TYPE Review
PUBLISHED 19 August 2024
DOI 10.3389/frai.2024.1398205

Check for updates

OPEN ACCESS

EDITED BY
Souptik Barua,
New York University, United States

REVIEWED BY
Ting Li,
National Center for Toxicological Research (FDA), United States
Jonathan Soldera,
University of Caxias do Sul, Brazil

*CORRESPONDENCE
Feras Al-Obeidat
✉ feras.al-obeidat@zu.ac.ae
Wael Hafez
✉ Wael.hafez@nmc.ae

RECEIVED 09 March 2024
ACCEPTED 26 July 2024
PUBLISHED 19 August 2024

CITATION
Al-Obeidat F, Hafez W, Gador M, Ahmed N, Abdeljawad MM, Yadav A and Rashed A (2024) Diagnostic performance of AI-based models versus physicians among patients with hepatocellular carcinoma: a systematic review and meta-analysis. *Front. Artif. Intell.* 7:1398205. doi: 10.3389/frai.2024.1398205

Diagnostic performance of AI-based models versus physicians among patients with hepatocellular carcinoma: a systematic review and meta-analysis

Feras Al-Obeidat[1]*, Wael Hafez[2,3]*, Muneir Gador[3], Nesma Ahmed[4], Marwa Muhammed Abdeljawad[4], Antesh Yadav[2] and Asrar Rashed[2,5]

[1]College of Technological Innovation, Zayed University, Abu Dubai, United Arab Emirates, [2]NMC Royal Hospital, Khalifa City, United Arab Emirates, [3]Internal Medicine Department, Medical Research and Clinical Studies Institute, The National Research Centre, Cairo, Egypt, [4]EMS, Dubai, United Arab Emirates, [5]Department of Computer Science, Edinburgh Napier University, Merchiston Campus, Edinburgh, United Kingdom

Background: Hepatocellular carcinoma (HCC) is a common primary liver cancer that requires early diagnosis due to its poor prognosis. Recent advances in artificial intelligence (AI) have facilitated hepatocellular carcinoma detection using multiple AI models; however, their performance is still uncertain.

Aim: This meta-analysis aimed to compare the diagnostic performance of different AI models with that of clinicians in the detection of hepatocellular carcinoma.

Methods: We searched the PubMed, Scopus, Cochrane Library, and Web of Science databases for eligible studies. The R package was used to synthesize the results. The outcomes of various studies were aggregated using fixed-effect and random-effects models. Statistical heterogeneity was evaluated using I-squared (I^2) and chi-square statistics.

Results: We included seven studies in our meta-analysis;. Both physicians and AI-based models scored an average sensitivity of 93%. Great variation in sensitivity, accuracy, and specificity was observed depending on the model and diagnostic technique used. The region-based convolutional neural network (RCNN) model showed high sensitivity (96%). Physicians had the highest specificity in diagnosing hepatocellular carcinoma(100%); furthermore, models-based convolutional neural networks achieved high sensitivity. Models based on AI-assisted Contrast-enhanced ultrasound (CEUS) showed poor accuracy (69.9%) compared to physicians and other models. The leave-one-out sensitivity revealed high heterogeneity among studies, which represented true differences among the studies.

Conclusion: Models based on Faster R-CNN excel in image classification and data extraction, while both CNN-based models and models combining contrast-enhanced ultrasound (CEUS) with artificial intelligence (AI) had good sensitivity. Although AI models outperform physicians in diagnosing HCC, they should be utilized as supportive tools to help make more accurate and timely decisions.

KEYWORDS

artificial intelligence, hepatocellular carcinoma, HCC, diagnostic performance, AI models

Introduction

Primary liver cancer is a challenging disease that was the second most common cause of cancer mortality globally in 2018, and it is the 7th most common type of cancer (Bray et al., 2018). Between 2020 and 2040, the number of new cases of liver cancer are predicted to increase by 55% per year (Rumgay et al., 2022). Liver cancer cells are differentiated into primary and secondary liver cancers according to the origin of the cancer cells. In primary liver cancer, the cancer originates within the liver itself, while secondary liver cancer is a result of the metastasis of other organs.

Primary liver cancer types include hepatocellular carcinoma (HCC), intrahepatic cholangiocarcinoma, hepatoblastoma, fibrolamellar carcinoma, angiosarcoma, and hemangiosarcoma (Astrologo et al., 2023). HCC accounts for approximately 75% of all liver cancer cases. Infection with hepatitis B and C is the major risk factor for HCC; however, other factors may play a role, such as aflatoxin exposure, alcohol consumption, and smoking (Chuang et al., 2009; Mittal and El-Serag, 2013). The treatment of hepatocellular carcinoma depends on several factors, including tumor size, cancer stage, extrahepatic metastasis, and the extent of vascular invasion. In general, patients with HCC have a poor prognosis, which is determined by the stage of liver disease, disease severity, and diagnosis timing. Therefore, early diagnosis is crucial for a better prognosis (Galle et al., 2018; Marrero et al., 2018).

Serum biomarkers, such as alpha-fetoprotein (AFP-L3), des-gamma-carboxy prothrombin (DCP), Golgi protein 73 (GP73), and glypican-3 (GPC3), have beneficial value for early diagnosis. Several trace chemicals, such as circulating tumor noncoding RNA (ct-ncRNA), cell-free DNA (cfDNA), circulating tumor DNA (ctDNA), and circulating tumor cells (CTCs), are released into biological fluids and could serve as valuable diagnostic agents (Di Tommaso et al., 2009; Toyoda et al., 2011; Guo et al., 2018; Choi et al., 2019). Fine-needle aspiration (FNA) biopsy is considered an additional confirmation test (Galle et al., 2018). Computed tomography (CT) and magnetic resonance imaging (MRI) are widely used in cancer monitoring and diagnosis. Nevertheless, their sensitivity and specificity for early HCC detection are relatively low, so liver-specific contrast agents are used to improve imaging accuracy. The combination of gadolinium-ethoxybenzyl diethylenetriamine pentaacetic acid (Gd-EOB-DTPA) and MRI improves the diagnosis of liver lesions, but this combination is not optimal for small lesion detection (Yu et al., 2014; Marrero et al., 2018). Furthermore, imaging techniques depend on observer interpretation, which represents a major source of error and misdiagnosis; therefore, artificial intelligence (AI)-based models have been developed to overcome this issue.

Radiomics analysis is a novel tool developed for extracting data from medical images and combined with imaging techniques for better performance (Lambin et al., 2012; Wang et al., 2023). Other AI models are based on machine learning, deep learning (DL), and convolutional neural networks (CNNs). The algorithms generated by machine learning must first undergo training on datasets to make predictions. Deep learning is a subset of machine learning that learns and extracts difficult data using multiple layers. Another technology is convolutional neural networks (CNNs), which are considered the ideal model for diagnosis because they can process complex visual data through multiple layers and filters. Different models have been developed in conjunction with traditional methods to optimize the diagnosis process. AI tools are unbiased, smart, cost-effective, and noninvasive, and their efficacy is comparable to that of humans (Yamashita et al., 2018; Saba et al., 2019; Awal et al., 2023). In this systematic review and meta-analysis, we aimed to evaluate the diagnostic performance of different AI models for the diagnosis of hepatocellular carcinoma in comparison with human expertise.

Methods

Literature search

This systematic review and meta-analysis was registered in PROSPERO; **CRD42024517634**; https://www.crd.york.ac.uk/PROSPERO/#recordDetails and conducted in compliance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement (Moher, 2009). We searched electronic databases, including PubMed, Scopus, Cochrane, and Web of Science, through the 15th of February 2024. We used the relevant keywords and MeSH terms for artificial intelligence, machine learning, liver cancer, hepatocellular carcinoma, and diagnosis.

Inclusion and exclusion criteria

Studies were included if they (1) compared the performance of the AI model with that of physicians and (2) reported the sensitivity of the model in diagnosing HCC. Studies were excluded if they (1) were non-English, (2) did not compare AI models with clinicians, (3) did not report the model's ability to differentiate between HCC and other types of liver cancer, or (4) did not report the outcomes of interest.

We excluded reviews, correspondences, editorials, errata, case reports, animal studies, and conference abstracts. No restrictions were applied to the publication year.

Study selection and data extraction

Two independent authors filtered the studies according to their titles and abstracts. The screening was assisted by Rayyan, an online software tool (Ouzzani et al., 2016). Disagreements were settled through discussions. The data were extracted by two independent authors using a standard data extraction sheet, and disagreements were resolved through discussion. We extracted the general characteristics of the included studies, such as the first author name, year of publication, country, sample size, aim of the study, model used, diagnostic technique used, and limitations of the study.

Quality assessment

We used the QUADAS-AI tool to assess bias in the included studies. Two independent authors assessed the quality of the included results, and any discrepancies were resolved through discussion (Sounderajah et al., 2021). This tool addresses four main domains:

1. Subject selection domain: signaling questions evaluate the quality of input data, patient eligibility criteria, source of datasets, image preprocessing, and information about the scanner model.
2. The external validation process is evaluated in the index test domain.
3. Reference standard domain: this domain assesses the ability of the reference standard to classify the target condition correctly.
4. Flow and timing domain: evaluate whether the time between index testing and reference standardization is reasonable.

A study is considered to be at low risk of bias if all signaling questions are answered with "yes," questions answered with "no" flag potential bias, and further discussion is required to reach a final decision. If sufficient data were not available, questions were answered with "unclear."

Statistical analysis

We used R version 4.2.2 (2022-10-31) and RStudio [version 2022.07.2 (2009–2022)] from RStudio, Inc. (R Core Team, 2022; RStudio Team, 2022). We conducted a meta-analysis of the sensitivity, specificity, and accuracy of AI models for the diagnosis of hepatocellular carcinoma (HCC) with the "metafor" package. The outcomes of the various studies were aggregated using fixed effects and random effects models because of the significant heterogeneity observed in the preliminary analysis. Statistical heterogeneity was evaluated using I-squared (I^2) and chi-square statistics. High I^2 values suggested considerable between-study variability, warranting the use of a random-effects model. Leave-one-out sensitivity analysis was conducted to assess the impact of individual studies on the overall meta-analysis.

Results

Search results and baseline characteristics of included studies

We retrieved 1,573 articles from database records. In total, 365 duplicates were detected. Title and abstract screening was performed for 1,172 records, and only 63 studies were eligible for full-text screening. Seven studies (Hamm et al., 2019; Kim et al., 2020; Zhen et al., 2020; Gao et al., 2021; Nishida et al., 2022; Liu et al., 2023; Urhuţ et al., 2023) were included in the meta-analysis, as shown in the PRISMA flow diagram (Figure 1) (Page et al., 2020).

General characteristics of the included studies

The general characteristics of the included studies are shown in Table 1. Seven studies demonstrated the potential of AI tools for improving the accuracy and efficiency of HCC diagnosis. Liu et al. (2023) established a Faster Region-based Convolutional Neural Network (RCNN) model for the differential diagnosis of primary clear cell carcinoma of the liver and common hepatocellular carcinoma (CHCC) (Liu et al., 2023). Similarly, Gao et al. (2021) developed an automatic diagnostic model to differentiate the types of malignant hepatic tumors based on multiphase contrast-enhanced computed tomography (CECT) and clinical data (Gao et al., 2021). Hamm et al. (2019) developed a custom convolutional neural network (CNN) model for classifying hepatic lesions on multiphasic MR images (Hamm et al., 2019). Kim et al. (2020) utilized a fine-tuned CNN to develop a deep learning model for detecting HCC using contrast-enhanced magnetic resonance imaging (MRI) (Kim et al., 2020). Urhuţ et al. (2023) evaluated the accuracy of an automated method for classifying liver lesions using contrast-enhanced ultrasound (CEUS) (Urhuţ et al., 2023). Nishida et al. (2022) constructed AI models for diagnosing liver tumors using B-mode ultrasonography, specifically CNNs, based on the visual geometry group network (VGGNet) model (Nishida et al., 2022). Finally, Zhen et al. (2020) developed a deep learning system (DLS) for classifying liver tumors based on enhanced and unenhanced MR and clinical data using CNNs based on the Inception-ResNet V2 network (Zhen et al., 2020).

Quality assessment

All studies had a low risk of bias in patient selection, index tests, and reference standards, except for Gao et al. (2021). Gao et al. (2021) reported a high risk of bias in the patient selection domain. All studies had concerns regarding the flow and timing domains, as the reported process was not sufficient to judge this domain. Overall, Gao et al. (2021) had a high risk of bias, while the other studies had a low risk (Figure 2).

Results

First of all, it is necessary to outline and explain the parameters of common effect and random effects models employed in the analysis. The fixed-effect model, also known as common-effect model assumes one true effect size underlying all studies included in this meta-analysis. In cases where studies are assumed to be similar enough in design and population, that's when this model is appropriate. The key parameter for this particular model is pooled effect size which represents average of the sizes of the effects obtained from all the studies.

Contrarily, random-effects model believes that there is a difference in true effect sizes among studies. It takes into consideration variability within and between variables across various research thus becoming an

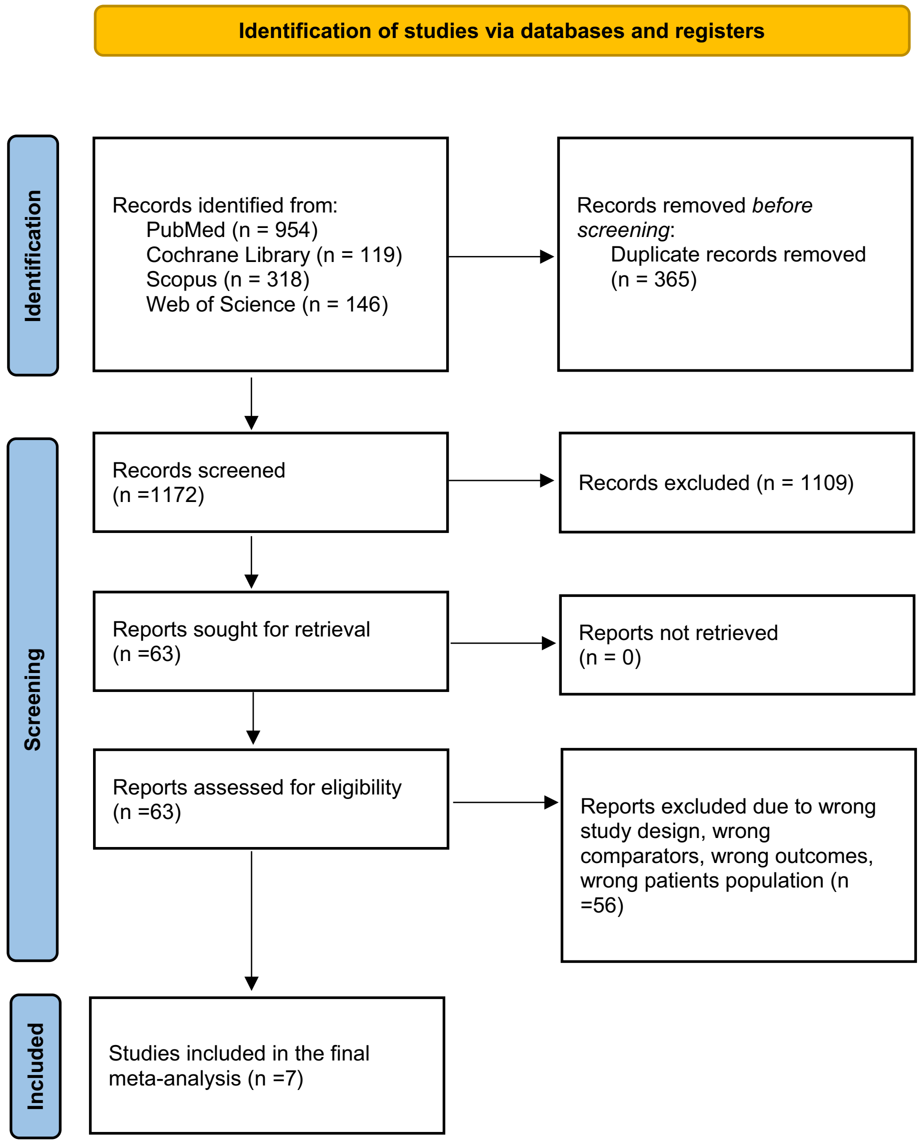

Identification of studies via databases and registers
Identification
Records identified from:
PubMed (n = 954)
Cochrane Library (n = 119)
Scopus (n = 318)
Web of Science (n = 146)
Records removed *before screening*:
Duplicate records removed
(n = 365)
Screening
Records screened
(n =1172)
Records excluded (n = 1109)
Reports sought for retrieval
(n =63)
Reports not retrieved
(n = 0)
Reports assessed for eligibility
(n =63)
Reports excluded due to wrong study design, wrong comparators, wrong outcomes, wrong patients population (n =56)
Included
Studies included in the final meta-analysis (n =7)

FIGURE 1
PRISMA flow diagram.

appropriate choice if considerable heterogeneity exists amongst several researches. Within this specific model the vital parameters include pooled effect size and between-study variance commonly referred to as Tau^2 indicating how much there exists variation amidst true effects.

The between-study variance in a meta-analysis is calculated by Tau^2 (Tau-squared) and it helps us understand how much the effect sizes vary across studies over and above what would be expected by chance alone. Although there is no exact cutoff, we can interpret small heterogeneity of values near 0, moderate heterogeneity ranging from 0.01 to 0.1, and large heterogeneity indicated by values greater than 0.1 (West et al., 2010). For example, in this context, a Tau^2 value of 0.0057 indicates that there is moderate heterogeneity which implies that while there may be some variation of the effect size found across the studies; it is not excessive enough to be called very high heterogeneity as such. In other words, differences between results of these studies are due to more than random occurrences but are not too far apart.

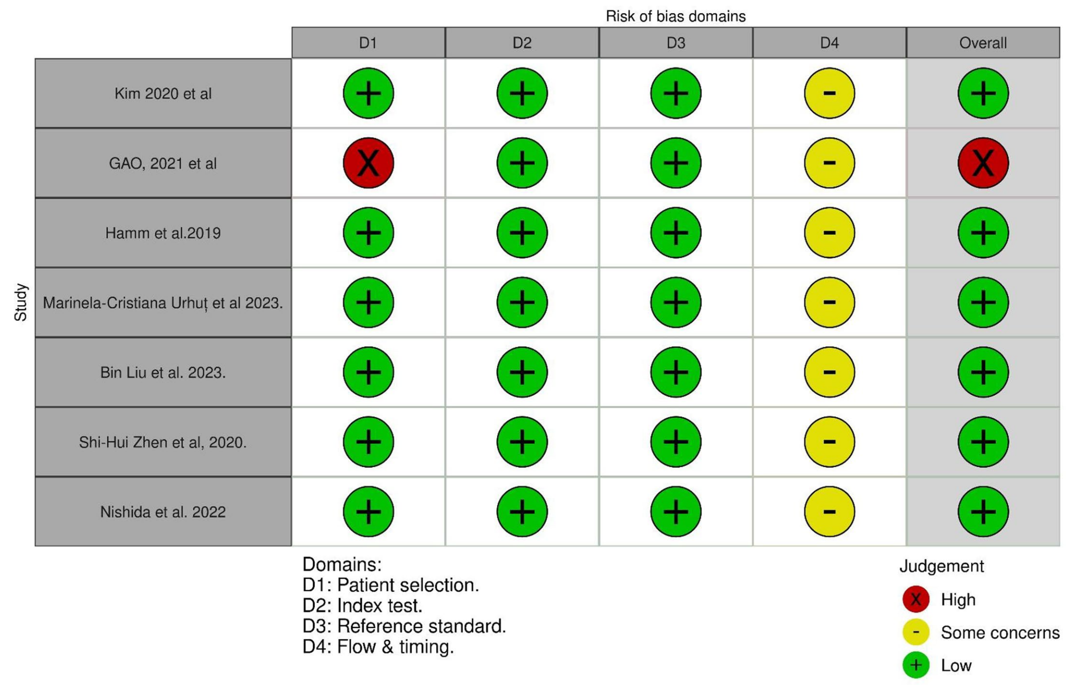
Sensitivity

We included the sensitivities reported in seven studies. Figure 3 shows the meta-analysis of 26 arm. We tested the included studies in the random and fixed effects models. According to the fixed effect model, the pooled sensitivity was 0.9317 (95% CI [0.9219, 0.9415]), suggesting high consistency across studies assuming a single underlying effect. The sensitivity of the random effects model was 0.8360 (95% CI [0.7909, 0.8811]), which accounts for the observed heterogeneity among studies and indicates a broader range of effect sizes. The average sensitivity of the AI-based models or physicians was 0.93, highlighting that the AI models, on average, perform similarly to physicians under the fixed effects model assumption, but these results showed more variability under the random effects model.

The performance of the AI models varied significantly depending on the specific model and diagnostic technique used. Some models,

TABLE 1 General characteristics of the included studies.

Ref.	Country	Sample size (total)	Title	Aim of the study	Diagnostic techniques used	AI tool used	Limitations of the study
Liu et al. (2023)	China	30 patients	Diagnosis of primary clear cell carcinoma of the liver based on Faster RCNN	Establish a Faster RCNN model for differential diagnosis of PCCCL and CHCC	Deep learning analysis of CT images	Faster RCNN	Single center study. -The sample size of the patients with PCCCL was small.
Gao et al. (2021)	China	159 patients	Deep learning for differential diagnosis of malignant hepatic tumors based on multiphase CECT and clinical data	Develop an automatic diagnostic model to differentiate types of malignant hepatic tumors	Multiphase CECT	Deep learning model (STIC)	Single-center study. -A limited number of imaging studies. Only typical lesions on MRI were used, excluding lesions with poor quality and more complex lesion types such as infiltrative HCC or complicated cysts. -Pathological proof was not available for all lesions.
Hamm et al. (2019)	USA	296 patients	Deep learning for liver tumor diagnosis part I: development of a CNN classifier for multiphasic MRI	Develop a CNN for classifying hepatic lesions on multiphasic MRI	Multiphasic MRI	Custom CNN	Relatively small training and testing dataset. -The included group was heterogeneous in terms of tumor types. -Insufficient number for some categories, such as focal nodular hyperplasia, liver abscess, liver adenoma, and cholangiocarcinoma. -Using only one type of ultrasound equipment and a single contrast agent for ultrasound. -Valuable information collected in daily clinical practice, such as tumoral markers, was not integrated.
Kim et al. (2020)	South Korea	950 images	Detection of Hepatocellular Carcinoma in Contrast-Enhanced MRI Using Deep Learning Classifier	Develop a deep learning model for detecting HCC using MRI	Contrast-enhanced MRI	Fine-tuned CNN	The image quality of the arterial phase was affected by transient severe motion artifacts. -The training dataset was obtained from a single vendor. -The study population had relatively good liver function. -The model is unable to detect atypical HCCs and low signal intensity in hepatobiliary phase MRI.
Urhuţ et al. (2023)	Romania	49 patients	Diagnostic Performance of an AI Model Based on CEUS in Patients with Liver Lesions	Evaluate the accuracy of an automated method for classifying liver lesions using CEUS	CEUS (contrast-enhance ultrasound)	AI system based on algorithms	Single center study. -Lesions segmentation in the training validation set was done manually by doctors.
Nishida et al. (2022)	Japan	55 patients	Artificial intelligence models for the ultrasonographic diagnosis of liver tumors	Construct AI models for diagnosing liver tumors using ultrasonography	B-mode ultrasonography	CNNs based on VGGNet	A single-center retrospective study. -Patients who have specific types of focal liver diseases (small HCC, HCC without pathology, inflammation, etc.) need to be included in future training.
Zhen et al. (2020)	China	201 patients	Deep Learning for Accurate Diagnosis of Liver Tumor Based on MRI and Clinical Data	Develop a DLS for classifying liver tumors based on MRI and clinical data	Enhanced and unenhanced MRI	CNNs based on Inception-ResNet V2	The AI model focuses on diagnosis not detection. -Other types of rare liver tumors were not involved in the training set.

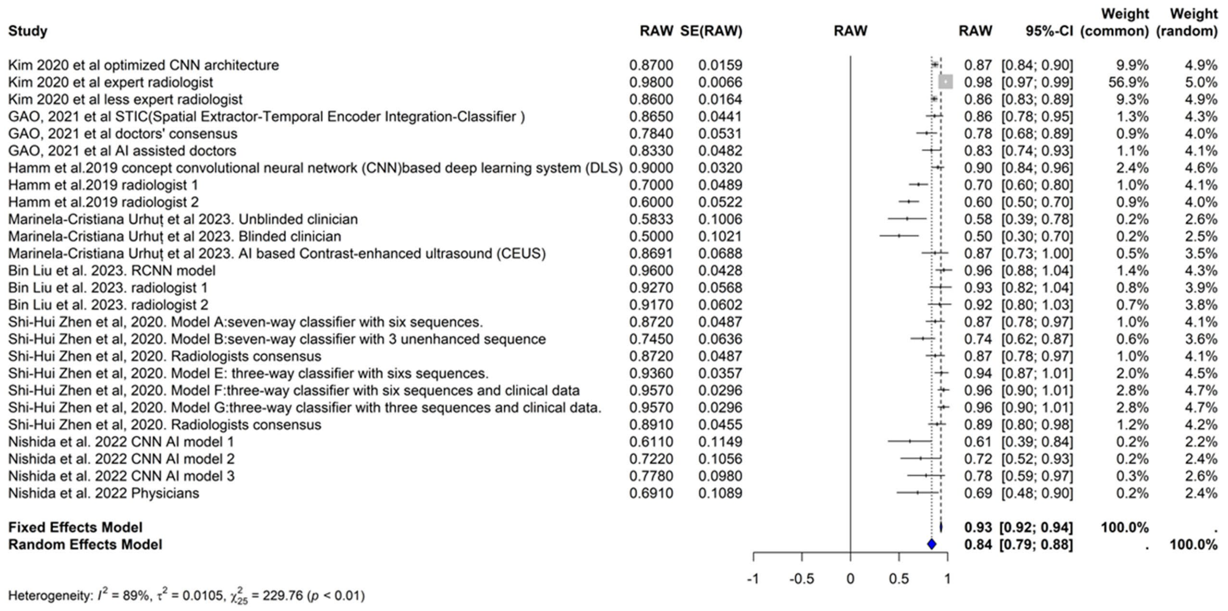


Study	D1	D2	D3	D4	Overall
Kim 2020 et al	+	+	+	-	+
GAO, 2021 et al	X	+	+	-	X
Hamm et al.2019	+	+	+	-	+
Marinela-Cristiana Urhuț et al 2023.	+	+	+	-	+
Bin Liu et al. 2023.	+	+	+	-	+
Shi-Hui Zhen et al, 2020.	+	+	+	-	+
Nishida et al. 2022	+	+	+	-	+

Domains:
D1: Patient selection.
D2: Index test.
D3: Reference standard.
D4: Flow & timing.

Judgement	
X	High
-	Some concerns
+	Low

FIGURE 2
Quality assessment of the included studies.



Study	RAW	SE(RAW)	RAW	95%-CI	Weight (common)	Weight (random)
Kim 2020 et al optimized CNN architecture	0.8700	0.0159	0.87	[0.84; 0.90]	9.9%	4.9%
Kim 2020 et al expert radiologist	0.9800	0.0066	0.98	[0.97; 0.99]	56.9%	5.0%
Kim 2020 et al less expert radiologist	0.8600	0.0164	0.86	[0.83; 0.89]	9.3%	4.9%
GAO, 2021 et al STIC(Spatial Extractor-Temporal Encoder Integration-Classifier)	0.8650	0.0441	0.86	[0.78; 0.95]	1.3%	4.3%
GAO, 2021 et al doctors' consensus	0.7840	0.0531	0.78	[0.68; 0.89]	0.9%	4.0%
GAO, 2021 et al AI assisted doctors	0.8330	0.0482	0.83	[0.74; 0.93]	1.1%	4.1%
Hamm et al.2019 concept convolutional neural network (CNN)based deep learning system (DLS)	0.9000	0.0320	0.90	[0.84; 0.96]	2.4%	4.6%
Hamm et al.2019 radiologist 1	0.7000	0.0489	0.70	[0.60; 0.80]	1.0%	4.1%
Hamm et al.2019 radiologist 2	0.6000	0.0522	0.60	[0.50; 0.70]	0.9%	4.0%
Marinela-Cristiana Urhuț et al 2023. Unblinded clinician	0.5833	0.1006	0.58	[0.39; 0.78]	0.2%	2.6%
Marinela-Cristiana Urhuț et al 2023. Blinded clinician	0.5000	0.1021	0.50	[0.30; 0.70]	0.2%	2.5%
Marinela-Cristiana Urhuț et al 2023. AI based Contrast-enhanced ultrasound (CEUS)	0.8691	0.0688	0.87	[0.73; 1.00]	0.5%	3.5%
Bin Liu et al. 2023. RCNN model	0.9600	0.0428	0.96	[0.88; 1.04]	1.4%	4.3%
Bin Liu et al. 2023. radiologist 1	0.9270	0.0568	0.93	[0.82; 1.04]	0.8%	3.9%
Bin Liu et al. 2023. radiologist 2	0.9170	0.0602	0.92	[0.80; 1.03]	0.7%	3.8%
Shi-Hui Zhen et al, 2020. Model A:seven-way classifier with six sequences.	0.8720	0.0487	0.87	[0.78; 0.97]	1.0%	4.1%
Shi-Hui Zhen et al, 2020. Model B:seven-way classifier with 3 unenhanced sequence	0.7450	0.0636	0.74	[0.62; 0.87]	0.6%	3.6%
Shi-Hui Zhen et al, 2020. Radiologists consensus	0.8720	0.0487	0.87	[0.78; 0.97]	1.0%	4.1%
Shi-Hui Zhen et al, 2020. Model E: three-way classifier with sixs sequences.	0.9360	0.0357	0.94	[0.87; 1.01]	2.0%	4.5%
Shi-Hui Zhen et al, 2020. Model F:three-way classifier with six sequences and clinical data	0.9570	0.0296	0.96	[0.90; 1.01]	2.8%	4.7%
Shi-Hui Zhen et al, 2020. Model G:three-way classifier with three sequences and clinical data.	0.9570	0.0296	0.96	[0.90; 1.01]	2.8%	4.7%
Shi-Hui Zhen et al, 2020. Radiologists consensus	0.8910	0.0455	0.89	[0.80; 0.98]	1.2%	4.2%
Nishida et al. 2022 CNN AI model 1	0.6110	0.1149	0.61	[0.39; 0.84]	0.2%	2.2%
Nishida et al. 2022 CNN AI model 2	0.7220	0.1056	0.72	[0.52; 0.93]	0.2%	2.4%
Nishida et al. 2022 CNN AI model 3	0.7780	0.0980	0.78	[0.59; 0.97]	0.3%	2.6%
Nishida et al. 2022 Physicians	0.6910	0.1089	0.69	[0.48; 0.90]	0.2%	2.4%
Fixed Effects Model			**0.93**	**[0.92; 0.94]**	**100.0%**	.
Random Effects Model			**0.84**	**[0.79; 0.88]**	.	**100.0%**

Heterogeneity: $I^2 = 89\%$, $\tau^2 = 0.0105$, $\chi^2_{25} = 229.76$ ($p < 0.01$)

FIGURE 3
Forst plot of sensitivity analysis for AI models in hepatocellular carcinoma diagnosis.

such as the RCNN model and the various classifiers used in Zhen et al. (2020), exhibited particularly high sensitivities. In contrast, the CNN-based models reported by Nishida et al. (2022) demonstrated lower sensitivities.

The heterogeneity of these studies was high. The I^2 statistic was 89.1%, indicating a large variation among the studies' estimates not only due to random error. The heterogeneity test was highly significant (Q = 229.76, df. = 25, p value <0.0001). The Tau^2 statistic was 0.0105 (95% CI [0.0058–0.0277]), supporting moderate heterogeneity and reinforcing the need for a random-effects model to accurately capture the variability in sensitivities among different studies. The leave-one-out sensitivity analysis revealed that the pooled sensitivity estimates of the fixed effects model ranged from 0.8678 to 0.9390, and those of the random effects model ranged from 0.8292 to 0.8506, which confirmed the robustness of our findings, with pooled sensitivity estimates remaining stable across different models. These results suggest that no single study unduly influenced the meta-analysis results. Figure 4 represents the funnel plot of the included studies, suggesting potential publication bias.

Specificity

The specificity of the AI models used in the diagnosis of hepatocellular carcinoma (HCC) was evaluated in six studies (Figure 5). The common effect model showed a perfect specificity of 1.0000 (95% CI [0.9999, 1.0001]). The specificity in the random effects model was 0.9252 (95% CI [0.8915, 0.9589]), reflecting the variability among the studies. In Kim et al. (2020), the optimized CNN model had the same specificity as the expert radiologist (93%). Model F in Zhen et al. (2020) had the highest specificity (96.2%), while the model in Urhuţ et al. (2023) had the lowest value (56.2%) (see Figure 6).

Heterogeneity among the studies was quantified using several statistical methods. The I^2 statistic was 89.5%, which indicates that the total variation in the study estimates was due to heterogeneity rather than chance. The Tau^2 statistic was 0.0035 (95% CI of [0.0024–0.0185]), which suggests moderate heterogeneity in specificity estimates across studies. This heterogeneity may be influenced by the varying odds of having HCC across these studies. Higher odds might reduce specificity due to the greater chance of false positives in a more homogenous patient population. The heterogeneity test was highly significant (Q = 209.54, df. = 22, p value <0.0001), confirming significant between-study variability, justifying the use of the random-effects model.

Accuracy

The accuracy of the AI models used in the diagnosis of hepatocellular carcinoma (HCC) was evaluated in four studies (Figure 7). The common effect model showed an accuracy of 0.9096 (95% CI [0.8958,0.9234]), suggesting high consistency across studies assuming a single underlying accuracy. The random effects model showed a slightly lower accuracy of 0.8423 (95% CI [0.7879–0.8966]). Most of the AI-based models achieved accuracies similar to or marginally better than those of physicians. Model 3 by Nishida et al. (2022) achieved the highest accuracy (92.7%), while the models of Kim et al. (2020) and Gao et al. (2021) had slightly better accuracy than did the other models (90% vs. 91 and 72.6% vs. 70.8%, respectively). The Urhuţ et al. (2023) model had the lowest accuracy (69.6%), which was consistent with the specificity value. The I^2 statistic was 74.1%, indicating that a large proportion of the total variation in the study estimates was due to heterogeneity rather than chance. The Tau^2 statistic was 0.0057 (95% CI [0.0013, 0.0211]), highlighting moderate heterogeneity in the accuracy estimates across studies. This variability may be influenced by the differing odds of HCC among the study populations. Higher odds might lead to greater accuracy due to the increased incidence of the condition, while lower odds might result in lower accuracy as the model encounters more non-HCC patients. The heterogeneity test result was highly significant (Q = 42.53, df. = 11, p value <0.0001), supporting the use of the random-effects model due to significant between-study differences.

Figure 8 illustrates a funnel plot utilized in the meta-analysis to examine potential publication bias; the funnel plot indicates potential publication bias among the included studies.

Odds of HCC in study populations

To provide context for evaluating the models' performance, the odds of having hepatocellular carcinoma (HCC) were calculated for each study population. In Liu et al. (2023), the odds of having HCC were infinite because the sample consisted entirely of HCC patients. According to Gao et al. (2021), the odds of having HCC are approximately 1.69. According to Hamm et al. (2019), the odds ratio was approximately 2.08. Kim et al. (2020) reported that the odds of having HCC were approximately 0.73. The odds ratio of Urhuţ et al. (2023) was approximately 0.69. Nishida et al. (2022) reported odds of approximately 1.2%. Finally, Zhen et al. (2020) reported that the odds of having HCC were approximately 2.94 (Table 2).

The varying odds of HCC across these studies could influence the reported sensitivities and specificities. For instance, Nishida

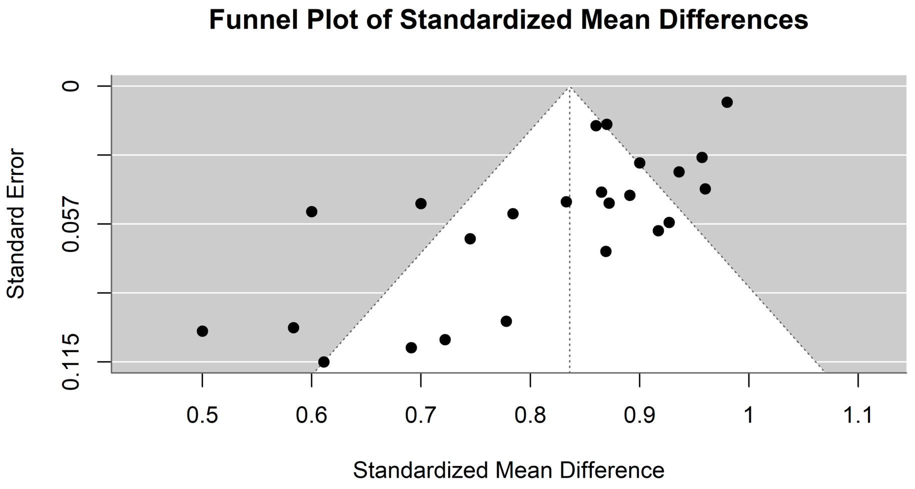
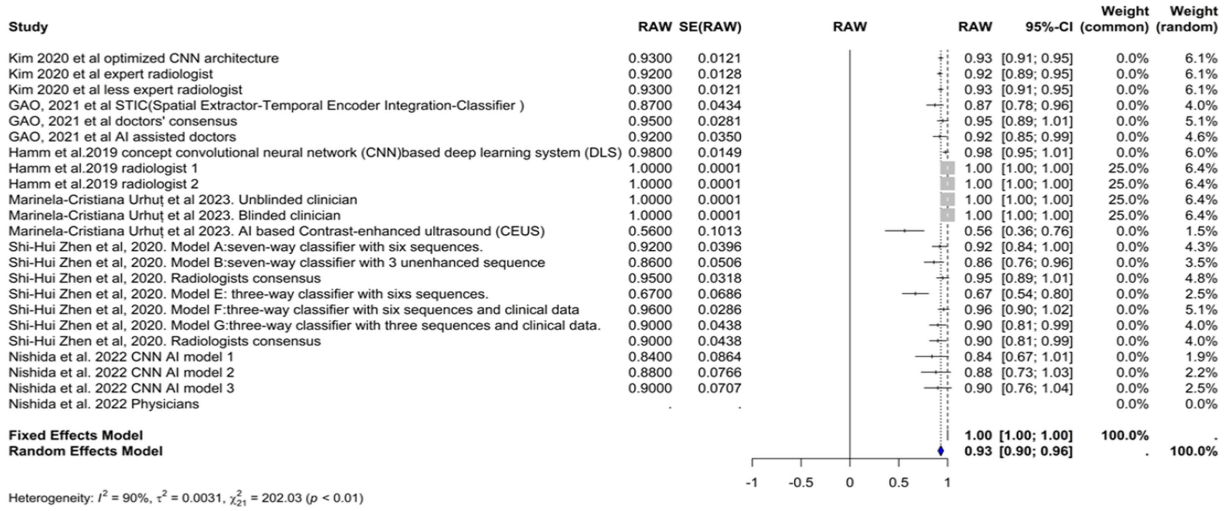


FIGURE 4
Funnel plot demonstrating publication bias in the meta-analysis of sensitivity across various models for hepatocellular carcinoma detection.



Study	RAW	SE(RAW)	RAW	95%-CI	Weight (common)	Weight (random)
Kim 2020 et al optimized CNN architecture	0.9300	0.0121	0.93	[0.91; 0.95]	0.0%	6.1%
Kim 2020 et al expert radiologist	0.9200	0.0128	0.92	[0.89; 0.95]	0.0%	6.1%
Kim 2020 et al less expert radiologist	0.9300	0.0121	0.93	[0.91; 0.95]	0.0%	6.1%
GAO, 2021 et al STIC(Spatial Extractor-Temporal Encoder Integration-Classifier)	0.8700	0.0434	0.87	[0.78; 0.96]	0.0%	4.0%
GAO, 2021 et al doctors' consensus	0.9500	0.0281	0.95	[0.89; 1.01]	0.0%	5.1%
GAO, 2021 et al AI assisted doctors	0.9200	0.0350	0.92	[0.85; 0.99]	0.0%	4.6%
Hamm et al.2019 concept convolutional neural network (CNN)based deep learning system (DLS)	0.9800	0.0149	0.98	[0.95; 1.01]	0.0%	6.0%
Hamm et al.2019 radiologist 1	1.0000	0.0001	1.00	[1.00; 1.00]	25.0%	6.4%
Hamm et al.2019 radiologist 2	1.0000	0.0001	1.00	[1.00; 1.00]	25.0%	6.4%
Marinela-Cristiana Urhuț et al 2023. Unblinded clinician	1.0000	0.0001	1.00	[1.00; 1.00]	25.0%	6.4%
Marinela-Cristiana Urhuț et al 2023. Blinded clinician	1.0000	0.0001	1.00	[1.00; 1.00]	25.0%	6.4%
Marinela-Cristiana Urhuț et al 2023. AI based Contrast-enhanced ultrasound (CEUS)	0.5600	0.1013	0.56	[0.36; 0.76]	0.0%	1.5%
Shi-Hui Zhen et al, 2020. Model A:seven-way classifier with six sequences.	0.9200	0.0396	0.92	[0.84; 1.00]	0.0%	4.3%
Shi-Hui Zhen et al, 2020. Model B:seven-way classifier with 3 unenhanced sequence	0.8600	0.0506	0.86	[0.76; 0.96]	0.0%	3.5%
Shi-Hui Zhen et al, 2020. Radiologists consensus	0.9500	0.0318	0.95	[0.89; 1.01]	0.0%	4.8%
Shi-Hui Zhen et al, 2020. Model E: three-way classifier with sixs sequences.	0.6700	0.0686	0.67	[0.54; 0.80]	0.0%	2.5%
Shi-Hui Zhen et al, 2020. Model F:three-way classifier with six sequences and clinical data	0.9600	0.0286	0.96	[0.90; 1.02]	0.0%	5.1%
Shi-Hui Zhen et al, 2020. Model G:three-way classifier with three sequences and clinical data.	0.9000	0.0438	0.90	[0.81; 0.99]	0.0%	4.0%
Shi-Hui Zhen et al, 2020. Radiologists consensus	0.9000	0.0438	0.90	[0.81; 0.99]	0.0%	4.0%
Nishida et al. 2022 CNN AI model 1	0.8400	0.0864	0.84	[0.67; 1.01]	0.0%	1.9%
Nishida et al. 2022 CNN AI model 2	0.8800	0.0766	0.88	[0.73; 1.03]	0.0%	2.2%
Nishida et al. 2022 CNN AI model 3	0.9000	0.0707	0.90	[0.76; 1.04]	0.0%	2.5%
Nishida et al. 2022 Physicians	.	.			0.0%	0.0%
Fixed Effects Model			**1.00**	**[1.00; 1.00]**	**100.0%**	**.**
Random Effects Model			**0.93**	**[0.90; 0.96]**	**.**	**100.0%**

Heterogeneity: $I^2 = 90\%$, $\tau^2 = 0.0031$, $\chi^2_{21} = 202.03$ ($p < 0.01$)

FIGURE 5
Forst plot of specificity analysis for AI models in hepatocellular carcinoma diagnosis.

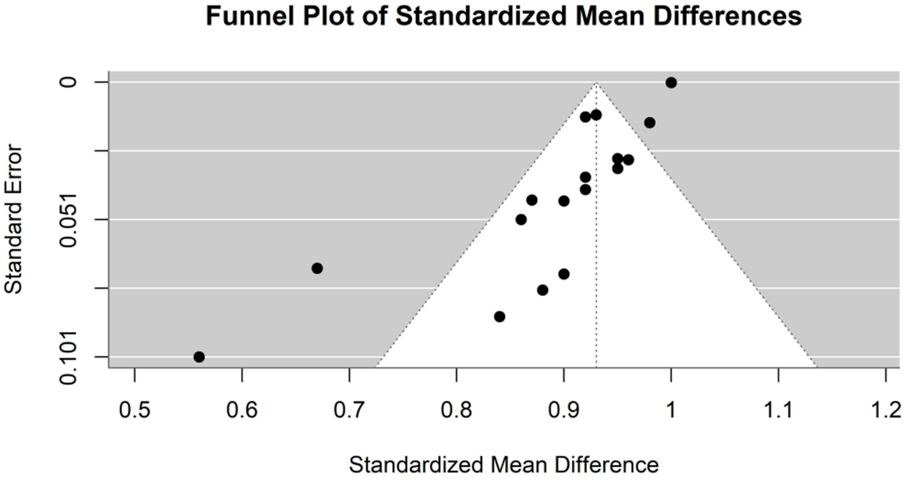


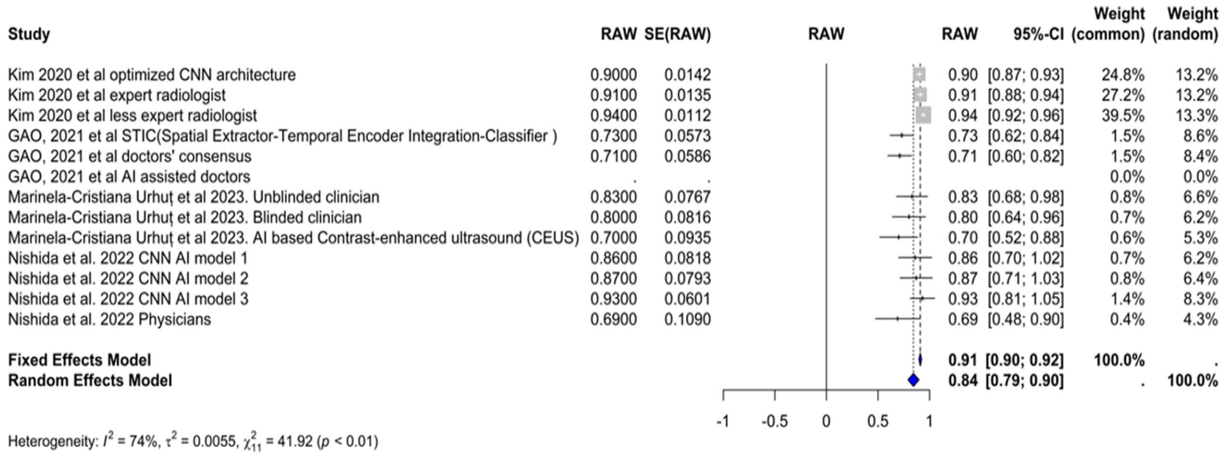
FIGURE 6
Funnel plot demonstrating publication bias in the meta-analysis of specificity across various models for hepatocellular carcinoma detection.

et al. (2022) demonstrated lower sensitivities, which may be influenced by the odds of having HCC in their study population (approximately 1.2%). A greater risk of HCC might lead to greater sensitivity due to the increased incidence of this condition. In comparison, lower odds might result in lower sensitivity as the model encounters more non-HCC patients. Specificity may also be affected similarly, with higher odds potentially reducing specificity due to the greater chance of false positives in a more homogenous patient population.

Discussion

In this systematic review and meta-analysis, we aimed to explore the potential of artificial intelligence tools for the diagnosis of hepatocellular carcinoma compared with human expertise. New models are being developed daily, which is considered an invaluable opportunity for advancing diagnostic accuracy and saving doctors' time. Several scientists have debated whether these tools can replace humans in the future. Our study is the first systematic review to compare AI-based tools with physicians in the diagnosis of HCC. Early diagnosis of HCC is pivotal because patient survival is linked to hepatocellular carcinoma staging; patients diagnosed in early stages have higher five-year survival rates than those diagnosed in late stages (70 and 20%, respectively). Therefore, there is an urgent need to develop accurate and sensitive tools to optimize the diagnostic process for HCC (American Cancer Society, 2020).

Our analysis showed that most AI models are more sensitive than physicians, except for Kim et al. (2020), in which the expert radiologist had greater sensitivity than the AI model (98 vs. 87, respectively); however, they had almost the same specificity (93%). The proposed model had a sensitivity similar to that of a nonexpert radiologist



Study	RAW	SE(RAW)	RAW	95%-CI	Weight (common)	Weight (random)
Kim 2020 et al optimized CNN architecture	0.9000	0.0142	0.90	[0.87; 0.93]	24.8%	13.2%
Kim 2020 et al expert radiologist	0.9100	0.0135	0.91	[0.88; 0.94]	27.2%	13.2%
Kim 2020 et al less expert radiologist	0.9400	0.0112	0.94	[0.92; 0.96]	39.5%	13.3%
GAO, 2021 et al STIC(Spatial Extractor-Temporal Encoder Integration-Classifier)	0.7300	0.0573	0.73	[0.62; 0.84]	1.5%	8.6%
GAO, 2021 et al doctors' consensus	0.7100	0.0586	0.71	[0.60; 0.82]	1.5%	8.4%
GAO, 2021 et al AI assisted doctors	.	.			0.0%	0.0%
Marinela-Cristiana Urhuţ et al 2023. Unblinded clinician	0.8300	0.0767	0.83	[0.68; 0.98]	0.8%	6.6%
Marinela-Cristiana Urhuţ et al 2023. Blinded clinician	0.8000	0.0816	0.80	[0.64; 0.96]	0.7%	6.2%
Marinela-Cristiana Urhuţ et al 2023. AI based Contrast-enhanced ultrasound (CEUS)	0.7000	0.0935	0.70	[0.52; 0.88]	0.6%	5.3%
Nishida et al. 2022 CNN AI model 1	0.8600	0.0818	0.86	[0.70; 1.02]	0.7%	6.2%
Nishida et al. 2022 CNN AI model 2	0.8700	0.0793	0.87	[0.71; 1.03]	0.8%	6.4%
Nishida et al. 2022 CNN AI model 3	0.9300	0.0601	0.93	[0.81; 1.05]	1.4%	8.3%
Nishida et al. 2022 Physicians	0.6900	0.1090	0.69	[0.48; 0.90]	0.4%	4.3%
Fixed Effects Model			**0.91**	**[0.90; 0.92]**	**100.0%**	.
Random Effects Model			**0.84**	**[0.79; 0.90]**	.	**100.0%**

Heterogeneity: $I^2 = 74\%$, $\tau^2 = 0.0055$, $\chi^2_{11} = 41.92$ ($p < 0.01$)

FIGURE 7
Forst plot of accuracy analysis for AI models in hepatocellular carcinoma diagnosis.

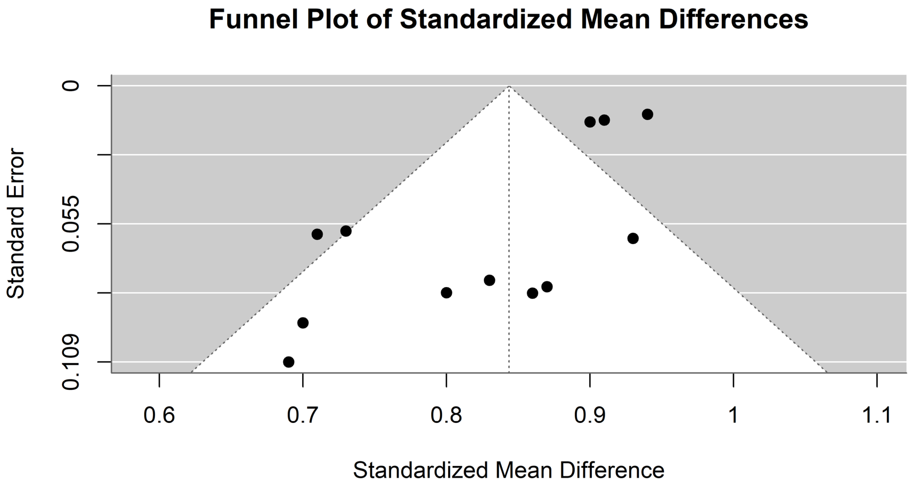


FIGURE 8
Funnel plot demonstrating publication bias in the meta-analysis of accuracy across various models for hepatocellular carcinoma detection.

(approximately 86.5%) (Kim et al., 2020). In some cases, the diagnosis of HCC using imaging techniques is quite challenging and requires strong experience, as there is great heterogeneity within HCC cells; different areas can have different growth patterns and levels of differentiation (Quaglia, 2018).

To address this issue, Zhen et al. (2020) developed an AI model that combines MR images and clinical data. This combination significantly improved the classification ability of the new models [AUC = 0.985 (95% CI, 0.960–1)]. Furthermore, Zhen et al. (2020) investigated the impact of sequence number on model performance. Different imaging sequences were used to train the model. Six sequences (contrast-enhanced T1 sequence, contrast-enhanced T1 sequence, diffusion, late arterial, portal venous, and equilibrium) or three sequences (T1, T2, and diffusion) as well as clinical data were utilized as model inputs. When clinical data were built into the model, the number of sequences did not have an impact on the performance. However, for models based on images only, model A, which had only three unenhanced sequences (AUC = 0.925, 95% CI = 0.871, 0.978), had better results than model B, which had 0.879, 95% CI = 0.813, 0.9452 (Zhen et al., 2020). Gao et al. (2021) measured the sensitivity and accuracy of AI-assisted physicians versus the AI model (STIC) or physicians alone. The AI-assisted physicians and the STIC model had approximately comparable sensitivities of 83.3 and 86.5%, respectively. Physicians had the lowest sensitivity (78.4%), but their accuracy did not significantly differ from that of the STIC model (70.8 and 72.6%, respectively) (Hamm et al., 2019; Kim et al., 2020; Zhen et al., 2020; Gao et al., 2021; Nishida et al., 2022; Liu et al., 2023; Urhuț et al., 2023).

The most sensitive model was the faster region-based convolutional neural network (RCNN) proposed by Liu et al. (2023). Models based on RCNNs exhibit a good ability to extract data from various images (Hamm et al., 2019; Kim et al., 2020; Zhen et al., 2020; Gao et al., 2021; Nishida et al., 2022; Liu et al., 2023; Urhuț et al., 2023). Raimundo et al. (2023) developed an AI model that combines RCCN and MRI for the diagnosis of breast cancer. Similarly, this model achieved a high accuracy of 94.4%. Another model by Liu et al.

TABLE 2 Odds of HCC in study populations.

Study	Arms	Odds ratio	Sample size	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)
Kim et al. (2020)	Optimized CNN architecture	0.73	549	0.8700	0.93	0.90
	Expert radiologist			0.9800	0.92	0.91
	Less expert radiologist			0.8600	0.93	0.94
Gao et al. (2021)	STIC (Spatial Extractor-Temporal Encoder Integration-Classifier)	1.69	60	0.8650	0.87	0.73
	Doctors' consensus			0.7840	0.95	0.71
	AI assisted doctors			0.8330	0.92	–
Hamm et al. (2019)	Concept convolutional neural network (CNN) based deep learning system (DLS)	2.08	88	0.9000	0.98	–
	Radiologist 1			0.7000	1.00	–
	Radiologist 2			0.6000	1.00	–
Urhuţ et al. (2023)	Unblinded clinician	0.69	24	0.5833	1.00	0.83
	Blinded clinician			0.5000	1.00	0.80
	AI based Contrast-enhanced ultrasound (CEUS)			0.8691	0.56	0.70
Liu et al. (2023)	RCNN model	Infinite	21	0.9600	–	–
	Radiologist 1			0.9270	–	–
	Radiologist 2			0.9170	–	–
Zhen et al. (2020)	Model A: seven-way classifier with six sequences.	2.94	47	0.8720	0.92	–
	Model B: seven-way classifier with 3 unenhanced sequences			0.7450	0.86	–
	Radiologists' consensus			0.8720	0.95	–
	Model E: three-way classifier with sixes sequences.			0.9360	0.67	–
	Model F: three-way classifier with six sequences and clinical data			0.9570	0.96	–
	Model G: three-way classifier with three sequences and clinical data.			0.9570	0.90	–
	Radiologists' consensus			0.8910	0.90	–
Nishida et al. (2022)	CNN AI model 1	1.2	18	0.6110	0.84	0.86
	CNN AI model 2			0.7220	0.88	0.87
	CNN AI model 3			0.7780	0.90	0.93
	Physicians			0.6910	–	0.69

(2022) was developed based on the RCNN and medical images for diagnosing bile duct tumor thrombi in patients with HCC. In addition, this model showed a high sensitivity of 94% and a good specificity of 78%, which confirms the efficiency of RCNN-based models in improving the diagnosis of medical images (Liu et al., 2022, 2023; Raimundo et al., 2023).

Several diagnostic AI models based on convolutional neural networks (CNNs) have evolved because they do not require a clear definition of the lesion to interpret the images. Yamashita et al. (2020) proposed a CNN-based model that could classify hepatic observations without predefining hand-crafted imaging features with 60.4% accuracy. Similarly, Wang et al. (2019) investigated a convolutional neural network (CNN)-based model for the differentiation of liver masses via dynamic contrast agent-enhanced computed tomography (CT). The model achieved a median diagnostic accuracy of 0.84 (Yasaka et al., 2018).

Nishida et al. (2022) developed three different models to establish the relationship between the training dataset size and CNN performance. Model 3, which had the most training data, performed better than the other two models and the physicians did. Increasing the training set size has a proven impact on improving the accuracy of CNN-based models; nevertheless, large training sets consume more money, time, and effort (Radiuk, 2017; Nishida et al., 2022).

Clinical data have a major influence on imaging interpretation. Urhuţ et al. (2023) compared the performance of novel AI-based contrast-enhanced ultrasound (CEUS) with that of two groups of clinicians: the first group knew all relevant clinical data, while the second group was blinded. The sensitivity and accuracy of the blinded group were lower than those of the unblinded group (58.3 and 83% vs. 50 and 79.6%, respectively). Furthermore, the AI-based contrast-enhanced ultrasound (CEUS) model had high sensitivity (86.9%) but less accuracy (69.9%) than did clinicians who were aware of the patient's case (Urhuț et al., 2023). The sensitivity of contrast-enhanced ultrasound varies across different studies. A meta-analysis involving studies conducted in 1996–2016 showed that the sensitivity of CEUS for detecting HCC was 85%, although Jiang et al. (2021) reported a lower sensitivity of 69% for lesions <20 mm and 75% for lesions ≥20 mm. The combination of CEUS and AFP levels showed promising results, and the sensitivity increased to 83.1%; therefore, there is a need to establish evidence of the efficiency of CEUS in the diagnosis of HCC and to investigate the performance of the triple combination of CEUS, AFP levels, and AI models (Zhang et al., 2017; Jiang et al., 2021). In addition to diagnosing HCC, machine learning has a promising role in liver transplantation. Machine learning-based models could provide excellent predictions of the risk of complications and short-and long-term mortality. Furthermore, they can predict posttransplant outcomes better than traditional scoring systems. The incorporation of AI into the transplantation process saves time and money and increases the rate of transportation success (Chongo and Soldera, 2024). In general, AI has great potential in diagnosing different gastrointestinal tract pathologies due to its ability to use complex mathematical data involving multiple parameters and sophisticated formulas to draw conclusions that would be challenging or unfeasible for humans to process alone. However, many concerns are associated with the potential of AI in prognostication, such as the need for high-quality images for model training, ethical considerations of data usage, and the cost–benefit ratio (Do and Gastrointestinal, 2023).

Limitations of the study

In our meta-analysis, high heterogeneity was observed, which may be attributed to the use of various AI models and diagnostic tools, in addition to differences in sample size. Some were single-center preliminary studies with small sample sizes, which is considered a major limitation. Another limitation is the inadequate inclusion of rare and complex liver lesions in the training and validation datasets of the AI models. There is a great need for randomized clinical trials with larger sample sizes. Moreover, developers should incorporate clinical history, laboratory findings, prior pathology reports, and immunohistochemistry tests to maximize model efficiency during the development of new AI models.

Conclusion

Compared with physicians, AI models have the ability to improve the diagnosis of hepatocellular carcinoma. Faster R-CNN-based models are excellent for imaging classification and data extraction. CNN-based models have high sensitivity, and increasing the size of the training database significantly augments the accuracy of CNN models. The sensitivity of CEUS in the diagnosis of HCC is debatable; however, the combination of CEUS and AI models yields high sensitivity. Despite the promising results, AI models should not entirely replace humans in the diagnostic process; rather, they should be used as an assistant tool for more accurate and less timed decisions. We need to conduct more studies on the performance of AI-assisted physicians versus physicians without assistance, considering physicians' level of experience.

Author contributions

FA-O: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. WH: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. MG: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. NA: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. MA: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. AY: Data curation, Formal analysis, Validation, Software, Writing – review and editing. AR: Data curation, Formal analysis, Methodology, Validation, Software, Writing – review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

American Cancer Society (2020). Erratum: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 70:313. doi: 10.3322/caac.21609

Astrologo, N. C. N., Gaudillo, J. D., Albia, J. R., and Roxas-Villanueva, R. M. L. (2023). Genetic risk assessment based on association and prediction studies. *Sci. Rep*. 13:15230. doi: 10.1038/s41598-023-41862-3

Awal, W., De, G. J., and Chan, J. (2023). Advances in artificial intelligence for the radiological diagnosis of hepatocellular carcinoma. *EAS J. Radiol Imaging Technol.* 5, 80–84. doi: 10.36349/easjrit.2023.v05i03.010

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

Choi, J., Kim, G., Han, S., Lee, W., Chun, S., and Lim, Y. (2019). Longitudinal assessment of three serum biomarkers to detect very early-stage hepatocellular carcinoma. *Hepatology* 69, 1983–1994. doi: 10.1002/hep.30233

Chongo, G., and Soldera, J. (2024). Use of machine learning models for the prognostication of liver transplantation: a systematic review. *World J. Transplant.* 14:88891. doi: 10.5500/wjt.v14.i1.88891

Chuang, S. C., La, V. C., and Boffetta, P. (2009). Liver cancer: descriptive epidemiology and risk factors other than HBV and HCV infection. *Cancer Lett.* 286, 9–14. doi: 10.1016/j.canlet.2008.10.040

Di Tommaso, L., Destro, A., Seok, J. Y., Balladore, E., Terracciano, L., Sangiovanni, A., et al. (2009). The application of markers (HSP70 GPC3 and GS) in liver biopsies is useful for detection of hepatocellular carcinoma. *J. Hepatol.* 50, 746–754. doi: 10.1016/j.jhep.2008.11.014

Do, P., and Gastrointestinal, T. (2023). Artificial intelligence as a prognostic tool for gastrointestinal tract pathologies. *Inteligência artificial como ferramenta prognóstica para* 34, 9–14.

Galle, P. R., Forner, A., Llovet, J. M., Mazzaferro, V., Piscaglia, F., Raoul, J. L., et al. (2018). EASL clinical practice guidelines: management of hepatocellular carcinoma. *J. Hepatol.* 69, 182–236. doi: 10.1016/j.jhep.2018.03.019

Gao, R., Zhao, S., Aishanjiang, K., Cai, H., Wei, T., Zhang, Y., et al. (2021). Deep learning for differential diagnosis of malignant hepatic tumors based on multiphase contrast-enhanced CT and clinical data. *J. Hematol. Oncol.* 14:154. doi: 10.1186/s13045-021-01167-2

Guo, W., Sun, Y. F., Shen, M. N., Ma, X. L., Wu, J., Zhang, C. Y., et al. (2018). Circulating tumor cells with stem-like phenotypes for diagnosis, prognosis, and therapeutic response evaluation in hepatocellular carcinoma. *Clin. Cancer Res.* 24, 2203–2213. doi: 10.1158/1078-0432.CCR-17-1753

Hamm, C. A., Wang, C. J., Savic, L. J., Ferrante, M., Schobert, I., Schlachter, T., et al. (2019). Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multiphasic MRI. *Eur. Radiol.* 29, 3338–3347. doi: 10.1007/s00330-019-06205-9

Jiang, Y., Zhang, M., Zhu, Y., and Zhu, D. (2021). Diagnostic role of contrast-enhanced ultrasonography versus conventional B-mode ultrasonography in cirrhotic patients with early hepatocellular carcinoma: a retrospective study. *J. Gastrointest. Oncol.* 12, 2403–2411. doi: 10.21037/jgo-21-611

Kim, J., Min, J. H., Kim, S. K., Shin, S. Y., and Lee, M. W. (2020). Detection of hepatocellular carcinoma in contrast-enhanced magnetic resonance imaging using deep learning classifier: a multi-center retrospective study. *Sci. Rep.* 10:9458. doi: 10.1038/s41598-020-65875-4

Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., van Stiphout, R. G. P. M., Granton, P., et al. (2012). Radiomics: extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* 48, 441–446. doi: 10.1016/j.ejca.2011.11.036

Liu, B., Li, J., Yang, X., Chen, F., Zhang, Y., and Li, H. (2023). Diagnosis of primary clear cell carcinoma of the liver based on faster region-based convolutional neural network. *Chin. Med. J.* 136, 2706–2711. doi: 10.1097/CM9.0000000000002853

Liu, J., Wu, J., Liu, A., Bai, Y., Zhang, H., and Yan, M. (2022). Preoperative diagnosis of hepatocellular carcinoma patients with bile duct tumor thrombus using the deep learning method. *JUSTC* 52:6. doi: 10.52396/JUSTC-2022-0057

Marrero, J. A., Kulik, L. M., Sirlin, C. B., Zhu, A. X., Finn, R. S., Abecassis, M. M., et al. (2018). Diagnosis, staging, and Management of Hepatocellular Carcinoma: 2018 practice guidance by the American Association for the Study of Liver Diseases. *Hepatology* 68, 723–750. doi: 10.1002/hep.29913

Mittal, S., and El-Serag, H. B. (2013). Epidemiology of hepatocellular carcinoma: consider the population. *J. Clin. Gastroenterol.* 47, S2–S6. doi: 10.1097/MCG.0b013e3182872f29

Moher, D. (2009). Preferred reporting items for systematic reviews and Meta-analyses: The PRISMA statement. *Ann. Intern. Med.* 151:264. doi: 10.7326/0003-4819-151-4-200908180-00135

Nishida, N., Yamakawa, M., Shiina, T., Mekada, Y., Nishida, M., Sakamoto, N., et al. (2022). Artificial intelligence (AI) models for the ultrasonographic diagnosis of liver tumors and comparison of diagnostic accuracies between AI and human experts. *J. Gastroenterol.* 57, 309–321. doi: 10.1007/s00535-022-01849-9

Ouzzani, M., Hammady, H., Fedorowicz, Z., and Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Syst. Rev.* 5:210. doi: 10.1186/s13643-016-0384-4

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2020). Statement: an updated guideline for reporting systematic reviews. *BMJ* 2021:n71. doi: 10.1136/bmj.n71

Quaglia, A. (2018). Hepatocellular carcinoma: a review of diagnostic challenges for the pathologist. *J. Hepatocell Carcin.* 5, 99–108. doi: 10.2147/JHC.S159808

R Core Team (2022). R: A language and environment for statistical computing (version 4.2.2). Vienna, Austria: R Foundation for Statistical Computing.

Radiuk, P. M. (2017). Impact of training set batch size on the performance of convolutional neural networks for diverse datasets. *Inf. Technol. Manag. Sci.* 20, 20–24. doi: 10.1515/itms-2017-0003

Raimundo, J. N. C., Fontes, J. P. P., Gonzaga Mendes Magalhães, L., and Guevara Lopez, M. A. (2023). An innovative faster R-CNN-based framework for breast Cancer detection in MRI. *J. Imaging.* 9:169. doi: 10.3390/jimaging9090169

RStudio Team (2022). RStudio: Integrated development environment for R. Boston, MA: RStudio, PBC.

Rumgay, H., Arnold, M., Ferlay, J., Lesi, O., Cabasag, C. J., Vignat, J., et al. (2022). Global burden of primary liver cancer in 2020 and predictions to 2040. *J. Hepatol.* 77, 1598–1606. doi: 10.1016/j.jhep.2022.08.021

Saba, L., Biswas, M., Kuppili, V., Cuadrado Godia, E., Suri, H. S., Edla, D. R., et al. (2019). The present and future of deep learning in radiology. *Eur. J. Radiol.* 114, 14–24. doi: 10.1016/j.ejrad.2019.02.038

Sounderajah, V., Ashrafian, H., Rose, S., Shah, N. H., Ghassemi, M., Golub, R., et al. (2021). A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. *Nat. Med.* 27, 1663–1665. doi: 10.1038/s41591-021-01517-0

Toyoda, H., Kumada, T., Tada, T., Kaneoka, Y., Maeda, A., Kanke, F., et al. (2011). Clinical utility of highly sensitive *Lens culinaris* agglutinin-reactive alpha-fetoprotein in hepatocellular carcinoma patients with alpha-fetoprotein <20 ng/mL. *Cancer Sci.* 102, 1025–1031. doi: 10.1111/j.1349-7006.2011.01875.x

Urhuț, M. C., Săndulescu, L. D., Streba, C. T., Mămuleanu, M., Ciocâlteu, A., Cazacu, S. M., et al. (2023). Diagnostic performance of an artificial intelligence model based on contrast-enhanced ultrasound in patients with liver lesions: a comparative study with clinicians. *Diagnostics* 13:3387. doi: 10.3390/diagnostics13213387

Wang, C. J., Hamm, C. A., Savic, L. J., Ferrante, M., Schobert, I., and Schlachter, T. (2019). Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. *Eur. Radiol.* 29, 3348–3357.

Wang, Z., Qin, H., Liu, S., Sheng, J., and Zhang, X. (2023). Precision diagnosis of hepatocellular carcinoma. *Chin. Med. J.* 136, 1155–1165. doi: 10.1097/CM9.0000000000002641

West, S. L., Gartlehner, G., Mansfield, A. J., Poole, C., and Tant, E. (2010). Comparative Effectiveness review Methods: Clinical Heterogeneity [Internet].

Yamashita, R., Mittendorf, A., Zhu, Z., Fowler, K. J., Santillan, C. S., Sirlin, C. B., et al. (2020). Deep convolutional neural network applied to the liver imaging reporting and data system (LI-RADS) version 2014 category classification: a pilot study. *Abdom. Radiol.* 45, 24–35. doi: 10.1007/s00261-019-02306-7

Yamashita, R., Nishio, M., Do, R. K. G., and Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights Imaging* 9, 611–629. doi: 10.1007/s13244-018-0639-9

Yasaka, K., Akai, H., Abe, O., and Kiryu, S. (2018). Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study. *Radiology* 286, 887–896. doi: 10.1148/radiol.2017170706

Yu, M. H., Kim, J. H., Yoon, J. H., Kim, H. C., Chung, J. W., Han, J. K., et al. (2014). Small (≤1-cm) hepatocellular carcinoma: diagnostic performance and imaging features at Gadoxetic acid–enhanced MR imaging. *Radiology* 271, 748–760. doi: 10.1148/radiol.14131996

Zhang, J., Yu, Y., Li, Y., and Wei, L. (2017). Diagnostic value of contrast-enhanced ultrasound in hepatocellular carcinoma: a meta-analysis with evidence from 1998 to 2016. *Oncotarget* 8, 75418–75426. doi: 10.18632/oncotarget.20049

Zhen, S. H., Cheng, M., Bo, T. Y., Fan, W. Y., Juengpanich, S., Yu, J. Z., et al. (2020). Deep learning for accurate diagnosis of liver tumor based on magnetic resonance imaging and clinical data. *Front. Oncol.* 10:680. doi: 10.3389/fonc.2020.00680

Glossary

HCC	Hepatocellular Carcinoma
AFP	Alpha-fetoprotein
DCP	Des-gamma-Carboxy Prothrombin
GP73	Golgi Protein 73
GPC3	Glypican-3
Ct-ncRNA	Circulating tumor noncoding RNA
CfDNA	Cell-free DNA
CtDNA	Circulating tumor DNA
CTC	Circulating tumor cell
FNA	Fine Needle Aspiration
WSI	Whole-Slide Imaging
CT	Computed Tomography
MRI	Magnetic Resonance Imaging
CNN	Convolutional Neural Network
AI	Artificial Intelligence
DL	Deep Learning
VGGNet	Visual Geometry Group Network
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analysis
RCNN	Region-based Convolutional Neural Network
CECT	Contrast-Enhanced Computed Tomography
PCCCL	Primary Clear Cell Carcinoma
CHCC	Common Hepatocellular Carcinoma
CEUS	Contrast-enhanced ultrasound