levels (≥3 domain low risk or < 3 domain low risk)

The methodological quality of included studies was evaluated using the QUADAS-AI by RevMan (Version 5.4). A cross-hairs plot was also produced (R V.4.2.1) to better display the variability between sensitivity/specificity estimates.[30] All other statistical analyses were conducted in Stata software (Version 15.0) with two-tailed probability of type I error of 0.05 ($\alpha = 0.05$).

Role of the funding source

Our study was funded by the Natural Science Foundation of China, the LiaoNing Revitalization Talents Program, and the 345 Talent Project of Shengjing Hospital

of China Medical University. The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

Study selection and characteristics of eligible studies

A total of 1212 records were retrieved on initial search and 513 duplicates were removed, and of these 642 studies were excluded based on screening of titles and abstracts, resulting in 57 studies for full-text review. Finally, 34 articles were included in the present systematic review and 28 had sufficient data for meta-analysis (Figure 1).

Majority of the studies (n = 31) were based on retrospective patient data except four studies. Only two studies using prospective data. One studies used images from public databases. Eight studies excluded low quality images, while twenty-six studies did not mention this process. Only three studies using out-of-sample dataset to perform external validation, of which two studies did not provide data of our concern for integrated analysis. Eight studies compared AI model with clinicians in the same dataset. Moreover, imaging modalities were classified as US (n = 19), MRI (n = 10), CT (n = 3), MRI and US (n = 1), and MRI and CT (n = 1). Furthermore, the distribution of the number of studies on AI algorithms in the present study is as follows: DL (11 studies) and ML (23 studies). Tables 1–4 show the detailed characteristics of these including studies.

Pooled performance of AI algorithms

The SROC curves for 28 included studies with 160 contingency tables are provided in Figure 2a, the combined SE and SP were 88% (95%CI: 85–90%) and 85% (82–88%), respectively, with an AUC of 0.93 (0.91–0.95) for all AI algorithms. When the highest accuracy

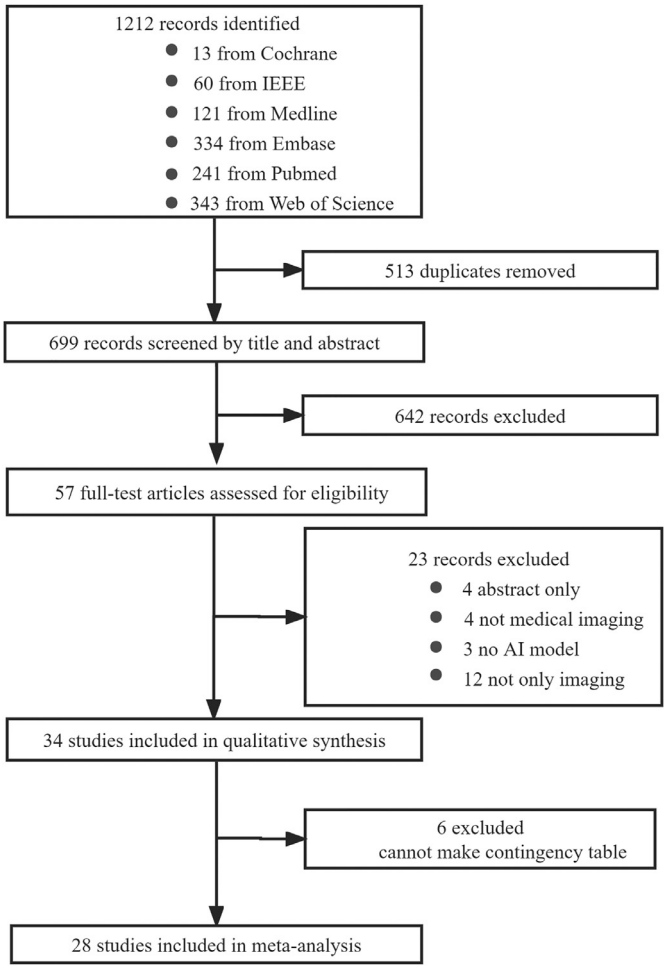

Figure 1. PRISMA flowchart of study selection.

Author [ref], year	Participants		*N*	Mean or median age (SD; range)
	Inclusion criteria	Exclusion criteria		
Liu et al,[31] 2022[a]	Patients with no previous pelvic surgery; patients with no previous gynecological disease history; patients who had MRI examinations performed at our institution before pelvic or laparoscopic surgery.	Patients with previous pelvic surgical history or radiation history; patients whose MRI data were unavailable either due to the examination being performed at another institution or due to claustrophobia; patients whose data lacked histological results.	196	46.3
Gao et al,[22] 2022[a]	Consecutive adult patients (aged ≥18 years) who presented with adnexal lesions in ultrasound in ten hospitals between September 2003, and May 2019.	Duplicated cases; postoperative patients who were deprived of adnexa; patients without histological diagnosis.	1,07,624	NR
Saida et al,[32] 2022[a]	Aged above 20 years for ethical reasons; pelvic MRI scan obtained as per the protocol followed at our hospital between January 2015 and December 2020; pathologically proven malignant epithelial tumors (i.e., carcinomas) or borderline tumors of the ovary for the malignant group; pathologically proven or clinically apparent benign lesions in the non-malignant group.	Malignant tumors in the pelvis other than the ovary; history of surgery of the uterus or ovaries other than caesarean section, chemotherapy, or radiation therapy of the pelvis; malignant ovarian epithelial tumors mixed with non-epithelial components.	465	50 (20–90)
Guo et al,[33] 2022[a]	Definite pathological diagnosis after operation; MRI and ultrasound were performed and the data were complete; the images could be used for diagnostic analysis; patient informed consent.	Incomplete ultrasound, MRI, or pathological data; combined with severe organic diseases, such as coagulation dysfunction, renal insufficiency, heart failure, and other surgical contraindications; history of ovarian surgery; combined with other pelvic diseases, such as endometrial cancer and rectal cancer.	207	NR
Li et al,[34] 2022[a]	Patients with ovarian tumor confirmed by histopathology; no history of malignant tumors other than ovarian tumor; patients who were undergoing pelvic CT examination within half a month before surgery.	Those who had received radiotherapy, chemotherapy, or radiotherapy–chemotherapy before CT examination; patients diagnosed with inflammatory diseases; patients with low image quality.	140	NR
Wang et al,[35] 2021[a]	A histologic diagnosis of benign, borderline, or malignant SOTs between March 2013 and December 2016; availability of diagnostic-quality preoperative US images; US scanning before neoadjuvant therapy or surgical resection.	No ultrasound results or the ovarian mass was not completely in the images; mucinous, clear cell, endometrioid, or metastatic cancer.	265	51 (15–79)
Chiappa et al,[36] 2021[a]	Diagnosis of OM; execution of a preoperative ultrasonographic examination within 2 weeks before surgery; surgery performed.	Age<18 years; absence of ultrasonographic images stored; consent withdrawn.	241	55 (18–84)

Table 1 *(Continued)*

Author [ref], year	Participants		N	Mean or median age (SD; range)
	Inclusion criteria	Exclusion criteria		
Jian et al,[37] 2021	All patients were histopathologically proven to have either BEOT (n = 165) or MEOT (n = 336).	NR	501	NR
Wang et al,[38] 2021[a]	Benign or malignant ovarian lesions confirmed by either pathology or imaging follow-up; available pre-operative MRI examination including T1C and T2WI; the quality of images was clear without motion or artifacts and were fit for analysis.	Lack pre-operative MRI; lack clear ovarian lesion; lack T1C images.	451	45.7
Hu et al,[39] 2021[a]	NR	Patients with poor image quality; patients without enhanced scanning; patients with unclear boundary and unable to outline	110	NR
Yu et al,[40] 2021[a]	SBOTs and SMOTs were diagnosed by postoperative pathology; SBOTs and SMOTs were in an early stage (I and II) according to the guideline of the FIGO; the images were of sufficient quality for radiomics analysis.	SBOTs and SMOTs which were in a late stage (III and IV) according to the FIGO guideline; patients who received any treatment before CT examination or were on treatment at the time of CT examination were also excluded to eliminate the effect of treatment on imaging features.	182	47.7
Ștefan et al,[41] 2021[a]	A lesion with a minimum diameter of at least 20 mm; the availability of conventional B-mode images; lack of imaging artifacts; and the existence of a patient's serial number.	No medical data corresponding to the PSN; the absence of a final pathological diagnosis to indicate the benign or malignant nature of the lesions; the pathological analysis performed at more than 30 days after the image acquisition; and no gynecological follow-up.	120	38.2
Christiansen et al,[42] 2021[a]	Surgery within 120 days after the ultrasound examination or ultrasound follow-up for a minimum of 3 years or until resolution of the lesion.	NR	758	NR
Akazawa et al,[43] 2020	Patients were ovarian tumors which had been diagnosed pathologically after surgical resection.	Lack of sufficient preoperative clinical data, such as tumor markers or the records of imaging tests.	202	51 (14—84)
Martínez et al,[44] 2019[a]	NR	NR	384	NR
Zhang et al,[20] 2019[a]	No previous pelvic surgery; no previous gynecological disease history; MRI examinations before pelvic or laparoscopic surgery were performed at our institution.	Previous pelvic surgical history or radiation history; MRI data were unavailable either for the examination performed at another institution or due to claustrophobia; no histological results.	438	52.7

Table 1 *(Continued)*

Author [ref], year	Participants		N	Mean or median age (SD; range)
	Inclusion criteria	Exclusion criteria		
Mol et al,[45] 2001[a]	Women who had surgery for an adnexal mass between January 1991 and December 1998 were included.	NR	170	46 (20−89)
Liu D et al,[46] 2017[a]	Patients with histologically proven diagnosis of EOCs; patients complete CT or MRI examination before operation in two weeks.	Surgery was performed outside our institution without definite histological diagnosis, incomplete clinical or CT and MRI records preoperatively.	65	56.4
Kazerooni et al,[47] 2017[a]	Patients were scheduled for surgical removal of suspicious ovarian masses and postoperative histopathological assessment within 2 weeks of MRI exam.	NR	55	38.4
Acharya et al,[48] 2014[a]	NR	Women with no anatomopathological evaluation.	20	49.5
Acharya et al,[49] 2013[a]	NR	Patients with no anatomopathological evaluation.	20	49.5
Acharya et al,[50] 2012[a]	NR	NR	20	49.5
Umar et al,[51] 2012	NR	NR	24	NR
Acharya et al,[52] 2012[a]	NR	Patients with no anatomopathological evaluation.	20	49.5
Al-Karawi et al,[53] 2021[a]	All ovarian tumors were given a histological diagnosis label.	NR	232	NR
Jian et al,[54] 2021	Histologically proven EOC; MRI performed within 1 month prior to gynecological operation; all four axial MRI sequences obtained: fast spin-echo T2-weighted imaging with fat saturation(T2WI FS), echo-planar DWI with gradient b factors of 0 and 600, 800, or 1000 s/mm^2, ADC map, and 2D volumetric interpolated breath hold examination (VIBE) contrast enhanced T1-weighted imaging with FS (CE-T1WI) in the late phase (150−190 s after the intravenous administration of contrast agent); absence of prior gynecological operation or chemotherapy prior to MRI scanning.	Patients without definitive histopathology or with poor MRI image quality (image has artifacts that cannot outline the tumor).	294	(51.2−57.2)
Li et al,[55] 2020	Histologically proven BEOT or MEOT from January 2010 to June 2018; MRI performed within 2 weeks prior to gynecological operation.	Lacking any one of these four axial MRI sequences; prior gynecological operation and/or chemotherapy before MRI scanning; poor MRI image quality with artifacts that affected the delineation of the tumor.	501	(47.2−51.6)
Acharya et al,[56] 2014[a]	NR	NR	20	NR
Pathak et al,[57] 2015[a]	NR	NR	120	NR

Table 1 *(Continued)*

Author [ref], year	Participants		N	Mean or median age (SD; range)
	Inclusion criteria	Exclusion criteria		
Ameye et al,[58] 2009[a]	NR	Exclusion criteria were pregnancy, inability to tolerate transvaginal sonography, and surgery performed more than 120 days after sonographic assessment.	1573	46 (9—94)
Jian et al,[59] 2022[a]	Inclusion criteria were as follows: patients with 1) BEOT or MEOT that was proven by surgery and histopathology from January 2010 to June 2018; 2) an MRI performed within 2 weeks before gynecological operation which included the following three axial MRI sequences: fast spin echo T2-weighted imaging with fat saturation (T2WI FS), echo planar diffusion-weighted imaging (DWI) with apparent diffusion coefficient (ADC) maps generated from maximum b-value imaging if images with multiple b-values available, and 2D volumetric interpolated breath-hold examination of contrast-enhanced T1-weighted imaging (CE-T1WI) with FS in the late phase (150— 190 seconds after the intravenous administration of contrast agent); and 3) no history of gynecological operations or chemotherapy prior to the MRI scan.	Patients with poor quality images were excluded (based on the evaluation of the radiologist with 10 years' experience in gynecological imaging) because artifacts could affect the observation of the tumor.	501	58.92 (14.05)
Alqasemi et al,[51] 2012[a]	NR	NR	24	NR
Chen et al,[60] 2012[a]	Inclusion criteria were as follows: patients with at least one persisting ovarian tumor detected at US (except for physiologic cysts) from January 2019 to November 2019, patients who underwent a surgical procedure with histopathologic results, an interval of 30 days between US examination and surgery, and patients who had no previous history of ovarian cancer.	Exclusion criteria were histopathologic analysis—confirmed uterine sarcomas or nongynecologic tumors, inconclusive histopathologic results, or poor US image quality.	422	46.4 (14.8)
Zheng et al,[61] 2022	Patients with either SBOTs or SMOTs, who underwent preoperative MRI scans and confirmed by postoperative pathology.	Exclusion criteria were as follows: (1) solid tissue <80% in lesion (25); (2) the tumor had significant metastases; (3) significant image artifacts.	1260	61 (20—79)

***Table 1*: Participant demographics for the 35 included studies.**

Abbreviation: BEOT: borderline epithelial ovarian tumor; CT: computed tomography; EOC: epithelial ovarian cancer; FIGO: International Federation of Gynecology and Obstetrics; MEOT: malignant epithelial ovarian tumors; NR=not reported; MRI: magnetic resonance imaging; OM: ovarian mass; SBOT; serous borderline ovarian tumors; SMOT: serous malignant ovarian tumors; SOT: serous ovarian tumors; T1C: T1-weighted contrast-enhanced sequence; T2WI: T2-weighted sequence; US: ultrasound.

[a] Studies (n = 28) included in the meta-analysis.

Author [ref], year	Reference standard	Type of internal validation	External validation	AI versus clinicians
Liu et al,[31] 2022[a]	Histopathology	NR	No	No
Gao et al,[22] 2022[a]	Histopathology	Random split sample validation	Yes	Yes
Saida et al,[32] 2022[a]	Histopathology	NR	No	Yes
Guo et al,[33] 2022[a]	Histopathology	K-fold cross validation	No	No
Li et al,[34] 2022[a]	Histopathology	Ten-fold cross-validation	No	No
Wang et al,[35] 2021[a]	Histopathology	Three-fold cross validation	No	No
Chiappa et al,[36] 2021[a]	Histopathology	Ten-fold cross validation	No	No
Jian et al,[37] 2021	Histopathology	Random split sample validation	No	No
Wang et al,[38] 2021[a]	Histopathology	Cross validation	No	Yes
Hu et al,[39] 2021[a]	NR	Ten-fold cross-validation	No	No
Yu et al,[40] 2021[a]	Histopathology	NR	No	No
Ștefan et al,[41] 2021[a]	Histopathology	NR	No	No
Christiansen et al,[42] 2021[a]	Histopathology	NR	No	Yes
Akazawa et al,[43] 2020	Histopathology	K-fold cross validation	No	No
Zhang et al, 2019[a]	Histopathology	Ten-fold cross validation	No	No
Martínez et al,[44] 2019[a]	Histopathology	Cross validation	No	No
Zhang et al,[20] 2019[a]	Histopathology	Leave-one-out cross-validation	No	Yes
Mol et al,[45] 2001[a]	Histopathology	Cross validation	No	No
Liu D et al,[46] 2017[a]	Histopathology	Cross validation	No	No
Kazerooni et al,[47] 2017[a]	Histopathology	Leave-one-out cross-validation	No	No
Acharya et al,[48] 2014[a]	Histopathology	Ten-fold cross validation	No	No
Acharya et al,[49] 2013[a]	Histopathology	Ten-fold cross validation	No	No
Acharya et al,[50] 2012[a]	NR	K-fold cross validation	No	No
Umar et al,[51] 2012	Histopathology	NR	No	No
Acharya et al,[52] 2012[a]	Histopathology	Ten-fold cross validation	No	No
Al-Karawi et al,[53] 2021[a]	Histopathology	Random split sample validation	No	No
Jian et al,[54] 2021	Histopathology	NR	Yes	Yes
Li et al,[55] 2020	Histopathology	NR	Yes	Yes
Acharya et al,[56] 2014[a]	NR	Ten-fold cross validation	No	No
Pathak et al,[57] 2015[a]	NR	Cross validation	No	No
Ameye et al,[58] 2009[a]	Histopathology	NR	No	Yes
Jian et al,[59] 2022[a]	Histopathology	NR	No	No
Alqasemi et al,[51] 2012[a]	Histopathology	NR	No	No
Chen et al,[60] 2012[a]	Histopathology	NR	No	Yes
Zheng et al,[61] 2022	Histopathology	Ten-fold cross validation	No	No

***Table 2*: Model training and validation for the 35 included studies.**
Abbreviation: AI: artificial intelligence; NR=not reported.
[a] Studies (n = 28) included in the meta-analysis.

contingency table was selected from these 28 studies, the pooled SE and SP were the same as 91% (84–95%) and 94% (89–97%), respectively (Figure 2b). A cross hairs plot shows reported point estimates and confidence intervals in Figure 3.

Quality assessment

The quality of included studies was determined by the QUADAS-AI (Supplementary figure 1). The detailed assessment results are presented with a diagram in Supplementary figure 2. Over half of the studies showed a high risk or an unclear risk of bias respectively for patient selections (n = 23) and index test (n = 31) because these studies did not clarify description of included patients detailing previous testing, presentation, setting, the intended use of the index test and lack of adequate external evaluation.

Subgroup meta-analyses

Considering the stage of development of the algorithm and the difference in nature, we categorized them into ML and DL algorithms and did a sub-analysis. The results demonstrated a pooled SE of 89% (95%CI: 85–92%) for ML and 88% (95%CI: 84–91%) for DL, and a pooled SP of 88% (95%CI: 82–92%) for ML and 84% (95%CI: 80–87%) for DL (Supplementary figure 3a, b).

Author [ref], year	Indicator definition			Algorithm		
	Device	Exclusion of poor-quality imaging	Heatmap provided	Algorithm architecture	ML/DL	Transfer learning applied
Liu et al,[31] 2022[a]	MRI	NR	No	LASSO	ML	No
Gao et al,[22] 2022[a]	US	Yes	No	DCNN	DL	No
Saida et al,[32] 2022[a]	MRI	NR	Yes	CNN	DL	No
Guo et al,[33] 2022[a]	MRI, US	NR	No	LR	ML	No
Li et al,[34] 2022[a]	CT	Yes	No	LR	ML	No
Wang et al,[35] 2021[a]	US	NR	Yes	DCNN	DL	No
Chiappa et al,[36] 2021[a]	US	NR	No	SVM	ML	No
Jian et al,[37] 2021	MRI	NR	No	MAC-Net	DL	No
Wang et al,[38] 2021[a]	MRI	Yes	No	CNN	DL	No
Hu et al,[39] 2021[a]	CT	Yes	No	LR	ML	No
Yu et al,[40] 2021[a]	CT	Yes	Yes	SVM	ML	No
Ștefan et al,[41] 2021[a]	US	NR	No	KNN	ML	No
Christiansen et al,[42] 2021[a]	US	NR	No	DNN	DL	No
Akazawa et al,[43] 2020	US	NR	No	SVM, KNN, RF, NB, XGBoost	ML	No
Martínez et al,[44] 2019[a]	US	NR	No	KNN, LD, SVM, ELM	ML	No
Zhang et al,[20] 2019[a]	MRI	NR	No	LASSO	ML	No
Mol et al,[45] 2001[a]	US	NR	No	LR, NN	ML	No
Liu D et al,[46] 2017[a]	CT, MRI	NR	No	RF	ML	No
Kazerooni et al,[47] 2017[a]	MRI	NR	No	SVM, LDA	DL	No
Acharya et al,[48] 2014[a]	US	NR	No	PNN	ML	No
Acharya et al,[49] 2013[a]	US	NR	No	DT	ML	No
Acharya et al,[50] 2012[a]	US	NR	No	SVM	ML	No
Umar et al,[51] 2012	US	NR	No	SVM	ML	No
Acharya et al,[52] 2012[a]	US	NR	No	DT	ML	No
Al-Karawi et al,[53] 2021[a]	US	NR	No	SVM	ML	No
Jian et al,[54] 2021	MRI	Yes	No	LASSO	ML	No
Li et al,[55] 2020	MRI	NR	No	LR	ML	No
Acharya et al,[56] 2014[a]	US	NR	No	PNN	ML	No
Pathak et al,[57] 2015[a]	US	NR	No	SVM	ML	No
Ameye et al,[58] 2009[a]	US	NR	No	LR	ML	No
Jian et al,[59] 2022[a]	MRI	Yes	No	MICNN	DL	No
Alqasemi et al,[51] 2012[a]	US	NR	No	SVM	ML	No
Chen et al,[60] 2012[a]	US	Yes	No	ResNet	DL	No
Zheng et al,[61] 2022	MRI	NR	No	LASSO	ML	No

***Table 3*: Indicator, algorithm, and data source for the 35 included studies.**

Abbreviation: AI: artificial intelligence; CNN: convolutional neural network; CT: computed tomography; DCNN: deep convolutional neural network; DL: deep learning; DT: decision tree; DNN: deep neural network; ELM: extreme learning machine; KNN: k-nearest neighbor; LASSO: least absolute shrinkage and selection operator method; LD: linear discriminant; LR: logistic regression; ML: machine learning; MRI: magnetic resonance imaging; NB: naïve bayes; NR=not reported; PNN: probabilistic neural networks; RF: random forest; SVM: support vector machine; US: ultrasound.

[a] Studies (n = 28) included in the meta-analysis.

Seventeen US studies had a pooled SE of 91% (87–93%), a pooled SP of 87% (82–91%), and with an AUC of 0.95 (0.93–0.97). Six MRI studies with a pooled SE of 83% (77–88%), pooled SP of 84% (80–87%), and an AUC of 0.90 (0.87–0.92). Three CT studies that had a pooled SE of 75% (68–81%), pooled SP of 75% (67–82%), and an AUC of 0.82 (0.78–0.85) (Supplementary figure 4a, b, c).

Eight studies presented the diagnostic accuracy between AI algorithms and human clinicians in the same dataset. The pooled SE was 82% (77–87%) for AI algorithms, and human clinicians had 77% (73–80%). The pooled SP was 86% (83–89%) for AI algorithms, and 80% (75–84%) in human clinicians. The AUC was 0.91 (0.88–0.93) and 0.85 (0.81–0.88) for AI algorithms and human clinicians, respectively (Supplementary figure 5a, b).

Fifteen studies had sample sizes ≤ 300 and thirteen studies had sample sizes > 300. The pooled SE was 85% (81–88%) for sample size ≤ 300, and 93% (89

Author [ref], year	Source of data	Number of images for training/ /testing	Data range	Open access data
Liu et al,[31] 2022[a]	Retrospective study, data from Gynecological and Obstetric Hospital, School of Medicine, Fudan University, Shanghai, China.	99/97	2014.01—2017.12	No
Gao et al,[22] 2022[a]	Retrospective study, data from Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, and seven other hospitals, Jingzhou First People's Hospital and Xiangyang Central Hospital.	575930/8416/7929	2003.09—2019.05	No
Saida et al,[32] 2022[a]	Retrospective study, data from Faculty of Medicine, University of Tsukuba.	3663/100	2015.01—2020.12	No
Guo et al,[33] 2022[a]	Retrospective study, data from Qilu Hospital.	138/69	2018.04—2021.04	No
Li et al,[34] 2022[a]	Retrospective study, data from the First Affiliated Hospital of Nanchang Medical College.	99/41	2017—2020	No
Wang et al,[35] 2021[a]	Retrospective study, data from Tianjin Medical University Cancer Institute and Hospital.	195/84	2013.03—2016.12	No
Chiappa et al,[36] 2021[a]	Retrospective study, data from Fondazione IRCCS Istituto Nazionale dei Tumori di Milano.	NR	2017.01—2019.12	No
Jian et al,[37] 2021	Retrospective, data from eight clinical centers in china.	282/119	NR	No
Wang et al,[38] 2021[a]	Retrospective study, data from one large academic center in the United States.	384/161	NR	No
Hu et al,[39] 2021[a]	Retrospective study, data from Lishui Hospital of Zhejiang University	76/34	2010.01—2018.12	No
Yu et al,[40] 2021[a]	Retrospective study, data from the Affiliated Hospital of Qingdao University.	127/55	2017.12—2020.06	No
Ștefan et al,[41] 2021[a]	Retrospective study, data from University of Medicine and Pharmacy	NR	2017.10—2019.02	No
Christiansen et al,[42] 2021[a]	Retrospective study, data from the Karolinska University Hospital(tertiary referral center)and Sodersjukhuset (secondary/tertiary referral center) in Stockholm, Sweden.	508/250	2010—2019	No
Akazawa et al,[43] 2020	Prospective study, date from Tokyo Women's Medical University Medical Center East.	141/61	2013.12—2019.01	No
Martínez et al,[44] 2019[a]	Retrospective study, data from the University Hospital of the Catholic University of Leuven.	NR	NR	No
Zhang et al,[20] 2019[a]	Retrospective study, data from Gynecological and Obstetric Hospital, School of Medicine, Fudan University, Shanghai, China.	NR	2014.01—2017.12	No
Mol et al,[45] 2001[a]	Prospective study, data from in the Saint Joseph Hospital in Veldhoven.	NR	1991.01—1998.12	No

Table 4 *(Continued)*

Author [ref], year	Source of data	Number of images for training/ /testing	Data range	Open access data
Liu D et al,[46] 2017[a]	Retrospective study, date from Department of Radiology, Shanghai Tenth People's hospital of Tongji University.	NR	2009.01−2015.10	No
Kazerooni et al,[47] 2017[a]	Prospectively study, NR.	NR	NR	No
Acharya et al,[48] 2014[a]	Retrospective study, NR.	2340/260	NR	No
Acharya et al,[49] 2013[a]	Retrospective study, NR.	1800/200	NR	No
Acharya et al,[50] 2012[a]	Retrospective study, NR.	1800/200	NR	No
Umar et al,[51] 2012	Retrospective study, NR.	NR	NR	No
Acharya et al,[52] 2012[a]	Retrospective study, NR.	1800/200	NR	No
Al-Karawi et al,[53] 2021[a]	Retrospective study, data from the IOTA research.	150/148 74/76	2005.11−2013.11	No
Jian et al,[54] 2021	Retrospective study, eight centers.	144/75/75	2010.01−2019.02	No
Li et al,[55] 2020	Retrospective study, NR.	250/92/159	2010.01−2018.06	No
Acharya et al,[56] 2014[a]	Retrospective study, NR.	2340/260	NR	No
Pathak et al,[57] 2015[a]	Retrospective study, NR.	70/50	NR	No
Ameye et al,[58] 2009[a]	Retrospective study, data from the IOTA research.	754/507	1999−2006	No
Jian et al,[59] 2022[a]	Retrospective study, NR.	342/159	20102018	No
Alqasemi et al,[51] 2012[a]	Retrospective study, NR.	400/95	NR	Yes
Chen et al,[60] 2012 [a]	Retrospective study, data from the Ruijin Hospital affiliated with Shanghai Jiaotong university School of Medicine.	296/41/85	2019.01−2019.11	No
Zheng et al,[61] 2022	Retrospective study, data from the Tianjin Medical University General Hospital from November 2010 to May 2020.	125/31	2010−2020	No

***Table 4*: Data source for the 35 included studies.**
[a] Studies (n = 28) included in the meta-analysis.

−95%) for sample size > 300. The SP for ≤ 300 was 82% (80−85%) and 91% (84−96%) for > 300. The AUC was 0.90 (0.87−0.92) for ≤ 300 and 0.97 (0.95 −0.98) for > 300 (Supplementary figure 6a, b).

Thirteen studies were published before 2020. Fifteen studies were published after 2020. The pooled SE was 89% (84−93%) for published before 2020, and 88 (85−90%) for published after 2020. The SP was 89% (83−93%) and 83% (80−85%), respectively. The AUC was 0.95 (0.93−0.97) and 0.92 (0.89−0.94), respectively ((Supplementary figure 7a, b).

Fifteen studies were geographically distributed in Asia and thirteen studies were geographically distributed outside Asia. The pooled SE was 87% (84−90%) and 90 (85−93%), respectively. The SP was 83% (80 −86%) and 89% (82−93%), respectively. The AUC was 0.92 (0.89−0.94) and 0.95 (0.93−0.97), respectively (Supplementary figure 8a, b).

There were ten studies with low risk in more than three evaluation domains and eighteen studies with high risk. The pooled SE was 86% (78−91%) and 89% (87−91%), respectively. The SP was 92% (88−95%) and 81% (76−85%), respectively. The AUC was 0.93 (0.90−0.95) and 0.93 (0.91−0.95), respectively (Supplementary figure 9a, b).

Heterogeneity analysis

The meta-analysis results of 28 studies suggested that AI algorithms were beneficial for the diagnosis of OC from medical imaging from random-effects model. However, there was substantial heterogeneity among the included studies, SE had an I^2 = 94.68%, while SP had I^2 = 97.50% (p < 0.01). The detailed results of subgroup and meta-regression analyses exploring the potential source of between-study heterogeneity are shown in Table. 5 and Supplementary figure 10-23. The results highlighted a statistically significant difference. Visual inspection of funnel plots suggested there was no publication bias (p = 0.83) (Supplementary figure 24).

Discussion

With the widespread application of AI in medical imaging during recent years, radiomics and AI models are now being actively evaluated for diagnostic accuracy in a variety of malignancy types. To our best knowledge, this is the first systematic review and meta-analysis specifically dedicated to AI system performance in the diagnosis of OC. We are strictly in line with the guidelines for diagnostic reviews,[62] and conducted a comprehensive

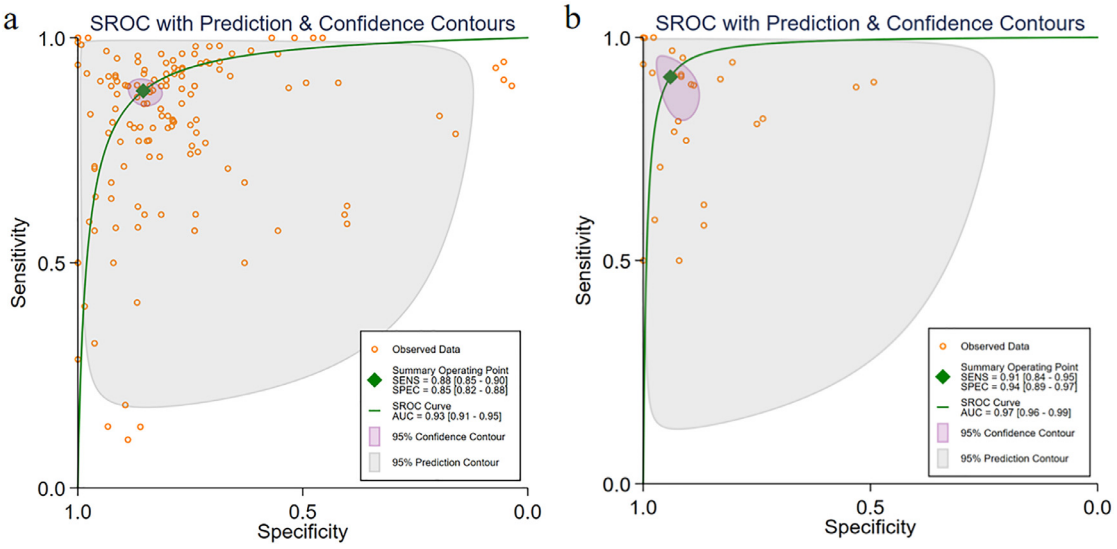

Figure 2. (a, b). SROC curves of all studies included in the meta-analysis (28 studies). a: SROC curves of ll studies included in the meta-analysis (28 studies with 160 tables). **b:** SROC curves of studies when selecting contingency tables reporting the highest accuracy (28 studies with 28 tables).

Abbreviations: AI: artificial intelligence; SROC = summary receiver operating characteristic; SENS = summary sensitivity; SPEC = summary specificity.

literature search in both medical databases and engineering and technology databases to ensure the rigor of the study.

After a careful selection of research on relevant topics, we found that AI algorithms excelled in the identification of OC using medical radiography imaging, which manifested an equivalent or even better performance than independent detection by human clinicians. This study also described the performance of the different imaging modalities, sample size, the year of publication, geographical distribution, and the different risk of bias levels. Potential sources of inter-study heterogeneity were identified based on the above subgroup and meta-regression analyses. More importantly, we rigorously rated study quality and risk of bias using an adapted QUADAS-AI assessment tool, which is a strength of this systematic review and will also better guide future related studies.

Advances in ML techniques may facilitate processing of large amounts of medical image data. Notwithstanding their utility, ML methods are known to have limitations[63] related to: manual extraction and selection of features, this is a fundamental task in order to find a group of significant variables to predict and correlate with outcome; Poor performance when dealing with imbalanced datasets. DL is the newest class of ML and

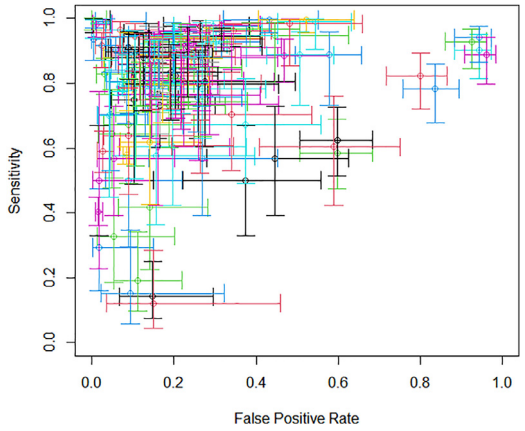

Figure 3. Cross-hair Plot of all studies included in the meta-analysis (28 studies with 160 tables).

	No. of studies	Sensitivity			*P* value[b]	Specificity			*P* value[b]
		Sensitivity	*P* value [a]	I^2 (95%CI)		Specificity	*P* value [a]	I^2 (95%CI)	
Overall	28	0.88 (0.85−0.90)	< 0.05	94.68 (94.16−95.19)		0.85 (0.82−0.88)	< 0.05	97.50 (97.31−97.69)	
Algorithm					< 0.05				< 0.05
Machine learning	19	0.89 (0.85−0.92)	< 0.05	95.11 (94.49−95.72)		0.88 (0.82−0.92)	< 0.05	97.69 (97.46−97.92)	
Deep learning	9	0.88 (0.84−0.91)	< 0.05	95.48 (94.84−96.11)		0.84 (0.80−0.87)	< 0.05	95.84 (95.28−96.41)	
Imaging modality					< 0.05				< 0.05
Ultrasound	17	0.91 (0.87−0.93)	< 0.05	96.58 (96.22−96.94)		0.87(0.82−0.91)	< 0.05	98.55 (98.43−98.66)	
Magnetic resonance imaging	6	0.83 (0.77−0.88)	< 0.05	85.72 (82.32−89.12)		0.84(0.80−0.87)	< 0.05	83.47 (79.37−87.58)	
Computed tomography	3	0.75 (0.68−0.81)	0.43	0.00 (0.00−100.00)		0.75 (0.67−0.82)	0.83	0.00 (0.00−100.00)	
Sample size					< 0.05				< 0.05
≤ 300	15	0.85 (0.81−0.88)	< 0.05	91.75 (90.61−92.90)		0.82 (0.80−0.85)	< 0.05	83.00 (80.08−84.93)	
> 300	13	0.93 (0.89−0.95)	< 0.05	97.96 (97.72−98.20)		0.91 (0.84−0.96)	< 0.05	99.42 (99.38−99.47)	
Risk of bias					< 0.05				< 0.05
Low	10	0.86 (0.78−0.91)	< 0.05	97.49 (97.14−97.84)		0.92 (0.88−0.95)	< 0.05	97.31 (96.92−97.69)	
High	18	0.89 (0.87−0.91)	< 0.05	91.78 (90.70−92.87)		0.81 (0.76−0.85)	< 0.05	95.94 (95.51−96.37)	
Geographical distribution					< 0.05				< 0.05
Asia	13	0.87 (0.84−0.90)	< 0.05	94.48 (93.74−95.22)		0.83 (0.80−0.86)	< 0.05	95.00 (94.35−95.65)	
Non Asia	15	0.90 (0.85−0.93)	< 0.05	96.36 (95.91−96.82)		0.89 (0.82−0.93)	< 0.05	98.17 (97.99−98.36)	
Year of publication					< 0.05				< 0.05
Before 2020	15	0.89 (0.84−0.93)	< 0.05	96.26 (95.81−96.71)		0.89 (0.83−0.93)	< 0.05	97.89 (97.68−98.10)	
After 2020	13	0.88 (0.85−0.90)	< 0.05	94.63 (93.87−95.39)		0.83 (0.80−0.85)	< 0.05	95.12 (94.45−95.79)	

***Table 5*: Summary estimate of pooled performance of artificial intelligence in image-based ovarian cancer detection.**

[a] *P*-Value for heterogeneity within each subgroup.

[b] *P*-Value for heterogeneity between subgroups with meta-regression analysis.

has been found to be advantageous to other forms of ML.[64] DL employs multiple layers of neural networks, leading to expanded 'neuronal' complexity, to significantly enhance computational power. However, with DL methods being more prone to overfitting and hence often requiring more data. Considering the stage of development of the algorithm and the difference in nature,[65,66] we also carried out a sub-analysis by the different algorithms, where no significant difference was observed. This may be attributed to the small dataset of included studies, most of which collected a few hundred data, limiting the advantages of DL.

Although great promise has been shown with AI algorithms in a variety of tasks across radiology and medicine as a whole, these systems are far from perfect, we should also critically consider some methodological issues:

First, data continues to be the most central and crucial constituent for learning AI systems.[67] Exploiting radiology report databases by using modern information processing technologies may improve report search and retrieval and help radiologists in diagnosis.[68] We need to call for advocacy for creating interconnected networks of identifying patient data from around the world and training AI on a large scale according to different patient demographics, geographic areas, diseases, etc. In addition, we emphasize that rare cancers, including OC, require more diverse image databases. In fact, maximization of the power of AI will require the deposition of medical data with sufficient annotation in large−scale databases.[69] However, such data are rarely curated, and this represents a major bottleneck in attempting to learn any AI model.[70] International collaborative projects (such as The Cancer Imaging Archive [http://www.cancerimagingarchive.net]) that build large, labeled datasets should make a substantial contribution to meeting this challenge. Curation can refer to patient cohort selection relevant for a specific AI task but can also refer to segmenting objects within images.[70,71] Curation ensures that training data adheres to a defined set of quality criteria and is clear of compromising artefacts. It can also help avoid unwanted variance in data owing to differences in data-acquisition standards and imaging protocols, especially across institutions, such as the time between contrast agent administration and actual imaging.[71−73] Only in this way can we create an AI that is socially responsible and benefits more people.

Second, as the advent of AI-based diagnostic test studies, there has been a parallel increase in the number of systematic reviews summarizing such findings.[7,74,75] Noteworthy, 94% studies have been performed in the absence of an AI-specific quality assessment criteria in those published systematic reviews.[27] During the past decade, the most frequently utilized tool is the QUADAS-2.[28] However, QUADAS-2 does not address the particular terminology that arises from AI diagnostic test studies, nor does it consider other issues that appear in AI research, such as the setting of the data set, sources of bias, etc.[27] Therefore, Sounderajah, V et al. proposed an AI-specific risk of bias tool, termed QUADAS-AI in 2021.[27] This tool provided us with a specific instruction to assess the risk of bias and applicability of the present study. Not surprisingly, most of the relevant studies were more often designed or conducted prior this guideline. We therefore accepted the low quality of some of the studies and the heterogeneity between the included studies. It also makes sense that we assume that patient selection, index test and flow and timing of studies used to evaluate the diagnostic performance of AI models will be optimized soon.

Third, although no publication bias was noticed in the present study, we must be honest about the fact that the available AI research is often a publication of positive results. We venture to guess that this phenomenon stems from reporting bias by researchers, which may have skewed the dataset and not conducive to the comparison between AI models and clinicians.[75,76] One more point, the extraordinary applications of AI technology in medicine will require healthcare workers to enhance their clinical workflow combination. Of the included studies, only two evaluated the performance of integrating AI with clinicians. It has been suggested that scientific research should shift from an AI-physician dichotomy to a combination of AI and clinicians, which would be more in line with realistic medical workflows.

Fourth, 28 of the 34 studies that met the inclusion criteria for this systematic review provided information of our concern for the development of contingency tables. There is a broad range of indicators employed in AI research to report diagnostic abilities. Metrics such as SE, SP, and accuracy are the most applied in numerous studies. If the number of subjects with/without disease is shown in the study, we can combine SE and SP to derive TP, TN, FP, and FN for the construction of the contingency table. Other metrics like precision, dice ratio, F1 score and recall, which are frequently used in computer science, also present as the default standard of measurement in some studies.[37] However, these metrics are not all-encompassing and alone we do not receive sufficient information to build a contingency table. Well-defined metrics at the intersection of health care systems and computer science are also prudent to consider for future research. Additionally, for AI based models, the obtained heatmaps show what aspects of the images are important for a given classification,[77] whereas few included studies provide such information. To reduce bias, we emphasize reporting information about segmentation properties or heat maps in AI model-based studies to draw conclusions about the elements of interest in AI models.

Fifth, there is a disagreement around the critical terminology applied in AI research. Different papers have

defined the same terminology in different ways. For example, for an AI-based model, the sample set is generally grouped into several separate sections, including a training set and a test set for evaluating the effectiveness of the model.[74] Although the term 'validation' is used in a causal sense, some researchers used this phrase to denote the dataset used to assess the diagnostic performance of the ultimate model.[22] Other investigations have described it as a dataset with a tweaking function in the exploitation process.[32] The inconsistency of naming renders it challenging to determine whether the set is independent. It is vital that the validation set comprises data isolated from training data and is exclusively dedicated to assess the eventual model. It has been proposed to classify the sample data set into a training set, a tuning set and a validation set, whose functions are to be applied for training the model, for tuning the parameters and for assessing the performance of the final model, respectively.[74] Considering the different sorts of validation sets, Altman et al.[78] designated the datasets used for in-sample validation as internal validation set and those for out-of-sample validation as external validation set, suggestions which are very realistic and contribute to the quality of the study. Researchers concerned with the application of AI in healthcare should be careful about the phenomena and optimize it for future research.

Sixth, within a purely image-based setting, AI can achieve on par or superior performance to physicians, thereby highlighting its potential as a decision support system with immediate clinical implications.[79] Although a fairly good evaluation can be made in this way, it does not take into account all the information that radiologists rely on when evaluating a difficult examination.[70] Nonimaging-based patient characteristics, such as demographic information, history of cancer, and genetic information, may be integrated into the model. Given a sufficiently large data set, AI could use these pieces of information in conjunction with the image data to identify women at high risk of cancer.

Seventh, the high performance of AI model comes at the cost of high complexity and vast number of parameters.[80] We may be unable to understand and explain why an AI model has made certain classifications in image analysis. This type of algorithm is often referred to as a "black box".[81] Compared with AI techniques, explainable artificial intelligence (XAI) can provide both decision-making and explanations of the model.[82] Some research have been conducted into XAI to overcome the limitation of the black-box nature of AI methods. For example, Laios et al.[83] have pioneered the implementation of XAI models in the field of gynecological oncology. They presented an ensemble AI-based model that predicted the outcomes following cytoreductive surgery for OC with high accuracy, and an XAI strategy that explained the patient and surgery-specific factors that led to that risk. The team also made a pioneering attempt to implement XAI models to explain the prediction of surgical effort at OC cytoreduction, by feeding the models with features that also include human factors.[84] However, most of radiomics extraction and imaging biomarkers analyses included in this review are used as "black box", and their application in clinical practice still lacks reliability and interpretability. This phenomenon is understandable given that the use of XAI in oncology is still in its infancy. Understanding the principles and applications of AI in medical imaging will facilitate assimilation and expedite advantages to practice.[85] We encourage future researches to consider the interpretability of AI models in modeling, to address challenges, and to find clinical approaches for the development of AI in the field of radiomics.

Eighth, most studies were carried out in a single center with limited data availability. Only three of the included studies have external validation, which refers to validating the performance of the model with out-of-sample datasets from other institutions. However, among the three included studies with external validation, only one was included in the meta-analysis. This precluded a subgroup meta-analysis in the present study, but emphasized the necessity for rigorous and reliable evaluation of AI performance in external datasets. The included studies are more likely to group an institution's dataset into a training set, a test set or an internal validation set. The performance was judged by the test set or internal validation. As the intention of the validation was to examine the performance of the model applied to patients from different populations, it is preferable to obtain a new dataset from a different organization. The lack of an external validation set may potentially lead to overestimation of the results, which could compromise the generalizability of the model.[78] Several reviews[7,75] in the AI field reported that studies with internally validated AI models outperform externally validated models in the detection of cervical cancer, breast cancer and tumor metastases. However, this is not surprising as the samples in the same dataset are often homogeneous and the diagnostic performance of the algorithm can easily be misjudged. Rigorous external validation is warranted in the design of AI-related diagnostic studies. Multicentric studies will have an significant role in this research field. The use of interoperable standards and uniform protocols will also be needed prior to conducting such a study. AI methods can provide valuable models for quality assurance, personalized and predictive medicine. For this purpose, the contribution of clinicians and researchers in the interpretation of models and their application has a crucial role in the daily clinical practice.

Additionally, limited prospective studies were carried out in real clinical environments. Most of the included studies were based on retrospective data and whose patients chosen from hospital medical records. It is well known that prospective studies would provide more

favorable evidence,[86] and we anticipate more prospective AI research to emerge in the future. And only considering the inclusion of English articles may omit important information from other language studies. Another limitation is that we did not contact the authors because most of the studies included in full-text screening (93%) provided the necessary data.

The present study represents a summary of the enormous potential of AI algorithms that are useful for detecting OC using medical radiology imaging. However, it is also acknowledged that this finding is derived from relatively low methodological quality research, which inevitably overestimates the accuracy of the algorithm. The research of AI-based systems in diagnosing OC needs to be further improved in terms of study design.

Contributors

H-LX, T-TG, H-ZS, YS, Y-HZ, and Q-JW contributed to the conception and design of the study. H-LX, F-HL, and H-YC contributed to the literature search and data extraction. H-LX, F-HL, and T-TG contributed to risk of bias evaluation. H-LX, T-TG, F-HL, H-YC, and Q-JW contributed to data analysis and interpretation. H-LX, F-HL, H-YC, QX, H-ZS, YS, SG, T-TG, and Q-JW wrote the first draft of the manuscript and edited the manuscript. All authors contributed to critical revision of the manuscript. All authors approved the manuscript. H-LX and T-TG contributed equally to this work.

Data sharing statement

The search strategy was shown in Supplementary Note 1, and the contingency tables of 28 studies included in the meta-analysis were shown in Supplementary Table 1. The results of risk of bias and publication bias were separately provided in the Supplementary Figure 8 and 9. Additional data are available on request.

Declaration of interests

All authors declare no competing interests.

Acknowledgments

This work was supported by the Natural Science Foundation of China, the LiaoNing Revitalization Talents Program, and the 345 Talent Project of Shengjing Hospital of China Medical University.

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.eclinm.2022.101662.

References

1 Jayson GC, Kohn EC, Kitchener HC, Ledermann JA. Ovarian cancer. *Lancet*. 2014;384(9951):1376–1388.
2 Auersperg N, Wong AS, Choi KC, Kang SK, Leung PC. Ovarian surface epithelium: biology, endocrinology, and pathology. *Endocr Rev*. 2001;22(2):255–288.
3 Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. *CA Cancer J Clin*. 2021;71(1):7–33.
4 Forstner R, Thomassin-Naggara I, Cunha TM, et al. ESUR recommendations for MR imaging of the sonographically indeterminate adnexal mass: an update. *Eur Radiol*. 2017;27(6):2248–2257.
5 Van Nimwegen LWE, Mavinkurve-Groothuis AMC, de Krijger RR, et al. MR imaging in discriminating between benign and malignant paediatric ovarian masses: a systematic review. *Eur Radiol*. 2020;30(2):1166–1181.
6 Ruytenberg T, Verbist BM, Vonk-Van Oosten J, Astreinidou E, Sjögren EV, Webb AG. Improvements in high resolution laryngeal magnetic resonance imaging for preoperative transoral laser microsurgery and radiotherapy considerations in early lesions. *Front Oncol*. 2018;8:216.
7 Zheng Q, Yang L, Zeng B, et al. Artificial intelligence performance in detecting tumor metastasis from medical radiology imaging: a systematic review and meta-analysis. *EClinicalMedicine*. 2020;31:100669.
8 Outwater EK, Dunton CJ. Imaging of the ovary and adnexa: clinical issues and applications of MR imaging. *Radiology*. 1995;194(1):1–18.
9 Medeiros LR, Freitas LB, Rosa DD, et al. Accuracy of magnetic resonance imaging in ovarian tumor: a systematic quantitative review. *Am J Obstet Gynecol*. 2011;204(1):67.e1–10.
10 Khiewvan B, Torigian DA, Emamzadehfard S, et al. An update on the role of PET/CT and PET/MRI in ovarian cancer. *Eur J Nucl Med Mol Imaging*. 2017;44(6):1079–1091.
11 Virarkar M, Ganeshan D, Gulati AT, Palmquist S, Iyer R, Bhosale P. Diagnostic performance of PET/CT and PET/MR in the management of ovarian carcinoma-a literature review. *Abdom Radiol (NY)*. 2021;46(6):2323–2349.
12 Li S, Liu J, Xiong Y, et al. A radiomics approach for automated diagnosis of ovarian neoplasm malignancy in computed tomography. *Sci Rep*. 2021;11(1):8730.
13 Lheureux S, Braunstein M, Oza AM. Epithelial ovarian cancer: evolution of management in the era of precision medicine. *CA Cancer J Clin*. 2019;69(4):280–304.
14 Zhang L, Wang H, Li Q, Zhao MH, Zhan QM. Big data and medical research in China. *BMJ*. 2018;360:j5910.
15 Mollura DJ, Culp MP, Pollack E, et al. Artificial intelligence in low- and middle-income countries: innovating global health radiology. *Radiology*. 2020;297(3):513–520.
16 Mookiah MR, Acharya UR, Chua CK, Lim CM, Ng EY, Laude A. Computer-aided diagnosis of diabetic retinopathy: a review. *Comput Biol Med*. 2013;43(12):2136–2155.
17 Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology*. 2016;278(2):563–577.
18 Tunali I, Gillies RJ, Schabath MB. Application of radiomics and artificial intelligence for lung cancer precision medicine. *Cold Spring Harb Perspect Med*. 2021;11(8):a039537. Published 2021 Aug 2.
19 Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol*. 2017;14(12):749–762.
20 Zhang H, Mao Y, Chen X, et al. Magnetic resonance imaging radiomics in categorizing ovarian masses and predicting clinical outcome: a preliminary study. *Eur Radiol*. 2019;29(7):3358–3371.
21 Aramendía-Vidaurreta V, Cabeza R, Villanueva A, Navallas J, Alcázar JL. Ultrasound image discrimination between benign and malignant adnexal masses based on a neural network approach. *Ultrasound Med Biol*. 2016;42(3):742–752.
22 Gao Y, Zeng S, Xu X, et al. Deep learning-enabled pelvic ultrasound images for accurate diagnosis of ovarian cancer in China: a retrospective, multicentre, diagnostic study. *Lancet Digit Health*. 2022;4(3):e179–e187.
23 Wang S, Zhang Y, Lei S, et al. Performance of deep neural network-based artificial intelligence method in diabetic retinopathy screening: a systematic review and meta-analysis of diagnostic test accuracy. *Eur J Endocrinol*. 2020;183(1):41–49.
24 Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71.

25 Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis of observational studies in epidemiology (MOOSE) group. *JAMA*. 2000;283(15):2008–2012.
26 Moons KG, de Groot JA, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med*. 2014;11(10):e1001744.
27 Sounderajah V, Ashrafian H, Rose S, et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. *Nat Med*. 2021;27(10):1663–1665.
28 Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8):529–536.
29 Yang B, Mallett S, Takwoingi Y, et al. QUADAS-C: a tool for assessing risk of bias in comparative diagnostic accuracy studies. *Ann Intern Med*. 2021;174(11):1592–1599.
30 Phillips B, Stewart LA, Sutton AJ. Cross hairs' plots for diagnostic meta-analysis. *Res Synth Methods*. 2010;1(3-4):308–315.
31 Liu X, Wang T, Zhang G, et al. 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*. 2022;15(1):22.
32 Saida T, Mori K, Hoshiai S, et al. Diagnosing ovarian cancer on MRI: a preliminary study comparing deep learning and radiologist assessments. *Cancers (Basel)*. 2022;14(4):987.
33 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.
34 Li S, Liu J, Xiong Y, et al. Application values of 2D and 3D radiomics models based on CT plain scan in differentiating benign from malignant ovarian tumors. *Biomed Res Int*. 2022;2022:5952296.
35 Wang H, Liu C, Zhao Z, 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.
36 Chiappa V, Bogani G, Interlenghi M, et al. The Adoption of radiomics and machine learning improves the diagnostic processes of women with ovarian masses (the AROMA pilot study). *J Ultrasound*. 2021;24(4):429–437.
37 Jian J, Xia W, Zhang R, et al. 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*. 2021;121:102194.
38 Wang R, Cai Y, Lee IK, et al. Evaluation of a convolutional neural network for ovarian tumor differentiation based on magnetic resonance imaging. *Eur Radiol*. 2021;31(7):4960–4971.
39 Hu Y, Weng Q, Xia H, et al. A radiomic nomogram based on arterial phase of CT for differential diagnosis of ovarian cancer. *Abdom Radiol (NY)*. 2021;46(6):2384–2392.
40 Yu XP, Wang L, Yu HY, et al. MDCT-based radiomics features for the differentiation of serous borderline ovarian tumors and serous malignant ovarian tumors. *Cancer Manag Res*. 2021;13:329–336.
41 Ștefan PA, Lupean RA, Mihu CM, et al. Ultrasonography in the diagnosis of adnexal lesions: the role of texture analysis. *Diagnostics (Basel)*. 2021;11(5):812.
42 Christiansen F, Epstein EL, Smedberg E, Åkerlund M, Smith K, Epstein E. Ultrasound image analysis using deep neural networks for discriminating between benign and malignant ovarian tumors: comparison with expert subjective assessment. *Ultrasound Obstet Gynecol*. 2021;57(1):155–163.
43 Akazawa M, Hashimoto K. Artificial intelligence in ovarian cancer diagnosis. *Anticancer Res*. 2020;40(8):4795–4800.
44 Martínez-Más J, Bueno-Crespo A, Khazendar S, et al. Evaluation of machine learning methods with Fourier transform features for classifying ovarian tumors based on ultrasound images. *PLoS One*. 2019;14(7):e0219388.
45 Mol BW, Boll D, De Kanter M, et al. Distinguishing the benign and malignant adnexal mass: an external validation of prognostic models. *Gynecol Oncol*. 2001;80(2):162–167.
46 Liu D, Zhang L, Indima N, et al. CT and MRI findings of type I and type II epithelial ovarian cancer. *Eur J Radiol*. 2017;90:225–233.
47 Kazerooni AF, Malek M, Haghighatkhah H, et al. Semiquantitative dynamic contrast-enhanced MRI for accurate classification of complex adnexal masses. *J Magn Reson Imaging*. 2017;45(2):418–427.
48 Acharya UR, Sree SV, Kulshreshtha S, et al. GyneScan: an improved online paradigm for screening of ovarian cancer via tissue characterization. *Technol Cancer Res Treat*. 2014;13(6):529–539.
49 Acharya UR, Sree SV, Saba L, Molinari F, Guerriero S, Suri JS. Ovarian tumor characterization and classification using ultrasound-a new online paradigm. *J Digit Imaging*. 2013;26(3):544–553.
50 Acharya UR, Sree SV, Krishnan MM, et al. Ovarian tumor characterization using 3D ultrasound. *Technol Cancer Res Treat*. 2012;11(6):543–552.
51 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(12):126003.
52 Acharya UR, Sree VS, Saba L, Molinari F, Guerriero S, Suri JS. Ovarian tumor characterization and classification: a class of GyneScan™ systems. *Annual international conference of the IEEE engineering in medicine and biology society*. 2012;2012:4446–4449.
53 Al-Karawi D, Al-Assam H, Du H, et al. An evaluation of the effectiveness of image-based texture features extracted from static B-mode ultrasound images in distinguishing between benign and malignant ovarian masses. *Ultrason Imaging*. 2021;43(3):124–138.
54 Jian J, Li Y, Pickhardt PJ, et al. MR image-based radiomics to differentiate type I and type II epithelial ovarian cancers. *Eur Radiol*. 2021;31(1):403–410.
55 Li Yong'ai, et al. MRI-based machine learning for differentiating borderline from malignant epithelial ovarian tumors: a multicenter study. *J Magn Reson Imaging : JMRI*. 2020;52(3):897–904.
56 Acharya UR, Mookiah MR, Vinitha Sree S, et al. Evolutionary algorithm-based classifier parameter tuning for automatic ovarian cancer tissue characterization and classification. *Ultraschall Med*. 2014;35(3):237–245.
57 Pathak H, Kulkarni V. Identification of ovarian mass through ultrasound images using machine learning techniques. *2015 IEEE international conference on research in computational intelligence and communication networks (ICRCICN)*. 2015137–140.
58 Ameye L, Valentin L, Testa AC, et al. A scoring system to differentiate malignant from benign masses in specific ultrasound-based subgroups of adnexal tumors. *Ultrasound Obstet Gynecol*. 2009;33(1):92–101.
59 Jian J, Li Y, Xia W, et al. MRI-based multiple instance convolutional neural network for increased accuracy in the differentiation of borderline and malignant epithelial ovarian tumors. *J Magn Reson Imaging*. 2022;56(1):173–181.
60 Chen H, Yang BW, Qian L, et al. Deep learning prediction of ovarian malignancy at US compared with O-RADS and expert assessment. *Radiology*. 2022;304(1):106–113.
61 Zheng Y, Wang H, Li Q, Sun H, Guo L. Discriminating between benign and malignant solid ovarian tumors based on clinical and radiomic features of MRI. *Acad Radiol*. 2022. S1076-6332(22)00331-2.
62 Deeks JJ, Bossuyt PM, Gatsonis C. *Handbook for DTA Reviews*. London: Cochrane Collaboration; 2011.
63 Litjens G, Ciompi F, Wolterink JM, et al. State-of-the-art deep learning in cardiovascular image analysis. *JACC Cardiovasc Imaging*. 2019;12(8 Pt 1):1549–1565.
64 Lee JG, Jun S, Cho YW, et al. Deep learning in medical imaging: general overview. *Korean J Radiol*. 2017;18(4):570–584.
65 Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. *Neural Comput*. 2006;18(7):1527–1554.
66 Manco L, Maffei N, Strolin S, Vichi S, Bottazzi L, Strigari L. Basic of machine learning and deep learning in imaging for medical physicists. *Phys Med*. 2021;83:194–205.
67 Hosny A, Parmar C, Quackenbush J, Schwartz LH. Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer*. 2018;18(8):500–510.
68 Zhou LQ, Wang JY, Yu SY, et al. Artificial intelligence in medical imaging of the liver. *World J Gastroenterol*. 2019;25(6):672–682.
69 Shimizu H, Nakayama KI. Artificial intelligence in oncology. *Cancer Sci*. 2020;111(5):1452–1460.
70 Geras KJ, Mann RM, Moy L. Artificial intelligence for mammography and digital breast tomosynthesis: current concepts and future perspectives. *Radiology*. 2019;293(2):246–259.
71 Ursprung S, Beer L, Bruining A, et al. Radiomics of computed tomography and magnetic resonance imaging in renal cell carcinoma-a systematic review and meta-analysis. *Eur Radiol*. 2020;30(6):3558–3566.
72 Wei J, Jiang H, Gu D, et al. Radiomics in liver diseases: current progress and future opportunities. *Liver Int*. 2020;40(9):2050–2063.

73 Bleker J, Kwee TC, Rouw D, et al. A deep learning masked segmentation alternative to manual segmentation in biparametric MRI prostate cancer radiomics. *Eur Radiol*. 2022;32(9):6526–6535.

74 Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health*. 2019;1(6):e271–e297.

75 Xue P, Wang J, Qin D, et al. Deep learning in image-based breast and cervical cancer detection: a systematic review and meta-analysis. *NPJ Digit Med*. 2022;5(1):19.

76 Simon AB, Vitzthum LK, Mell LK. Challenge of directly comparing imaging-based diagnoses made by machine learning algorithms with those made by human clinicians. *J Clin Oncol*. 2020;38(16):1868–1869.

77 Guimarães P, Batista A, Zieger M, Kaatz M, Koenig K. Artificial intelligence in multiphoton tomography: atopic dermatitis diagnosis. *Sci Rep*. 2020;10(1):7968. Published 14 May 2020.

78 Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med*. 2000;19(4):453–473.

79 Wu YT, Wei J, Hadjiiski LM, et al. Bilateral analysis based false positive reduction for computer-aided mass detection. *Med Phys*. 2007;34(8):3334–3344.

80 Papadimitroulas P, Brocki L, Christopher Chung N, et al. Artificial intelligence: deep learning in oncological radiomics and challenges of interpretability and data harmonization. *Phys Med*. 2021;83:108–121.

81 Yang G, Ye Q, Xia J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: a mini-review, two showcases and beyond. *Inf Fusion*. 2022;77:29–52.

82 Zhang Y, Weng Y, Lund J. Applications of explainable artificial intelligence in diagnosis and surgery. *Diagnostics (Basel)*. 2022;12(2):237. Published 19 Jan 2022.

83 Laios A, Kalampokis E, Johnson R, et al. Explainable artificial intelligence for prediction of complete surgical cytoreduction in advanced-stage epithelial ovarian cancer. *J Pers Med*. 2022;12(4):607. Published 10 Apr 2022.

84 Laios A, Kalampokis E, Johnson R, et al. Factors predicting surgical effort using explainable artificial intelligence in advanced stage epithelial ovarian cancer. *Cancers (Basel)*. 2022;14(14):3447. Published 15 Jul 2022.

85 Currie G, Hawk KE, Rohren E, Vial A, Klein R. Machine learning and deep learning in medical imaging: intelligent imaging. *J Med Imaging Radiat Sci*. 2019;50(4):477–487.

86 Seidelmann SB, Claggett B, Cheng S, et al. Dietary carbohydrate intake and mortality: a prospective cohort study and meta-analysis. *Lancet Public Health*. 2018;3(9):e419–e428.