



Deep learning and machine learning in image-based hepatocellular carcinoma detection: a systematic review and meta-analysis

Jingchao Wang¹ · Jingjie Bai¹ · Yanyang Wang¹ · Cong Wang¹ · Heping Deng¹

Received: 8 February 2026 / Revised: 3 June 2026 / Accepted: 8 June 2026

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2026

Abstract

Background Early detection and management are essential in hepatocellular carcinoma (HCC) care. Artificial intelligence (AI) is enhancing the precision and efficiency of diagnostic imaging for liver tumor assessment.

Objective This study aimed to characterize the diagnostic performance of AI algorithms for detecting HCC via systematic review and meta-analysis.

Methods A systematic literature search was conducted in PubMed, Embase, Cochrane, Web of Science, and IEEE Xplore Digital Library in November 2024 to identify studies evaluating AI algorithms for detecting HCC. Diagnostic performance data from the included studies were pooled via meta-analysis and compared against human clinicians (11 studies). Subgroup analyses were conducted based on classification task (binary classification tasks (BCTs) vs. multi-class classification tasks (MCCTs)), validation type, algorithmic approach (deep learning (DL) vs. machine learning (ML)), and imaging modality (ultrasound (US), computed tomography (CT), or magnetic resonance imaging (MRI)).

Results Forty-one studies evaluating AI-based diagnostic models were included. The pooled sensitivity (SEN), specificity (SPC), and area under the curve (AUC) were 87% (95% CI 84–90%), 91% (95% CI 89–93%), and 0.95 (95% CI 0.92–0.96). In comparison, 11 studies assessing human clinicians yielded a pooled SEN, SPC, and AUC of 78% (95% CI 73–83%), 91% (95% CI 86–94%), and 0.91 (95% CI 0.88–0.93). These findings indicated that AI algorithms were related to high diagnostic accuracy and exhibited numerically higher performance metrics than human clinicians. Among the encompassed studies, 31 addressed BCTs and ten addressed MCCTs. Subgroup analyses across validation type, algorithmic approach, and imaging modality consistently demonstrated high diagnostic performance for AI. However, substantial heterogeneity was observed overall and within subgroups, with no clear source identified.

Conclusions AI algorithms are associated with high diagnostic accuracy for image-based detection of HCC, potentially surpassing human performance. However, additional prospective investigations with rigorous external validation and larger sample sizes are still necessary to detect HCC in medical imaging (MI).

Keywords Hepatocellular carcinoma · Deep learning · Machine learning · Artificial intelligence · Diagnostic imaging

Introduction

Primary liver cancer, most notably hepatocellular carcinoma (HCC), is the third predominant cause of cancer-related mortality globally, reflecting the significance of early detection and timely clinical management [1]. Chronic liver disease, especially cirrhosis, represents the strongest risk factor for the development of HCC, accounting for approximately 80% of cases [2]. Hepatocarcinogenesis in the cirrhotic liver progresses from nonmalignant precursor nodules - such as regenerative nodules and low- or high-grade dysplastic nodules - to overt HCC [3, 4]. In high-risk patients with chronic liver disease, non-invasive imaging modalities, such as

✉ Heping Deng
37500402@hebmu.edu.cn
Jingchao Wang
wangjingchao0123@163.com
Jingjie Bai
1192289864@qq.com
Yanyang Wang
15612199068@163.com
Cong Wang
wangcong0207@163.com

¹ Department of ultrasound, Hebei Medical University Third Hospital, Shijiazhuang, China

ultrasound (US), computed tomography (CT), and magnetic resonance imaging (MRI), have replaced biopsy as the primary diagnostic reference standard for HCC when lesions demonstrate the typical imaging characteristics. Biopsy is now primarily indicated when imaging features are atypical or when additional pathological information is required [5]. Nevertheless, despite substantial improvements in imaging resolution and technological sophistication, accurate HCC detection remains dependent on radiologist expertise, leading to variability and diagnostic inconsistency, especially in resource-limited settings. The foregoing limitations impede diagnostic reliability when using imaging alone, highlighting an urgent need for diagnostic approaches that reduce dependence on subjective interpretation while sustaining high diagnostic accuracy.

In liver tumor assessment, artificial intelligence (AI)-based analytical systems have demonstrated diagnostic performance that is comparable to, or, in certain studies, exceeds that of experienced clinicians across multiple imaging modalities [6–10]. By leveraging complex machine learning (ML) and deep learning (DL) architectures, these algorithms can extract high-order, non-intuitive imaging features, thereby enhancing diagnostic discrimination. In a meta-analysis of 44 studies, Salehi et al. reported pooled sensitivity (SEN) and specificity (SPC) of 85% and 84% for AI-based HCC assessment, versus 70% and 85% for clinicians [11].

However, existing studies and prior meta-analyses exhibit important methodological limitations [11, 12] - including heterogeneous study quality, inconsistent results, and a lack of granular subgroup analyses - that hinder the formation of unified conclusions and constrain clinical adoption. Therefore, we conducted a systematic review and meta-analysis to evaluate the diagnostic performance of AI algorithms for HCC detection and to explore sources of heterogeneity across studies.

Methods

Protocol registration and study design

The study protocol was registered with PROSPERO (Registration No: CRD420251070494). Our study followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [13]. Ethical approval or informed consent was exempted.

Search strategy and eligibility criteria

PubMed, Embase, Cochrane, Web of Science, and IEEE Xplore Digital Library were retrieved until November

2024. No restrictions were placed on geographic location or publication type. Nevertheless, non-English studies, letters, scientific reports, conference abstracts, as well as narrative reviews, were removed. The search strategy was formulated with experienced clinicians and medical researchers, and the strings are provided in Online Appendix 1.

Two independent reviewers screened titles, abstracts, and full texts using a standardized eligibility form. Disagreements were first resolved through discussion; if consensus could not be reached, a third reviewer made the final decision. This process was predefined before screening. Studies were eligible if they reported the diagnostic performance of AI algorithms for hepatic cancer detection via medical imaging (MI), including outcomes like accuracy, SEN, and SPC. There were no restrictions on participant characteristics, imaging approach, or the intended clinical AI application context.

Histopathological confirmation or imaging findings consistent with established diagnostic criteria were the reference standard. Studies with imperfect ground truths (e.g., low SEN or SPC) were excluded. Investigations focusing on image segmentation performance were excluded due to the absence of diagnostic classification. Animal or non-human studies, as well as duplicate reports, were excluded. Primary outcome assessment centered on diagnostic performance, assessed using SEN, SPC, and area under the curve (AUC) derived from the confusion matrix. The secondary evaluation focused on methodological quality using QUADAS-AI and adherence to reporting standards.

Data extraction

Two reviewers independently extracted study characteristics and diagnostic performance information through a standard form. The two reviewers both have backgrounds in radiology. One reviewer is a radiologist with 10 years of radiological experience, and the other is a radiologist with 8 years of radiological experience, specializing in diagnostic imaging and AI.

Dissents were addressed via consensus or adjudicated by a third reviewer. Binary diagnostic accuracy data were extracted into contingency tables, including true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for computing pooled SEN, SPC, and others. When multiple contingency tables were reported for varied AI models or subgroups in the same study, we selected only one model per dataset (the model with the highest diagnostic accuracy) for the primary analysis. Studies that did not report sufficient data to construct a complete 2×2 table were excluded from the quantitative synthesis.

Quality assessment

The risk of bias (RoB) and applicability of the encompassed studies were examined independently by three reviewers utilizing QUADAS-AI [14], which is an adaptation of QUADAS-2 specifically designed for studies evaluating AI-based diagnostic interventions. The tool assesses four domains: patient selection, index test (AI algorithm), reference standard, and flow and timing. Each domain is rated as “low risk,” “high risk,” or “unclear risk” based on predefined signaling questions. Disagreements among reviewers were resolved through discussion or, if necessary, by consensus with a fourth reviewer.

Statistical analysis

This study employed a bivariate random-effects model (Reitsma model) to jointly model SEN and SPC while accounting for the correlation between them. Based on this model, a hierarchical summary receiver operating characteristic (HSROC) curve was constructed to evaluate overall diagnostic performance. Pooled SEN, SPC, and AUC were estimated using the HSROC model. Additionally, the 95% confidence region and 95% prediction region were calculated, where the confidence region reflects the uncertainty of the pooled estimates, whereas the prediction region indicates the range of diagnostic performance that may be observed in future new studies. Meta-analyses were carried out to report the best accuracy in studies reporting many AI models. Heterogeneity was examined via the I^2 statistic, and subgroup and regression analyses were executed to find

possible sources. Publication bias was visually assessed via funnel plots.

Separate meta-analyses were carried out: 1. BCTs: (i) validation type (internal vs. external), (ii) AI algorithm type (DL vs. ML), (iii) imaging modality (US, CT, MRI, etc.); 2. MCCTs: (i) validation type (internal vs. external), (ii) AI algorithm type, (iii) imaging modality; 3. Pooled performance comparison of AI algorithms versus human clinicians on the same dataset.

Meta-analyses were performed only when ≥ 2 original studies were available. Data were analyzed via STATA MP 18.0, and quality was assessed through RevMan 5.4. $P < 0.05$ signified statistical significance, and all tests were two-tailed.

Data availability

The search strategy and aggregated data are displayed in the supplementary documents.

Result

Study selection and characteristics

12,818 records were initially identified, of which 9,729 were screened after removing 3,089 duplicates. Following title and abstract checking, 9,295 articles were excluded for failing to meet predefined eligibility criteria. 434 full-text articles were retrieved and reviewed, among which 393 were excluded based on a detailed eligibility assessment. Ultimately, 41 studies on HCC were encompassed (Fig. 1). Study characteristics are displayed in Tables 1, 2 and 3 [6, 8–10, 15–51].

Regarding the study design, 31 studies used retrospective data, one adopted a prospective design, and nine did not specify whether the data collection method was prospective or retrospective. Three studies utilized open-access image sources; 22 explicitly removed low-quality images, whereas 19 did not report image quality criteria. In terms of classification tasks, thirty-one studies employed binary classification tasks (BCTs) (HCC vs. non-HCC lesions), and ten used multi-class classification frameworks. For model validation, 11 studies involved external validation with out-of-sample datasets, whereas the rest employed internal validation. The types of external validation used included: 4 studies with independent institutions, 3 studies with different time periods, 3 studies with temporal separation, and 1 study with a public dataset. Regarding AI algorithms, nineteen studies only focused on DL algorithms and eight on ML algorithms. Seven studies compared DL models with human clinicians, four compared ML models with human clinicians, and three

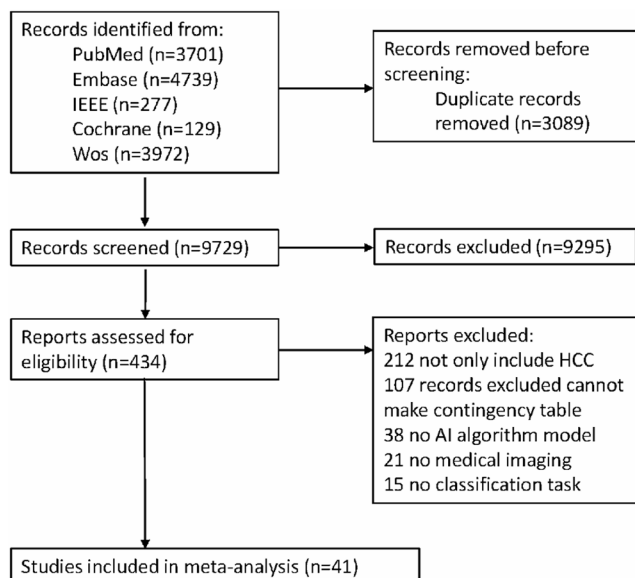


Fig. 1 PRISMA flowchart of study selection. Displayed is the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow of search methodology and literature selection process

Table 1 Study design and basic demographics

First author	Participants	Inclusion criteria	Exclusion criteria	N(patient)	Mean or median age (SD; range)
Zossou et al. [15]	NR	NR	NR	300	NR
Xiao et al. [16]	(1) patients at risk for HCC as defined by the CEUS LI-RADS 2017 criteria, including those with cirrhosis, chronic hepatitis B virus infection without cirrhosis, a current or past history of HCC, and includes adult candidates for liver transplantation and adult liver transplant recipients. (2) at least one liver local lesion untreated of LR-M [LR-M = probably or definitely malignant but not HCC specific, the criteria is nodule of any size and washout (< 60 s) or marked washout within 2 min or rim enhancement and washout]; (3) 18 years old. (4) visible at pre-contrast ultrasound. (5) utilizing Sonovue as the contrast agent. (6) untreated lesions. (7) complete images, particularly the dynamic video data of the arterial phase. (8) availability of pathological or clinical diagnosis results.	(1) without the above HCC risk factors. (2) <18 years old. (3) with cirrhosis due to congenital hepatic fibrosis. (4) with cirrhosis due to a vascular disorder such as hereditary hemorrhagic telangiectasia, Budd-Chiari syndrome, etc. (5) tumor within the portal vein, hepatic vein, or both (classified as CEUS LR-TIV). (6) severe cardiovascular dysfunction. (7) diffuse HCC. (8) lesions previously treated or those with local recurrence after treatment. (9) CEUS date missing.	268	Training and internal validation sets: 53 (NR, NR) External validation sets: 54 (NR, NR)	
Li et al. [17]	NR	NR	NR	86	NR
Chen et al. [18]	(1) The diagnosis of HCC, ICC, and cHCC-ICC were confirmed by postoperative pathology. (2) BMUS examination was performed within 14 days before hepatic resection.	(1) patients with multiple lesions. (2) patients who received anticancer treatment (trans-arterial chemoembolization, radiofrequency ablation, etc.) before the BMUS examination. (3) patients with incomplete clinical data or poor imaging qualities.	465	Modeling cohort: HCC 57.39 (11.23, NR) ICC 60.79 (12.22, NR) cHCC-ICC 57.36 (11.32, NR) Independent test cohort: HCC 57.00 (10.87, NR) ICC 65.26 (5.48, NR) cHCC-ICC 55.38 (8.48, NR) NR NR	
Zhao et al. [19]	NR	NR	NR	1822 cases not patients (245 benign cases and 1577 malignant cases) case 317	NR
Zhang et al. [20]	NR	NR	NR	NR	NR

Table 1 (continued)

First author	Participants	Inclusion criteria	Exclusion criteria	N(patient)	Mean or median age (SD; range)
Xu et al. [21]	Newly pathologically confirmed HCC or non-HCC within 1 month after undergoing abdominal CECT.		Did not comply with LI-RADS v2018 inclusion criteria; lesions classified as benign, probably benign, unable to define as benign or malignant on imaging; individual lesion size ≤ 1 cm and ≥ 10 cm or diffuse lesions; portal or hepatic vein thrombosis; prior treatment for liver lesions, such as transarterial chemoembolization, high-intensity focused ultrasound, etc.; lack of complete clinical data; and incomplete or poor image quality.	1049	HCC: 53 (NR) Non-HCC: 51 (NR)
Wan et al. [22]	NR		NR	174	HCC: 62 (NR) ICC: 63 (NR)
Roy et al. [23]	NR		NR	1080	NR
Park et al. [24]	(1) age ≥ 18 years. (2) high risk for HCC according to the LI-RADS v2018 (presence of cirrhosis or chronic hepatitis B infection regardless of the presence of cirrhosis). (3) MRI findings of a focal hepatic solid nodule.		(1) inadequate final diagnosis such as unknown final diagnosis of malignancy as a result of immediate locoregional therapy or insufficient follow-up (< 2 years) for benign lesions to determine size stability. (2) poor quality of images for interpretation. (3) only observations categorized as LR1, LR2, LR5, LR-TIV, or LR-M, according to the LI-RADS v2018 MFs and the algorithm.	187	Development: 62.8 (NR, 33–84) Test: 64.4 (NR, 37–81)
Midya et al. [9]	Patients who had abdominal CT scans within three months of operation with a routine clinical imaging protocol of contrast-enhanced portal venous phase CT were included		The patients who had no contrast-enhanced CT scans; had metal artifacts infiltrating the tumor on CT imaging; had prior ablation, embolization, resection, or transplantation, as these prior treatments would alter the appearance of the tumors on imaging and compromise the quantitative image analysis; had tumors that were ruptured; or had tumors with a diffuse infiltrative pattern (as tumor borders were challenging to determine for analysis).	814	NR
Liu et al. [25]	(1) underwent a preoperative MRI examination. (2) had no history of treatment for hepatic tumor prior to the study. (3) pathologically confirmed HCC or MF- ICC.		(1) image quality was insufficient for further analysis. (2) T2WI-MRI was incomplete.	112	MF-ICC: 58.45 (NR, 27–78) HCC: 52.68 (NR, 31–92)
Li et al. [10]	(1) the presence of chronic hepatitis B (2) the presence of visible nodules on CEUS. (3) confirmation of all visible nodules by postoperative pathology.		(1) any treatment before CEUS. (2) those that did not undergo CEUS at the corresponding center within three months before surgery. (3) absence of standard images available for review by the investigators.	454	Training Cohort: HCC: 57.26 (11.16, NR) LM: 59.15 (9.67, NR) Validation cohort: HCC: 61.52 (10.84, NR) LM: 59.75 (11.27, NR) 55.0 (12.0, 20–81)
Huang et al. [26]	Consecutive patients with pathologically confirmed HCC or ICC who underwent hepatectomy		(1) Pathologically diagnosed as cHCC-CC (2) Without preoperative CT images (3) CT images with artifacts, blurs or faults (4) Lesions unable to be accurately segmented	494	

Table 1 (continued)

First author	Participants Inclusion criteria	Exclusion criteria	N(patient)	Mean or median age (SD; range)
Fotouhi et al. [8]	(1) 140 LR-5 definitely HCC nodules, which were evaluated by post-operative sample pathology results; (2) 115 LR-3 intermediate probability of malignancy liver nodules, which were evaluated by last follow-up imaging. (3) 137 LR-4 the probable HCC nodules, which were evaluated by core needle biopsy and pathology results.	(1) prior surgery (2) prior chemoembolization (3) Imaging other than MRI (4) Nodules with indeterminate diagnosis	245	Benign: 63.2.4 (10.2, NR) HCC: 69.2 (13.5, NR)
Chen et al. [27]	(1) Patients with pathologically confirmed diagnosis of IMCC or poorly differentiated HCC based on surgical resection or percutaneous biopsy. Combined hepatocellular-cholangiocarcinomas have been excluded. (2) Gd-EOB-DTPA-enhanced MRI performed within 1 month before first invasive procedure or any other treatment.	(1) History of liver surgery or interventional therapy. (2) Poor imaging quality. (3) Concomitant with other liver malignancies. (4) The appearance of a classic IMCC in the hepatobiliary phase of Gd-EOB-DTPA-enhanced MRI, which is hyperintensity at the central region with a peripheral hypointense rim defect in the hepatobiliary phase	134	aIMCC: 59.49 (9.46, 37–73) pHCC: 56.99 (11.27, 26–79)
Zhang et al. [28]	(1) patients with HCC and FNH were all pathologically confirmed after surgical resection. (2) all enrolled patients underwent US examination before surgery. (3) patients with multiple lesions had pathologically confirmed ones enrolled	(1) patients have complicated clinical conditions such as pregnancy and taking medication for collagen diseases. (2) patients received additional treatment before examination such as chemotherapy, radiofrequency ablation (RFA), or transcatheter arterial chemoembolization (TACE).	407	Model: 44.51 (16.34, NR) External test: 48.12 (13.68, NR)
Xu et al. [29]	All patients with pathologic results of liver cancer underwent non-contrast CT at our institution	(1) lack of abdominal noncontrast CT image at 1.5 mm thickness. (2) poor image quality. (3) an interval between pathological results and CT examination of more than 3 months. (4) coinfection with virus such as HBV or HCV. (5) incomplete clinical data	211	NR

Table 1 (continued)

First author	Participants Inclusion criteria	Exclusion criteria	N(patient)	Mean or median age (SD; range)
Sun et al. [30]	Inclusion criteria for SHCC patients were as follows: (1) patients were diagnosed with SHCC by a radiologist with more than ten years of experience. (2) Patients received two MR examinations, of which SHCC was not apparent on the former but was diagnosed on the latter. The interval between the two examinations was not more than one year. (3) Patients had never been treated with transcatheter arterial chemoembolization (TACE) in the newly developed SHCC lesion. (4) Patients' MR imaging at the newly developed SHCC lesion site was consistent with the 'wash-in and wash-out' phenomenon described in the guidelines. (5) Patients had complete T1-weighted, arterial-phase, portal venous-phase, and delayed-phase MR images at both time-points. The inclusion criteria for external-control cohort are as follows: (1) patients were diagnosed with hepatic cyst or haemangioma by a radiologist with more than ten years of experience. (2) There were one to three lesions, and each one measured less than 2.3 cm in diameter. (3) Patients had no history of hepatitis, drug-induced liver damage, or alcohol abuse. (4) Patients had complete T1-weighted, arterial-phase, portal venous-phase, and delayed-phase MR images NR	(1) the tumour outline was unclear on MR images in the SHCC cohort. (2) The quality of the MR images was poor in the internal- or external-control cohort.	124	SHCC: 56.9 (9.35, 36–73) Hepatic cyst or haemangioma: 48.5 (14.88, 36–73)
Ling et al. [31]	NR	NR	479	56 (11, NR)
Hu et al. [32]	Only patients with diagnosis confirmed by pathology were included in the study	(1) 48 lesions with a size of less than 2 cm in all dimensions. (2) liver tumors with mixed-cell carcinoma. (3) time from MRI examination to pathological diagnosis of greater than one year. (4) images with incomplete or inappropriate imaging protocols. (5) images with motion or other artifact impeding interpretability. NR	514	NR
Cao et al. [33]	(1) diagnosed HCC or HCH based on liver biopsy or clinical findings (based on clinical signs such as laboratory tests, imaging examinations) (2) no contraindications to contrast medium and has undergone upper abdominal contrast-enhanced CT scans	NR	50	NR
Zhou et al. [34]	a consecutive of patients with definite pathological results of non-cystic FLL were registered	(1) benign lesions. (2) without cirrhosis. (3) with previous treatment. (4) without US images. (5) lesion size < 1.0 cm. (6) unsatisfied US image quality.	172	Training Cohort: HCC: 58.9 (11.2, NR) OM: 60.2 (11.1, NR) Test cohort: HCC: 59.8 (10.4, NR) OM: 63.4 (8.5, NR)

Table 1 (continued)

First author	Participants Inclusion criteria	Exclusion criteria	N(patient)	Mean or median age (SD; range)
Zheng et al. [35]	(1) presence of cirrhosis. (2) lesion size ≤ 2 cm (3) <1-month interval between MRI and pathological examination	(1) examinations had not been performed using Philips Ingenia equipment; (2) history of extrahepatic malignant tumors; (3) history of local treatment for HCC; (4) severe motion artifacts detected between DCE-MRI and DWI (> 5 slices of misalignment).	120	Training-validation cohort: sHCC: 60.2 (8.9, NR) Cirrhosis: 56.0 (13.7, NR) External test cohort: sHCC: 57.1 (8.8, NR) Cirrhosis 56.3 (11.9, NR)
Wang et al. [36]	NR	NR	8453	HPPH training set: HCC: 57.36 (NR, 8–86) Non-HCC: 54.19 (NR, 2–93) HPPH test set: HCC: 56.38 (NR, 17–87) Non-HCC: 52.42 (20–84) HPCH test set: HCC: 55.43 (NR, 28–82) Non-HCC: 53.74 (NR, 17–90) NR
Tiyarat-tanachai et al. [37]	NR	USG studies without further investigation for definitive diagnoses of FLLs and USG studies in which the lesion characteristics were altered by prior treatments.	4,808	
Stoll-mayer et al. [38]	Included lesions were either histologically confirmed or exhibited typical characteristics of the given lesion type with MRI.	Patients younger than 18 years of age at the time of imaging	216	FNH 45 (12, NR) HCC 66 (5, NR) MET 57 (10, NR) Total 54 (14, NR) 61.0 (11.6, 14–94)
Ryu et al. [6]	NR	NR	3873	
Ren et al. [39]	(1) being pathologically confirmed as HCC or ICC. (2) having received liver ultrasound within 1 month before surgery and the ultrasound images information being intact. (3) having not received antitumor treatments before, including liver transplantation (LT), microwave ablation (MWA), radiofrequency ablation (RFA), and transcatheter arterial chemoembolization (TACE). (4) the ultrasound images satisfying the analytical requirements and the target lesions being totally visible on the ultrasound images. (5) no history of concurrent malignancies.	(1) History of other cancer. (2) previous treatment history of HCC or ICC. (3) Liver ultrasonography was performed over 1 month before surgery. (4) Incomplete ultrasound and clinical data are available. (5) Image quality dissatisfy analysis. (6) The target lesion was unclearly visible.	226	Training set: 57.2 (11.1, NR) Internal Test set: 58.7 (9.3, NR) External test: 59.1 (11.2, NR)

Table 1 (continued)

First author	Participants Inclusion criteria	Exclusion criteria	N(patient)	Mean or median age (SD; range)
Oestmann et al. [40]	(1) Only patients with histopathological diagnosis were included. (2) In the non-HCC class, only primary liver neoplasms were included. (3) HCC lesions with locoregional therapy performed between MR imaging and resection/transplantation were included only if residual viable tumor was present on histology that would allow confirmation of etiology.	(1) If more than one lesion was visible on MRI in the segment described by the pathologist, images of CT-guided biopsy were used to ascertain the biopsied lesion. If these were unavailable, all lesions in the segment were excluded. (2) Lesions that were biopsied before the MRI scan were excluded if procedure-related hemorrhage was leading to significant alteration of T1 signal. (3) HCC lesions with locoregional therapy performed between MR imaging and resection/transplantation were excluded when tumors with complete necrosis.	118	HCC: 61 (8, NR) Non-HCC: 59 (13, NR)
Nakai et al. [41]	patients who underwent initial liver surgery for suspected primary liver cancer at our hospital from January 2004 through September 2019 were included	(1) preoperative history of transcatheter arterial chemoembolization and radiofrequency ablation for primary liver cancer. (2) tumor size ≤ 20 mm. (3) no suitable triple-phase CT within 100 days before surgery. (4) inappropriate artifact on the preoperative CT. (5) unclear tumor on the preoperative CT. (6) pathological diagnosis of the surgical specimen was not primary liver cancer. (7) massive intra-abdominal bleeding on the preoperative CT. (8) patients ≤ 20 years.	617	Training: mHCC: 68.4 (9.9, NR) pHCC: 67.2 (10.2, NR) ICC: 68.0 (10.1, NR) Validation: mHCC: 67.4 (9.1, NR) pHCC: 61.8 (11.4, NR) ICC: 67.2 (12.7, NR) Test: mHCC: 67.4 (9.0, NR) pHCC: 65.6 (11.5, NR) ICC: 62.7 (12.9, NR) NR
Menegotto et al. [42]	NR	NR	174	NR
Zhou et al. [43]	at least one FLL appeared in the CT image; malignant tumors had corresponding pathological diagnosis results; patients had not undergone any disease-related treatment before CT inspection, including transarterial chemoembolization therapy, surgery, radiofrequency ablation, and systemic chemotherapy; image quality was high and clear.	NR	435	NR
Shi et al. [44]	patients (age ≥ 18 years) with FLLs other than cysts underwent four-phase CT exams in our hospital	(1) lesions that could not be reliably classified by the best available reference standard as HCC or non-HCC. (2) lesion sizes below 1 cm. (3) CT exams with fewer than four phases or with severe image artifacts. (4) previous transcatheter arterial chemoembolization or other previous locoregional therapy. (5) loss to follow-up.	342	49.1 (10.5, 19–86)

Table 1 (continued)

First author	Participants Inclusion criteria	Exclusion criteria	N(patient)	Mean or median age (SD; range)
Kim et al. [45]	hepatobiliary phase images of the pre-operative gadoxetic acid-enhanced liver MRI	The exclusion criteria for internal-control cohort are as follows: (1) severe motion artifacts. (2) missing images, low image quality. (3) absence of preoperative MR images The exclusion criteria for external-control cohort are as follows: (1) motion artifact. (2) missing image. (3) low image quality NR	500	Internal set: 56 (9.7, NR) External set: 57 (9.6, NR)
Huang et al. [46]	NR		NR	NR
Cao et al. [47]	NR	Lesions larger than 10 cm; images with prominent artifacts; and prior local-regional therapy prior to the CT examination.	375	Category A: HCC Training set: 49 (NR, 24–81), Test set: 49.5 (NR, 33–70) Category B: Metastases Training set: 58.5 (NR, 23–79) Test set: 58 (NR, 23–79) Category C: Benign non-inflammatory FLLs Training set: 34 (NR, 17–82) Test set: 34 (10–74) Category D: Hepatic abscesses Training set: 54 (NR, 4–82) Test set: 55 (NR, 25–82) NR
Özyurt et al. [48]	NR	NR	75	
Oyama et al. [49]	NR	NR	93	HCC: NR (NR, 55–85) MT: NR (NR, 43–79) HH: NR (NR, 31–81)

Table 1 (continued)

First author	Participants	Inclusion criteria	Exclusion criteria	N(patient)	Mean or median age (SD; range)
Hamm et al. [50]	Patients' abdominal MRI examinations with one of the following hepatic lesions: simple cyst, cavernous hemangioma, focal nodular hyperplasia (FNH), HCC, intrahepatic cholangiocarcinoma (ICC), and colorectal cancer (CRC) metastasis were included in this study. Up to three imaging studies per patient were included as long as studies were more than 3 months apart. Up to nine different lesions were used in each study. The majority of included lesions were untreated; treated lesions were only included if the selected lesion showed progression, or the patient underwent loco-regional therapy more than 1 year ago and now presented with residual tumor.		The images were excluded if two radiologist contained discrepancies or displayed poor image quality. Patients younger than 18 years were excluded.	296	57 (14, NR)
Gatos et al. [51]	NR		NR	71	NR

directly compared DL and ML algorithms using identical datasets. Imaging modalities encompassed US (B-mode and contrast-enhanced ultrasound (CEUS)) in 12 studies, CT (including contrast-enhanced computed tomography (CECT)) in 16 studies, as well as MRI (including contrast-enhanced MRI) in 13 studies.

Pooled performance of studies

All 41 studies contributed sufficient quantitative data for diagnostic performance analysis, and 50 contingency tables were generated. The pooled SEN, SPC, and AUC were 87% (95% CI 84–89%), 91% (95% CI 89–93%), and 0.95 (95% CI 0.92–0.96) (Fig. 2a).

Among the 28 studies evaluating DL algorithms (35 contingency tables), the pooled SEN, SPC, and AUC were 86% (95% CI 83–89%), 92% (95% CI 90–94%), and 0.95 (95% CI 0.93–0.97), respectively (Fig. 2b). The 13 studies assessing ML algorithms (15 contingency tables) yielded a pooled SEN of 88% (95% CI 83–91%), SPC of 84% (95% CI 79–88%), and AUC of 0.92 (95% CI 0.90–0.94) (Fig. 2c). In comparison, 11 studies evaluating clinician performance (29 contingency tables) demonstrated pooled SEN, SPC, and AUC of 78% (95% CI 73–83%), 91% (95% CI 86–94%), and 0.91 (95% CI 0.88–0.93), respectively (Fig. 2d).

Subgroup meta-analyses

Our meta-analysis encompassed 31 studies employing BCTs and 9 studies employing multi-class classification tasks (MCCTs). The 31 BCT studies, which generated 37 contingency tables, yielded pooled SEN, SPC, and AUC of 88% (95% CI 85–91%), 88% (95% CI 85–91%), and 0.94 (95% CI 0.92–0.96), respectively (Fig. 3a). The 10 MCCT studies, generating 13 contingency tables, demonstrated pooled SEN, SPC, and AUC of 82% (95% CI 78–86%), 95% (95% CI 93–96%), and 0.96 (95% CI 0.94–0.97), respectively (Fig. 3b).

Three separate meta-analyses were carried out in the BCTs group

Validation types

Among the 27 studies (28 contingency tables) using internal validation datasets, the pooled SEN, SPC, and AUC were 89% (95% CI 86–92%), 89% (95% CI 86–92%), and 0.95 (95% CI 0.93–0.97), respectively (Fig. 4a). In contrast, nine studies (nine contingency tables) employing external validation datasets yielded pooled SEN, SPC, and AUC of 85% (95% CI 79–90%), 84% (95% CI 77–89%), and 0.92 (95% CI 0.89–0.94), respectively (Fig. 4b).

Table 2 Model training and validation methods

Author [ref]	Reference standard	Type of internal validation	External validation
Zossou et al. [15]	Consistent label, expert consensus	Hand-out cross validation	No
Xiao et al. [16]	Histopathology or the combination of clinical and imaging reference standards	Hand-out cross validation	Yes
Li et al. [17]	Histopathology	Hand-out cross validation:	No
Chen et al. [18]	Histopathology	Five-fold cross-validation	No
Zhao et al. [19]	NR	Hand-out cross validation	No
Zhang et al. [20]	NR	Hand-out cross validation	No
Xu et al. [21]	Histopathology	Hand-out cross validation	Yes
Wan et al. [22]	Histopathology	Five-fold cross validation	No
Roy et al. [23]	NR	Hand-out cross validation	No
Park et al. [24]	Histopathology or diagnosis of follow-up imaging	Hand-out cross validation	Yes
Midya et al. [9]	Histopathology	Hand-out cross validation	Yes
Liu et al. [25]	Histopathology	Hand-out cross validation	No
Li et al. [10]	Histopathology	Five-fold cross-validation	Yes
Huang et al. [26]	Histopathology	Hand-out cross validation	No
Fotouhi et al. [8]	Histopathology or diagnosis of follow-up imaging	Hand-out cross validation	Yes
Chen et al. [27]	Histopathology	Hand-out cross validation	No
Zhang et al. [28]	Histopathology	Hand-out cross validation	Yes
Xu et al. [29]	Histopathology	Hand-out cross validation	No
Sun et al. [30]	Histopathology	Hand-out cross validation	No
Ling et al. [31]	Histopathology	Five-fold cross-validation	No
Hu et al. [32]	Histopathology	Hand-out cross validation	No
Cao et al. [33]	Histopathology or clinical findings (based on clinical signs such as laboratory tests, imaging examinations)	Hand-out cross validation	No
Zhou et al. [34]	Histopathology	Hand-out cross validation	No
Zheng et al. [35]	Histopathology and cirrhosis was diagnosed based on typical imaging features, including surface nodularity, hypertrophy of left hemiliver and caudate lobe, splenomegaly, esophageal, or gastric varices and ascites.	Five-fold cross-validation	Yes
Wang et al. [36]	In the training set, 425 (65.7%) HCC cases and 1184 (17.2%) non-HCC cases underwent Histopathology. Histopathology in two test sets.	Hand-out cross validation	Yes
Tiyarattanachai et al. [37]	Histopathology and/or MRI/CT reports	Hand-out cross validation	Yes
Stollmayer et al. [38]	Histopathology or exhibited typical characteristics of the given lesion type with MRI.	Hand-out cross validation	No
Ryu et al. [6]	Histopathology or exhibited typical characteristics of image.	Hand-out cross validation	No
Ren et al. [39]	Histopathology	Hand-out cross validation	Yes
Oestmann et al. [40]	Histopathology	Monte Carlo cross-validation	No
Nakai et al. [41]	Histopathology	Hand-out cross validation	No
Menegotto et al. [42]	NR	Hand-out cross validation	No
Zhou et al. [43]	Histopathology for malignant tumors, 336 benign nodules through histopathology or a combination of typical image performance and clinical data	Hand-out cross validation	No
Shi et al. [44]	Histopathology or combination of imaging and clinical follow-up	Hand-out cross validation:	No
Kim et al. [45]	Histopathology	Hand-out cross validation	Yes
Huang et al. [46]	Histopathology	Hand-out cross validation	No
Cao et al. [47]	Histopathology or combination of clinical and radiological findings with follow-up	Hand-out cross validation	No
Özyurt et al. [48]	NR	Hand-out cross validation	No
Oyama et al. [49]	HCC group: Histopathology; MT group: Histopathology, radiologically features or clinical history of colorectal cancer; HH group: radiologically features	Leave-one-out validation	No
Hamm et al. [50]	Histopathology or exhibited typical characteristics of radiologically features	Hand-out cross validation	No
Gatos et al. [51]	Histopathology or exhibited typical characteristics of radiologically features	Leave-one-out validation	No

Table 3 Indicators, algorithms, and data sources

Author [ref]	Indicator definition		Algorithm	Data source		Number of features for training/ test/internal/external	Data range	Open access data
	Device	Exclusion of poor- quality cases		ML/DL	Transfer learning applied			
Zossou et al. [15]	CECT	NR	CNN	DL	No	The data from Centre National Hospitalier Universitaire Hubert Koutoukou Maga (CNHU-HKM) hospital.	images: training 13,658, validation 3416 and test 950	No
Xiao et al. [16]	CEUS	Yes	RF	ML	No	The data in the training and internal test set from Sun Yat-Sen University Cancer Center (SYSUCC)	lesions: training 140, validation 35 and test 113	No
Li et al. [17]	CEUS	NR	CNNs	DL	No	The data in the independent external test set from the First Affiliated Hospital of Sun Yat-sen University (FAH-SYSU), and Chongqing University Cancer Hospital (CUCH)	independent external test set: July 2014 to November 2018 (FAH-SYSU) and March 2020 to June 2022 (CUCH))	No
Chen et al. [18]	US	Yes	Resnet18	DL	No	The data from First Affiliated Hospital of Sun Yat-sen University	images: training 852 and test 295	No
Zhao et al. [19]	CEUS	NR	UEE model	DL	No	The data from General Hospital of the People's Liberation Army Hospital	patients: training 415 and test 50	No
Zhang et al. [20]	CECT	NR	MIDC-net	DL	No	The data from Shanghai Dongfang Hepatobiliary Hospital	cases: training 753, validation 126 and test 126	No
Xu et al. [21]	CECT	Yes	Transfer Swin-Transformer Architecture	DL	No	The data from Second Affiliated Hospital of Chongqing Medical University or the First Affiliated Hospital of Army Military Medical University	cases: training 222, test 95	No
Wan et al. [22]	CEUS	NR	T-AFML method	ML	No	The data from Nanjing Drum Tower Hospital	lesion: training 708, internal validation 178 and external validation 196	No
Roy et al. [23]	CT	NR	Hybrid CNN model	DL	yes	The data from https://www.ncbi.nlm.nih.gov/public/LIDC-IDRI and http://biogps.org/dataset .	training 139, test 35	No
Park et al. [24]	CE-MRI	Yes	Decision tree	ML	No	NR	training 852, test 228	YES
Midya et al. [9]	CECT	Yes	Modified Inception v3	DL	yes	The data from Memorial Sloan Kettering Cancer Center (MSK) (New York, NY, USA), Washington University (St. Louis, MO, USA), and Erasmus Medical Center (Rotterdam, The Netherlands).	feature: development 245, test 30 patient: training 491, validation 162, test 161	No

Table 3 (continued)

Author [ref]	Indicator definition		Algorithm	Data source			Number of features for training/ test/internal/external	Data range	Open access data
	Device	Exclusion of poor- quality cases		ML/DL	Transfer learning applied	Source of data			
Liu et al. [25]	MRI	Yes	SFFNet	DL	yes	The data from Second Affiliated Hospital of Kunming Medical University	training 68, validation 23, test 35	July 2015 to September 2021	No
Li et al. [10]	CEUS	Yes	Generalized linear models	ML	No	The data from multicenter liver cancer database (http://www.usliver.org/home.html)	features: training 366, external validation 88	July 2017 to January 2022	YES
Huang et al. [26]	CECT	Yes	CSAM-Net	DL	No	The data from First Affiliated Hospital of Nanchang University	images, training 346, validation 98, test 50	January 2015 to December 2020	No
Fotouhi et al. [8]	CE-MRI	NR	SVM	ML	No	NR	lesion: training+validation 314, external test 78	April 2019 to December 2022	No
Chen et al. [27]	CE-MRI	Yes	SVM	ML	No	NR	patient: training 93, test 41	March 2016 to March 2022	No
Zhang et al. [28]	US	Yes	Xception	DL	No	The data from Zhongshan Hospital of Fudan University.	lesions: model 310 and external 103	HCC during January 2016 to December 2020 and FNH during January 2013 to December 2020	No
Xu et al. [29]	CT	Yes	SVM	ML	No	NR	patient: training 122, validation 89	August 2018 to November 2019	No
Sun et al. [30]	CE-MRI	Yes	SVM	ML	No	The data from Shandong Provincial Hospital affiliated to Shandong First Medical University	feature: intra-group: training 92, test 30 inter-group: training 104, test 34	September 2018 to June 2021	No
Ling et al. [31]	CECT	NR	MExPaLe model	DL	No	NR	lesion: training 481, validation 120.	2014 to 2017	No
Hu et al. [32]	CE-MRI	Yes	TPOT	ML	No	The data from Second Xiangya Hospital, Hospital of the University of Pennsylvania, and Rhode Island Hospital	independent training 344, validation 97 and test 48	2008 to 2018	No
Cao et al. [33]	CT	NR	CNN	DL	No	NR	images: training set <i>n</i> 559 and test set 215	NR	No
Zhou et al. [34]	US	Yes	Resnet18	DL	No	NR	images: training set 144 and test set 37	April 2015 to September 2020	No
Zheng et al. [35]	CE-MRI	Yes	Pattern matching and deep learning (PM-DL) model	DL	No	NR	patient: Training-validation cohort 56; external test cohort 64	training and validation June 2016 to October 2019; external test November 2019 to February 2021	No

Table 3 (continued)

Author [ref]	Indicator definition		Algorithm		Data source		Number of features for training/ test/internal/external		Data range	Open access data
	Device	Exclusion of poor-quality cases	Algorithm architecture	ML/DL	Transfer learning applied	Source of data				
Wang et al. [36]	CECT	NR	HCCNet	DL	No	The date in the training set from Henan Provincial Peoples' Hospital The date in the internal test set from Henan Provincial Peoples' Hospital The date in the external test set from Henan Provincial Cancer Hospital	images: training : internal test : external test = 8,082,076 : 413,251 : 674,869	training: November 2016 to March 2019 internal test set: December 2016 to September 2019 external test set: February 2018 to April 2019	No	No
Tiyarat-tanachai et al. [37]	US	NR	CNN	DL	No	The data in the training and internal test from King Chulalongkorn Memorial Hospital, Bangkok, Thailand. The data in the external validation test from Mahachai Hospital, Samut Sakhon, Thailand and Queen Savang Vadhana Memorial Hospital, Chonburi, Thailand	images: training : internal test : external test = 40,397 : 6,191 : 18,922 lesion: training : internal test : external test = 20,432 : 845 : 1,195	training: November 2017 to October 2020	No	No
Stollmayer et al. [38]	CE-MRI	NR	2D DenseNet	DL	No	NR	training : validation: test = 156 : 30 : 30	November 2017 to October 2020	No	No
Ryu et al. [6]	US	NR	CNN	DL	No	NR	images: training 3909 and test 400	January 2004 to February 2018	No	No
Ren et al. [39]	US	Yes	SVM	ML	No	The data in the training and internal validation from Henan Provincial Peoples Hospital and the First Affiliated Hospital of Zhengzhou University, The data in the external validation from Henan Cancer Hospital	patient: training set 149, test set 38, and independent validation set 39	January 2019 to March 2021	No	No
Oestmann et al. [40]	CE-MRI	Yes	3D CNN	DL	No	NR	lesions: training set 140, testing set 10	2010 to 2018	No	No
Nakai et al. 2021 [41]	CECT	Yes	CNN	DL	No	NR	patient: training set 493, validation set 62, and test validation set 62	January 2004 to September 2019	No	No
Menegotto et al. [42]	CT	NR	VGG19	DL	No	The data of patients with HCC from TCGA-LIHC database The data of patients without HCC from TCGA-STAD, TCGA-KIRP and CPTAC-PDA databases.	images: training 29,104, validation 8,316 and test 4,164	NR	Yes	Yes
Zhou et al. [43]	CECT	NR	US-DLM+Clin model	DL	yes	NR	lesions: training set 359, testing set 90	January 2016 to December 2018	No	No
Shi et al. [44]	CECT	Yes	MP-CDN	DL	No	The date from Third Affiliated Hospital, Sun Yat-Sen University	lesions: training set 462, testing set 155	January 2009 to January 2016	No	No

Table 3 (continued)

Author [ref]	Indicator definition Device	Algorithm		Data source		Number of features for training/ test/internal/external	Data range	Open access data
		Exclusion of poor- quality cases	Algorithm architecture	ML/DL	Transfer learning applied			
Kim et al. [45]	CE-MRI	Yes	CNN	DL	No	NR	images: training 64,853, internal validation 13,896, internal test 13,896, external set 950	internal 2010 to 2014, external 2015 to 2017
Huang et al. [46]	CEUS	NR	CAD approach	ML	No	The data from First Affiliated Hospital of Sun Yat-sen University	lesion: Dataset 1: training data 30 Dataset 2: training data 387, testing data 30	NR
Cao et al. [47]	CECT	Yes	MP-CDN	DL	No	NR	lesions: training set 410, testing set 107	2012 to 2017
Özyurt et al. [48]	CECT	NR	F-PH- CNN	DL	No	The data from Elazig Education and Research Hospital	samples: training set 78, testing set 34	NR
Oyama et al. [49]	MRI	NR	Texture analysis and topological data analysis using persistent homology	ML	No	NR	full	HCC patient group: November 2009 to April 2017 MT patient group: December 2010 to September 2017 HH patient group: May 2015 to September 2017
Hamm et al. [50]	CE-MRI	Yes	Proof-of-concept CNN	DL	No	NR	lesions: training 434 and test 60	2010 to 2017
Gatos et al. [51]	MRI	NR	PNN	ML	No	NR	lesions: test 70	January 2013 to January 2015

Fig. 2 Pooled performance of AI algorithms and human clinicians. **a** Receiver operating characteristic (ROC) curves for studies rating AI algorithms (41 studies with 50 tables). **b** ROC curves for studies rating DL algorithms (28 studies with 35 tables). **c** ROC curves for studies evaluating ML algorithms (13 studies with 15 tables). **d** ROC curves for studies evaluating human clinicians (11 studies with 29 tables)

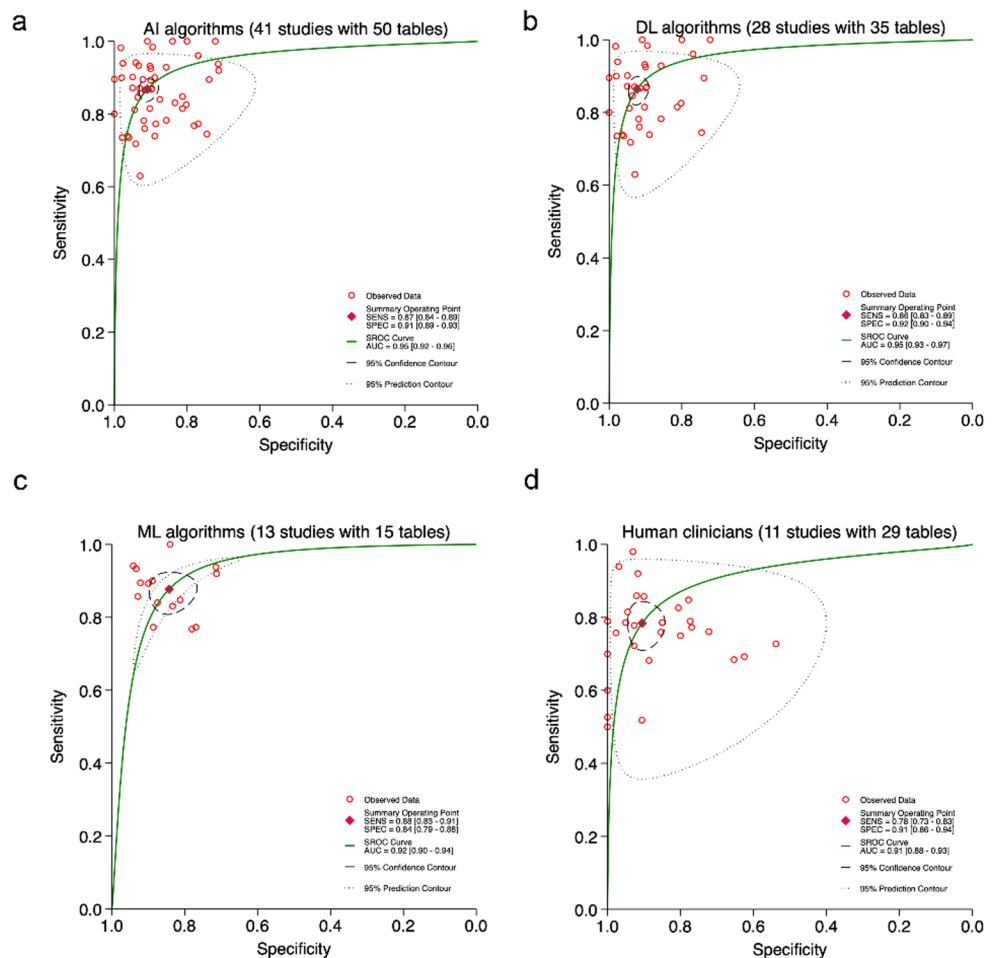
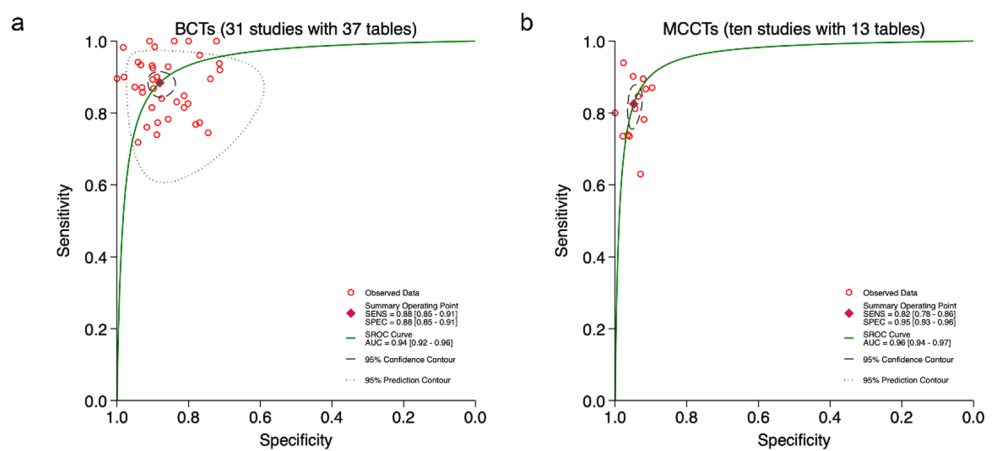


Fig. 3 Pooled performance of AI algorithms by classification type. **a** ROC curves for studies evaluating BCTs (31 studies with 37 tables). **b** ROC curves for studies rating MCCTs (ten studies with 13 tables)



DL versus ML

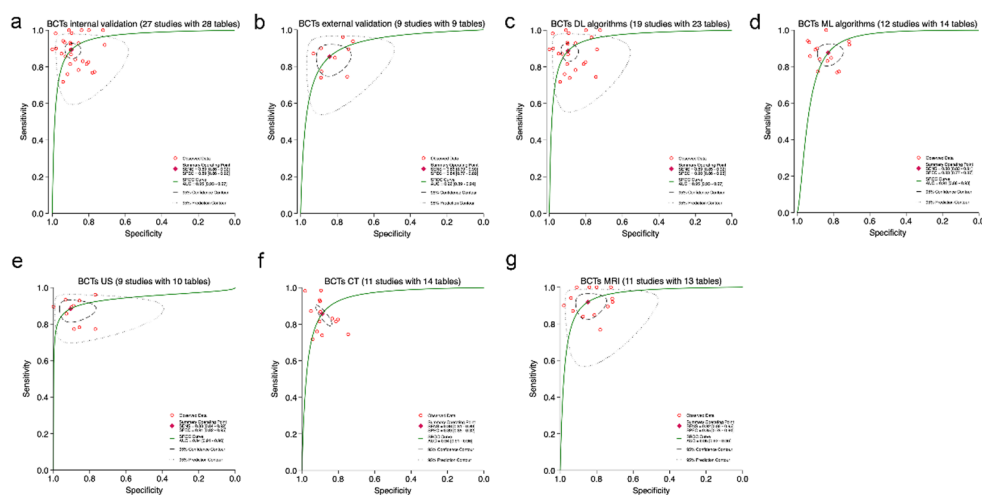
Nineteen studies applying DL algorithms (23 contingency tables) demonstrated pooled SEN, SPC, and AUC of 89% (95% CI 85–92%), 90% (95% CI 86–93%), and 0.95 (95% CI 0.93–0.97), respectively. In comparison, 12 studies using ML algorithms (14 contingency tables) yielded pooled SEN, SPC, and AUC of 88% (95% CI 83–91%), 83% (95%

CI 77–87%), and 0.91 (95% CI 0.88–0.93), respectively (Fig. 4c, d).

Imaging modalities

Nine US-based studies (10 contingency tables) demonstrated pooled SEN, SPC, and AUC of 88% (95% CI 84–92%), 91% (95% CI 82–95%), and 0.94 (95% CI

Fig. 4 Pooled performance of AI algorithms in BCTs. **a** ROC curves for studies utilizing internal validation (27 studies with 28 tables). **b** ROC curves for studies adopting external validation (9 studies with 9 tables). **c** ROC curves for studies using DL algorithms (19 studies with 23 tables). **d** ROC curves for studies using ML algorithms (12 studies with 14 tables). **e** ROC curves for studies using US (9 studies with 10 tables). **f** ROC curves for studies using CT (11 studies with 14 tables). **g** ROC curves for studies using MRI (11 studies with 13 tables)



0.91–0.96), respectively (Fig. 4e). Eleven CT-based studies (14 contingency tables) yielded pooled SEN, SPC, and AUC of 86% (95% CI 81–89%), 89% (95% CI 85–92%), and 0.94 (95% CI 0.91–0.96), respectively (Fig. 4f). Eleven MRI-based studies (13 contingency tables) achieved pooled SEN, SPC, and AUC of 92% (95% CI 86–95%), 85% (95% CI 78–90%), and 0.95 (95% CI 0.92–0.96), respectively (Fig. 4g).

Three separate meta-analyses were conducted in the MCCTs groups

Validation types

Nine studies (nine contingency tables) using internal validation datasets demonstrated pooled SEN, SPC, and AUC of 82% (95% CI 75–87%), 95% (95% CI 92–96%), and 0.95 (95% CI 0.93–0.97), respectively (Fig. 5a). In contrast, two studies comprising four contingency tables and employing external validation datasets yielded pooled SEN, SPC, and AUC of 83% (95% CI 78–87%), 95% (95% CI 94–95%), and 0.96 (95% CI 0.94–0.97), respectively (Fig. 5b).

DL versus ML

Nine DL-based studies (12 contingency tables) were eligible for meta-analysis and demonstrated pooled SEN, SPC, and AUC of 82% (95% CI 77–86%), 95% (95% CI 93–96%), and 0.96 (95% CI 0.94–0.97), respectively (Fig. 5c). One ML-based studies (one contingency table) lacked sufficient information to determine pooled SEN and SPC and were therefore removed.

Imaging modalities

Five CT-based studies (five contingency tables) demonstrated strong diagnostic performance, with pooled SEN, SPC, and AUC of 83% (95% CI 71–91%), 94% (95% CI 92–96%), and 0.96 (95% CI 0.94–0.97), respectively (Fig. 5d). The three US-based studies (six contingency tables) and two MRI-based studies (two contingency tables) did not provide sufficient data for pooled analyses and were therefore excluded.

Heterogeneity analysis

All included studies demonstrated that AI algorithms represent a valuable tool for the detection of hepatic cancer using MI in comparison with human clinicians. Heterogeneity was assessed using the I^2 statistic. Considerable heterogeneity was observed for both SEN ($I^2=68.19\%$) and SPC ($I^2=82.06\%$) (Fig. 6), reflecting methodological and population differences rather than selective publication. The Deeks' test ($P=0.83$) indicated no significant publication bias (Fig. S1). Given the substantial heterogeneity, a random-effects model was applied. To identify the source/sources of such extreme heterogeneity, we conducted a subgroup analysis, and see the supplementary (Table S1 and Fig. S2–8). The heterogeneity was not attributable to any specific subgroup, with the majority of subgroup I^2 values ranging from 50% to 86%.

Quality assessment

Study quality was rated via QUADAS-AI [14]. Findings are provided in Fig. S9. An item-wise assessment based on risk-of-bias domains and concerns over applicability is detailed in Fig. S10. In patient selection, 14 studies displayed a high or unclear RoB owing to unreported eligibility (inclusion or

Fig. 5 Pooled performance of AI algorithms in MCCTs. **a** ROC curves for studies utilizing internal validation (nine studies with 9 tables). **b** ROC curves for studies utilizing external validation (two studies with 4 tables). **c** ROC curves for studies utilizing DL algorithms (9 studies with 12 tables). **d** ROC curves for studies utilizing CT (5 studies with 5 tables)

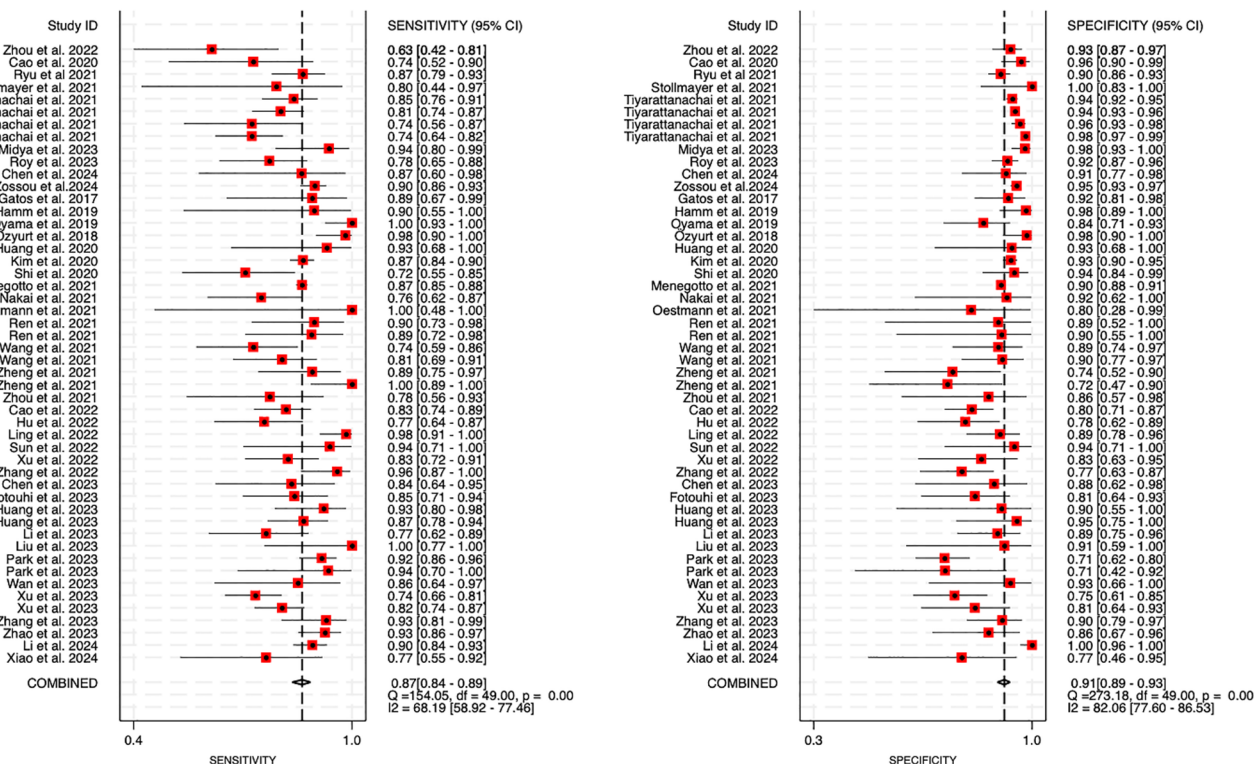
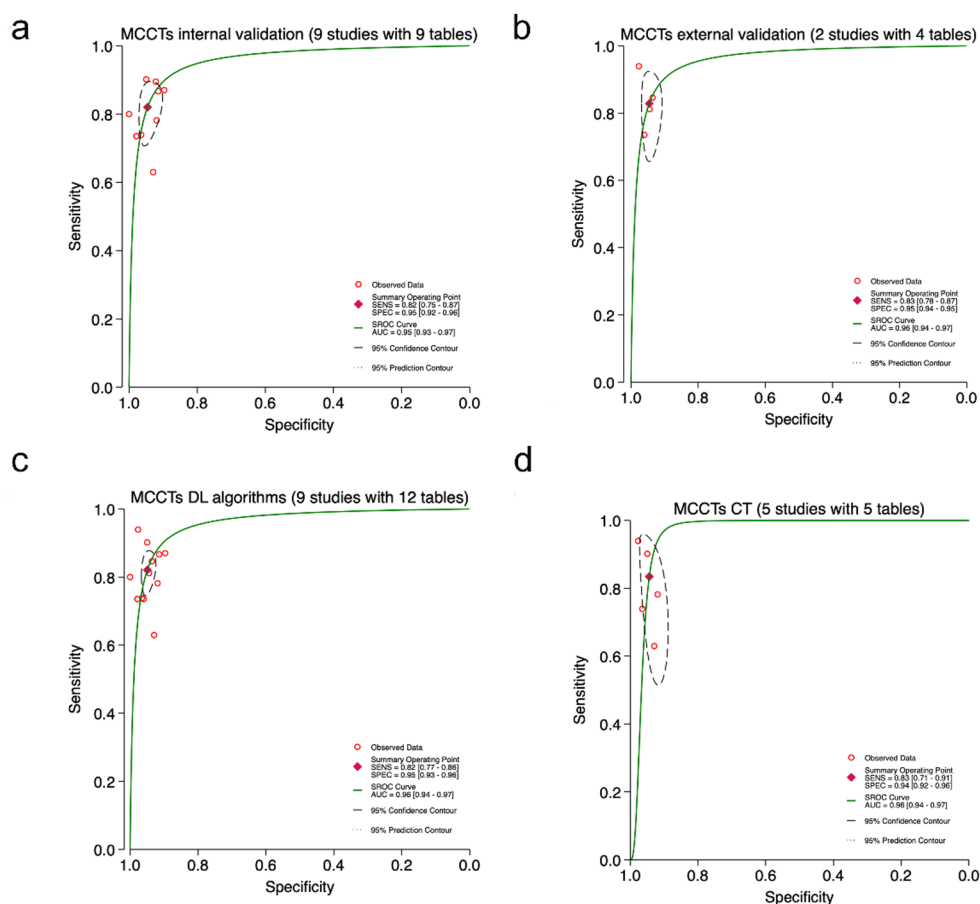


Fig. 6 Summary estimates of pooled performance presented as a forest plot. Data are derived from 41 studies encompassed in the meta-analysis

exclusion) criteria or open-access datasets. Regarding the index test, no study exhibited a high or unclear RoB. Concerning the reference standard, five studies displayed a high or unclear RoB due to unreported reference standards. As for flow and timing, two studies displayed a high or unclear RoB because they lacked information regarding appropriate intervals or the application of interventions between the index test and reference standard. Regarding applicability concerns, two studies exhibited high or unclear applicability in patient selection, and five studies presented unclear concerns in reference standard, with no concerns identified in index test. Overall, the included studies exhibited mild methodological quality, with the main concerns arising from patient selection.

Discussion

MI is crucial in diagnosing HCC, but manual interpretation is time-consuming and over-reliant on expert radiologists. Our study incorporated 41 studies assessing the diagnostic performance of AI algorithms in detecting HCC. AI algorithms exhibited promising diagnostic performance for HCC detection, with pooled AUC estimates reflecting high discriminative ability. Notably, DL algorithms demonstrated marginally higher diagnostic performance than traditional ML algorithms (AUC: 0.95, 95% CI 0.93–0.97 vs. 0.92, 95% CI 0.90–0.94). Compared with radiologists or clinicians (AUC: 0.91, 95% CI 0.88–0.93), AI algorithms (AUC: 0.95, 95% CI 0.92–0.96) showed superior discriminative performance for HCC detection. These findings suggest that AI algorithms, particularly DL approaches, have considerable potential to enhance diagnostic accuracy, streamline workflow efficiency, and mitigate reliance on highly experienced clinicians. However, the clinical application of AI algorithms should be approached with caution, given the significant heterogeneity observed in pooled studies. This heterogeneity could not be attributed to any specific subgroup, as all subgroup analyses revealed high I^2 values.

A systematic search identified two systematic reviews with meta-analyses evaluating AI algorithms in MI for HCC detection. Consistent with our results, Salehi et al. reported that AI algorithms exhibited high diagnostic performance for HCC detection [11]. For internally validated AI algorithms, the pooled SEN, SPC, and AUC were 84% (95% CI 81–87%), 92% (95% CI 90–94%), and 0.94 (95% CI 0.92–0.96). In contrast, externally validated AI algorithms demonstrated pooled SEN, SPC, and AUC of 85% (95% CI 78–89%), 84% (95% CI 72–91%), and 0.91 (95% CI 0.88–0.93). Furthermore, the diagnostic performance of AI algorithms exceeded that of clinicians, who exhibited pooled SEN, SPC, and AUC of 70% (95% CI 60–78%), 85% (95%

CI 77–90%), and 0.84 (95% CI 0.81–0.87). Nonetheless, significant heterogeneity was present in these meta-analyses, attributable to the absence of necessary data and heterogeneity in hepatic lesion classification and comparator methods. Further analyses were limited to meta-regression based on imaging modality and use of data augmentation due to the insufficient number of studies. Moreover, AI algorithms were primarily categorized as DL or ML during development. Nevertheless, these studies evaluated overall diagnostic performance without separate analyses for DL and ML. In addition, prior studies predominantly focused on BCTs, limiting exploration of MCCT and constraining the generalizability of AI algorithms. Limiting AI evaluation to BCTs undermines clinical utility, as MCCT aligns more closely with differential diagnosis in clinical practice, enabling the identification of multiple similar imaging features, supporting precise diagnoses, and facilitating personalized precision medicine. Campello et al. analyzed 20 studies encompassing 34,245 focal liver lesions (FLLs) to rate the performance of ML and DL algorithms in differentiating malignant from benign FLLs using US and CEUS [12]. ML algorithms demonstrated high diagnostic performance for identifying malignant FLLs on US and CEUS, with similar pooled sensitivities (81.7% and 87.1%) and pooled specificities (84.8% and 87.0%). For CEUS, DL algorithms achieved pooled SEN and SPC of 92.4% (95% CI, 88.5–95.0%) and 88.2% (95% CI, 81.1–92.9%). These findings underscore the potential of AI algorithms as effective tools in detecting malignant liver lesions. However, substantial heterogeneity was observed. Two studies leveraged external validation datasets, and several were retrospective with limited sample sizes, potentially leading to overestimation of AI performance. Moreover, due to the limited number of studies, direct comparisons between AI algorithms and clinicians were not performed. The analyses were further restricted to BCT of malignant versus benign FLLs, without consideration of specific liver cancer subtypes. Given the distinct treatment strategies and prognostic implications of different liver lesions, BCT alone does not fully satisfy clinical requirements. Therefore, these limitations constrain the generalizability of their conclusions and highlight the need for further investigation into their diagnostic performance in MI. In contrast, our meta-analysis specifically focused on HCC detection, a distinct subtype of malignant liver nodules. Extrapolating diagnostic performance from general malignant liver lesions to HCC may overestimate accuracy, as HCC identification relies on characteristic imaging features, like arterial phase hyperenhancement and early mild wash-out on CEUS [52]. Additionally, our study separately evaluated the performance of DL and ML algorithms and compared AI performance across BCT and

MCCT. Therefore, our study provides a more comprehensive assessment of AI algorithms for HCC detection.

MCCTs more closely reflect real-world clinical practice, where the diagnostic task extends beyond simply distinguishing HCC from non-HCC lesions. In routine clinical settings, physicians are often required to identify specific hepatic tumor subtypes to guide individualized therapeutic decision-making and prognostic evaluation. To explore this issue, we performed a subgroup analysis by categorizing the 41 included studies into BCTs and MCCTs. Ten studies focused on MCCTs for the classification of various liver lesions, including HCC, intrahepatic cholangiocarcinoma (ICC), colorectal liver metastasis (CRLM), and other tumor types. The pooled analysis demonstrated that MCCTs achieved a slightly higher AUC than BCTs (0.96, 95% CI 0.94–0.97 vs. 0.94, 95% CI 0.92–0.96). However, compared with BCTs, MCCTs exhibited lower SEN (82% vs. 88%) but substantially higher SPC (95% vs. 88%). These findings likely reflect the inherent complexity of multi-class classification. Unlike binary models, multi-class algorithms must discriminate among multiple lesion categories that often share overlapping imaging characteristics while simultaneously capturing subtle class-specific features. The notably high SPC observed in MCCTs suggests a strong ability to correctly exclude non-target lesions, which may help reduce false-positive diagnoses and unnecessary clinical interventions. However, the relatively lower SEN (82%) indicates a non-negligible risk of missed diagnoses in MCCTs. First, the same lesion type may exhibit substantial imaging heterogeneity across patients and disease stages, making it challenging for models to learn robust and representative features. Second, the development of MCCT models relies heavily on accurate lesion annotation by experienced radiologists, and variability in annotation quality may adversely affect model SEN. Third, class imbalance is often more pronounced in MCCT datasets, potentially biasing models toward more prevalent lesion categories, such as HCC, while compromising performance for less common entities, including ICC and CRLM. Consequently, future studies should prioritize the establishment of large-scale prospective multicenter cohorts with standardized annotation protocols, balanced class distributions, and rigorous external validation to improve SEN while maintaining the excellent SPC demonstrated by current MCCT models.

Our meta-analysis identified 11 studies that incorporated external validation datasets, including nine BCT studies and two MCCT studies. In the BCT subgroup, models evaluated using internal validation demonstrated superior diagnostic performance compared with those assessed using external validation (AUC: 0.95, 95% CI 0.93–0.97 vs. 0.92, 95% CI 0.89–0.94). This performance decline in external validation may indicate a degree of overfitting to the development

dataset, thereby limiting model generalizability across independent populations and imaging environments [53]. Furthermore, minor information leakage introduced during preprocessing or feature extraction cannot be entirely excluded in complex imaging-analysis pipelines [54]. MCCT studies showed a marginally higher pooled AUC for external validation than for internal validation (0.96, 95% CI 0.94–0.97 vs. 0.95, 95% CI 0.93–0.97). However, this finding should be interpreted cautiously because only two MCCT studies included external validation datasets. The resulting pooled estimate is therefore less stable and more vulnerable to the influence of individual study characteristics, dataset composition, and potential publication bias. Nonetheless, external validation remains the gold standard for assessing model credibility, and excellent internal validation results require confirmation through rigorous external validation [55, 56]. Future MCCT studies incorporating external validation are warranted to establish more reliable estimates of diagnostic performance and generalizability.

In our subgroup analysis, DL demonstrated higher diagnostic performance than ML in BCTs, with AUCs of 0.95 (95% CI 0.93–0.97; 19 studies) and 0.91 (95% CI 0.88–0.93; 12 studies). These findings suggest the potential advantages of DL for HCC detection. As a subfield of ML, DL employs multilayer neural networks to extract complex patterns directly from raw data, enabling automated feature representation and facilitating image-classification tasks. This data-driven paradigm enhances its clinical practicality [57]. Nonetheless, caution remains warranted when considering its clinical application, as DL models are more susceptible to overfitting and generally exhibit poorer interpretability in comparison to ML models [58]. ML, in contrast, does not require large training datasets and yields more interpretable outputs. Nevertheless, it relies on manual feature extraction, and its capacity for data-driven learning is comparatively limited [59]. Therefore, future work should not only evaluate DL and ML models independently for HCC diagnosis but also explore hybrid frameworks that employ DL for automated feature extraction and ML for classification, thereby determining whether such hybrid models demonstrate superior diagnostic performance. In the present study, the diagnostic performance between DL and ML for MCCTs cannot be compared because limited ML studies were available. Therefore, additional research is needed to assess and compare their performance across BCT and MCCT.

Multiple imaging modalities are widely utilized for HCC detection, including US, CT, and MRI. Our meta-analysis demonstrated that, in BCTs, AI algorithms achieved slightly superior diagnostic performance with MRI compared with US and CT, although the differences across imaging modalities were modest. Due to insufficient data, only CT-based studies were eligible for pooled analysis in the MCCT

subgroup, and CT demonstrated high diagnostic performance for HCC detection. Overall, AI algorithms appear to be applicable for HCC detection across various imaging modalities. However, these results should be interpreted with prudence for two reasons. First, because only three US studies and two MRI studies were available, US-based and MRI-based analyses were not performed for MCCT. Second, in the subgroup analysis, the diagnostic performance related to each imaging modality reflected a combination of non-contrast and contrast-enhanced techniques. For instance, the US subgroup encompassed non-contrast US and CEUS. Liver Imaging Reporting and Data System (LI-RADS) is essential in characterizing liver lesions and supporting diagnostic interpretation across US, CT, and MRI [60–62]. In LI-RADS-based AI research, the classification categories and the imaging features used to determine these categories (e.g., perfusion characteristics or blood-flow patterns) may serve as reliable labels to enhance the diagnostic performance of AI models. As a result, most studies developed diagnostic models using contrast-enhanced imaging modalities. The development of AI algorithms for non-contrast imaging remains highly valuable, given its lower cost and avoidance of contrast-agent-related risks. Our meta-analysis found that limited studies assessed the diagnostic performance of AI algorithms using non-contrast imaging. Therefore, there is an urgent need for prospective investigations employing standardized imaging protocols to evaluate AI algorithms for HCC diagnosis using non-contrast US, CT, or MRI. Such studies will be essential for generating reliable, high-quality evidence to support the assessment of their diagnostic performance.

Overall, only one study adopted a prospective design, whereas nine studies did not specify whether their datasets were collected prospectively or retrospectively. The data sources of retrospective investigations were derived from pre-existing electronic medical records or open online databases, which were not originally designed for algorithmic evaluation in real-world clinical settings, and the quality of imaging data could not be standardized. Although prior studies indicate that AI algorithms are related to high diagnostic accuracy, retrospective designs are subject to substantial RoB and do not characterize the diagnostic performance of AI algorithms within current clinical screening workflows. Accordingly, prospective, multi-center studies with pre-specified protocols and reader-blinded comparisons remain essential for the evaluation of AI algorithms. In addition, 22 studies excluded low-quality images, whereas 19 studies did not report image quality. High diagnostic performance of AI algorithms in MI is generally related to reliable, high-quality images and appropriate annotation [63]. Low-quality images, like those affected by acquisition deficiencies, motion artifacts, or low resolution, may

lead to overestimation or underestimation of diagnostic performance. Thus, AI algorithms require high-quality imaging data to support the development of more robust and accurate models and to mitigate the high heterogeneity observed across studies. Patients with cirrhosis have a substantially higher pretest probability of HCC compared to those without underlying liver disease [2]. Studies that utilized datasets with cirrhotic patients may have resulted in an elevated pretest probability relative to general populations. This scenario may overestimate the clinical utility of AI-based diagnostic algorithms when applied to lower-risk settings. Future studies should explicitly report the proportion of patients with cirrhosis and other liver disease etiologies to enable more nuanced subgroup analyses and facilitate the translation of AI diagnostic tools into diverse clinical settings.

Nevertheless, there are limitations. First, most encompassing studies were retrospective, and low-quality images were not removed, making it difficult to control for many confounding factors. Additionally, nine of the included studies failed to report their basic retrospective or prospective design. This omission constitutes a drag on the overall evidence quality and limits the assessment of potential design-related bias. Second, only 11 studies performed external validation, leaving uncertainty regarding the degree to which reported diagnostic performance may be generalizable to clinical practice. Third, owing to the limited eligible studies, the subgroups evaluating ML and US, MRI-based approaches in MCCTs could not be analyzed. Moreover, non-contrast and contrast-enhanced imaging modalities were merged into a single subgroup, which may have obscured potential differences in diagnostic performance and reduced the generalizability of AI models. Fourth, substantial heterogeneity was identified and not attenuated in subgroup analyses, suggesting that no single subgroup accounted for this variability. Potential contributing factors may include inconsistencies in image labeling and reference standards, variations in patient population spectra, differences in diagnostic thresholds, and divergence in image preprocessing workflows. Fifth, AI and clinician assessments were not consistently performed on the same patient populations. Therefore, our comparison between AI and clinicians is descriptive rather than inferential. Given these limitations, future research should prioritize large-scale, prospective studies with rigorous external validation to more reliably characterize the diagnostic performance of AI-based models. Concurrently, additional primary research is needed to facilitate more granular analyses of specific computational methodologies and imaging modalities. Standardization of key methodological factors, including image annotation, patient spectrum considerations, and preprocessing pipelines, will also be essential to reduce

unexplained heterogeneity and to advance the development of reliable, generalizable AI models. Additionally, our review focused on diagnostic performance metrics and did not systematically evaluate factors critical to clinical implementation, including model explainability, clinician trust, and reader-AI interaction. These dimensions were infrequently reported in the included studies and warrant dedicated investigation in future research.

Conclusion

AI algorithms were associated with high diagnostic performance in MI for detecting HCC, with a pooled AUC of 0.95 (95% CI 0.92–0.96), and demonstrated performance comparable to that of clinicians across both BCT and MCCT settings. While these findings suggest AI holds promise as a supportive tool, the results should be interpreted as exploratory. Prospective studies with external validation, large sample sizes, and dedicated evaluation of human-AI collaboration are needed to translate this potential into reliable clinical utility. In addition, Future research must extend beyond performance metrics to address critical gaps in explainability, clinician trust, and reader-AI interaction, which are essential for safe and meaningful clinical integration.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00261-026-05627-6>.

Author contributions Jingchao Wang: Conceptualization, Methodology, Writing - Original Draft. Jingjie Bai: Resources, Data Curation. Yanyang Wang: Investigation, Resources. Cong Wang: Software, Visualization. Heping Deng: Supervision, Writing- Reviewing and Editing, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data availability The search strategy and aggregated data are displayed in the supplementary documents.

Declarations

Competing interests The authors declare no competing interests.

References

- Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*. 2021;71(3):209–49. <https://doi.org/10.3322/caac.21660>.
- Fattovich G, Stroffolini T, Zagni I, Donato F. Hepatocellular carcinoma in cirrhosis: incidence and risk factors. *Gastroenterology*. 2004;127(5 Suppl 1):S35–50. <https://doi.org/10.1053/j.gastro.2004.09.014>.
- Choi JY, Lee JM, Sirlin CB. CT and MR imaging diagnosis and staging of hepatocellular carcinoma: part I. Development, growth, and spread: key pathologic and imaging aspects. *Radiology*. 2014;272(3):635–54. <https://doi.org/10.1148/radiol.14132361>.
- Choi JY, Lee JM, Sirlin CB. CT and MR imaging diagnosis and staging of hepatocellular carcinoma: part II. Extracellular agents, hepatobiliary agents, and ancillary imaging features. *Radiology*. 2014;273(1):30–50. <https://doi.org/10.1148/radiol.14132362>.
- Marrero JA, Kulik LM, Sirlin CB, Zhu AX, Finn RS, Abecassis MM, et al. Diagnosis, Staging, and Management of Hepatocellular Carcinoma: 2018 Practice Guidance by the American Association for the Study of Liver Diseases. *Hepatology*. 2018;68(2):723–50. <https://doi.org/10.1002/hep.29913>.
- Ryu H, Shin SY, Lee JY, Lee KM, Kang HJ, Yi J. Joint segmentation and classification of hepatic lesions in ultrasound images using deep learning. *Eur Radiol*. 2021;31(11):8733–42. <https://doi.org/10.1007/s00330-021-07850-9>.
- Li M, Fan Y, You H, Li C, Luo M, Zhou J, et al. Dual-Energy CT Deep Learning Radiomics to Predict Macrotrabecular-Massive Hepatocellular Carcinoma. *Radiology*. 2023;308(2):e230255. <https://doi.org/10.1148/radiol.230255>.
- Fotouhi M, Samadi Khoshe Mehr F, Delazar S, Shahidi R, Setayeshpour B, Toosi MN et al (2023) Assessment of LI-RADS efficacy in classification of hepatocellular carcinoma and benign liver nodules using DCE-MRI features and machine learning. *European Journal of Radiology Open*. <https://doi.org/10.1016/j.ejro.2023.100535>
- Midya A, Chakraborty J, Srouji R, Narayan RR, Boerner T, Zheng J, et al. Computerized Diagnosis of Liver Tumors From CT Scans Using a Deep Neural Network Approach. *IEEE J Biomed Health Inform*. 2023;27(5):2456–64. <https://doi.org/10.1109/jbhi.2023.248489>.
- Li J, Li H, Xiao F, Liu R, Chen Y, Xue M, et al. Comparison of machine learning models and CEUS LI-RADS in differentiation of hepatic carcinoma and liver metastases in patients at risk of both hepatitis and extrahepatic malignancy. *Cancer Imaging*. 2023;23(1):63. <https://doi.org/10.1186/s40644-023-00573-8>.
- Salehi MA, Harandi H, Mohammadi S, Shahrabi Farahani M, Shojaei S, Saleh RR. Diagnostic Performance of Artificial Intelligence in Detection of Hepatocellular Carcinoma: A Meta-analysis. *J Imaging Inform Med*. 2024;37(4):1297–311. <https://doi.org/10.1007/s10278-024-01058-1>.
- Campello CA, Castanha EB, Vilardo M, Staziaki PV, Francisco MZ, Mohajer B, et al. Machine learning for malignant versus benign focal liver lesions on US and CEUS: a meta-analysis. *Abdom Radiol (NY)*. 2023;48(10):3114–26. <https://doi.org/10.1007/s00261-023-03984-0>.
- Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA Statement. *Open Med*. 2009;3(3):e123–30.
- Sounderajah V, Ashrafian H, Rose S, Shah NH, Ghassemi M, Golub R, et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. *Nat Med*. 2021;27(10):1663–5. <https://doi.org/10.1038/s41591-021-01517-0>.
- Zossou VS, Rodrigue Gnangnon FH, Biaou O, de Vathaire F, Allodji RS, Ezin EC. Automatic Diagnosis of Hepatocellular Carcinoma and Metastases Based on Computed Tomography Images. *J Imaging Inform Med*. 2024. <https://doi.org/10.1007/s10278-024-01192-w>.
- Xiao M, Deng Y, Zheng W, Huang L, Wang W, Yang H, et al. Machine learning model based on dynamic contrast-enhanced ultrasound assisting LI-RADS diagnosis of HCC: A multicenter diagnostic study. *Heliyon*. 2024;10(20):e38850. <https://doi.org/10.1016/j.heliyon.2024.e38850>.

17. Li MD, Li W, Lin MX, Lin XX, Hu HT, Wang YC et al (2024) Systematic comparison of deep-learning based fusion strategies for multi-modal ultrasound in diagnosis of liver cancer. *Neurocomputing*. <https://doi.org/10.1016/j.neucom.2024.128257>
18. Chen J, Zhang W, Bao J, Wang K, Zhao Q, Zhu Y, et al. Implications of ultrasound-based deep learning model for preoperatively differentiating combined hepatocellular-cholangiocarcinoma from hepatocellular carcinoma and intrahepatic cholangiocarcinoma. *Abdominal radiology* (New York). 2024;49(1):93–102. <https://doi.org/10.1007/s00261-023-04089-4>.
19. Zhao L, Liu S, An Y, Cai W, Li B, Wang SH et al (2023) A unified end-to-end classification model for focal liver lesions. *Biomedical Signal Processing and Control*. <https://doi.org/10.1016/j.bspc.2023.105260>
20. Zhang X, Jia N, Wang Y (2023) Multi-input dense convolutional network for classification of hepatocellular carcinoma and intrahepatic cholangiocarcinoma. *Biomedical Signal Processing and Control*. <https://doi.org/10.1016/j.bspc.2022.104226>
21. Xu Y, Zhou C, He X, Song R, Liu Y, Zhang H, et al. Deep learning-assisted LI-RADS grading and distinguishing hepatocellular carcinoma (HCC) from non-HCC based on multiphase CT: a two-center study. *European Radiology*. 2023;33(12):8879–88. <https://doi.org/10.1007/s00330-023-09857-w>.
22. Wan P, Xue H, Liu C, Chen F, Shao W, Qin J, et al. Transport-Based Anatomical-Functional Metric Learning for Liver Tumor Recognition Using Dual-View Dynamic CEUS Imaging. *IEEE Transactions on Biomedical Engineering*. 2023;70(3):1012–23. <https://doi.org/10.1109/TBME.2022.3207473>.
23. Saha Roy S, Roy S, Mukherjee P, Halder Roy A. An automated liver tumour segmentation and classification model by deep learning based approaches. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization*. 2023;11(3):638–50. <https://doi.org/10.1080/21681163.2022.2099300>.
24. Park S, Byun J, Hwang SM. Utilization of a Machine Learning Algorithm for the Application of Ancillary Features to LI-RADS Categories LR3 and LR4 on Gadoxetate Disodium-Enhanced MRI. *Cancers*. 2023;15(5). <https://doi.org/10.3390/cancers15051361>.
25. Liu Y, Wang B, Mo X, Tang K, He J, Hao J. A Deep Learning Workflow for Mass-Forming Intrahepatic Cholangiocarcinoma and Hepatocellular Carcinoma Classification Based on MRI. *Current Oncology*. 2023;30(1):529–44. <https://doi.org/10.3390/currocol30010042>.
26. Huang JL, Sun Y, Wu ZH, Zhu HJ, Xia GJ, Zhu XS, et al. Differential diagnosis of hepatocellular carcinoma and intrahepatic cholangiocarcinoma based on spatial and channel attention mechanisms. *J Cancer Res Clin Oncol*. 2023;149(12):10161–8. <https://doi.org/10.1007/s00432-023-04935-4>.
27. Chen X, Chen Y, Chen H, Zhu J, Huang R, Xie J, et al. Machine learning based on gadoxetic acid-enhanced MRI for differentiating atypical intrahepatic mass-forming cholangiocarcinoma from poorly differentiated hepatocellular carcinoma. *Abdom Radiol (NY)*. 2023;48(8):2525–36. <https://doi.org/10.1007/s00261-023-03870-9>.
28. Zhang WB, Hou SZ, Chen YL, Mao F, Dong Y, Chen JG, et al. Deep Learning for Approaching Hepatocellular Carcinoma Ultrasound Screening Dilemma: Identification of α -Fetoprotein-Negative Hepatocellular Carcinoma From Focal Liver Lesion Found in High-Risk Patients. *Frontiers in Oncology*. 2022;12. <https://doi.org/10.3389/fonc.2022.862297>.
29. Xu X, Mao Y, Tang Y, Liu Y, Xue C, Yue Q, et al. Classification of Hepatocellular Carcinoma and Intrahepatic Cholangiocarcinoma Based on Radiomic Analysis. *Comput Math Methods Med*. 2022;2022:5334095. <https://doi.org/10.1155/2022/5334095>.
30. Sun K, Shi L, Qiu J, Pan Y, Wang X, Wang H. Multi-phase contrast-enhanced magnetic resonance image-based radiomics-combined machine learning reveals microscopic ultra-early hepatocellular carcinoma lesions. *Eur J Nucl Med Mol Imaging*. 2022;49(8):2917–28. <https://doi.org/10.1007/s00259-022-05742-8>.
31. Ling Y, Ying S, Xu L, Peng Z, Mao X, Chen Z, et al. Automatic volumetric diagnosis of hepatocellular carcinoma based on four-phase CT scans with minimum extra information. *Frontiers in Oncology*. 2022;12. <https://doi.org/10.3389/fonc.2022.960178>.
32. Hu R, Li H, Horng H, Thomasian NM, Jiao Z, Zhu C, et al. Automated machine learning for differentiation of hepatocellular carcinoma from intrahepatic cholangiocarcinoma on multiphasic MRI. *Sci Rep*. 2022;12(1):7924. <https://doi.org/10.1038/s41598-022-11997-w>.
33. Cao Y, Yu J, Zhang H, Xiong J, Luo Z. Classification of hepatic cavernous hemangioma or hepatocellular carcinoma using a convolutional neural network model. *Journal of Gastrointestinal Oncology*. 2022;13(2):787–91. <https://doi.org/10.21037/jgo-22-197>.
34. Zhou H, Jiang T, Li Q, Zhang C, Zhang C, Liu Y, et al. US-Based Deep Learning Model for Differentiating Hepatocellular Carcinoma (HCC) From Other Malignancy in Cirrhotic Patients. *Front Oncol*. 2021;11:672055. <https://doi.org/10.3389/fonc.2021.672055>.
35. Zheng R, Wang L, Wang C, Yu X, Chen W, Li Y, et al. Feasibility of automatic detection of small hepatocellular carcinoma (≤ 2 cm) in cirrhotic liver based on pattern matching and deep learning. *Phys Med Biol*. 2021;66(8). <https://doi.org/10.1088/1361-6560/abf2f8>.
36. Wang M, Fu F, Zheng B, Bai Y, Wu Q, Wu J, et al. Development of an AI system for accurately diagnose hepatocellular carcinoma from computed tomography imaging data. *Br J Cancer*. 2021;125(8):1111–21. <https://doi.org/10.1038/s41416-021-01511-w>.
37. Tiyyarattanachai T, Apiparakoon T, Marukatat S, Sukcharoen S, Geratikornsupuk N, Anukulkarnkusol N, et al. Development and validation of artificial intelligence to detect and diagnose liver lesions from ultrasound images. *PLoS ONE*. 2021;16(6 June). <https://doi.org/10.1371/journal.pone.0252882>.
38. Stollmayer R, Budai BK, Tóth A, Kalina I, Hartmann E, Szoldán P, et al. Diagnosis of focal liver lesions with deep learning-based multichannel analysis of hepatocyte-specific contrast-enhanced magnetic resonance imaging. *World Journal of Gastroenterology*. 2021;27(35):5978–88. <https://doi.org/10.3748/wjg.v27.i35.5978>.
39. Ren S, Li Q, Liu S, Qi Q, Duan S, Mao B, et al. Clinical Value of Machine Learning-Based Ultrasomics in Preoperative Differentiation Between Hepatocellular Carcinoma and Intrahepatic Cholangiocarcinoma: A Multicenter Study. *Front Oncol*. 2021;11:749137. <https://doi.org/10.3389/fonc.2021.749137>.
40. Oestmann PM, Wang CJ, Savic LJ, Hamm CA, Stark S, Schobert I, et al. Deep learning-assisted differentiation of pathologically proven atypical and typical hepatocellular carcinoma (HCC) versus non-HCC on contrast-enhanced MRI of the liver. *European Radiology*. 2021;31(7):4981–90. <https://doi.org/10.1007/s00330-020-07559-1>.
41. Nakai H, Fujimoto K, Yamashita R, Sato T, Someya Y, Taura K, et al. Convolutional neural network for classifying primary liver cancer based on triple-phase CT and tumor marker information: a pilot study. *Jpn J Radiol*. 2021;39(7):690–702. <https://doi.org/10.1007/s11604-021-01106-8>.
42. Menegotto AB, Becker CDL, Cazella SC. Computer-aided diagnosis of hepatocellular carcinoma fusing imaging and structured health data. *Health Inf Sci Syst*. 2021;9(1):20. <https://doi.org/10.1007/s13755-021-00151-x>.

43. Zhou J, Wang W, Lei B, Ge W, Huang Y, Zhang L, et al. Automatic Detection and Classification of Focal Liver Lesions Based on Deep Convolutional Neural Networks: A Preliminary Study. *Frontiers in Oncology*. 2020;10. <https://doi.org/10.3389/fonc.2020.581210>.
44. Shi W, Kuang S, Cao S, Hu B, Xie S, Chen S, et al. Deep learning assisted differentiation of hepatocellular carcinoma from focal liver lesions: choice of four-phase and three-phase CT imaging protocol. *Abdom Radiol (NY)*. 2020;45(9):2688–97. <https://doi.org/10.1007/s00261-020-02485-8>.
45. Kim J, Min JH, Kim SK, Shin SY, Lee MW. Detection of Hepatocellular Carcinoma in Contrast-Enhanced Magnetic Resonance Imaging Using Deep Learning Classifier: A Multi-Center Retrospective Study. *Sci Rep*. 2020;10(1):9458. <https://doi.org/10.1038/s41598-020-65875-4>.
46. Huang Q, Pan F, Li W, Yuan F, Hu H, Huang J, et al. Differential Diagnosis of Atypical Hepatocellular Carcinoma in Contrast-Enhanced Ultrasound Using Spatio-Temporal Diagnostic Semantics. *IEEE Journal of Biomedical and Health Informatics*. 2020;24(10):2860–9. <https://doi.org/10.1109/JBHI.2020.2977937>.
47. Cao SE, Zhang LQ, Kuang SC, Shi WQ, Hu B, Xie SD, et al. Multiphase convolutional dense network for the classification of focal liver lesions on dynamic contrast-enhanced computed tomography. *World J Gastroenterol*. 2020;26(25):3660–72. <https://doi.org/10.3748/wjg.v26.i25.3660>.
48. Ozyurt F, Tuncer T, Avci E, Koc M, Serhatlioglu I. A Novel Liver Image Classification Method Using Perceptual Hash-Based Convolutional Neural Network. *Arabian Journal for Science and Engineering*. 2019;44(4):3173–82. <https://doi.org/10.1007/s13369-018-3454-1>.
49. Oyama A, Hiraoka Y, Obayashi I, Saikawa Y, Furui S, Shiraishi K, et al. Hepatic tumor classification using texture and topology analysis of non-contrast-enhanced three-dimensional T1-weighted MR images with a radiomics approach. *Sci Rep*. 2019;9(1):8764. <https://doi.org/10.1038/s41598-019-45283-z>.
50. Hamm CA, Wang CJ, Savic LJ, Ferrante M, Schobert I, Schlachter T, et al. Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI. *European Radiology*. 2019;29(7):3338–47. <https://doi.org/10.1007/s00330-019-06205-9>.
51. Gatos I, Tsantis S, Karamesini M, Spiliopoulos S, Karnabatidis D, Hazle JD, et al. Focal liver lesions segmentation and classification in nonenhanced T2-weighted MRI. *Med Phys*. 2017;44(7):3695–705. <https://doi.org/10.1002/mp.12291>.
52. Zheng W, Li Q, Zou XB, Wang JW, Han F, Li F, et al. Evaluation of Contrast-enhanced US LI-RADS version 2017: Application on 2020 Liver Nodules in Patients with Hepatitis B Infection. *Radiology*. 2020;294(2):299–307. <https://doi.org/10.1148/radiol.2019.190878>.
53. Steyerberg EW, Harrell FE, Jr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol*. 2016;69:245–7. <https://doi.org/10.1016/j.jclinepi.2015.04.005>.
54. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns (N Y)*. 2023;4(9):100804. <https://doi.org/10.1016/j.patter.2023.100804>.
55. Riley RD, Ensor J, Snell KI, Debray TP, Altman DG, Moons KG, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ*. 2016;353:i3140. <https://doi.org/10.1136/bmj.i3140>.
56. Strandberg R, Jepsen P, Hagstrom H. Developing and validating clinical prediction models in hepatology - An overview for clinicians. *J Hepatol*. 2024;81(1):149–62. <https://doi.org/10.1016/j.jhep.2024.03.030>.
57. 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):477–87. <https://doi.org/10.1016/j.jmir.2019.09.005>.
58. Lee JG, Jun S, Cho YW, Lee H, Kim GB, Seo JB, et al. Deep Learning in Medical Imaging: General Overview. *Korean J Radiol*. 2017;18(4):570–84. <https://doi.org/10.3348/kjr.2017.18.4.570>.
59. Spann A, Yasodhara A, Kang J, Watt K, Wang B, Goldenberg A, et al. Applying Machine Learning in Liver Disease and Transplantation: A Comprehensive Review. *Hepatology*. 2020;71(3):1093–105. <https://doi.org/10.1002/hep.31103>.
60. European Association for the Study of the L. EASL Clinical Practice Guidelines on the management of hepatocellular carcinoma. *J Hepatol*. 2025;82(2):315–74. <https://doi.org/10.1016/j.jhep.2024.08.028>.
61. Lyschchik A, Fetzer DT, Kono Y, Wilson SR, Dietrich CF, Clevert DA, et al. Liver Imaging Reporting and Data System Contrast-Enhanced US Nonradiation Treatment Response Assessment Version 2024. *Radiology*. 2024;311(2):e232369. I think there are something wrong with this reference, it seems two same reference for No. 61, please delete one? <https://doi.org/10.1148/radiol.232369>
62. van der Pol CB, McInnes MDF, Salameh JP, Levis B, Chernyak V, Sirlin CB, et al. CT/MRI and CEUS LI-RADS Major Features Association with Hepatocellular Carcinoma: Individual Patient Data Meta-Analysis. *Radiology*. 2022;302(2):326–35. <https://doi.org/10.1148/radiol.202121244>.
63. Flory MN, Napel S, Tsai EB. Artificial Intelligence in Radiology: Opportunities and Challenges. *Semin Ultrasound CT MR*. 2024;45(2):152–60. <https://doi.org/10.1053/j.sult.2024.02.004>.

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.