Development of fully automated models for staging liver fibrosis using non-contrast MRI and artificial intelligence: a retrospective multicenter study

Chunli Li,[a,i] Yuan Wang,[b,i] Ruobing Bai,[a,i] Zhiyong Zhao,[c] Wenjuan Li,[d] Qianqian Zhang,[d] Chaoya Zhang,[e] Wei Yang,[b] Qi Liu,[f] Na Su,[g] Yueyue Lu,[a] Xiaoli Yin,[a] Fan Wang,[a] Chengli Gu,[a] Aoran Yang,[a] Baihe Luo,[a] Minghui Zhou,[a] Liuhanxu Shen,[a] Chen Pan,[a] Zhiying Wang,[a] Qijun Wu,[h] Jiandong Yin,[a] Yang Hou,[a,**] and Yu Shi[a,*]

[a]Department of Radiology, Shengjing Hospital of China Medical University, Shenyang, Liaoning, China
[b]Department of Radiology, Cancer Hospital of China Medical University, Liaoning Cancer Hospital & Institute, Shenyang, Liaoning, China
[c]Department of Medical Imaging, Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, Shandong, China
[d]Department of Radiology, Yantai Yuhuangding Hospital, Qingdao University, Yantai, Shandong, China
[e]Department of Radiology, Hubei Cancer Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, China
[f]Department of Radiology, The Second Affiliated Hospital of Baotou Medical College, Baotou, Neimenggu, China
[g]Department of Radiology, The Sixth People's Hospital of Shenyang, Shenyang, Liaoning, China
[h]Department of Clinical Epidemiology, Shengjing Hospital of China Medical University, Shenyang, Liaoning, China

eClinicalMedicine
2024;77: 102881

Published Online xxx
https://doi.org/10.1016/j.eclinm.2024.102881

Summary

Background Accurate staging of liver fibrosis (LF) is essential for clinical management in chronic liver disease. While non-contrast MRI (NC-MRI) yields valuable information for liver assessment, its effectiveness in predicting LF remains underexplored. This study aimed to develop and validate artificial intelligence (AI)-powered models utilizing NC-MRI for staging LF.

Methods A total of 1726 patients from Shengjing Hospital of China Medical University, registered between October 2003 and October 2022, were retrospectively collected, and divided into development (n = 1208) and internal test (n = 518) cohorts. An external test cohort consisting of 337 individuals from six centers, registered between June 2015 and November 2022, were also included. All participants underwent NC-MRI (T1-weighted imaging, T1WI; and T2-fat-suppressed imaging, T2FS) and liver biopsies. Two classification models (CMs), named T1 and T2FS, were trained on respective image types using 3D contextual transformer networks and evaluated on both test cohorts. Additionally, three CMs—Clinic, Image, and Fusion—were developed using clinical features, T1 and T2FS scores, and their integration via logistic regression. Classification effectiveness of CMs was assessed using the area under the receiver operating characteristic curve (AUC). A comparison was conducted between the optimal models (OMs) with highest AUC and other methods (transient elastography, five serum biomarkers, and six radiologists).

Findings Fusion models (i.e., OM) yielded the highest AUC among the CMs, achieving AUCs of 0.810 for significant fibrosis, 0.881 for advanced fibrosis, and 0.918 for cirrhosis in the internal test cohort, and 0.808, 0.868, and 0.925, respectively, in the external test cohort. The OMs demonstrated superior performance in AUC, significantly surpassing transient elastography (only for staging ≥ F2 and ≥ F3 grades), serum biomarkers, and three junior radiologists for staging LF. Radiologists, with the aid of the OMs, can achieve a higher AUC in LF assessment.

Interpretation AI-powered models utilizing NC-MRI, including T1WI and T2FS, accurately stage LF.

Funding National Natural Science Foundation of China (No. 82071885); General Program of the Liaoning Provincial Department of Education (LJKMZ20221160); Liaoning Province Science and Technology Joint Plan (2023JH2/101700127); the Leading Young Talent Program of Xingliao Yingcai in Liaoning Province (XLYC2203037).

*Corresponding author.
**Corresponding author.
E-mail addresses: 18940259980@163.com (Y. Shi), houyang1973@163.com (Y. Hou).
[i]Contributed equally and share the first authorship.

Keywords: Liver fibrosis; Non-contrast MRI; Artificial intelligence; Multicenter study

Research in context

Evidence before this study

We conducted a search on PubMed for publications without language restrictions, published before June 22, 2024. The search terms included "liver fibrosis" or "hepatic fibrosis" and "MRI" combined with "deep learning" or "radiomics" or "artificial intelligence" with these terms appearing in the abstract, title, or MESH headings. Although there have been studies on MRI for liver fibrosis assessment, no fully automated model for staging liver fibrosis based on non-contrast MRI (NC-MRI: T1-weighted imaging, T1WI; and T2-fat-suppressed imaging, T2FS) utilizing large-scale multicenter data has been established.

Added value of this study

This study presented an innovative method for staging liver fibrosis using NC-MRI integrated with advanced artificial intelligence (AI). Our AI-powered models, developed with a substantial dataset of 1208 patients and validated using an internal dataset of 518 patients and an external dataset of 337 patients, demonstrated superior diagnostic performance compared to radiologists and traditional methods such as transient elastography and serum biomarkers. Notably, our models, which incorporated clinical features (liver volume, spleen volume, age, and sex) and NC-MRI scores, obtained using 3D contextual transformer networks, demonstrated potential for superior diagnostic performance compared to radiologists and traditional methods such as transient elastography and serum biomarkers.

Implications of all the available evidence

The integration of our AI-powered models into NC-MRI can assist radiologists in the rapid and accurate staging of liver fibrosis in clinical practice. This advancement has the potential to significantly enhance diagnostic and therapeutic decision-making for patients with chronic liver diseases.

Introduction

Liver fibrosis, characterized by excessive collagen deposition and scar tissue formation due to chronic inflammation, markedly impairs hepatic circulation and alters liver architecture.[1] Leveraging the liver's regenerative capacity, early-stage fibrosis can often be reversed with proper treatment.[1] However, without intervention, ongoing inflammation and fibrosis may progress to cirrhosis, presenting greater treatment challenges.[1,2] This stage enhances the likelihood of hepatocellular carcinoma and hepatic decompensation, alongside additional complications such as ascites, hepatic encephalopathy, and variceal bleeding.[3–5] Notably, cirrhosis has ascended as a significant contributor to global mortality, accounting for 2.4% of all deaths worldwide in 2019.[6] Liver biopsy, the current reference method for fibrosis assessment in clinical practice, is unsuitable for routine fibrosis screening due to its invasiveness, susceptibility to sampling error, associated morbidity, complication risks, and interpretative variability.[7,8] Considering the silent and non-specific symptomatology of early-stage liver fibrosis, its detection presents significant challenges. Thus, it is critical to develop a highly accurate and efficient non-invasive diagnostic approach for the standard management protocols of chronic liver diseases.

A variety of evaluation methodologies have been proposed, showing enhanced diagnostic performance in the assessment of liver fibrosis. Among these, magnetic resonance elastography (MRE) has been established as the reference standard for non-invasive assessment.[9–12] Nevertheless, MRE's application is constrained by the need for specialized technical expertise and equipment, along with associated higher costs.[13] Considering these limitations, ultrasound elastography and serological assessments are acknowledged for their practical application and economic viability.[13–15] However, these alternative methods have their own shortcomings. Ultrasound elastography, for instance, may yield a higher failure rate in specific groups, particularly obese patients, while serological tests, despite less dependency on patient physique, are affected by individual biological variability and are unable to offer direct visualization of the liver.

Magnetic resonance imaging (MRI) is renowned for its high-resolution and exceptional soft tissue contrast, vital for precise liver tissue delineation.[16] Gadoxetic acid-enhanced MRI (GA-MRI) further amplifies this with dynamic contrast enhancement.[17–19] The implementation of advanced artificial intelligence (AI) technologies, including deep learning and radiomics, has proven highly effective at uncovering nuanced details in medical images and notably improving disease diagnosis accuracy,[20–22] especially in liver fibrosis assessment when combined with GA-MRI.[23–25] However, the limitations posed by potential contrast agent allergies and high costs necessitate exploration into AI-powered methods using non-contrast MRI (NC-MRI) for fibrosis assessment. Studies using radiomic approaches with T1-weighted imaging (T1WI) in predicting liver fibrosis grades or integrating T2-weighted fat-suppressed imaging (T2FS) radiomic features with clinical data have demonstrated promising results in assessing liver stiffness with reasonable accuracy,[26,27] despite facing

challenges such as limited sample sizes and the need for robust validation. Impressively, the fusion of multiparametric MRI with radiomics has unveiled significant promise in providing preoperative survival estimates for intrahepatic cholangiocarcinoma surgeries and identifying microvascular invasion in intrahepatic cholangiocarcinoma.[28,29] In this context, we propose a plausible hypothesis that integrating AI with NC-MRI (T1WI and T2FS) and clinical indicators from large-scale datasets could simplify and refine fibrosis evaluations, achieving both efficiency and accuracy.

Therefore, our study aimed to develop and validate AI-powered models for staging liver fibrosis, utilizing NC-MRI (T1WI and T2FS) and clinical data from large-scale multicenter datasets.

Methods

Ethics statement

Approval for this retrospective study was obtained from the Institutional Ethics Review Board at the participating hospitals (Shengjing Hospital of China Medical University: 2024PS863K; Shandong Provincial Hospital Affiliated to Shandong First Medical University: SWYX: NO.2024-483; The Second Affiliated Hospital of Baotou Medical College: 2024-ZX-051; Yantai Yuhuangding Hospital, Qingdao University: 2024-624; Hubei Cancer Hospital, Tongji Medical College: LLHBCH2024YN-074; Liaoning Cancer Hospital & Institute: 20210247X; The Sixth People's Hospital of Shenyang: 2024-09-002-01), with the need for informed consent waived due to its retrospective nature.

Study population

This study preliminarily encompassed 3141 consecutive patients from Shengjing Hospital of China Medical University between October 2003 and October 2022. These patients met the inclusion criteria, having undergone preliminary MRI scans and liver biopsy before receiving systemic medical treatment. A total of 1415 patients were excluded according to the following exclusion criteria: (1) Age below 18 (n = 38); (2) Poor image quality (n = 18); (3) Biopsy missing fibrosis grades (n = 982); (4) Presence of multiple (five or more) or large (≥10 cm) hepatic masses (n = 126); (5) Liver biopsy and MRI scans conducted with intervals exceeding 3 months (n = 251). Finally, 1726 patients were selected for further analysis.

In accordance with the timing of their MRI examinations, 1208 patients were included in the development cohort from October 2003 to February 2019, whereas 518 patients were designated to the internal test cohort between March 2019 and October 2022. Additionally, 337 patients from six independent hospitals (Liaoning Cancer Hospital & Institute; Shandong Provincial Hospital Affiliated to Shandong First Medical University; The Second Affiliated Hospital of Baotou Medical College; Yantai Yuhuangding Hospital, Qingdao University; Hubei Cancer Hospital, Tongji Medical College; and The Sixth People's Hospital of Shenyang), following the same inclusion and exclusion criteria as Shengjing Hospital of China Medical University, were chosen as the external test cohort to assess the developed models. The recruitment flowchart for all patients in this study is depicted in Fig. S1.

Liver biopsy

Liver fibrosis grades were assessed by expert pathologists using liver biopsies in accordance with the METAVIR scoring system.[30] Grades of ≥ F2, ≥ F3, and F4 correspond to significant fibrosis, advanced fibrosis, and cirrhosis, respectively.

To assess inter-rater agreement, specimens of 113 patients were selected and independently reviewed by two experienced pathologists, who have 14 and 18 years of experience, respectively. Each specimen's fibrosis grade was evaluated using the METAVIR scoring system.

Image acquisition and ROI annotation

The MRI images of all patients in this study were acquired using 1.5- or 3.0-T units equipped with dedicated body coils. All MRI scans comprised two standard non-contrast sequences: axial T1WI and T2FS. Detailed information regarding the scans is summarized in the Table S1.

From the development cohort, 120 patients with axial T1WI and T2FS images were randomly selected, with equal representation from each fibrosis stage (F0–1: n = 30, F2: n = 30, F3: n = 30, and F4: n = 30), for the development of the liver-spleen segmentation model. Annotation of 3D liver and spleen regions of interest (ROIs) in axial T1WI and T2FS images was conducted manually by experienced radiologists with over 5 years of expertise, in a blinded manner, utilizing ITK-SNAP.

To explore the variability in the annotated liver and spleen ROIs, another radiologist with over 3 years of experience also performed independent annotations on the images of these patients.

Establishment of segmentation models

The 3D full resolution framework of nnUNet (https://github.com/MIC-DKFZ/nnUNet), an adaptive medical image segmentation approach, was employed for establishing the segmentation model.[31] To achieve universal segmentation of liver-spleen in both T1WI and T2FS images, the two types of images were combined and input into the segmentation model for training. During the model training process, the networks underwent 300 epochs of training using the stochastic gradient descent (SGD) optimizer. A combination of dice and cross-entropy loss functions was employed during training. The initial learning rate was set to 1×10^{-2}, and after each epoch, the learning rate decayed

by 3×10^{-5}. Additionally, the momentum parameter was set to 0.99, and nesterov momentum update was utilized to expedite convergence. The model was trained on an RTX 3090 with 24 GB of memory and implemented using PyTorch (version 2.3.0) with Python version 3.9.17.

Development of classification models

In this study, T1 and T2FS models were developed utilizing 3D contextual transformer networks (CoTNet), an adaptation of the previously published 2D CoTNet architecture (https://github.com/JDAI-CV/CoTNet).[32] A detailed description of the data preprocessing, data augmentation, and 3D CoTNet structure is provided in the Supplementary Methods. The training protocol harnessed the adam optimizer, setting the learning rate to 1×10^{-4} and the weight decay to 1×10^{-5}. To rigorously assess model performance, ordinal regression loss was implemented, ensuring a detailed quantification of the models' ability to predict and rank fibrosis grades. Then, ReduceLROnPlateau scheduler was employed to dynamically adjust the learning rate based on the validation loss, with a reduction factor of 0.5 and a patience of 10 epochs. The training spanned 300 epochs in total. All models were trained in the same development environment as the segmentation network.

To improve fibrosis prediction, three models were developed: a Clinic model, an Image model, and a Fusion model. The Clinic models for predicting ≥ F2, ≥ F3, and F4 grades utilized logistic regression, incorporating four clinical features (age, sex, liver volume, and spleen volume). Image models were established using logistic regression by combining two scores derived from both the T1 and T2FS models. The Fusion models were developed utilizing logistic regression, integrating two scores from both the T1 and T2FS models with four clinical features for the staging of fibrosis grades. Five-fold cross-validation was utilized for the development cohort to enhance the model's generalization ability. The results for both test cohorts were computed using the mean scores from the five models obtained through five-fold cross-validation. The optimal models (OMs) for identifying ≥ F2, ≥ F3, and F4 grades were selected based on their classification performance on both test cohorts.

Interpretability of classification models

Two representative cases were selected from the MRE-based subgroup, with MRE used as the reference standard. Gradient-weighted class activation mapping (Grad-CAM) was utilized to highlight important ROIs for classification targets in the T1 and T2FS models.[33] Additionally, the key ROIs highlighted in the heatmaps of the T1 and T2FS models were compared with the stiffness ROIs measured by MRE to evaluate their consistency.The Shapley Additive Explanations method was employed to enhance the interpretability and transparency of the OMs.[34] This method provides a more precise understanding of the individual contributions of each variable to the model's predictions. Shapley values were calculated to evaluate the contribution of each parameter to the overall performance of the OMs for ≥ F2, ≥ F3, and F4 assessment.

Reader study

In the internal test cohort, 259 cases were randomly selected for two reader studies to stage fibrosis grades using NC-MRI, including T1WI and T2FS. Firstly, the performance of the OMs was compared with that of three junior radiologists (1–3 years of experience) and three senior radiologists (5–10 years of experience) in diagnosing liver fibrosis. Details related to radiologists for fibrosis assessment[35–39] are provided in the Supplementary Methods. To simulate real clinical diagnostic scenarios, apart from being informed about the age and sex, all other relevant information was masked during the evaluation process. Subsequently, an investigation was conducted to determine whether the OMs could assist radiologists. Each radiologist was required to have a washout period of at least 1 month between the two rounds of assessment. All assessment was performed using a digital imaging and communications in medicine viewer.

Serum fibrosis tests, transient elastography, and MRE

Five serum biomarkers, including the aspartate aminotransferase and platelet ratio index (APRI),[40] liver fibrosis factor 4 index (FIB-4),[41] red blood cell volume distribution width platelet ratio (RPR),[42] γ-glutamyl transpeptidase to platelet ratio (GPR),[43] and king score (K-S)[44] were calculated based on previous studies. The corresponding formulas for these biomarkers are provided in the Supplementary Methods. Examination results containing the relevant serum indicators were collected from the internal test cohort. A comparison was made between the OMs developed in this study and each individual serum biomarker in the cohorts where all serum biomarkers were available. In the serum-based subgroup, Complex Clinic models aimed at assessing ≥ F2, ≥ F3, and F4 grades were developed using logistic regression. These models incorporated features such as age, sex, liver volume, spleen volume, aspartate aminotransferase (AST), alanine transaminase (ALT), and platelet counts (PLT). Additionally, Complex Fusion models were created based on logistic regression, leveraging the same seven features along with two scores derived from the T1 and T2FS models.

Liver stiffness measurement data obtained via transient elastography (TE) and MRE was collected from the internal test cohort. In cohorts where liver stiffness measurement data was available, we compared the performance of the OMs developed in this study with that of TE and MRE.

Statistics

To describe the distribution of samples in different cohorts, the variables are respectively documented using the appropriate mean ± standard deviation (SD) for continuous variables, and numbers and percentages for categorical variables.

Linearly weighted κ statistics for categorical variables were used to evaluate the agreement between the two pathologists for fibrosis grades. The level of agreement was defined as follows: κ = 0–0.20, poor agreement; κ = 0.21–0.40, fair agreement; κ = 0.41–0.60, moderate agreement; κ = 0.61–0.80, good agreement; and κ = 0.81–1.00, very good agreement.[45]

The Dice Similarity Coefficient (DSC) was utilized to evaluate the performance of the segmentation models by measuring the overlap between the predicted masks and the ground truth masks.[46,47] The DSC value ranges from 0 to 1, where values closer to 1 indicate better consistency between the two segmentations, and values closer to 0 indicate poorer segmentation performance.[47]

The area under the receiver operating characteristic (ROC) curve (AUC) was employed to assess the diagnostic performance of the classification models. DeLong's test was used to compare the AUCs between OMs and other methods.[48] A *P* value of less than 0.05 was recognized as statistically significant. Calibration curve[49] and decision curve analysis (DCA)[50] were separately employed to evaluate the calibration and clinical usefulness of the OMs for staging fibrosis grades.

The DSC and AUC were calculated, and ROC curves were drawn using Python software (version 3.9.17). Calibration curves and decision curve analysis were conducted using R software (version 4.2.2).

Role of funding source

The funders of this study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. All authors had full access to the data in the study and had final responsibility for the decision to submit for publication.

Results

Study design and patient characteristics

The comprehensive study design is depicted in Fig. 1. We formulated five predictive models for staging fibrosis grades (Clinic, T1, T2FS, Image, and Fusion), employing five-fold cross-validation on a development cohort. Subsequently, these models underwent validation on both test cohorts. A total of 1208 participants (mean age: 44.76 ± 12.05 years; 763 [63.16%] males and 445 [36.84%] females) constituted the development cohort. In the internal test cohort, the mean age of participants was 46.93 ± 13.06 years, with 254 (49.03%) males and 264 (50.97%) females. As for the external test cohort, participants had a mean age of 46.94 ± 12.87 years, comprising 175 (51.93%) males and 162 (48.07%) females. Demographic characteristics of the three cohorts are detailed in Table 1.

Agreement evaluation

A good agreement (κ = 0.76) between the two independent pathologists was observed in the assessment of the fibrosis grades. The mean DSC scores of the annotated liver and spleen ROIs between the two radiologists were 0.990 ± 0.012 and 0.991 ± 0.006, respectively.

Performance of segmentation models

In liver segmentation, nnUNet consistently achieved mean DSC scores of 0.990 ± 0.002 across all five folds. Similarly, for spleen segmentation, the mean DSC scores were 0.991 ± 0.002, indicating a high degree of precision and reliability. Furthermore, our investigation revealed consistent performance in liver and spleen segmentation across different MRI images (liver at T1: 0.990 ± 0.002, at T2FS: 0.991 ± 0.002; spleen at T1: 0.987 ± 0.003, at T2FS: 0.994 ± 0.001), with minimal performance differences. We additionally investigated the segmentation performance of nnUNet in different units and indicated similar performance between 1.5- and 3.0-T units (liver at 1.5-T unit: 0.990 ± 0.002, at 3.0-T unit: 0.990 ± 0.002; spleen at 1.5-T unit: 0.991 ± 0.002, at 3.0-T unit: 0.991 ± 0.002). The bar charts displaying the mean DSC scores for liver and spleen are presented in Fig. 2 A-D. Original images and segmentation masks from nnUNet for two representative cases (F1 and F4 grades) are shown in Fig. 2E and F.

Performance of classification models

The cutoff values for predicting ≥ F2 grades were 0.522, 0.511, 0.629, 0.461 and 0.544 for the Fusion, Image, T2FS, T1, and Clinic models, respectively. The Fusion model, identified as the OM, excelled, showing a higher AUC compared to the other four models in both test cohorts, with an AUC of 0.810 internally and 0.808 externally. The AUCs for the other four models in the internal test cohort were as follows: 0.808 for Image, 0.803 for T2FS, 0.761 for T1, and 0.748 for Clinic. Similarly, in the external test cohort, the AUCs were 0.805 for Image, 0.793 for T2FS, 0.761 for T1, and 0.781 for Clinic. Moreover, the OM for predicting ≥ F2 grades, demonstrated the highest accuracy in both test cohorts (internal: 0.734, external: 0.718). Fig. 3A and C display the ROC curves of the five models in both test cohorts. Confusion matrices for the results of the ≥ F2 and F0–1 grades are presented in Fig. 3B and D. The corresponding sensitivity and specificity were 0.749 and 0.714 for the internal test cohort, and 0.775 and 0.632 for the external test cohort. A detailed description of the model performance for predicting ≥ F2 grades is outlined in Table 2.

The cutoff values for assessing ≥ F3 grades were 0.689, 0.583, 0.501, 0.469 and 0.451 for the Fusion, Image, T2FS, T1, and Clinic models, respectively. The

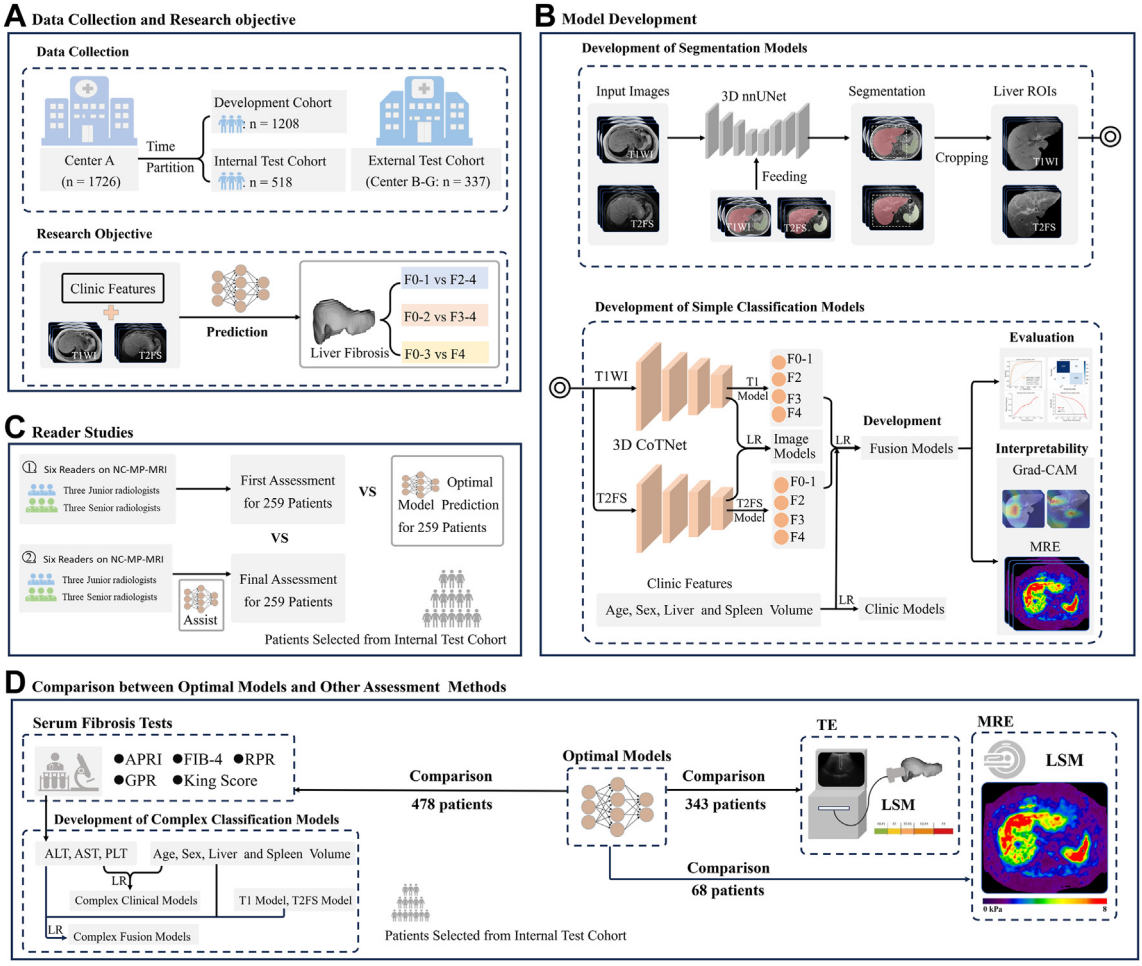

Fig. 1: Overview of the study design. (A) Data collection and research objective; (B) Development of segmentation models for the liver and spleen on NC-MRI, and five classification models for staging fibrosis grades; (C) Two reader studies on 259 cases (without and with the assistance of OMs) to assess fibrosis grades using NC-MRI; (D) A comparison between the OMs and other methods (serum fibrosis tests of 478 patients, TE of 343 patients, and MRE of 68 patients). Abbreviations: OM, optimal models; NC-MRI, non-contrast MRI; LR, logistic regression; TE, transient elastography; MRE, magnetic resonance elastography; LSM, liver stiffness measurement; Center A: Shengjing Hospital of China Medical University; Center B: Liaoning Cancer Hospital & Institute; Center C: Shandong Provincial Hospital Affiliated to Shandong First Medical University; Center D: The Second Affiliated Hospital of Baotou Medical College; Center E: Yantai Yuhuangding Hospital, Qingdao University; Center F: Hubei Cancer Hospital, Tongji Medical College; Center G: The Sixth People's Hospital of Shenyang.

Fusion model, recognized as the OM for ≥ F3 grades, demonstrated the highest AUC among all models, with an AUC of 0.881 for the internal test cohort and 0.868 for the external test cohort. In both test cohorts, the AUC values for the other four models were as follows: 0.879 and 0.863 for the Image, 0.870 and 0.848 for T2FS, 0.845 and 0.818 for T1, and 0.815 and 0.846 for the Clinic. Additionally, the Fusion model achieved the highest accuracy and specificity in both cohorts, with 0.824 and 0.903 internally, and 0.813 and 0.874 externally. The ROC curves of the five models in both test cohorts are provided in Fig. 3E and G. Fig. 3F and H present the confusion matrices for the results of the ≥ F3 and F0–2 grades. The performance of the models for ≥ F3 grades are detailed in Table 3.

The cutoff values for identifying F4 grade were 0.568, 0.593, 0.426, 0.572 and 0.373 for the Fusion, Image, T2FS, T1, and Clinic models, respectively. In the internal test cohort, the Fusion model, used as OM, achieved the highest AUC of 0.918, which is higher than the other models (0.912 for Image, 0.906 for T2FS, 0.874 for T1, and 0.883 for Clinic). The OM showed slight improvements in the external test cohort, with an AUC increase of 0.007, while still maintaining diagnostic superiority over the other models. The AUCs of the other four models were 0.920 for Image, 0.900 for T2FS, 0.883 for T1, and 0.869 for Clinic. Furthermore, for predicting F4 grade, the OM exhibited comparable performance in terms of accuracy, achieving 0.867 in the internal test cohort and 0.866 in the external test

cohort. Fig. 3I and K depict the ROC curves of the five models in both test cohorts. Fig. 3J and L display the confusion matrices presenting results for the F4 and F0–3 grades. The performance details of the five models for F4 grade are summarized in Table 4.

Notable differences ($P < 0.01$) between the OMs and the Clinic models for predicting ≥ F2, ≥ F3, and F4 grades were observed in the internal test cohort, whereas in the external test cohort, these differences were significant ($P < 0.001$) only for the F4 grade. The OMs for ≥ F2, ≥ F3, and F4 prediction exhibited substantial disparities ($P < 0.01$) when compared to the T1 model across both test cohorts. In the internal test cohort, no notable differences ($P > 0.05$) were observed between the OMs for assessing ≥ F2, ≥ F3, and F4 grades and the T2FS model; however, in the external test cohort, these differences were significant for identifying ≥ F3 and F4 grades ($P < 0.05$). There were no significant differences ($P > 0.05$) between OMs and Image models for predicting ≥ F2, ≥ F3, and F4 grades in both test cohorts. The detailed descriptions are provided in Tables 2–4

In a detailed stratification by etiology, the performance of the OM varied across different types of liver diseases in predicting ≥ F2 grades. The OM exhibited relatively good performance for chronic hepatitis C (CHC) and drug-induced liver disease (DILI), moderate performance for autoimmune hepatitis (AH) and metabolic dysfunction-associated steatotic liver disease (MASLD), and relatively poorer performance for chronic hepatitis B (CHB). The performance for alcoholic liver disease (ALD) was inconsistent: poorer internally but better externally. For predicting ≥ F3 grades, the OM

Characteristic	Development cohort (n = 1208)	Internal test cohort (n = 518)	External test cohort (n = 337)
Age (years)	44.76 ± 12.05	46.93 ± 13.06	46.94 ± 12.87
Sex			
Male	763 (63.16%)	254 (49.03%)	175 (51.93%)
Female	445 (36.84%)	264 (50.97%)	162 (48.07%)
Liver volume (cm^3)	1148.34 ± 329.96	1083.37 ± 345.12	1138.91 ± 385.57
Spleen volume (cm^3)	335.59 ± 254.78	257.27 ± 173.62	294.26 ± 199.66
Underlying liver disease			
CHB	541 (44.78%)	278 (53.67%)	165 (48.96%)
CHC	123 (10.18%)	69 (13.32%)	40 (11.87%)
AH	122 (10.10%)	74 (14.29%)	43 (12.76%)
ALD	58 (4.80%)	27 (5.21%)	29 (8.60%)
DILI	36 (2.98%)	27 (5.21%)	25 (7.42%)
MASLD	92 (7.62%)	43 (8.30%)	35 (10.39%)
None	236 (19.54%)	0 (0.00%)	0 (0.00%)
Fibrosis grades			
F0–1	461 (38.16%)	227 (43.82%)	133 (39.47%)
F2	155 (12.83%)	103 (19.88%)	66 (19.58%)
F3	121 (10.02%)	54 (10.42%)	30 (8.90%)
F4	471 (38.99%)	134 (25.87%)	108 (32.05%)

Abbreviation: CHB, chronic hepatitis B; CHC, chronic hepatitis C; AH, autoimmune hepatitis; DILI, drug-induced liver injury; MASLD, metabolic dysfunction-associated steatotic liver disease.

Table 1: **Demographic characteristics of three cohorts for staging fibrosis grades.**

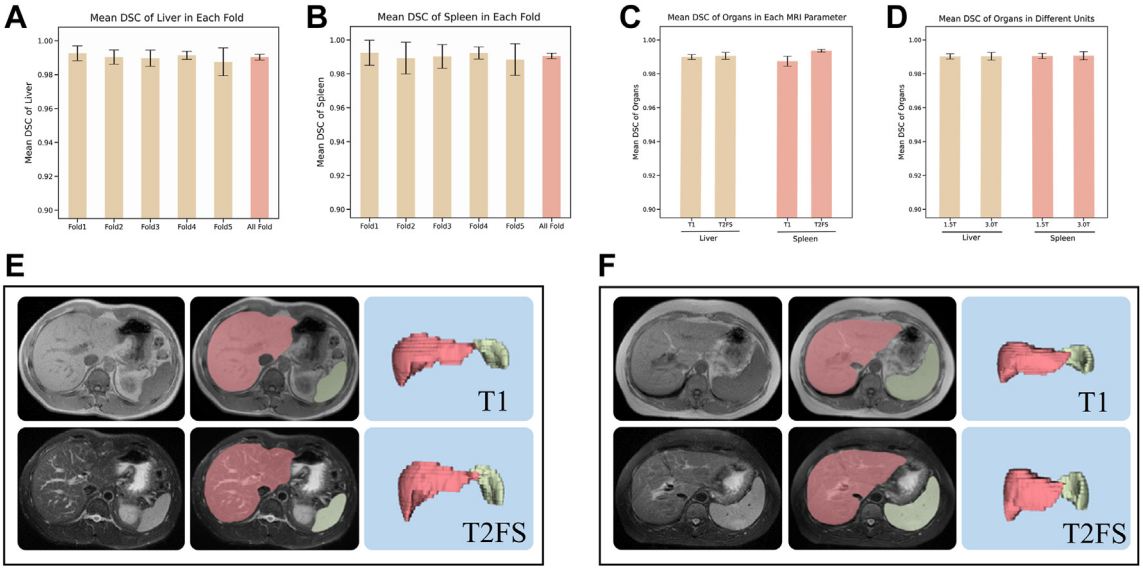

Fig. 2: Model performance for liver and spleen segmentation. (A) Box plot illustrating the mean DSC of the liver across each fold.; (B) Box plot illustrating mean DSC of spleen in each fold; (C) Box plot illustrating mean DSC of liver and spleen in each MRI parameter; (D) Box plot illustrating mean DSC of liver and spleen in different units; A–D, Error bars indicate 95% confidence interval; (E) Origin images and segmentation masks of F1 grade; (F) Origin images and segmentation masks of F4 grade. Abbreviation: DSC, dice Similarity Coefficient; T2FS, T2-weighted fat-suppressed imaging.

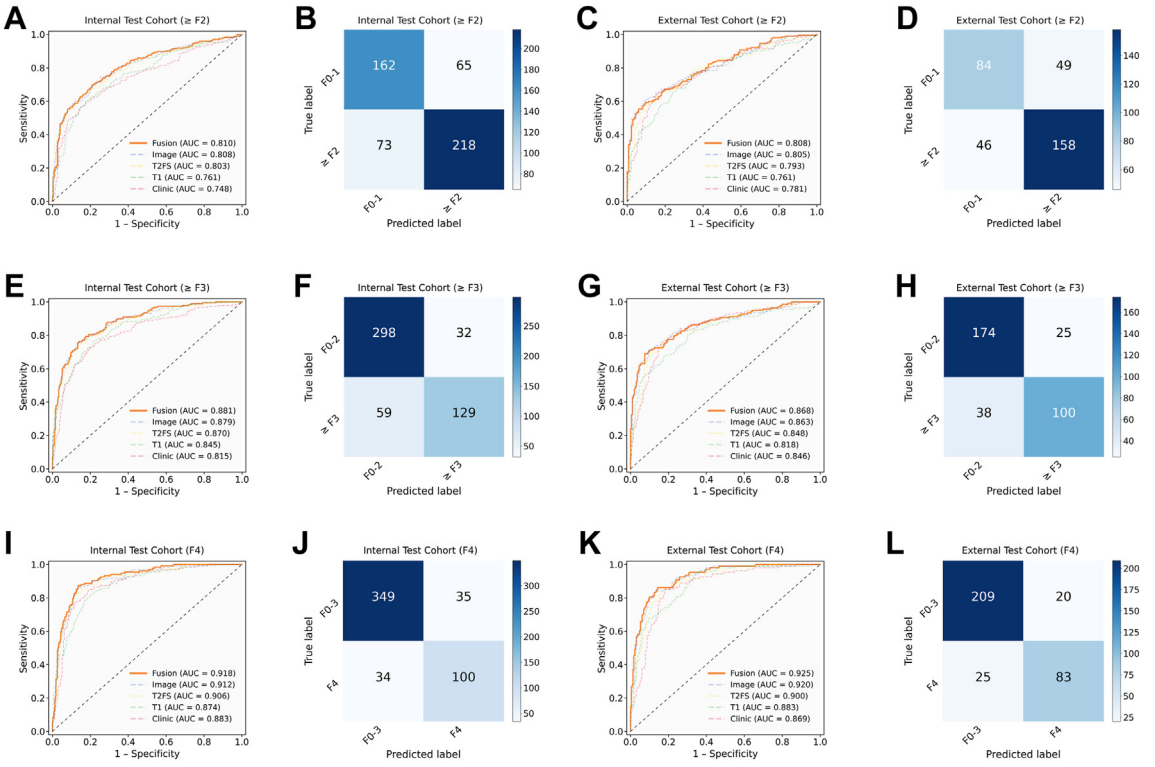

Fig. 3: Performance of classification models for staging fibrosis grades. The ROC curves of five models in the internal (≥F2 at A; ≥ F3 at E; F4 at I) and the external test cohorts (≥F2 at C; ≥F3 at G; F4 at K). The confusion matrices of the OMs in the internal (≥F2 at B; ≥ F3 at F; F4 at J) and the external test cohorts (≥F2 at D; ≥ F3 at H; F4 at L). Abbreviation: OM, optimal model; ROC, receiver operating characteristic; AUC, area under the receiver operating characteristic curve.

demonstrated strong performance for ALD, followed by DILI and CHC, with intermediate performance for MASLD and CHB. The performance for AH fluctuated, showing poorer results internally but better results externally. For identifying the F4 grade, the OM showed relatively good performance for DILI, CHB, and AH, while the performance for ALD was acceptable. The performance for MASLD and CHC varied, performing worse in internal tests but better in external tests. Detailed performance for the staging of liver fibrosis in a detailed stratification of causes are summarized in Table S3. The distribution of fibrosis grades within different etiological subgroups are provided in Fig. S3.

In a sex-based subgroup analysis, the OMs for identifying ≥ F2 and ≥ F3 grades demonstrated higher AUC values in males than in females. However, for assessing F4 grade, the AUC difference between the two sex subgroups was minimal, and the OM performance was relatively good. The consistent performance observed in the external cohort mirrored the trend in the internal cohort, indicating that the OMs had strong generalization ability. Additionally, both subgroups showed a gradual increase in AUC values with higher fibrosis grades. The detailed descriptions regarding the subgroup analysis based on sex are encapsulated in Table S4.

We observed that the OMs for predicting ≥ F2, ≥ F3, and F4 grades in the 3.0-T MRI subgroup exhibited higher AUC values compared to the 1.5-T MRI subgroup. Additionally, the performance of the OMs in

Models	AUC (95% CI)	Accuracy	Sensitivity	Specificity	*P* value[a]
Internal test cohort					
Fusion	0.810 (0.773, 0.847)	0.734	0.749	0.714	Reference
Image	0.808 (0.771, 0.845)	0.730	0.763	0.687	0.618
T2FS	0.803 (0.766, 0.840)	0.710	0.584	0.872	0.499
T1	0.761 (0.720, 0.802)	0.703	0.656	0.762	**<0.001**
Clinic	0.748 (0.706, 0.790)	0.685	0.691	0.678	**0.001**
External test cohort					
Fusion	0.808 (0.764, 0.853)	0.718	0.775	0.632	Reference
Image	0.805 (0.760, 0.850)	0.700	0.779	0.579	0.499
T2FS	0.793 (0.746, 0.840)	0.709	0.667	0.774	0.221
T1	0.761 (0.710, 0.811)	0.697	0.706	0.684	**0.002**
Clinic	0.781 (0.732, 0.830)	0.709	0.740	0.662	0.176

Abbreviations: AUC, area under the receiver operating characteristic curve; CI, confidence interval. [a]*P* value was calculated using the DeLong test. Bold text indicates that the *P* value is less than 0.05.

Table 2: **The performance of classification models for predicting ≥F2 grades.**

staging fibrosis grades was relatively consistent across the subgroups from the three different manufacturers. Among the various subgroups of two magnetic field strengths and the three main manufacturers, a gradual increase in AUC values correlating with higher grades of fibrosis was also observed. The similar findings were observed in the Image models. OMs and Image models for assessing ≥ F3, and F4 grades yielded the slightly higher performance in Siemens Healthineers compared with GE Healthcare and Philips Healthcare. The distribution of participants and ROC curves of OMs and Image models in various subgroups are provided in Fig. S4.

The calibration curves and the DCA for assessing different fibrosis grades in both test cohorts are presented in Fig. 4. The calibration curves for the OMs of ≥ F2, ≥ F3, and F4 prediction showed strong concordance between predicted probabilities and observed outcomes in both test cohorts. The DCA revealed that when the threshold probability exceeded 30% for both patients and doctors, employing the OMs for staging fibrosis grades offered greater benefits compared to either treating all patients or treating none.

Interpretability of classification models

To determine whether the T1 and T2FS models target the correct areas, Grad-CAM was used to visualize the internal features of the neural network. Our analysis indicated that, in two representative cases, the liver ROIs focused on by both models partially overlapped with the ROIs of higher stiffness as revealed by MRE stiffness maps (Fig. 5). These findings demonstrated that the T1 and T2FS models had acquired the ability to capture certain valuable information relevant to liver fibrosis assessment. Moreover, consistent findings were observed in the subgroup of 68 participants with MRE.

In the OMs for ≥ F2, ≥ F3, and F4 assessments, the scores derived from the T2FS model contributed significantly more to fibrosis prediction compared to those from the T1 model. Furthermore, both T2FS and T1 model scores provided greater predictive contributions than individual clinical characteristics, including age, sex, liver volume, and spleen volume (Fig. S5). Age and spleen volume were particularly noteworthy, as they contributed more effectively to F4 prediction compared to other clinical features. Among these clinical characteristics, liver and spleen volume demonstrated a significant contribution in the prediction of ≥ F2 and ≥ F3 grades.

The influence of these features on the performance of OMs, as illustrated in Fig. S5, was observed. For predictions of ≥ F2, ≥ F3, and F4 grades, T2FS score, T1 score, and spleen volume had a positive impact, whereas liver volume had a negative impact. Additionally, age and sex showed a positive influence in the prediction of ≥ F2 and F4 grades.

Models	AUC (95%CI)	Accuracy	Sensitivity	Specificity	*P* value[a]
Internal test cohort					
Fusion	0.881 (0.851, 0.912)	0.824	0.686	0.903	Reference
Image	0.879 (0.849, 0.910)	0.822	0.734	0.873	0.400
T2FS	0.870 (0.837, 0.902)	0.815	0.739	0.858	0.087
T1	0.845 (0.811, 0.880)	0.786	0.633	0.873	**<0.001**
Clinic	0.815 (0.775, 0.855)	0.763	0.750	0.770	**<0.001**
External test cohort					
Fusion	0.868 (0.827, 0.908)	0.813	0.725	0.874	Reference
Image	0.863 (0.821, 0.905)	0.792	0.783	0.799	0.166
T2FS	0.848 (0.805, 0.891)	0.783	0.804	0.769	**0.040**
T1	0.818 (0.771, 0.866)	0.763	0.674	0.824	**<0.001**
Clinic	0.846 (0.803, 0.889)	0.769	0.812	0.739	0.223

Abbreviations: AUC, area under the receiver operating characteristic curve; CI, confidence interval. [a]P value was calculated using the DeLong test. Bold text indicates that the *P* value is less than 0.05.

Table 3: **The performance of classification models for assessing ≥ F3 grades.**

Performance of reader assessment

Fig. 6 illustrates the performance of two reader studies conducted using 259 samples selected from the internal test cohort. The first reader study was conducted without the assistance of the OMs. For ≥ F2, ≥ F3 and F4 prediction, OMs (AUC range: 0.817–0.903) significantly outperformed the three junior (AUC range: 0.641–0.757; $P < 0.01$) and yielded higher AUCs compared with three senior (AUC range: 0.631–0.840) radiologists in first reader study.

In second reader study, after a washout period of at least one month, radiologists were provided with the primary prediction probabilities from the OMs, along with measurements of liver and spleen volumes, leading to a significant improvement in AUCs for staging liver fibrosis (AUC range for ≥ F2 grades: 0.715–0.733; ≥ F3 grades: 0.771–0.815; and F4 grade: 0.784–0.851). A similar performance for ≥ F2 grades (First: 0.649–0.676,

Models	AUC (95% CI)	Accuracy	Sensitivity	Specificity	*P* value[a]
Internal test cohort					
Fusion	0.918 (0.892, 0.944)	0.867	0.746	0.909	Reference
Image	0.912 (0.884, 0.939)	0.859	0.716	0.909	0.082
T2FS	0.906 (0.877, 0.935)	0.847	0.858	0.844	0.161
T1	0.874 (0.841, 0.908)	0.819	0.500	0.930	**<0.001**
Clinic	0.883 (0.851, 0.916)	0.807	0.851	0.792	**0.006**
External test cohort					
Fusion	0.925 (0.897, 0.953)	0.866	0.769	0.913	Reference
Image	0.920 (0.890, 0.949)	0.855	0.731	0.913	0.130
T2FS	0.900 (0.867, 0.934)	0.816	0.880	0.786	**0.015**
T1	0.883 (0.846, 0.920)	0.825	0.583	0.939	**<0.001**
Clinic	0.869 (0.829, 0.909)	0.789	0.861	0.755	**<0.001**

Abbreviations: AUC, area under the receiver operating characteristic curve; CI, confidence interval. [a]P value was calculated using the DeLong test. Bold text indicates that the *P* value is less than 0.05.

Table 4: **The performance of classification models for identifying F4 grade.**

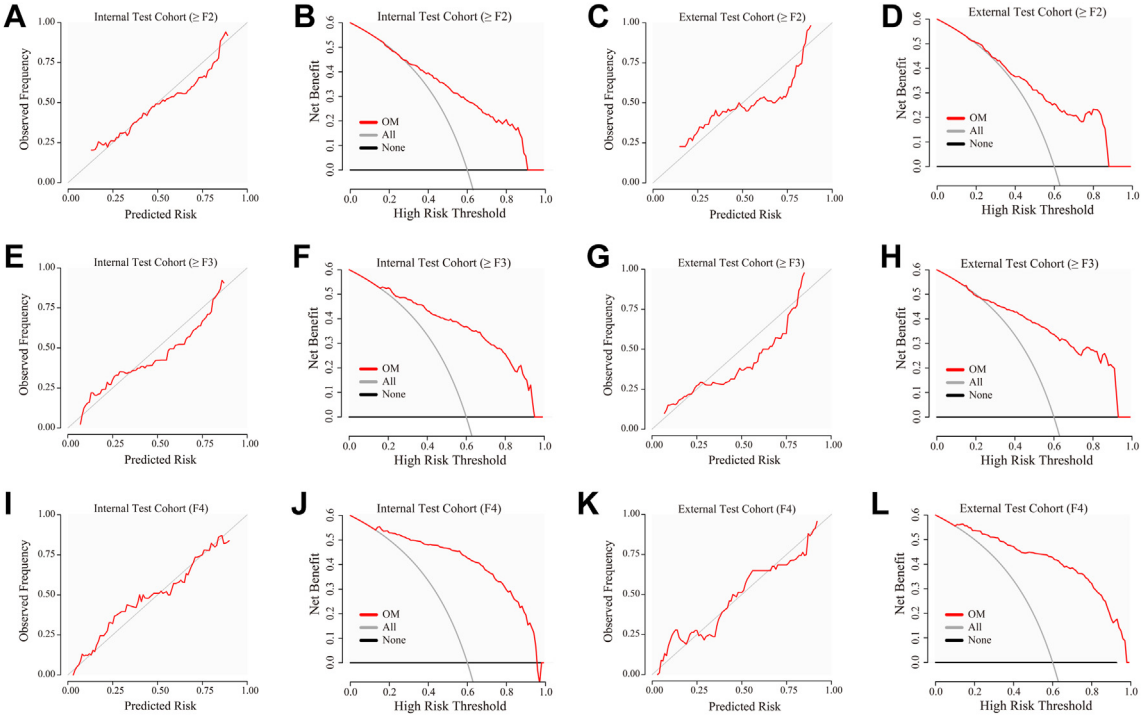

Fig. 4: Calibration and DCA curves for staging fibrosis grades. The OM for the evaluation of ≥ F2 grades in the internal (Calibration at A; DCA at B) and the external test cohorts (Calibration at C; DCA at D); the OM for the assessment of ≥ F3 grades in the internal (Calibration at E; DCA at F) and the external test cohorts (Calibration at G; DCA at H); the OM for the assessment of F4 grade in the internal (Calibration at I; DCA at J) and the external test cohorts (Calibration at K; DCA at L). Abbreviation: OM, optimal model; DCA, decision curve analysis.

and second: 0.726–0.737), ≥ F3 grades (First: 0.741–0.768, and second: 0.799–0.834), and F4 grade (First: 0.784–0.849, and second: 0.846–0.873) assessment was observed in terms of accuracy (Table S5).

Comparison between OMs and other methods

In the MRE-based subgroup, which comprised 68 patients from the internal test cohort, the OM for assessing the F4 (0.946 vs. 0.969) grade showed a slightly lower AUC compared to MRE. However, the AUCs of the OMs for identifying ≥ F2 (0.888 vs. 0.955) and ≥ F3 (0.842 vs. 0.930) grades were obviously lower compared to those achieved by MRE. Detailed descriptions about two methods are summarized in Table S6. The ROC curves of OMs and MRE are presented in Fig. 7A–C.

A total of 343 patients selected from an internal test cohort were used to compare the diagnostic performance of OMs with that of TE for staging liver fibrosis. The distribution of fibrosis grades and etiology within TE subgroups was provided in Fig. S6A and B. The OMs for predicting ≥ F2, ≥ F3, and F4 grades achieved higher AUCs of 0.827, 0.906, and 0.922, respectively, surpassing those of TE, which were 0.741, 0.812, and 0.878, as shown in Fig. 7D–F. Moreover, OMs for assessing ≥ F2, ≥ F3, and F4 grades exhibited comparable performance in terms of accuracy, outperforming TE by 0.131, 0.085, and 0.047, respectively. Detailed descriptions of the two methods are provided in Table S7. The performance disparity between the two methods for predicting ≥ F2 and ≥ F3 grades was significant ($P = 0.011$ and $P = 0.009$, respectively), whereas no notable difference was observed in predicting F4 grade ($P = 0.236$) (Table S9).

The diagnostic performance of OMs was compared with that of five serum biomarkers in a sample of 478 patients from the internal test cohort (Fig. 7G–I). The distribution of fibrosis grades and etiology within serum subgroups was presented in Fig. S6C and D. The OMs for predicting ≥ F2, ≥ F3, and F4 grades demonstrated superior diagnostic performance, achieving the highest AUCs of 0.814, 0.902, and 0.926, respectively. These results surpassed those obtained by APRI (0.764, 0.771, and 0.767), FIB-4 (0.770, 0.795, and 0.815), GPR (0.724, 0.761, and 0.762), RPR (0.764, 0.815, and 0.842), and K-S (0.787, 0.811, and 0.822). The Complex Clinic models for identifying ≥ F2 (0.786), ≥ F3 (0.855), and F4 (0.911) grades achieved higher AUC values compared to the Clinic models, APRI, FIB-4, GPR, and RPR. Furthermore, the Complex Clinic models for assessing ≥ F3 and F4 grades also yielded higher AUC values than K-S. We observed that the Complex Fusion model for predicting F4 grade achieved a higher AUC compared to

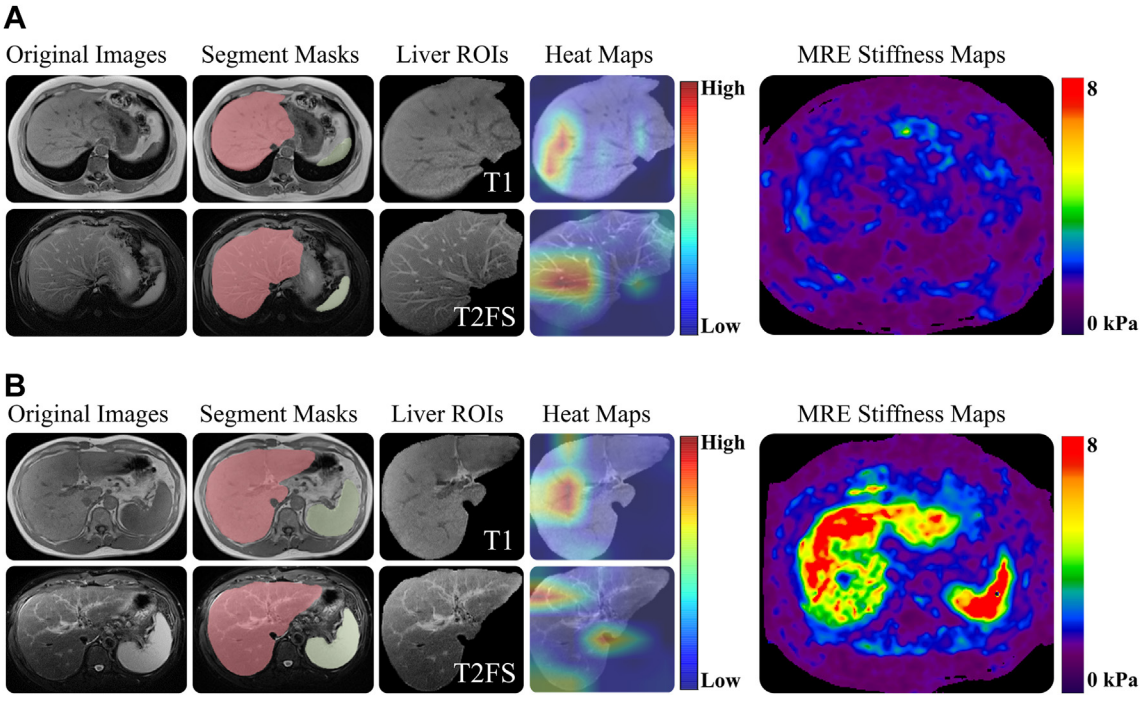

Fig. 5: Comparison between MRE stiffness maps and the visualization heat maps of T1 and T2FS models using Grad-CAM in different fibrosis grades. (A) Original T1 and T2FS images, segmented masks, live ROIs, heat maps, and MRE stiffness map of a case with F1 grade; (B) Original T1 and T2FS images, segmented masks, live ROIs, heat maps, and MRE stiffness map of a case with F4 grade. The fibrosis grades of the cases were confirmed using MRE. In the heat maps of the liver ROIs, red signifies higher activation, while blue represents lower activation. Abbreviation: ROIs, regions of interest; T2FS, T2-weighted fat-suppressed imaging; Grad-CAM, gradient-weighted class activation mapping; MRE, magnetic resonance elastography.

the Fusion model. However, for assessing ≥ F2 and ≥ F3 grades, the Complex Fusion models demonstrated a lower AUC. Table S8 delineates comprehensive details concerning the performance of these methodologies. For the prediction of ≥ F2, ≥ F3, and F4 grades, significant differences ($P < 0.05$) were observed between OMs and four methods: APRI, FIB-4, GPR, and RPR. No significant difference ($P = 0.215$) was found between OM and KS for ≥ F2 prediction; however, notable disparities ($P < 0.01$) were observed for ≥ F3 and F4 grades. The detailed descriptions are summarized in Table S9.

Discussion

This study sought to develop and validate AI-powered models for the automated staging of liver fibrosis using clinical features and NC-MRI from multicenter data. Consistent with our hypothesis, AI's capability to effectively capture valuable diagnostic information from NC-MRI and integrate it with clinical features can enhance the performance for staging liver fibrosis, achieving AUC values in the range of 0.808–0.925.

Previous studies have rarely conducted detailed stratified investigations based on etiology.[23–27] In this study, we performed a comprehensive stratified analysis of different etiologies, and found that OMs demonstrated robust and consistent performance in staging liver fibrosis for CHB and DILI subgroups of both test cohorts, particularly excelling in predicting the F4 grade. This finding is partially consistent with previous studies.[51,52] However, variability in performance was observed among other etiologies, such as CHC, AH, ALD, and MASLD, likely due to limited sample sizes and differences in sample distribution within these subgroups. Additionally, variations in fibrosis and inflammation across different etiologies may collectively influence the model's prediction outcomes. Although fibrosis is staged similarly, the characteristics of fibrosis at the same grade can vary among different etiologies. For instance, fibrosis in ALD and MASLD typically presents as micronodular, whereas fibrosis due to CHB and CHC often presents as macronodular. Furthermore, the inflammatory pathological features vary significantly among different etiologies of liver disease, such as portal inflammation and interface hepatitis in CHC,[30] interface hepatitis and periportal necrosis in AH,[53] hepatocyte ballooning and Mallory-Denk bodies in ALD,[54] and steatosis, hepatocellular ballooning and lobular inflammation in MASLD.[55] From an overall cohort perspective, OMs displayed consistent and

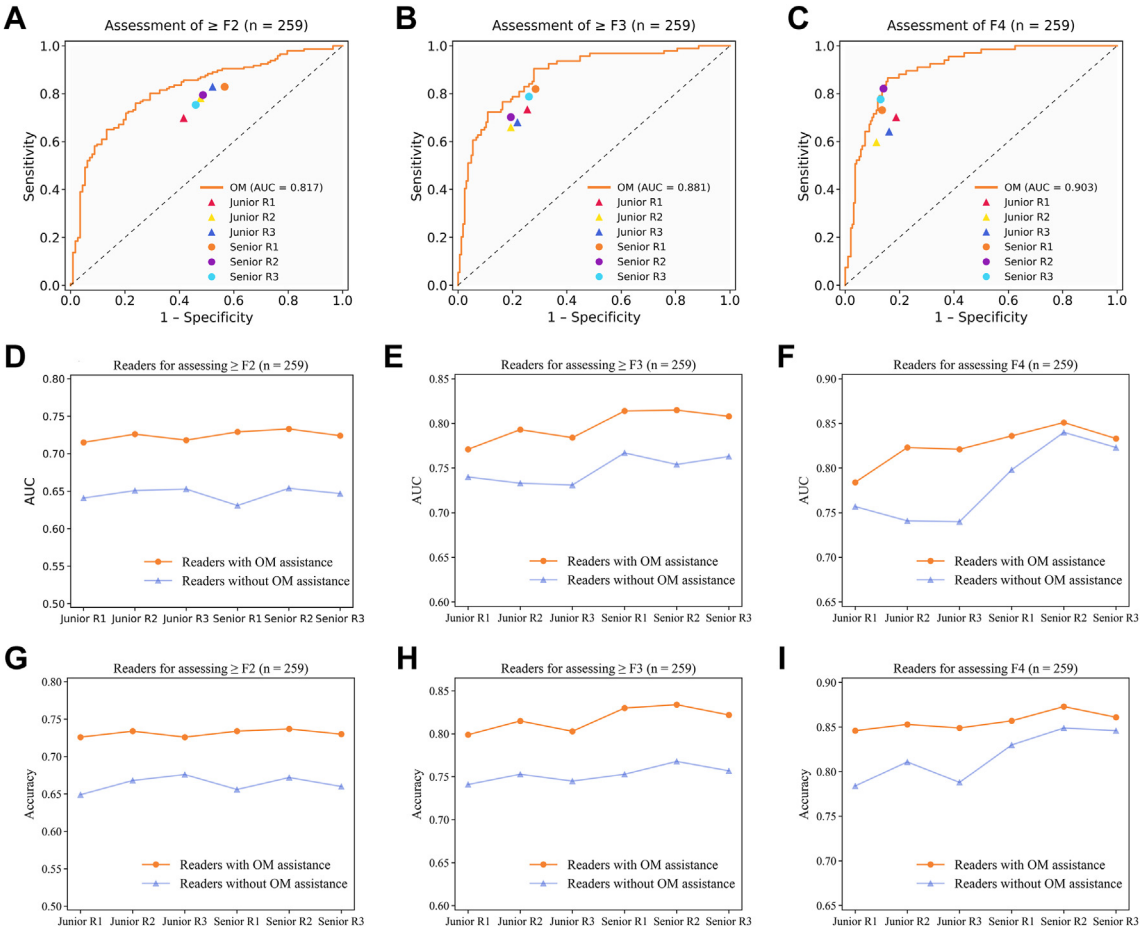

Fig. 6: Performance of two reader studies using 259 samples selected from the internal test cohort, both without and with the assistance of OMs. (A–C) The readers without the assistance of OMs for ≥ F2, ≥ F3 and F4 prediction; (D–F) The AUCs of readers with the assistance of OMs for ≥ F2, ≥ F3 and F4 prediction; (G–I) The accuracy of readers with the assistance of OMs for ≥ F2, ≥ F3 and F4 prediction. Abbreviation: OM, optimal model; R1-6, radiologists with varying levels of experience; AUC, area under the receiver operating characteristic curve.

reliable generalization capabilities across both test cohorts. However, further extensive validation is necessary to comprehensively evaluate its performance across different etiological subgroups.

Existing non-invasive methods, such as TE and serum biomarkers, are widely recommended for liver fibrosis screening in clinical practice.[13–15,56] However, the AI-powered models developed in this study yielded higher AUCs compared to these traditional methods in subgroup analysis. Notably, we observed that in our subgroup, the performance of LSM obtained through TE was not as robust as previously reported by Boursier et al.[57] in the MASLD population,[58] especially in predicting ≥ F2 grades (AUC: 0.741 in our study vs. 0.842 in the study by Boursier et al.[57]). Furthermore, a previous study by Degos et al.[59] involving patients with chronic viral hepatitis found that the performance of LSM measured by FibroScan for predicting ≥ F2 grades had an AUC of only 0.76. When comparing the results of these two studies, we discovered that the performance of TE for predicting ≥ F2 grades generally decreased in populations with chronic viral hepatitis. In the subgroup analyzed in our study, more than half of the participants had chronic viral hepatitis, which could partly explain the lower performance of TE observed. Furthermore, both previous studies indicated that TE was more accurate in diagnosing more advanced fibrosis. In our study, the higher proportion of individuals with F0–1 and F2 grades may also contribute to the comparatively lower performance of TE. The AI-powered models developed in this study demonstrated comparable performance to FibroScan from two previous studies in predicting ≥ F2 grades (AUC: 0.810 in our study vs. 0.842 in Boursier et al.[57] and 0.76 in Degos et al.[59]) in the internal test cohort, while also exceeding the performance of FibroScan in predicting ≥ F3 (AUC: 0.881 in our study vs. 0.831 in Boursier et al.[57]) and F4 grades (AUC: 0.918 in our study vs. 0.864 in Boursier et al.[57]

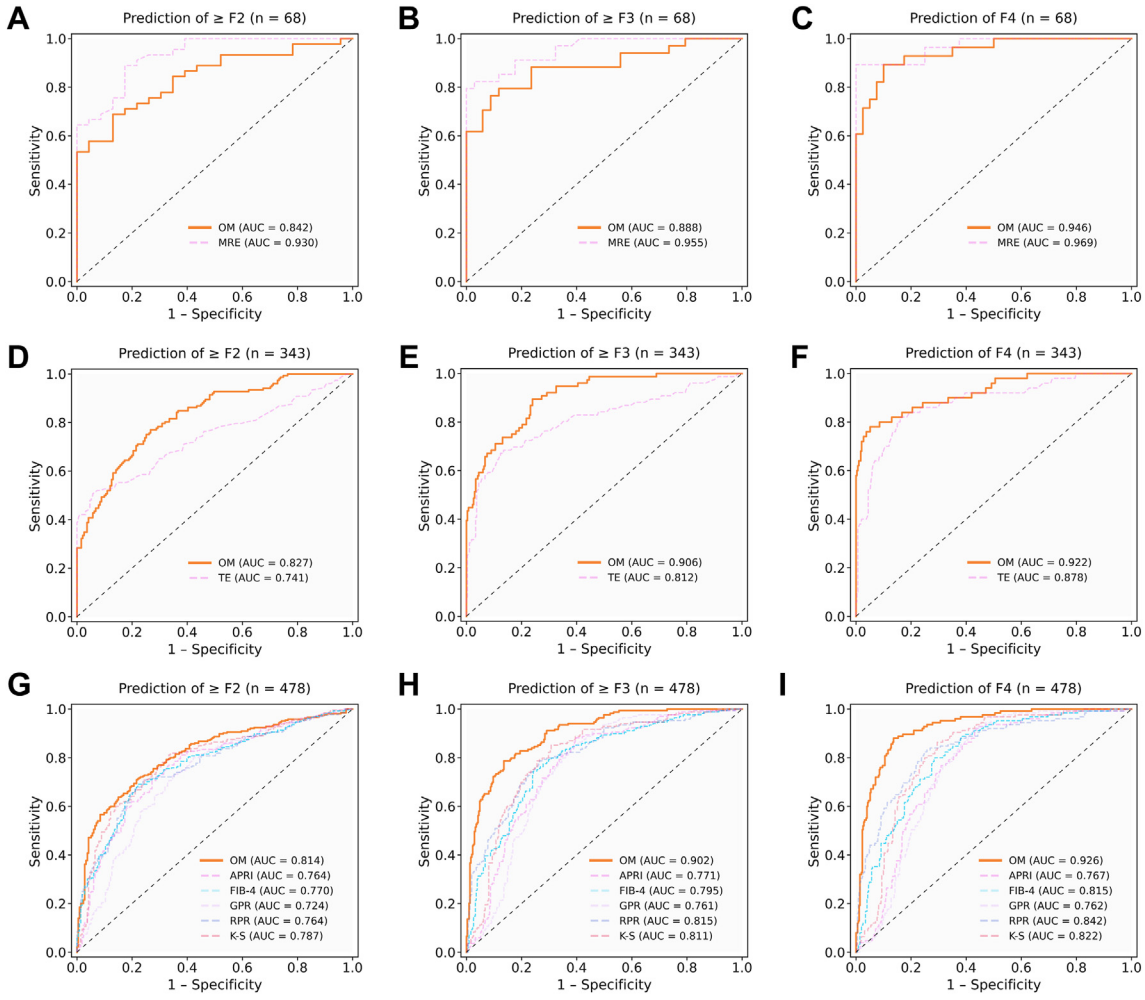

Fig. 7: ROC curves of OMs and other methods (MRE: n = 68; TE: n = 343; and five serum methods: n = 478) using samples selected from the internal test cohort. (A–C) ROC curves of OMs and MRE for ≥ F2, ≥ F3 and F4 prediction; (D–F) ROC curves of OMs and TE for ≥ F2, ≥ F3 and F4 prediction; (G–I) ROC curves of OMs and five serum methods for ≥ F2, ≥ F3 and F4 prediction. Abbreviation: OM, optimal model; MRE, magnetic resonance elastography; TE, transient elastography; APRI, aspartate aminotransferase and platelet ratio index; FIB-4, liver fibrosis factor 4 index; RPR, red blood cell volume distribution width platelet ratio; GPR, γ-glutamyl transpeptidase to platelet ratio; K-S, king score; AUC, area under the receiver operating characteristic curve.

and 0.90 Degos et al.[59]). Additionally, our AI-powered models maintained robust and similar performance in an external test cohort.

Several studies have demonstrated that MRE has a high diagnostic performance in assessing liver fibrosis and can be used as a noninvasive reference standard.[9–12] To further explore the clinical utility of our developed models, we compared OMs and MRE for staging liver fibrosis in a subgroup of 68 patients. We found that the OM for assessing the F4 grade showed a slightly lower AUC compared to MRE, while the AUCs of the OMs for identifying ≥ F2 and ≥ F3 grades were significantly lower compared to those achieved by MRE. This indicated that a certain gap existed between our developed models and MRE in the evaluation of ≥ F2 and ≥ F3 grades. However, a limited number of 3D-MRE were included and analyzed for the staging of fibrosis grades. In future studies, we will include more MRE data for comparative analysis with our models. Furthermore, in the serum subgroups, Complex Fusion models for assessing ≥ F2, ≥ F3, and F4 grades were established, which additionally incorporated three serum features—PLT, AST, and ALT—compared to the OMs. We observed that the AUC value of the Complex Fusion model was higher than that of the OM only in predicting the F4 grade. This could be due to the lower performance of these features in evaluating early fibrosis, resulting in no enhancement in the performance of the Complex Fusion models for predicting ≥ F2 and ≥ F3 grades. Furthermore, in serum-based subgroup, the

higher proportion of individuals with F0–1 and F2 grades may also be another reason. DeLong test indicated that no significant differences between OMs and Image models for staging fibrosis grades were observed in both test cohorts. The incorporation of five additional clinical features into the OMs failed to yield significant performance improvements compared to the Image models. This indicated that the Image models effectively encapsulated essential information regarding the staging of fibrosis, already including the majority of insights derived from clinical features. Moreover, the Image models achieved similar performance to OMs and were therefore sufficient for staging fibrosis grades. However, our study involved a limited number of clinical features, which may be a reason why the inclusion of clinical features produced no significant performance enhancements. Recent studies have indicated that clinical features, including body mass index, albumin, and prothrombin time, among others, were valuable for assessing ≥ F2 grades.[60,61] However, these features were not incorporated into our models. To enhance the model's performance and robustness, future studies will incorporate more valuable clinical features and validate the model within a larger, multicenter cohort to confirm whether incorporating clinical features can further significantly improve the performance of the models. We additionally observed that our developed OMs and Image models demonstrated higher AUC values for predicting ≥ F2, ≥ F3, and F4 grades in the 3.0-T MRI subgroup compared to the 1.5-T MRI subgroup, and their performance in staging fibrosis grades was relatively consistent across subgroups from different manufacturers. Moreover, OMs and Image models for the prediction of ≥ F3 and F4 grades exhibited slightly higher performance with Siemens Healthineers compared to GE Healthcare and Philips Healthcare. This may be due to the fact that the subgroup cohort with Siemens Healthineers did not include a 1.5T device. Future research will include more samples to further validate the performance of our developed models across different manufacturers.

The AI-powered models developed in this study significantly outperformed six radiologists of varying seniority in staging liver fibrosis. In clinical practice, the NC-MRI features of early liver fibrosis are often subtle and difficult for radiologists to detect with the naked eye.[38] Consequently, radiologists typically do not perform routine staging of liver fibrosis when interpreting NC-MRI, contributing to their lower performance. AI-powered models excelled at detecting subtle variations by analyzing extensive NC-MRI (T1WI and T2FS) datasets, utilizing this valuable latent information to enhance the performance of fibrosis classification. However, due to the more pronounced changes in liver texture and morphological structure as liver fibrosis progresses,[35,62,63] our models demonstrated lower performance in detecting early-stage fibrosis (≥F2 and ≥ F3) compared to that of F4 grade. To elucidate the opaque nature of AI decision processes, we employed Grad-CAM to visualize our model's feature maps and compared them to the non-invasive reference standard of MRE stiffness maps. The results indicated that the T1 and T2FS models effectively captured partial yet valuable diagnostic information related to tissue stiffness, thereby enhancing their interpretability. Furthermore, similar results were noted within the subgroup of 68 participants who underwent MRE. We also found that in OMs for staging liver fibrosis, T2FS model scores contributed more to fibrosis prediction than T1 model scores. This result is partially consistent with previous evidence, which indicated that high signals observed in T2FS images reflect the reticular structure of fibrous bands surrounding regenerative nodules, further validating the effectiveness of the T2FS model in evaluating liver fibrosis.[39] Moreover, both model scores offered greater predictive value than individual clinical characteristics and positively impacted the accuracy of predictions. This also supports our hypothesis that deep learning can capture more valuable information from T1WI and T2FS images. Among the clinical features, we found that spleen volume had a positive impact, whereas liver volume had a negative impact, consistent with the previous findings by Pickhardt et al.[64] Furthermore, we additionally found that radiologists assisted by the AI powered models exhibited improved diagnostic performance, further demonstrating the practical value of our model in clinical applications.

Previous studies have explored the viability of leveraging advanced AI technology based on MRI for staging liver fibrosis.[23–27] Yasaka et al.[23] and Hectors et al.[24] demonstrated that deep learning models utilizing GA-MRI achieved superior performance in predicting ≥F2 grades (AUC: 0.85 and 0.91), with slightly reduced accuracy for ≥F3 grades (AUC: 0.84 and 0.90) and F4 (AUC: 0.84 and 0.85). Additionally, Park et al.[25] conducted a study using radiomics analysis of GA-MRI, yielding similar results for ≥F2 and ≥F3 assessment (AUC: 0.90 and 0.89, respectively), while demonstrating slightly better diagnostic performance for F4 prediction (AUC: 0.91) compared to the Yasaka et al.[23] and Hectors et al.[24] studies. Our developed model exhibited lower performance (AUC: 0.81) for predicting ≥F2 grades in comparison to these previous studies. This may be attributed to the use of GA-MRI in these previous studies, which highlight subtle changes in liver tissue by using the contrast agent, thereby improving the performance of early fibrosis assessment.[18,19] Moreover, due to the incorporation of a substantial number of NC-MRI in this study for model training, our developed model achieved superior performance (AUC: 0.918) in predicting cirrhosis. Our model consistently maintained robust performance (AUC: 0.925) in an external validation cohort comprising six centers, highlighting its stability and

reliability. In two recent studies, Ni et al.[26] demonstrated that a radiomics strategy utilizing non-contrast T1WI within a rodent model proficiently staged liver fibrosis, while He et al.[27] found that integrating clinical data with radiomic features of T2FS resulted in fair-to-good diagnostic precision for categorically assessing liver stiffness measured by MRE. However, these two studies were conducted with a smaller sample size from a single center, akin to previous GA-MRI studies, lacking further multi-center validation. When applying AI methods, particularly deep learning models, a larger sample size is generally advantageous for capturing a wide range of variability in the data and improving the model's generalizability. Compared to previous studies, the sample size included in our study is significantly larger, providing a robust foundation for training and validating our models.

There are limitations that require acknowledgment in this study. Primarily, it was a retrospective analysis utilizing datasets from six centers, introducing risks of selection bias and inherent biases. Future prospective studies with larger-scale datasets will be necessary to further validate our models. Secondly, the dataset used for model development was unbalanced with respect to pathologic fibrosis grades and included data from patients with liver tumors. This imbalance could potentially affect the performance of our models.[65] Ideally, employing a development dataset that features a large volume of MRI data balanced across different fibrosis grades could further enhance model performance. Therefore, we will consider utilizing more advanced algorithms or acquiring more ideal data in future studies to address this issue. Thirdly, a recent study demonstrated that Couinaud's liver segmentation is feasible using non-contrast T1-VIBE Dixon imaging.[66] Nonetheless, our current study did not include these features, which have proven valuable for fibrosis assessment.[64,67] To improve the performance of our model, we will incorporate liver segment features into subsequent research. Finally, the performance of AI models can further benefit from even larger and more diverse datasets. Future studies could expand on our work by incorporating additional patient data from varied geographic and clinical settings to further strengthen the model's generalizability.

In conclusion, our proposed models, integrating clinical features with NC-MRI, including T1WI and T2FS, have demonstrated promising diagnostic performance for staging liver fibrosis. While extensive clinical validation is still needed, our study offers valuable insights for the screening and regular management of fibrosis in chronic liver disease.

Contributors

Chunli Li, Yuan Wang, Ruobing Bai, and Yu Shi conceived and designed the study. Zhiyong Zhao, Qi Liu, Wenjuan Li, Qianqian Zhang, Chaoya Zhang, Wei Yang, Na Su, Fan Wang, Liuhanxu Shen, Chengli Gu, Baihe Luo, Minghui Zhou and Aoran Yang did the data collection. Ruobing Bai and Yueyue Lu performed image annotations. Zhiying Wang, Chengli Gu, Chen Pan, Xiaoli Yin, Ruobing Bai and Yu Shi reviewed images and performed diagnostic analysis. Chunli Li, Jiandong Yin, Yang Hou, and Yu Shi contributed to development of methodology. Qijun Wu performed statistical and computational analysis of data. Chunli Li, Yuan Wang, and Ruobing Bai wrote the original draft manuscript. Yang Hou and Yu Shi reviewed and performed revision of the manuscript. All authors had full access to all the data and verified the underlying data. All authors read and approved the final manuscript for publication.

Data sharing statement

To protect patient privacy, MRI data and other patient-related information are not publicly accessible. Nonetheless, all data and algorithm source codes used in this study are available upon reasonable request by contacting the corresponding author. The program codes are accessible online (https://github.com/CLL-CMU/StagingLiverFibrosis/).

Declaration of interests

We declare no competing interests related to this study.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 82071885); General Program of the Liaoning Provincial Department of Education (LJKMZ20221160); Liaoning Province Science and Technology Joint Plan (2023JH2/101700127); the Leading Young Talent Program of Xingliao Yingcai in Liaoning Province (XLYC2203037).

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2024.102881.

References

1 Caligiuri A, Gentilini A, Pastore M, Gitto S, Marra F. Cellular and molecular mechanisms underlying liver fibrosis regression. *Cells*. 2021;10(10):2759.
2 Kisseleva T, Brenner D. Molecular and cellular mechanisms of liver fibrosis and its regression. *Nat Rev Gastroenterol Hepatol*. 2021;18(3):151–166.
3 Ginès P, Krag A, Abraldes JG, Solà E, Fabrellas N, Kamath PS. Liver cirrhosis. *Lancet*. 2021;398(10308):1359–1376.
4 Tapper EB, Ufere NN, Huang DQ, Loomba R. Review article: current and emerging therapies for the management of cirrhosis and its complications. *Aliment Pharmacol Ther*. 2022;55(9):1099–1115.
5 Huang DQ, Tan DJH, Ng CH, et al. Hepatocellular carcinoma incidence in alcohol-associated cirrhosis: systematic review and meta-analysis. *Clin Gastroenterol Hepatol*. 2023;21(5):1169–1177.
6 Huang DQ, Terrault NA, Tacke F, et al. Global epidemiology of cirrhosis - aetiology, trends and predictions. *Nat Rev Gastroenterol Hepatol*. 2023;20(6):388–398.
7 Standish RA, Cholongitas E, Dhillon A, Burroughs AK, Dhillon AP. An appraisal of the histopathological assessment of liver fibrosis. *Gut*. 2006;55(4):569–578.
8 Maharaj B, Maharaj RJ, Leary WP, et al. Sampling variability and its influence on the diagnostic yield of percutaneous needle biopsy of the liver. *Lancet*. 1986;1(8480):523–525.
9 Pepin KM, Welle CL, Guglielmo FF, Dillman JR, Venkatesh SK. Magnetic resonance elastography of the liver: everything you need to know to get started. *Abdom Radiol*. 2022;47(1):94–114.
10 Lefebvre T, Wartelle-Bladou C, Wong P, et al. Prospective comparison of transient, point shear wave, and magnetic resonance elastography for staging liver fibrosis. *Eur Radiol*. 2019;29(12):6477–6488.
11 Selvaraj EA, Mózes FE, Jayaswal ANA, et al. Diagnostic accuracy of elastography and magnetic resonance imaging in patients with NAFLD: a systematic review and meta-analysis. *J Hepatol*. 2021;75(4):770–785.

12 Li C, Zhu H, Wang Y, Gu Y, Shi Y. Three-dimensional magnetic resonance elastography in chronic liver disease. *Port Hypertens Cirrhos*. 2023;2(1):32–39.
13 Zheng T, Qu Y, Chen J, et al. Noninvasive diagnosis of liver cirrhosis: qualitative and quantitative imaging biomarkers. *Abdom Radiol*. 2024;49(6):2098–2115.
14 Ginès P, Castera L, Lammert F, et al. Population screening for liver fibrosis: toward early diagnosis and intervention for chronic liver diseases. *Hepatology*. 2022;75(1):219–228.
15 Patel K, Asrani SK, Fiel MI, et al. Accuracy of blood-based biomarkers for staging liver fibrosis in chronic liver disease: a systematic review supporting the AASLD practice guideline. *Hepatology*. 2024.
16 Alzoubi O, Arar A, Singh V, Erturk SM, Mozes F, Pavlides M. MRI in liver cirrhosis. *Port Hypertens Cirrhos*. 2022;1(1):23–41.
17 Kim JY, Lee SS, Byun JH, et al. Biologic factors affecting HCC conspicuity in hepatobiliary phase imaging with liver-specific contrast agents. *AJR Am J Roentgenol*. 2013;201(2):322–331.
18 Aguirre DA, Behling CA, Alpert E, Hassanein TI, Sirlin CB. Liver fibrosis: noninvasive diagnosis with double contrast material-enhanced MR imaging. *Radiology*. 2006;239(2):425–437.
19 Feier D, Balassy C, Bastati N, Stift J, Badea R, Ba-Ssalamah A. Liver fibrosis: histopathologic and biochemical influences on diagnostic efficacy of hepatobiliary contrast-enhanced MR imaging in staging. *Radiology*. 2013;269(2):460–468.
20 Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology*. 2016;278(2):563–577.
21 Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014;5:4006.
22 Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts H. Artificial intelligence in radiology. *Nat Rev Cancer*. 2018;18(8):500–510.
23 Yasaka K, Akai H, Kunimatsu A, Abe O, Kiryu S. Liver fibrosis: deep convolutional neural network for staging by using Gadoxetic acid-enhanced hepatobiliary phase MR images. *Radiology*. 2018;287(1):146–155.
24 Hectors SJ, Kennedy P, Huang KH, et al. Fully automated prediction of liver fibrosis using deep learning analysis of gadoxetic acid-enhanced MRI. *Eur Radiol*. 2021;31(6):3805–3814.
25 Park HJ, Lee SS, Park B, et al. Radiomics analysis of Gadoxetic acid-enhanced MRI for staging liver fibrosis. *Radiology*. 2019;290(2):380–387.
26 Ni M, Wang L, Yu H, et al. Radiomics approaches for predicting liver fibrosis with nonenhanced T_1-weighted imaging: comparison of different radiomics models. *J Magn Reson Imaging*. 2021;53(4):1090–1091.
27 He L, Li H, Dudley JA, et al. Machine learning prediction of liver stiffness using clinical and T2-weighted MRI radiomic data. *AJR Am J Roentgenol*. 2019;213(3):592–601.
28 Yang Y, Zou X, Zhou W, et al. Multiparametric MRI-based radiomic signature for preoperative evaluation of overall survival in intrahepatic cholangiocarcinoma after partial hepatectomy. *J Magn Reson Imaging*. 2022;56(3):739–751.
29 Gao W, Wang W, Song D, et al. A multiparametric fusion deep learning model based on DCE-MRI for preoperative prediction of microvascular invasion in intrahepatic cholangiocarcinoma. *J Magn Reson Imaging*. 2022;56(4):1029–1039.
30 Bedossa P, Poynard T. An algorithm for the grading of activity in chronic hepatitis C. The METAVIR cooperative study group. *Hepatology*. 1996;24(2):289–293.
31 Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18(2):203–211.
32 Li Y, Yao T, Pan Y, Mei T. Contextual transformer networks for visual recognition. *IEEE Trans Pattern Anal Mach Intell*. 2023;45(2):1489–1500.
33 Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. *IEEE Conf Comput Vis Pattern Recognit*. 2016:2921–2929.
34 Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30:4768–4777.
35 Bonekamp S, Kamel I, Solga S, Clark J. Can imaging modalities diagnose and stage hepatic fibrosis and cirrhosis accurately? *J Hepatol*. 2009;50(1):17–35.
36 Rustogi R, Horowitz J, Harmath C, et al. Accuracy of MR elastography and anatomic MR imaging features in the diagnosis of severe hepatic fibrosis and cirrhosis. *J Magn Reson Imaging*. 2012;35(6):1356–1364.
37 Choi KJ, Jang JK, Lee SS, et al. Development and validation of a deep learning system for staging liver fibrosis by using contrast agent–enhanced CT images in the liver. *Radiology*. 2018;289(3):688–697.
38 Bashir MR, Horowitz JM, Kamel IR, et al. ACR appropriateness criteria® chronic liver disease. *J Am Coll Radiol*. 2020;17(5):S70–S80.
39 Martí-Bonmatí L, Delgado F. MR imaging in liver cirrhosis: classical and new approaches. *Insights Imaging*. 2010;1:233–244.
40 Wai CT, Greenson JK, Fontana RJ, et al. A simple noninvasive index can predict both significant fibrosis and cirrhosis in patients with chronic hepatitis C. *Hepatology*. 2003;38(2):518–526.
41 Sterling RK, Lissen E, Clumeck N, et al. Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. *Hepatology*. 2006;43(6):1317–1325.
42 Chen B, Ye B, Zhang J, Ying L, Chen Y. RDW to platelet ratio: a novel noninvasive index for predicting hepatic fibrosis and cirrhosis in chronic hepatitis B. *PLoS One*. 2013;8(7):e68780.
43 Lemoine M, Shimakawa Y, Nayagam S, et al. The gamma-glutamyl transpeptidase to platelet ratio (GPR) predicts significant liver fibrosis and cirrhosis in patients with chronic HBV infection in West Africa. *Gut*. 2016;65(8):1369–1376.
44 Cross TJ, Rizzi P, Berry PA, Bruce M, Portmann B, Harrison PM. King's score: an accurate marker of cirrhosis in chronic hepatitis C. *Eur J Gastroenterol Hepatol*. 2009;21(7):730–738.
45 Kundel HL, Polansky M. Measurement of observer agreement. *Radiology*. 2003;228(2):303–308.
46 Dice LR. Measures of the amount of ecologic association between species. *Ecology*. 1945;26(3):297–302.
47 Han X, Wu X, Wang S, et al. Automated segmentation of liver segment on portal venous phase MR images using a 3D convolutional neural network. *Insights Imaging*. 2022;13(1):26.
48 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837–845.
49 Kramer AA, Zimmerman JE. Assessing the calibration of mortality benchmarks in critical care: the hosmer-lemeshow test revisited. *Crit Care Med*. 2007;35(9):2052–2056.
50 Vickers AJ, Cronin AM, Elkin EB, Gonen M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Med Inform Decis Mak*. 2008;8(1):53.
51 Xu X, Su Y, Song R, et al. Performance of transient elastography assessing fibrosis of single hepatitis B virus infection: a systematic review and meta-analysis of a diagnostic test. *Hepatol Int*. 2015;9(4):558–566.
52 Marcellin P, Ziol M, Bedossa P, et al. Non-invasive assessment of liver fibrosis by stiffness measurement in patients with chronic hepatitis B. *Liver Int*. 2009;29(2):242–247.
53 Muratori L, Lohse AW, Lenzi M. Diagnosis and management of autoimmune hepatitis. *BMJ*. 2023;380:e070201.
54 Tannapfel A, Denk H, Dienes HP, et al. Histopathological diagnosis of non-alcoholic and alcoholic fatty liver disease. *Virchows Arch*. 2011;458(5):511–523.
55 Kleiner DE, Makhlouf HR. Histology of nonalcoholic fatty liver disease and nonalcoholic steatohepatitis in adults and children. *Clin Liver Dis*. 2016;20(2):293–312.
56 Fang XH, Xie M, Zhao Y, et al. Both APRI and FIB-4 could effectively evaluate liver fibrosis in liver transplantation recipients. *Port Hypertens Cirrhos*. 2022;1(3):197–199.
57 Boursier J, Vergniol J, Guillet A, et al. Diagnostic accuracy and prognostic significance of blood fibrosis tests and liver stiffness measurement by FibroScan in non-alcoholic fatty liver disease. *J Hepatol*. 2016;65(3):570–578.
58 Targher G, Byrne CD, Tilg H. MASLD: a systemic metabolic disorder with cardiovascular and malignant complications. *Gut*. 2024;73(4):691–702.
59 Degos F, Perez P, Roche B, et al. Diagnostic accuracy of FibroScan and comparison to liver fibrosis biomarkers in chronic viral hepatitis: a multicenter prospective study (the FIBROSTIC study). *J Hepatol*. 2010;53(6):1013–1021.
60 Rui F, Xu L, Yeo YH, et al. Machine learning-based models for advanced fibrosis and cirrhosis diagnosis in chronic hepatitis B

patients with hepatic steatosis. *Clin Gastroenterol Hepatol*. 2024;S1542-3565(24):00553–005536.
61 Sarvestany SS, Kwong JC, Azhie A, et al. Development and validation of an ensemble machine learning framework for detection of all-cause advanced hepatic fibrosis: a retrospective cohort study. *Lancet Digit Health*. 2022;4(3):e188–e199.
62 Smith AD, Branch CR, Zand K, et al. Liver surface nodularity quantification from routine CT images as a biomarker for detection and evaluation of cirrhosis. *Radiology*. 2016;280(3):771–781.
63 Lo GC, Besa C, King MJ, et al. Feasibility and reproducibility of liver surface nodularity quantification for the assessment of liver cirrhosis using CT and MRI. *Eur J Radiol Open*. 2017;4:95–100.
64 Pickhardt PJ, Malecki K, Hunt OF, et al. Hepatosplenic volumetric assessment at MDCT for staging liver fibrosis. *Eur Radiol*. 2017;27(7):3060–3068.
65 He H, Garcia EA. Learning from imbalanced data. *IEEE T Knowl Data En*. 2009;21(9):1263–1284.
66 Zbinden L, Catucci D, Suter Y, et al. Automated liver segmental volume ratio quantification on non-contrast T1–vibe dixon liver MRI using deep learning. *Eur J Radiol*. 2023;167:111047.
67 Lee S, Elton DC, Yang AH, et al. Fully automated and explainable liver segmental volume ratio and spleen segmentation at CT for diagnosing cirrhosis. *Radiol Artif Intell*. 2022;4(5):e210268.