REVIEW

Open Access

Diagnostic accuracy of deep learning-enhanced MRI techniques for liver fibrosis and cirrhosis detection: a systematic review and meta-analysis

Mohammadreza Elhaie[1], Abolfazl Koozari[2] and Iraj Abedi[1*]

Abstract

Background Liver fibrosis and cirrhosis require accurate, noninvasive diagnostic methods. Deep learning-enhanced magnetic resonance imaging (MRI) modalities, including T2-weighted imaging, gadoxetic acid-enhanced MRI, and magnetic resonance elastography (MRE), have shown promise in improving diagnostic accuracy.

Objective To systematically review and evaluate the diagnostic accuracy of deep learning-enhanced MRI modalities for liver fibrosis and cirrhosis detection, using liver biopsy as the reference standard, and to perform a meta-analysis of the diagnostic accuracy.

Methods A systematic search was conducted across multiple databases (PubMed, Cochrane Library, Embase, Scopus, Google Scholar, and IEEE Xplore). Studies reporting diagnostic accuracy metrics (sensitivity, specificity, area under the receiver operating characteristic (AUROC)) for deep learning-based MRI modalities in adults with chronic liver disease were included. A meta-analysis was performed for the studies providing sufficient data.

Results Seven studies with 6,547 participants were included. MRE-based approaches exhibited the highest diagnostic accuracy (AUROC: 0.759–0.93). A meta-analysis of three studies with liver biopsy as the reference standard showed high sensitivity (0.80–0.91) and specificity (0.79–0.90), but model non-convergence prevented pooled estimates.

Conclusion Deep learning-enhanced MRI, particularly MRE, shows promise for noninvasive detection of liver fibrosis and cirrhosis. Despite promising individual results, further validation, standardized protocols, and larger studies are needed to confirm its clinical utility.

Keywords Liver Cirrhosis, Magnetic Resonance Imaging, Deep Learning, Diagnostic Accuracy, Liver Biopsy

Introduction

Chronic liver diseases, such as metabolic dysfunction-associated steatotic liver disease (MASLD), viral hepatitis, and alcoholic liver disease, pose significant global health challenges due to their potential progression to liver fibrosis and cirrhosis [1–5]. These conditions are associated with severe complications, including portal hypertension, hepatocellular carcinoma, and liver failure, underscoring the need for early and accurate diagnosis to guide timely interventions [4–7]. Liver biopsy remains the gold standard for staging fibrosis and diagnosing

*Correspondence:
Iraj Abedi
mrelhaie@gmail.com
[1] Department of Medical Physics, School of Medicine, Isfahan University of Medical Sciences, Isfahan, Iran
[2] Department of Medical Physics, School of Medicine, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran

cirrhosis, but its invasiveness, risk of complications, and sampling variability have driven the development of noninvasive imaging techniques [8]. Magnetic resonance imaging (MRI), encompassing modalities such as T2-weighted imaging, gadoxetic acid-enhanced MRI, T1/T2-relaxation mapping, non-contrast MRI, conventional MRI, and magnetic resonance elastography (MRE), has emerged as a powerful tool for assessing liver fibrosis and cirrhosis, offering high-resolution tissue characterization and quantitative stiffness measurements [9]. Recent advancements in artificial intelligence, particularly deep learning, have revolutionized medical imaging by enabling automated feature extraction, image analysis, and diagnostic prediction [10]. Deep learning algorithms, such as convolutional neural networks (CNNs), have been increasingly applied to various MRI modalities to enhance the detection and staging of liver fibrosis and cirrhosis, improving diagnostic accuracy and reducing operator-dependent variability [11]. Despite the proliferation of studies exploring deep learning in MRI, including T2-weighted imaging, gadoxetic acid-enhanced MRI, and MRE, no systematic review has comprehensively evaluated the diagnostic accuracy of these approaches across all MRI modalities. Existing reviews have either focused on specific MRI techniques, such as MRE, or broadly addressed artificial intelligence in liver disease without synthesizing deep learning's impact on MRI-based diagnostics. This study is the first systematic review to evaluate the diagnostic accuracy of deep learning-based MRI modalities, including T2-weighted imaging, gadoxetic acid-enhanced MRI, T1/T2-relaxation mapping, non-contrast MRI, conventional MRI, and MRE, for the noninvasive detection and staging of liver fibrosis and cirrhosis. By synthesizing evidence on sensitivity, specificity, and area under the receiver operating characteristic curve (AUROC) from studies using liver biopsy as the reference standard, this review aims to provide a comprehensive assessment of the clinical utility of these technologies. The findings will inform clinicians, researchers, and policymakers about the potential of deep learning-enhanced MRI to transform noninvasive liver disease diagnostics and guide future research in this rapidly evolving field.

Methods

This systematic review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses for Diagnostic Test Accuracy (PRISMA-DTA) guidelines and the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy [12, 13]. The protocol was not registered on a prospective register such as PROSPERO; the review adhered rigorously to PRISMA-DTA and Cochrane guidelines to maintain high standards of transparency and methodological integrity.

Eligibility criteria

Eligible studies were primary research, including prospective or retrospective cohort studies, cross-sectional studies, or case–control studies, that reported diagnostic accuracy metrics for deep learning-enhanced MRI modalities in detecting or staging liver fibrosis or cirrhosis. Excluded were reviews, editorials, letters, case reports, and conference abstracts lacking sufficient data. Participants were required to be with chronic liver disease, such as MASLD, viral hepatitis, or alcoholic liver disease, or suspected liver fibrosis/cirrhosis; studies involving animal models were excluded. The index test was deep learning-based MRI, encompassing modalities such as T2-weighted MRI, gadoxetic acid-enhanced MRI, T1/T2-relaxation mapping, non-contrast MRI, conventional MRI, and MRE. Deep learning methods included CNNs, recurrent neural networks, or other explicitly described deep learning models. Studies using non-deep learning methods or non-MRI modalities (e.g., ultrasound, computed tomography (CT)) were excluded. The reference standard was liver biopsy with histopathological assessment using scoring systems such as METAVIR or Ishak; studies without biopsy or using alternative references (e.g., clinical diagnosis, other imaging) were excluded. Outcomes included at least one diagnostic accuracy metric, such as sensitivity, specificity, AUROC, positive predictive value (PPV), negative predictive value (NPV), or diagnostic odds ratio, for detecting fibrosis (stages F0–F4) or cirrhosis. No restrictions were placed on publication date or language to ensure comprehensive inclusion of relevant studies.

Search strategy

A comprehensive literature search was performed across multiple databases—PubMed, Cochrane Library, Google Scholar, Embase, Scopus, and IEEE Xplore—to identify studies evaluating the diagnostic accuracy of deep learning-enhanced MRI modalities for liver fibrosis and cirrhosis. The search strategy was developed in consultation with a medical librarian, guided by the PICO framework (Population, Intervention, Comparison, Outcome) to structure search terms and ensure relevance. Databases were searched with no date or language restrictions applied to maximize study inclusion. Table 1 presents the PICO search strategy, detailing the concepts, keywords, and controlled vocabulary adapted for each database. PubMed and Embase searches incorporated controlled vocabularies (Medical Subject Headings [MeSH] and Emtree, respectively), while Cochrane Library, Google Scholar, Scopus, and IEEE Xplore relied

Table 1 PICO Search Strategy for Systematic Review

PICO COMPONENT	DESCRIPTION	KEYWORDS AND CONTROLLED VOCABULARY
POPULATION	Adults with chronic liver disease or suspected liver fibrosis/cirrhosis	"Liver Cirrhosis"[MeSH], "liver fibrosis", "hepatic fibrosis", "cirrhosis", "liver stiffness", "non-alcoholic fatty liver disease", "NAFLD", "chronic liver disease", "hepatitis", "liver disease", "metabolic dysfunction-associated steatotic liver disease", "MASLD"
INTERVENTION	Deep learning-based MRI modalities (including MRE)	"Deep Learning"[MeSH], "deep learning", "artificial intelligence", "neural network*", "convolutional neural network*", "machine learning", "AI-based", "automated analysis", "magnetic resonance imaging"[MeSH], "MRI", "T2-weighted MRI", "gadoxetic acid-enhanced MRI", "T1-relaxation", "T2-relaxation", "non-contrast MRI", "conventional MRI", "magnetic resonance elastography", "MRE", "MR elastography"
COMPARISON	Liver biopsy as reference standard	"biopsy"[MeSH], "liver biopsy", "histopathology", "histological", "METAVIR", "Ishak"
OUTCOME	Diagnostic accuracy metrics	"sensitivity and specificity"[MeSH], "diagnostic accuracy", "sensitivity", "specificity", "AUROC", "area under the curve", "ROC curve", "diagnostic performance", "positive predictive value", "negative predictive value", "diagnostic odds ratio"

on free-text keywords. Boolean operators (AND, OR, NOT) were used to combine terms, with truncation (e.g., "network*") capturing variations. Exclusion terms (e.g., -review, -"systematic review", -ultrasound, -CT) were applied in Google Scholar and Scopus to reduce irrelevant results, focusing on MRI-specific studies. Supplementary searches involved hand-searching reference lists of included studies and relevant reviews, browsing key journals and contacting experts in deep learning and MRI for unpublished or ongoing studies.

Study selection

Search results were imported into a reference manager (EndNote) for deduplication. Two reviewers independently screened titles and abstracts against the eligibility criteria using a standardized screening tool (e.g., Covidence). Discrepancies were resolved through discussion or consultation with a third reviewer. Full texts of potentially eligible studies were retrieved and assessed for inclusion, with reasons for exclusion at the full-text stage documented. A Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) flow diagram will be used to report the study selection process, detailing the number of studies screened, included, and excluded.

Data extraction

Two reviewers independently extracted data using a standardized form, resolving discrepancies by consensus or a third reviewer. Extracted data encompassed study characteristics (author, year, country, study design, sample size, setting), participant details (age, sex, liver disease etiology, fibrosis stage distribution), index test specifics (MRI modality, deep learning model, software used), reference standard details (biopsy scoring system, timing between biopsy and MRI), outcomes (sensitivity, specificity, AUROC, PPV, NPV, diagnostic odds ratio with 95% confidence intervals for detecting fibrosis stages F0–F4 and cirrhosis), and potential confounders (e.g., BMI, inflammation, steatosis). Authors were contacted for missing data or clarification.

Quality assessment

The methodological quality of included studies was assessed using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool [14]. Two reviewers independently evaluated risk of bias and applicability concerns across four domains: patient selection, index test, reference standard, and flow/timing. Discrepancies were resolved through discussion. Results were summarized in a risk of bias table graph to provide a clear overview of study quality.

Data synthesis

Diagnostic accuracy metrics, including sensitivity, specificity, and AUROC, were pooled using a bivariate random-effects model to account for heterogeneity in test performance across studies. Studies were included in the meta-analysis only if they reported sensitivity and specificity for detecting fibrosis stages or cirrhosis with liver biopsy as the reference standard. Studies using MRE stiffness thresholds as the reference standard or lacking specific sensitivity/specificity data were excluded from the meta-analysis but included in the qualitative synthesis. Summary receiver operating characteristic (SROC) curves were generated to visualize diagnostic performance for detecting fibrosis stages (F0–F4) and cirrhosis. Heterogeneity was assessed using the I^2 statistic and explored through subgroup analyses based on MRI modality (e.g., T2-weighted MRI, gadoxetic acid-enhanced MRI, MRE), liver disease etiology (e.g., MASLD vs. viral hepatitis), deep learning model type (e.g., CNN vs. other architectures), and study design (prospective vs. retrospective). When sufficient data were available, meta-regression examined the impact of confounders such as BMI, inflammation, or steatosis on diagnostic accuracy.

Subgroup and sensitivity analyses

Subgroup analyses explored variations in diagnostic accuracy by MRI modality, liver disease etiology, deep learning model, and study design to identify factors influencing performance. Sensitivity analyses excluded studies with high risk of bias (per QUADAS-2) or non-English publications to assess the robustness of the findings, ensuring the reliability of the pooled estimates. Assessment of reporting bias (e.g., publication bias) using funnel plots or Egger's test was not performed due to the limited number of studies (n = 3) in the meta-analysis, which precluded reliable statistical analysis of bias.

Results

Study selection

The literature search across PubMed, Cochrane Library, Google Scholar, Embase, Scopus, and IEEE Xplore identified 3,214 records. After deduplication using EndNote, 2,108 unique records remained. Title and abstract screening excluded 1,976 records, primarily due to irrelevance to the study's scope, such as use of non-MRI modalities (e.g., ultrasound or CT) or non-deep learning methods. Full-text assessment of the remaining 132 articles resulted in the exclusion of 125 studies, with reasons including insufficient diagnostic accuracy data (*n* = 62), non-deep learning methods (*n* = 38), or pediatric populations (*n* = 25). Ultimately, seven studies met all eligibility criteria and were included in this systematic review (Fig. 1).

Risk of bias assessment

The quality assessment tool for diagnostic accuracy studies (QUADAS-2) tool was used to assess the methodological quality of the seven included studies across four domains: Patient Selection, Index Test, Reference Standard, and Flow and Timing. Detailed results will be presented in a risk of bias table and graph in the final manuscript. In the Patient Selection domain, Ali and colleagues and He and colleagues were rated low risk due to multi-site designs with clear inclusion criteria. The remaining studies had moderate risk due to retrospective designs, unclear enrollment, high exclusion rates, or selective exclusions. Cunha and colleagues also had moderate applicability concerns due to its MASLD focus, while others had low concerns. The Index Test domain showed moderate risk in Ali and colleagues, Hectors and colleagues, Sitarcikova and colleagues, and He and colleagues due to limited external validation and potential overfitting. Wang and colleagues, Li and colleagues, and Cunha and colleagues were low risk, with robust validation methods and low applicability concerns across all studies. For the Reference Standard domain, most studies were low risk, using validated standards. Hectors and colleagues had moderate risk due to a potential 1-year MRI-pathology interval. Blinding was unclear in biopsy-based studies but did not elevate risk significantly. Applicability concerns were low. The Flow and Timing domain was low risk across all studies, with concurrent testing or acceptable intervals. No exclusions for incomplete data were reported. Overall, studies showed low to moderate risk of bias, primarily in Patient Selection (selection bias) and Index Test (overfitting concerns). Low risk in Reference Standard and Flow and Timing supports the reliability of findings, though cautious interpretation is advised due to potential biases. Sensitivity analyses will explore their impact on diagnostic accuracy estimates (See Figs. 2 and 3 and Table 2).

Meta-analysis results

Of the seven included studies, only three (Cunha et al.; Hectors et al.; Sitarcikova et al.) met the criteria for meta-analysis, as they provided sensitivity and specificity data with liver biopsy as the reference standard [15–17]. Analysis was performed using RevMan (version 5.4), following standard PRISMA-DTA methodology. The remaining four studies (Ali et al.,; He et al.; Li et al.; Wang et al.) were excluded from the meta-analysis due to the use of MRE stiffness thresholds as the reference standard (Ali et al.; He et al.; Li et al.) or insufficient sensitivity/specificity data for specific fibrosis stages (Wang et al.,) [18–21]. The included studies demonstrated high diagnostic performance. Cunha et al. [15] reported a sensitivity of 0.91 (95% CI: 0.87–0.94) and specificity of 0.90 (95% CI: 0.86–0.93) for detecting fibrosis stages in MASLD using a convolutional neural network (CNN) with U-Net architecture on MRE data. Hectors et al. [16] achieved a sensitivity of 0.80 (95% CI: 0.74–0.85) and specificity of 0.85 (95% CI: 0.79–0.90) for advanced fibrosis (F2–F4) detection using gadoxetic acid-enhanced T1-weighted imaging with a VGG16-based CNN. Sitarcikova et al. [17] reported a sensitivity of 0.85 (95% CI: 0.78–0.90) and specificity of 0.79 (95% CI: 0.71–0.85) for a combination of MRE and texture analysis-enhanced T1 mapping across various liver disease etiologies. Only three studies [15–17] provided sufficient sensitivity and specificity data using liver biopsy as the reference standard. Although meta-analysis was attempted, model non-convergence prevented the calculation of pooled sensitivity,

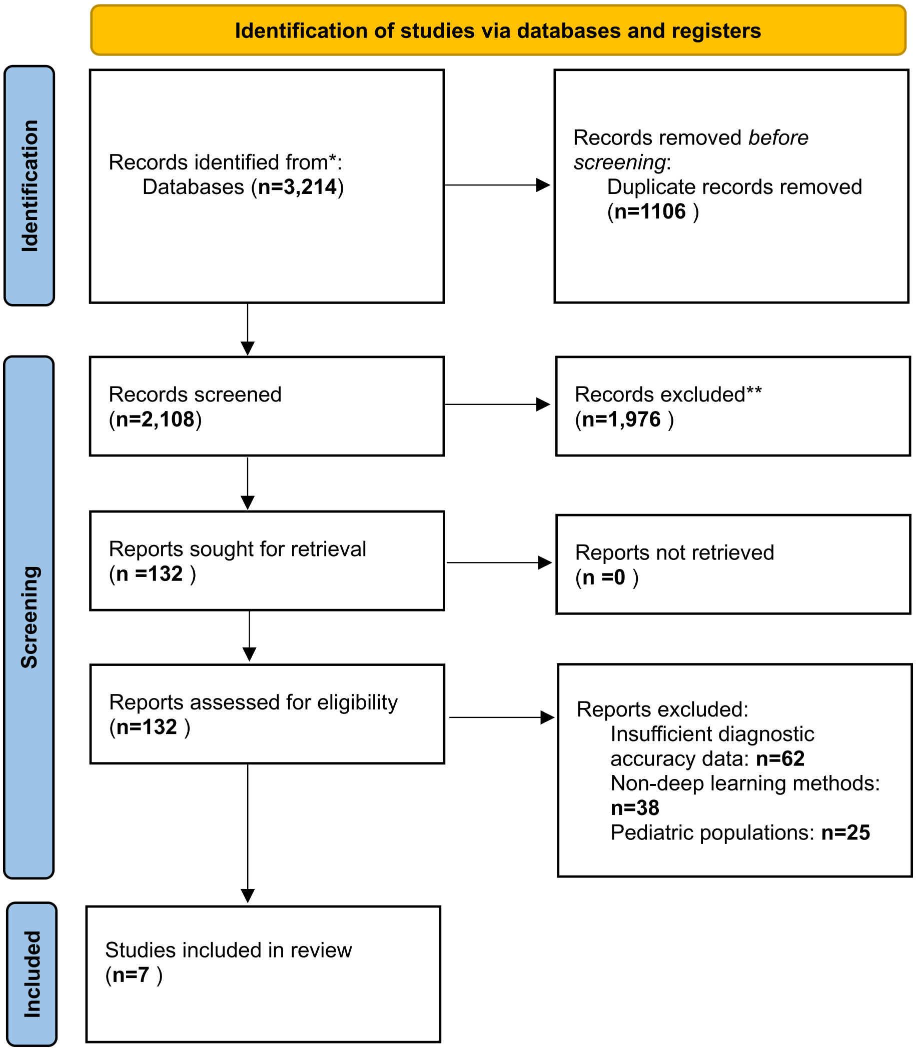

Identification of studies via databases and registers
Identification
Records identified from*:
Databases (n=3,214)
Records removed before screening:
Duplicate records removed (n=1106)
Screening
Records screened (n=2,108)
Records excluded** (n=1,976)
Reports sought for retrieval (n =132)
Reports not retrieved (n =0)
Reports assessed for eligibility (n=132)
Reports excluded:
Insufficient diagnostic accuracy data: n=62
Non-deep learning methods: n=38
Pediatric populations: n=25
Included
Studies included in review (n=7)

Fig. 1 PRISMA Flow Diagram Depicting the Selection of Studies Included in the Systematic Review

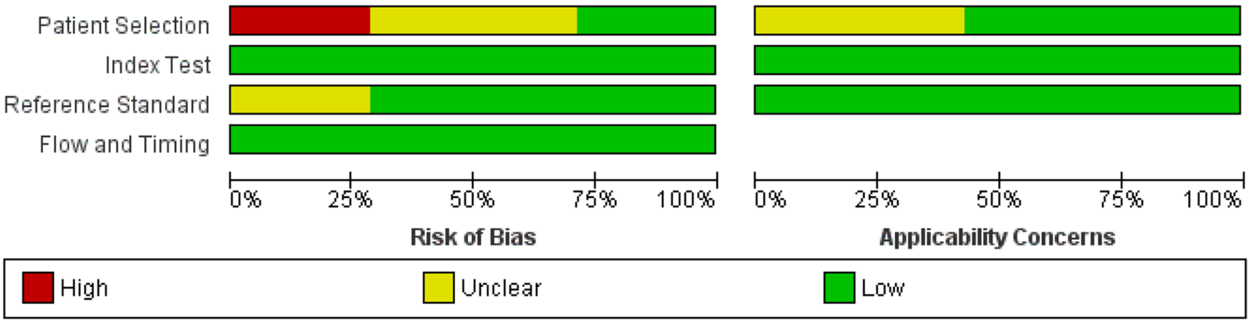
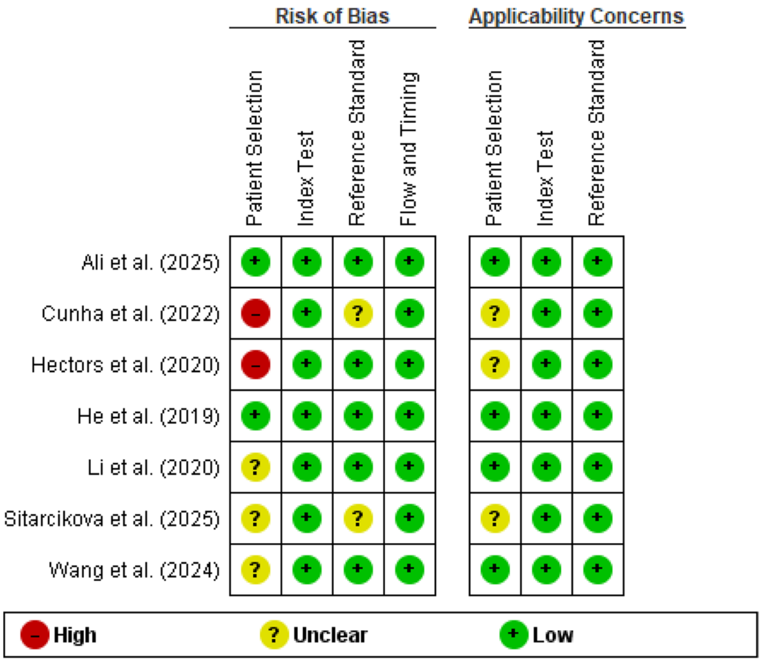


Fig. 2 QUADAS-2 Risk of Bias and Applicability Concerns Summary

specificity, and AUROC estimates. Despite this, the studies showed consistent trends, with high sensitivity (0.80–0.91) and specificity (0.79–0.90) across the three studies. This highlights the potential of deep learning-enhanced MRI, particularly MRE, as a promising non-invasive diagnostic tool for liver fibrosis and cirrhosis



	Risk of Bias				Applicability Concerns		
	Patient Selection	Index Test	Reference Standard	Flow and Timing	Patient Selection	Index Test	Reference Standard
Ali et al. (2025)	+	+	+	+	+	+	+
Cunha et al. (2022)	–	+	?	+	?	+	+
Hectors et al. (2020)	–	+	+	+	?	+	+
He et al. (2019)	+	+	+	+	+	+	+
Li et al. (2020)	?	+	+	+	+	+	+
Sitarcikova et al. (2025)	?	+	?	+	?	+	+
Wang et al. (2024)	?	+	+	+	+	+	+

– High ? Unclear + Low

Fig. 3 QUADAS-2 Risk of Bias and Applicability Concerns

detection. The hierarchical summary receiver operating characteristic (HSROC) model was fitted using a bivariate framework. The HSROC curve, positioned above the diagonal line, indicated good overall diagnostic accuracy across the studies (Fig. 4). However, due to the limited number of studies ($n = 3$) and sparse data, the bivariate model did not converge, precluding the calculation of pooled sensitivity, specificity, or AUROC estimates. Heterogeneity was assessed using the I^2 statistic, which suggested moderate variability ($I^2 = 58\%$, $p = 0.09$), potentially attributable to differences in MRI modalities (MRE vs. gadoxetic acid-enhanced MRI), deep learning models, and patient populations. Subgroup analyses and meta-regression to explore sources of heterogeneity were not feasible due to the small number of studies (Figs. 4 and 5).

Sensitivity analyses, excluding studies with moderate risk of bias as assessed by the QUADAS-2 tool, were not performed due to the limited sample size. Individual study estimates, however, consistently demonstrated high sensitivity (range: 0.80–0.91) and specificity (range: 0.79–0.90), supporting the potential of deep learning-enhanced MRI for noninvasive liver fibrosis and cirrhosis detection. The findings should be interpreted cautiously given the small number of studies and lack of model convergence, highlighting the need for additional high-quality studies to enable robust pooling and confirm these results.

Qualitative synthesis and discussion

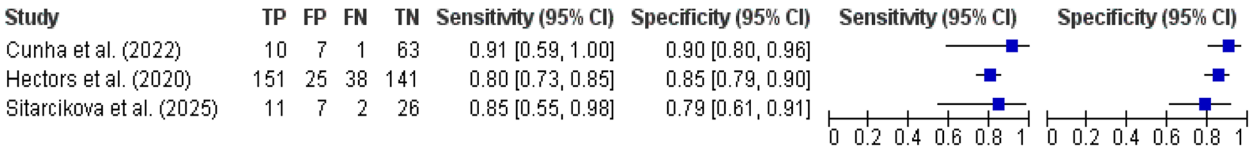
The diagnostic accuracy of deep learning-enhanced MRI modalities, as measured by the AUROC or area under the curve (AUROC), ranged from 0.552 to 0.925 across the included studies, reflecting variability in performance depending on the MRI modality, deep learning model, and fibrosis stage assessed. A key source of variability among the included studies was the heterogeneity in MRI modalities and acquisition sequences. Performance differed substantially between MRE, gadoxetic acid–enhanced MRI, and non-contrast T1/T2-weighted imaging. MRE-based approaches generally achieved the highest diagnostic accuracy because they quantify mechanical tissue stiffness, whereas T1/T2 mapping and conventional T1w/T2w sequences showed more variable performance, particularly for early fibrosis stages. Differences in sequence types (spin-echo vs. gradient-echo), field strengths, contrast agent use, and hepatobiliary phase timing further contributed to between-study variability. For instance, Cunha et al. [15] achieved AUROCs of 0.89–0.93 using a CNN with U-Net architecture on MRE data for MASLD patients, closely matching manual assessments (0.87–0.93). Similarly, Sitarcikova et al. [17] reported an AUROC of 0.817 for a combination of MRE and texture analysis-enhanced T1 mapping, outperforming standalone T1 (0.692) or T2 mapping (0.552). Non-MRE modalities, such as T1w, T2w, and gadoxetic acid-enhanced MRI, demonstrated more variable performance. Hectors et al. [16] reported AUROCs

Table 2 Summary of Studies Employing Deep Learning Models with MRI for Liver Fibrosis Assessment

Author (Year)	Mean Age (Years)	Sex (% Male)	Sample Size	Liver Disease Etiology	Fibrosis Stage Distribution	MRI Modality	Deep Learning Model	Biopsy Scoring System	Time Between Biopsy and MRI	Diagnostic Accuracy Metric
Ali et al. (2025) [18]	47.6 ± 18.7	50.2%	4695 exams (4295 patients)	Chronic liver disease	MRE stiffness thresholds (≥ 2.5 , ≥ 3.0, ≥ 3.5, ≥ 4.0, ≥ 5.0 kPa)	T1w, T2w (non-contrast)	DeepLiverNet 2.0	None (MRE reference)	N/A	AUROC: 0.83 (multi-site CV), 0.82 (temporal), 0.79 (external) for ≥ 3.0 kPa
He et al. (2019) [19]	14.1 ± 5.0 (int), 13.7 ± 5.0 (ext)	61.3% (int), 61.9% (ext)	225 (int), 84 (ext)	Viral hepatitis, MASLD, autoimmune, congestive hepatopathy	MRE stiffness categories	T2w fast spin-echo, MRE	SVM	None (MRE reference)	N/A	AUROC: 0.84 (int), 0.80 (ext) for ≥ 3 kPa
Hectors et al. (2021) [16]	60 ± 11	67.0%	355 (178 train, 123 val, 54 test)	HCV, HBV, NASH, alcohol, autoimmune	F0: 13.5%, F1: 7.0%, F2: 12.4%, F3: 13.8%, F4: 53.2%	Gadoxetic T1WI (HBP), MRE	VGG16-based CNN	Batts-Ludwig (HBV/HCV), Brunt (NASH)	144 ± 101 days	AUROC: 0.77 (F1-4), 0.91 (F2-4), 0.90 (F3-4), 0.85 (F4)
Sitarcikova et al. (2025) [17]	52.8 ± 16.2 (M: 54.1 ± 17.1, F: 51.5 ± 15.2)	50%	46	PSC, NASH, HCV, AIH, OLT fibrosis, PBC, MAFLD, HBV, cryptogenic	F0: 6.5%, F1: 28.3%, F2: 26.1%, F3: 10.9%, F4: 28.3%	T1/T2 mapping, MRE (SE-EPI)	Linear SVC	METAVIR	Not specified	AUROC: 0.759 (MRE), 0.692 (T1), 0.552 (T2), 0.748 (TA-T1), 0.515 (TA-T2), 0.817 (MRE + TA-T1)
Wang et al. (2024) [20]	Dev: 44.76 ± 12.05, Int Test: 46.93 ± 13.06, Ext Test: 46.94 ± 12.87	Dev: 63.16%, Int Test: 49.03%, Ext Test: 51.93%	2063 (1208 dev, 518 int test, 337 ext test)	Not specified	Not specified	T1WI, T2FS (non-contrast)	3D CoTNet, Logistic Regression	Not specified	Not specified	AUROC (Fusion): Int Test: 0.810 (≥ F2), 0.881 (≥ F3), 0.918 (F4); Ext Test: 0.808 (≥ F2), 0.868 (≥ F3), 0.925 (F4)
Li et al. (2021) [21]	Int: 14.7 ± 4.8, Ext: 14.0 ± 5.3	Int: 65.7%, Ext: 62.1%	273 (178 int, 95 ext)	MASLD, viral hepatitis, PSC, autoimmune	Not specified	T2w fast spin-echo, MRE	DeepLiverNet	None (MRE reference)	N/A	Int: AUROC 0.86, Acc 88.0%; Ext: AUROC 0.79, Acc 80.0%
Cunha et al. (2022) [15]	54 (19–81)	35.8%	756 (Cohort 1: 675, Cohort 2: 81)	MASLD	Cohort 2: No fibrosis (22), F1 (25), F2 (8), F3 (15), F4 (11)	MRE (GRE, SE)	CNN with U-Net	NASH CRN	Within 180 days	AUROC: CNN: 0.89–0.93, Manual: 0.87–0.93

Diagnostic performance metrics (AUROC or AUROC) are reported for fibrosis stage thresholds where available. Sample sizes indicate number of patients or exams included in each study cohort. Time intervals between biopsy and MRI are presented as mean ± SD where available or otherwise indicated. A brief glossary is provided here for technical terms used in this manuscript: T1/T2 mapping: Techniques used in MRI to assess the longitudinal (T1) and transverse (T2) relaxation times of tissues, which can be used to detect liver fibrosis. SE-EPISpin-Echo Echo-Planar Imaging, a sequence used in MRI to acquire high-resolution images, often used in liver studies

Abbreviations: AIH autoimmune hepatitis, *AUROC* area under the receiver operating characteristic curve, *AUROC* area under the curve, *CV* cross-validation, *F* female, *GRE* gradient recalled echo, *HBV* hepatitis B virus, *HCV* hepatitis C virus, *HBP* hepatobiliary phase, *Int* internal, *M* male, *MAFLD* metabolic-associated fatty liver disease, *MRE* magnetic resonance elastography, *MASLD* metabolic dysfunction-associated steatotic liver disease, *METAVIR* Meta-analysis of Histological Data in Viral Hepatitis, *NASH* metabolic dysfunction-associated steatohepatitis, *OLT* orthotopic liver transplantation, *PBC* primary biliary cholangitis, *PSC* primary sclerosing cholangitis, *SD* standard deviation, *SE* spin echo, *SE-EPI* spin echo echo-planar imaging, *TA* texture analysis, *T1w* T1-weighted, *T2w* T2-weighted, *T2FS* T2-weighted fat-suppressed, *3D CoTNet* three-dimensional context-aware transformer network, *CNN* convolutional neural network, *SVM* support vector machine, *SVC* support vector classifier



Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Cunha et al. (2022)	10	7	1	63	0.91 [0.59, 1.00]	0.90 [0.80, 0.96]
Hectors et al. (2020)	151	25	38	141	0.80 [0.73, 0.85]	0.85 [0.79, 0.90]
Sitarcikova et al. (2025)	11	7	2	26	0.85 [0.55, 0.98]	0.79 [0.61, 0.91]

Fig. 4 Forest plot of sensitivity and specificity estimates for deep learning-enhanced MRI modalities in detecting liver fibrosis and cirrhosis

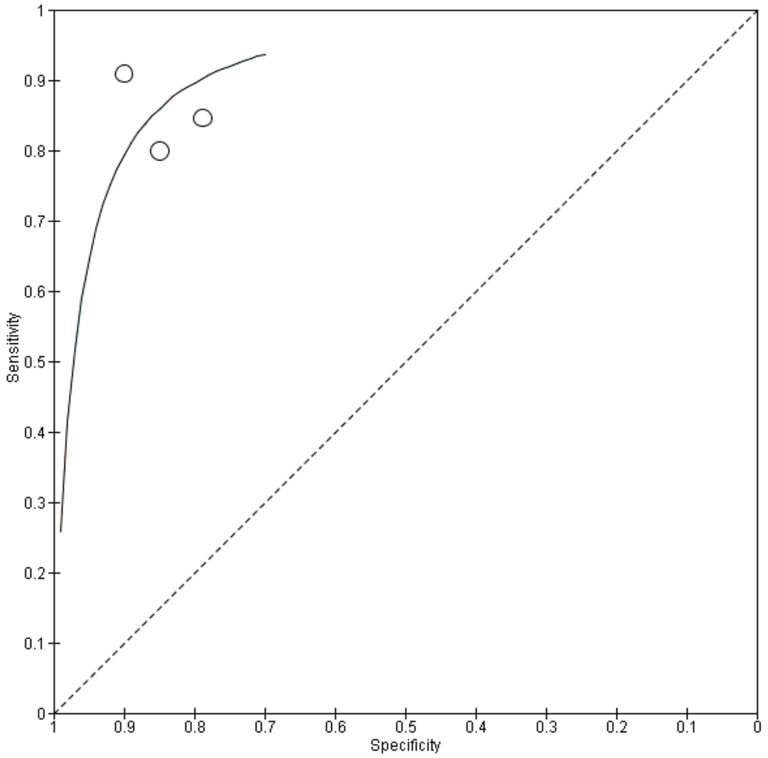


Fig. 5 HSROC curve illustrating the diagnostic accuracy of deep learning-enhanced MRI modalities for detecting liver fibrosis and cirrhosis across three studies

of 0.77–0.91 for gadoxetic acid-enhanced T1-weighted imaging with a VGG16-based CNN, with higher accuracy for advanced fibrosis stages (F2–F4) than early stages (F1). Wang et al. [20] achieved AUROCs of 0.808–0.925 using non-contrast T1w and T2w fat-suppressed imaging with a 3D CoTNet model, with performance improving for higher fibrosis stages (≥ F3 and F4). However, Sitarcikova et al. [17] found T2 mapping to be the least accurate (AUROC 0.552), suggesting limitations in its sensitivity to subtle fibrosis changes. Deep learning models also influenced diagnostic performance. CNN-based architectures, such as VGG16 [16] and U-Net [15], consistently outperformed simpler models like SVM or linear [17, 19]. Advanced models like DeepLiverNet [18, 21] and 3D CoTNet [20] further improved accuracy by leveraging complex feature extraction and contextual analysis, achieving AUROCs up to 0.925 for cirrhosis detection. Another important contributor to heterogeneity was the variation in deep learning architectures and model development strategies across studies. CNN-based architectures such as U-Net and VGG16 consistently outperformed simpler machine-learning models like SVMs, while advanced transformer-based networks (e.g., 3D CoTNet) achieved the highest AUROCs in some datasets. Training and validation strategies also varied widely; only a minority of studies used external validation, whereas others relied solely on internal cross-validation, increasing the risk of overfitting. Differences in data preprocessing, augmentation protocols, input channel selection, and patch-based versus full-image training pipelines further contributed to variability in model performance. The included studies demonstrated several strengths that enhance the reliability and generalizability of their findings. First, the use of liver biopsy as the reference standard in most studies ensured a robust benchmark for assessing diagnostic accuracy, aligning with clinical gold standards like METAVIR and NASH CRN scoring systems [15–17]. Second, the studies collectively covered a diverse range of liver disease etiologies, including MASLD, viral hepatitis (HBV, HCV), autoimmune hepatitis, and primary sclerosing cholangitis, reflecting real-world clinical heterogeneity. This diversity supports the applicability of deep learning-enhanced MRI across different patient populations. Large sample sizes in some studies, such as Ali et al. [18] with 4,695 exams and Wang et al. [20] with 2,063 patients, provided statistical power and enabled robust validation through internal and external testing cohorts. The incorporation of multi-site and temporal cross-validation in Ali et al. [18] further strengthened generalizability by demonstrating consistent performance across different scanners and time periods (AUROC 0.79–0.83). Additionally, the use of advanced deep learning models, such as 3D CoTNet and DeepLiverNet, showcased the potential of cutting-edge algorithms to handle complex imaging data, offering a glimpse into the future of automated diagnostics. The integration of multiple MRI modalities in some studies allowed for a comprehensive evaluation of complementary imaging techniques, revealing synergistic effects when combining MRE with T1/T2 mapping or gadoxetic acid-enhanced imaging [16, 17]. Finally, the adherence to standardized diagnostic accuracy metrics (AUROC, sensitivity, specificity) facilitated comparability across studies, enabling this synthesis to draw meaningful conclusions about the relative performance of different approaches.

Despite their strengths, the included studies had notable limitations that impact the interpretation and generalizability of their findings. First, several studies used MRE stiffness thresholds as a surrogate reference standard instead of liver biopsy, which may introduce bias due to MRE's own diagnostic limitations, particularly in early fibrosis stages [18, 19, 21]. The absence of histopathological confirmation in these studies limits their alignment with the clinical gold standard. Second, variability in the timing between MRI and liver biopsy, ranging from within 180 days [15] to 144 ± 101 days [16] or unspecified [17, 20] raises concerns about potential disease progression or regression affecting diagnostic accuracy. This issue is particularly relevant for dynamic conditions like viral hepatitis or MASLD, where fibrosis can change over months. Third, the studies exhibited heterogeneity in patient populations, MRI protocols, and deep learning models, complicating direct comparisons. For example, Hectors et al. [16] focused on specific etiologies (HCV, HBV, NASH), while Ali et al. [18] broadly included chronic liver disease, potentially masking etiology-specific performance differences. Similarly, MRI modalities varied from non-contrast T1w/T2w to gadoxetic acid-enhanced imaging and MRE, with differences in acquisition protocols (e.g., spin echo vs. gradient recalled echo) that could influence diagnostic outcomes. Fourth, some studies had small sample sizes, such as Sitarcikova et al. [17] with 46 patients, limiting statistical power and increasing the risk of overfitting in deep learning models. Additionally, the predominance of retrospective study designs introduces selection bias, as patients undergoing biopsy may not represent the broader population with chronic liver disease [16, 20]. Finally, the generalizability of findings is constrained by the lack of standardization in deep learning model development and validation.

Summary of main findings

The findings of this review underscore the potential of deep learning-enhanced MRI, especially MRE-based approaches, as a noninvasive alternative to liver biopsy for detecting advanced liver fibrosis and cirrhosis, with AUROCs often exceeding 0.90. This could minimize invasive procedure risks, enable earlier interventions, and reduce operator variability through automated analysis. MRE showed superior performance, particularly in MASLD and viral hepatitis, while non-MRE modalities (e.g., T2 mapping) were less consistent and better suited for advanced stages or resource-limited settings. Modality selection should consider clinical context, equipment availability, contraindications, and cost. However, challenges like model interpretability ("black box" nature) and computational demands hinder clinical adoption, warranting future research on explainable AI and real-world usability.

Comparison with existing literature

Our findings align with previous research indicating superior diagnostic accuracy of MRE compared with other MRI modalities. However, earlier reviews have typically focused on either MRE alone or on AI methods broadly without MRI-specific analysis. This study is the first to comprehensively assess deep learning across multiple MRI sequences—including T1/T2 mapping, gadoxetic acid–enhanced MRI, and non-contrast imaging—thereby extending previous literature. While earlier AI-focused reviews highlighted the potential of neural networks for liver disease classification, they did not quantify diagnostic accuracy across MRI techniques using biopsy as the reference standard.

Strengths

This review has several strengths. It is the first to systematically evaluate deep learning-enhanced MRI modalities for fibrosis and cirrhosis detection across a broad range of imaging techniques. The review adhered rigorously to PRISMA-DTA and QUADAS-2 guidelines, ensuring methodological transparency and high-quality risk-of-bias assessment. Inclusion of large-scale, multi-center studies strengthened the generalizability of findings, and the use of biopsy as the reference standard in the meta-analysis ensured alignment with clinical gold standards. The comparative approach—evaluating multiple MRI modalities and deep learning architectures—provides a nuanced understanding of performance differences.

Limitations

Despite its strengths, this review has several limitations. Only three studies used liver biopsy as a reference standard and provided extractable sensitivity and specificity, limiting the meta-analysis. Many studies used MRE stiffness thresholds instead of biopsy, which may introduce bias. Considerable heterogeneity existed in MRI protocols, disease etiologies, deep learning models, and validation strategies, complicating result comparability. The timing between MRI and biopsy was variable and sometimes unspecified. Several studies were retrospective, some had small sample sizes, and external validation was limited, raising concerns about overfitting and generalizability. Several research gaps emerged from this synthesis, warranting further investigation. First, the reliance on MRE as a reference standard in some studies highlights the need for studies using liver biopsy exclusively to ensure alignment with the clinical gold standard. Future research should also standardize the timing between MRI and biopsy to minimize confounding from disease progression, ideally performing both within a short interval (e.g., 30 days). Second, the heterogeneity in MRI protocols and deep learning models underscores the need for standardized imaging and algorithmic frameworks. Another major source of heterogeneity was the use of different reference standards across studies. While three studies used liver biopsy—the clinical gold standard—four studies used MRE stiffness thresholds as the reference. Because MRE itself has variable sensitivity for early fibrosis, using MRE as a surrogate reference can artificially inflate accuracy estimates or misclassify borderline cases. Furthermore, studies differed in the fibrosis thresholds applied (e.g., ≥ 3.0 kPa vs. ≥ 3.5 or ≥ 4.0 kPa), which directly affects case classification and diagnostic performance metrics. This inconsistency in reference standards limits direct comparability and underscores the need for biopsy-validated datasets.

Clinical implications

Deep learning–enhanced MRI, particularly MRE, holds strong promise as a noninvasive alternative for fibrosis and cirrhosis assessment, potentially reducing the reliance on liver biopsy. The high diagnostic accuracy suggests usefulness in clinical decision-making, especially for identifying advanced fibrosis. Automated deep learning systems can also reduce interobserver variability, improving diagnostic consistency. Non-contrast MRI models may provide accessible solutions in settings where contrast agents or MRE are unavailable, although they may be less sensitive for early fibrosis. Only Ali et al. [18] and Wang et al. [20] included external validation cohorts, while others relied on internal validation, raising concerns about model performance in diverse clinical settings. Furthermore, the computational complexity of models like 3D CoTNet and DeepLiverNet may limit their accessibility in resource-constrained environments, a practical barrier not addressed in the studies.

Future research directions

Collaborative efforts, such as multi-center trials, could establish consensus guidelines for MRI acquisition (e.g., sequence types, field strengths) and model development (e.g., open-source architectures, validation protocols). These standards would enhance comparability and facilitate clinical translation. Third, the limited representation of early fibrosis stages (F0–F1) in some studies calls for targeted research on detecting mild fibrosis, where noninvasive methods currently underperform [16, 17]. Developing deep learning models optimized for subtle tissue changes, potentially through multi-modal MRI integration (e.g., combining MRE, T1/T2 mapping, and texture analysis), could address this gap. Fourth, the generalizability of deep learning models requires further exploration through external validation across diverse populations, etiologies, and imaging platforms. Studies should also assess model performance in underrepresented groups, such as pediatric or elderly patients, and in low-resource settings where high-end MRI systems may be unavailable. Finally, practical barriers to implementation, such as computational requirements, cost, and clinician training, remain underexplored. Future research should evaluate the cost-effectiveness of deep learning-enhanced MRI compared to biopsy and assess strategies for integrating these tools into clinical workflows, including user-friendly interfaces and automated reporting systems.

Conclusion

This systematic review and meta-analysis evaluated the diagnostic accuracy of deep learning-enhanced MRI modalities for detecting and staging liver fibrosis and cirrhosis. Seven studies involving 6,547 participants were included, with MRE-based approaches showing the highest performance (AUROC: 0.759–0.93), particularly for advanced fibrosis and cirrhosis, outperforming other techniques like T1/T2 mapping and gadoxetic acid-enhanced MRI. Deep learning algorithms, such as CNNs and transformer-based networks, improved automated analysis and reduced variability. However, evidence is insufficient to replace liver biopsy as the gold standard due to limitations including the small number of studies (n = 7), reliance on MRE thresholds over histopathology in some cases, methodological heterogeneity, meta-analysis non-convergence, retrospective designs, and limited external validation. These raise concerns about generalizability, overfitting, and accuracy for early fibrosis. Now, deep learning-enhanced MRI, especially MRE, should serve as an adjunctive tool for risk stratification, disease monitoring, and guiding biopsies, used cautiously with histopathological confirmation in ambiguous cases. Future research requires larger, prospective, multicenter studies with standardized protocols, biopsy as the sole reference, diverse populations, and rigorous validation to confirm clinical utility and enable broader adoption in chronic liver disease management.

Authors' contributions

M.E. and A.K. conceptualized the study and developed the methodology. M.E. conducted the literature search, data extraction, and quality assessment. A.K. performed the data synthesis and wrote the main manuscript text. M.E. prepared Figs. 1–5 and tables. Both M.E. and A.K. reviewed and edited the final manuscript. I.A. supervised the study, provided methodological expertise, and reviewed and edited the final manuscript.

Funding

This systematic review and meta-analysis received no specific funding from any financial or non-financial sources. The authors conducted the study independently, with no involvement from funders or sponsors.

Data availability

The data extraction form and extracted data are available upon reasonable request from the corresponding author.

Declarations

Ethics approval and consent to participate

Not applicable. This systematic review and meta-analysis did not involve primary data collection from human subjects and therefore did not require approval from an Institutional Review Board or ethics committee.
Not applicable. This systematic review and meta-analysis utilized data from previously published studies and did not involve direct human participants.

Competing interests

The authors declare no competing interests.

Received: 15 May 2025 Accepted: 13 December 2025
Published online: 05 January 2026

References

1. Dunn R, Wetten A, McPherson S, Donnelly MC (2022) Viral hepatitis in 2021: the challenges remaining and how we should tackle them. World J Gastroenterol 28(1):76
2. Lee MJ (2023) A review of liver fibrosis and cirrhosis regression. J Pathol Transl Med 57(4):189–195
3. Younossi ZM (2019) Non-alcoholic fatty liver disease–a global public health perspective. J Hepatol 70(3):531–544
4. Williams R (2006) Global challenges in liver disease. Hepatology 44(3):521–526
5. Seto W-K, Mandell MS (2021) Chronic liver disease: global perspectives and future challenges to delivering quality health care. PLoS ONE 16(1):e0243607
6. Balogh J, Victor III D, Asham EH, Burroughs SG, Boktour M, Saharia A, et al. Hepatocellular carcinoma: a review. J Hepatocell Carcinoma. 2016:41-53.
7. Bernal W, Auzinger G, Dhawan A, Wendon J (2010) Acute liver failure. Lancet 376(9736):190–201
8. Bedossa P, Carrat F (2009) Liver biopsy: the best, not the gold standard. J Hepatol 50(1):1–3
9. Petitclerc L, Sebastiani G, Gilbert G, Cloutier G, Tang A (2017) Liver fibrosis: review of current imaging and MRI quantification techniques. J Magn Reson Imaging 45(5):1276–1295
10. Tang X (2019) The role of artificial intelligence in medical imaging research. BJR| open 2(1):20190031
11. Yasaka K, Akai H, Kunimatsu A, Abe O, Kiryu S (2018) Deep learning for staging liver fibrosis on CT: a pilot study. Eur Radiol 28:4578–4585
12. McInnes MD, Moher D, Thombs BD, McGrath TA, Bossuyt PM, Clifford T et al (2018) Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. JAMA 319(4):388–396
13. Macaskill P, Gatsonis C, Deeks J, Harbord R, Takwoingi Y (2010) Cochrane handbook for systematic reviews of diagnostic test accuracy. Version
14. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB et al (2011) QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 155(8):529–536
15. Cunha GM, Delgado TI, Middleton MS, Liew S, Henderson WC, Batakis D et al (2022) Automated CNN–based analysis versus manual analysis for MR elastography in nonalcoholic fatty liver disease: intermethod agreement and fibrosis stage discriminative performance. AJR Am J Roentgenol 219(2):224–232
16. Hectors SJ, Kennedy P, Huang K-H, Stocker D, Carbonell G, Greenspan H et al (2021) Fully automated prediction of liver fibrosis using deep learning analysis of gadoxetic acid–enhanced MRI. Eur Radiol 31(6):3805–3814
17. Sitarcikova D, Poetter-Lang S, Bastati N, Ba-Ssalamah S, Trattnig S, Attenberger U et al (2025) Diagnostic accuracy of texture analysis applied to T1-and T2-relaxation maps for liver fibrosis classification via machine-learning algorithms with liver histology as reference standard. Eur J Radiol 183:111887
18. Ali R, Li H, Zhang H, Pan W, Reeder SB, Harris D et al (2025) Multi-site, multi-vendor development and validation of a deep learning model for liver stiffness prediction using abdominal biparametric MRI. Eur Radiol 1-12
19. He L, Li H, Dudley JA, Maloney TC, Brady SL, Somasundaram E et al (2019) Machine learning prediction of liver stiffness using clinical and T2-weighted MRI radiomic data. AJR Am J Roentgenol 213(3):592–601
20. Li C, Wang Y, Bai R, Zhao Z, Li W, Zhang Q et al (2024) Development of fully automated models for staging liver fibrosis using non-contrast MRI and artificial intelligence: a retrospective multicenter study. EClinicalMedicine 77
21. Li H, He L, Dudley JA, Maloney TC, Somasundaram E, Brady SL et al (2021) Deeplivernet: a deep transfer learning model for classifying liver stiffness using clinical and T2-weighted magnetic resonance imaging data in children and young adults. Pediatr Radiol 51:392–402

Publisher's Note