Original Article

Deep learning-based automated assessment of hepatic fibrosis via magnetic resonance images and nonimage data

Weixia Li[1#^], Yajing Zhu[2#], Gangde Zhao[3#], Xiaoyan Chen[4], Xiangtian Zhao[5], Haimin Xu[4], Yingyu Che[6], Yinan Chen[2,7], Yuxiang Ye[2], Xin Dou[8], Hui Wang[3], Jingliang Cheng[6], Qing Xie[3], Kemin Chen[1]

[1]Department of Radiology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China; [2]SenseTime Research, SenseTime, Shanghai, China; [3]Department of Infectious Diseases, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China; [4]Department of Pathology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China; [5]Department of Radiology, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangdong, China; [6]Department of Magnetic Resonance Imaging, the First Affiliated Hospital of Zhengzhou University, Zhengzhou, China; [7]WCH-SenseTime Joint Lab, SenseTime, Chengdu, China; [8]SenseBrain Technology, SenseTime, Princeton, NJ, USA

Contributions: (I) Conception and design: W Li, K Chen; (II) Administrative support: K Chen, J Cheng, Q Xie; (III) Provision of study materials or patients: X Chen, H Xu, H Wang; (IV) Collection and assembly of data: W Li, G Zhao, X Zhao, Y Che; (V) Data analysis and interpretation: W Li, Y Zhu, Y Chen, Y Ye, X Dou; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Kemin Chen, MD. Department of Radiology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, No. 197 Ruijin Er Road, Huangpu District, Shanghai 200025, China. Email: keminchenrj@163.com; Qing Xie, MD. Department of Infectious Diseases, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, No. 197 Ruijin Er Road, Huangpu District, Shanghai 200025, China. Email: xq10403@rjh.com.cn.

Background: Accurate staging of hepatic fibrosis is critical for prognostication and management among patients with chronic liver disease, and noninvasive, efficient alternatives to biopsy are urgently needed. This study aimed to evaluate the performance of an automated deep learning (DL) algorithm for fibrosis staging and for differentiating patients with hepatic fibrosis from healthy individuals via magnetic resonance (MR) images with and without additional clinical data.

Methods: A total of 500 patients from two medical centers were retrospectively analyzed. DL models were developed based on delayed-phase MR images to predict fibrosis stages. Additional models were constructed by integrating the DL algorithm with nonimaging variables, including serologic biomarkers [aminotransferase-to-platelet ratio index (APRI) and fibrosis index based on four factors (FIB-4)], viral status (hepatitis B and C), and MR scanner parameters. Diagnostic performance, was assessed via the area under the receiver operating characteristic curve (AUROC), and comparisons were through use of the DeLong test. Sensitivity and specificity of the DL and full models (DL plus all clinical features) were compared with those of experienced radiologists and serologic biomarkers via the McNemar test.

Results: In the test set, the full model achieved AUROC values of 0.99 [95% confidence interval (CI): 0.94–1.00], 0.98 (95% CI: 0.93–0.99), 0.90 (95% CI: 0.83–0.95), 0.81 (95% CI: 0.73–0.88), and 0.84 (95% CI: 0.76–0.90) for staging F0–4, F1–4, F2–4, F3–4, and F4, respectively. This model significantly outperformed the DL model in early-stage classification (F0–4 and F1–4). Compared with expert radiologists, it showed superior specificity for F0–4 and higher sensitivity across the other four classification tasks. Both the DL and full models showed significantly greater specificity than did the biomarkers for staging advanced fibrosis (F3–4 and F4).

^ ORCID: 0000-0001-5945-6350.

Conclusions: The proposed DL algorithm provides a noninvasive method for hepatic fibrosis staging and screening, outperforming both radiologists and conventional biomarkers, and may facilitate improved clinical decision-making.

Keywords: Liver; fibrosis; magnetic resonance imaging (MRI); deep learning (DL); biomarker

Submitted Nov 10, 2024. Accepted for publication May 16, 2025. Published online Aug 18, 2025.
doi: 10.21037/qims-2024-2506
View this article at: https://dx.doi.org/10.21037/qims-2024-2506

Introduction

Chronic liver disease (CLD), a highly prevalent disease, with significant morbidity and mortality, leverages a massive burden in terms of lost lives and medical costs worldwide (1). CLD eventually results in hepatic fibrosis, which in turn can progress to cirrhosis and ultimately to end-stage liver disease (2). Studies suggest that the prognosis and management of patients with CLD can differ widely based on the progression of hepatic fibrosis. Therefore, accurate staging of hepatic fibrosis is an important clinical issue (3,4). Although liver biopsy is the current gold standard for the assessment of hepatic fibrosis, it is invasive and involves several drawbacks and complications (5,6). Consequently, it is necessary to find less invasive and more feasible methods for fibrosis assessment.

Several noninvasive methods have been investigated as alternatives for liver biopsy, including aminotransferase-to-platelet ratio index (APRI) and fibrosis index based on the four factors (FIB-4), but these serum fibrosis markers have low sensitivity in detecting the early stages of fibrosis (7). Meanwhile, ultrasound (US) elastography and magnetic resonance elastography (MRE) have been successfully implemented in staging hepatic fibrosis. However, these noninvasive assessments require well-trained professionals, whose failure rate has been reported to range from 6% to 23% (8), and dedicated equipment that is not readily available (9).

Recently, deep learning (DL) has attracted substantial attention as an artificial intelligence strategy (10). DL for automated liver segmentation might eliminate inter-reader variability in the assessment of liver fibrosis as compared with other quantitative methods that involve the selection of regions of interest by radiologists (11-13). There has also been a degree of initial success in the application of DL to the assessment of liver fibrosis on the basis of radiological images (12-20). However, to our knowledge, there is no literature on the performance of three-dimensional (3D) fully automated DL (FADL) in noninvasively staging hepatic fibrosis and distinguishing fibrosis from nonfibrosis based on dynamic contrast-enhanced magnetic resonance (MR) images. Given that MR does not involve ionizing radiation compared in contrast with computed tomography (CT), a 3D FADL model that enables liver fibrosis staging based on MR would benefit a large number of patients.

In our study, we hypothesized that 3D FADL could serve as a technique for the staging of hepatic fibrosis. Thus, this study aimed to evaluate the severity of hepatic fibrosis and distinguish patients with hepatic fibrosis from healthy individuals directly from delayed MR images via 3D FADL. In addition, the outputs of 3D FADL and certain nonimage information were combined to further improve the diagnostic performance. We present this article in accordance with the STARD reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-2024-2506/rc).

Methods

Patients

This two-center retrospective study was conducted in accordance with the Declaration of Helsinki and its subsequent amendments and was approved by the Ethics Committee of Ruijin Hospital, Shanghai Jiao Tong University School of Medicine (approval No. RJ2018-209), and by the Ethics Committee of the First Affiliated Hospital of Zhengzhou University (approval No. 2017-KY-01). The requirement for informed consent was waived due to the retrospective nature of the study. A total of 921 potentially eligible consecutive patients were enrolled from Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, and the First Affiliated Hospital of Zhengzhou University between March 2012 and April 2020. The patient enrollment procedure is displayed in *Figure 1*. The inclusion criteria were as follows: (I) patients with fibrosis who underwent dynamic contrast liver magnetic resonance imaging (MRI) completed within 6 months

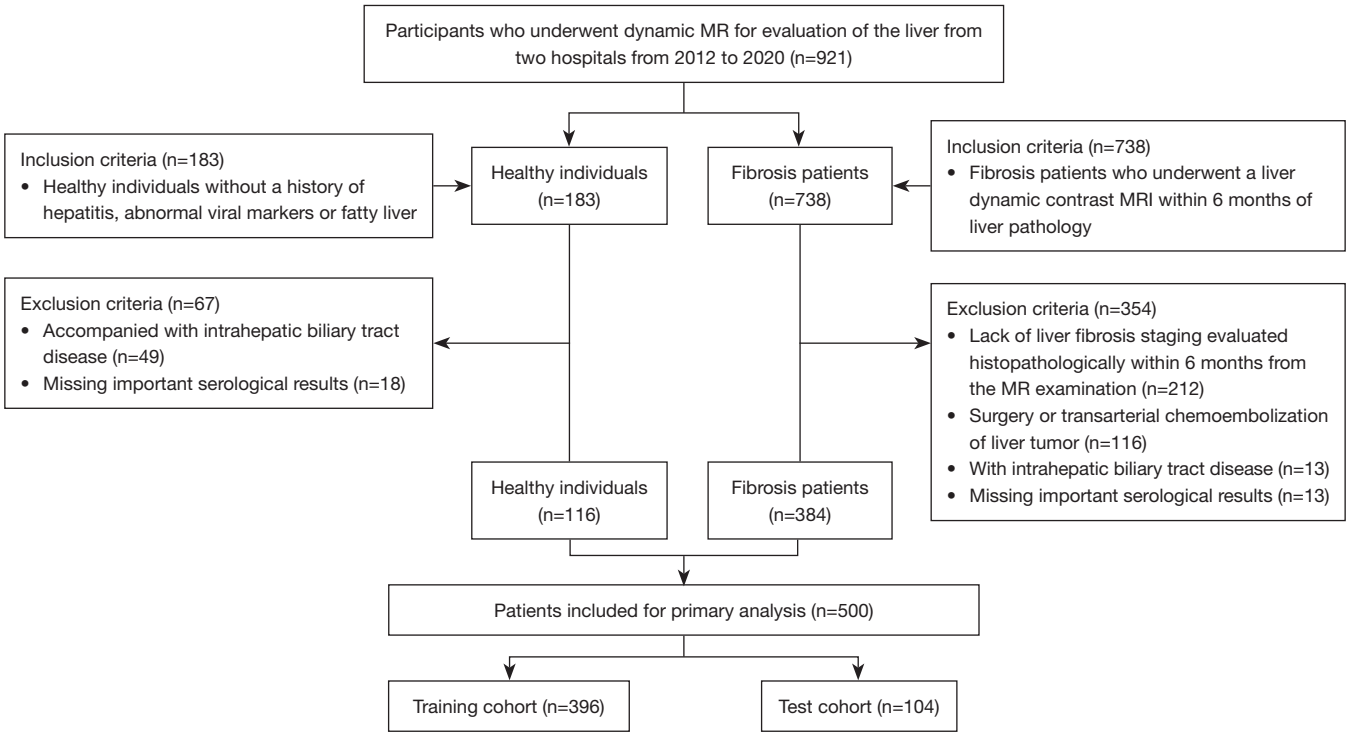

Figure 1 Flowchart of patient enrollment. In total, 500 out of 921 patients were enrolled in this study. MR, magnetic resonance; MRI, magnetic resonance imaging.

Table 1 Baseline characteristics of patients

Variables	Total patients (N=500)	Training cohort, N=396 (79.2%)	Test cohort, N=104 (20.8%)	P value
Age (years)	52.0 [40.0, 62.0]	52.0 [40.0, 61.0]	52.5 [43.5, 63.0]	0.704
Gender (female)	198 (39.6)	158 (39.8)	40 (38.5)	0.878
BMI (kg/m^2)	23.2 [21.1, 25.4]	23.2 [21.0, 25.7]	23.2 [21.3, 24.9]	0.628
HBV†	289 (57.8)	228 (57.6)	61 (58.7)	0.931
HCV	16 (3.2)	14 (3.5)	2 (1.9)	0.543
APRI	0.46 [0.28, 0.92]	0.45 [0.27, 0.90]	0.50 [0.29, 1.09]	0.401
FIB-4	1.71 [1.08, 3.01]	1.69 [1.05, 3.13]	1.76 [1.17, 2.71]	0.740

Qualitative variables are expressed as N (%), and quantitative variables are expressed as the median [IQR]. P values were calculated between the training and test cohorts. †, hepatitis B virus includes coinfection of HCV. APRI, aminotransferase-to-platelet ratio index; BMI, body mass index; FIB-4, fibrosis index based on four factors; HBV, hepatitis B virus; HCV, hepatitis C virus; IQR, interquartile range.

of diagnosis liver pathology and (II) healthy individuals without history of hepatitis, abnormal viral markers, or fatty liver. The exclusion criteria for the patients with fibrosis were as follows: (I) no confirmed fibrosis stage within 6 months of an MR examination; (II) chemotherapy or surgery performed before MR examination; (III) lack of key serological results; and (IV) intrahepatic biliary tract disease. The exclusion criteria for the control participants were those with intrahepatic biliary tract disease.

The patients who were ultimately enrolled were randomly divided into a training cohort and test cohort at a ratio of approximately 4:1 at each fibrosis stage among the patients. The detailed participant characteristics are shown in *Table 1*.

Table 2 Representative MR imaging information used for the training and test datasets

Parameter	Total patients	Training cohort	Test cohort	P value
Static magnetic field				0.406
1.5 T	280 (56.0)	226 (57.1)	54 (51.9)	
3.0 T	220 (44.0)	170 (42.9)	50 (48.1)	
MR manufacturer				0.867
GE	121 (24.2)	98 (24.7)	23 (22.1)	
Philips	195 (39.0)	151 (38.1)	44 (42.3)	
Siemens	81 (16.2)	64 (16.2)	17 (16.3)	
UIH uMR	103 (20.6)	83 (21.0)	20 (19.2)	
TR (ms)	3.74 [3.65, 4.70]	3.74 [3.65, 4.70]	3.73 [3.65, 4.70]	0.838
TE (ms)	1.74 [1.32, 2.20]	1.75 [1.32, 2.20]	1.74 [1.32, 2.20]	0.413
Matrix	258 [217, 288] × 200 [195, 252]	256 [217, 288] × 200 [195, 252]	264 [252, 288] × 224 [195, 252]	–
Flip angle (degree)	10 [10, 10]	10 [10, 10]	10 [10, 11]	0.442
Section thickness (mm)	3.00 [2.00, 4.40]	3.00 [2.00, 4.40]	3.00 [2.00, 4.40]	0.163
Section interval (mm)	2.50 [2.00, 3.00]	2.50 [2.00, 3.00]	2.50 [2.00, 3.00]	0.849
Bandwidth (Hz)	360 [345, 1,340]	360 [345, 1,340]	360 [345, 1,340]	0.224

Qualitative variables are expressed as N (%), and quantitative variables are expressed as the median [IQR]. P values were calculated between the training and test cohorts. IQR, interquartile range; MR, magnetic resonance; TE, echo time; TR, repetition time.

Pathology reference standards

The reference standard was obtained from biopsy or surgery in 109 patients and 270 patients in the training group, respectively, and in 27 and 69 patients in the test group, respectively. The median time between the MR examination and biopsy was 3 [IQR, interquartile range (IQR), 1–15] days for the training group and 12 (IQR, 3–34) days for the test group. The median time between the MR examination and surgery was 5 (IQR, 4–8) and 6 (IQR, 3–11) days for the training and test groups, respectively. Pathologic assessment of liver fibrosis was performed according to Scheuer Scoring System (Appendix 1) (21). In addition, we enrolled 116 control participants [grouped as normal (N)] who differed from patients with F0, because some of these patients still had liver inflammation (with A ≥1) or abnormal liver function even when the fibrosis stage was F0 (Table S1), indicating that there may be some damage to the liver. The healthy individuals had no history of hepatitis, abnormal viral hepatitis markers, or fatty liver detected on MRI-derived proton density fat fraction (MRI-PDFF <5%) (22). They had liver metastasis from colorectal cancer, definitively benign liver tumor, and pancreatic or duodenal tumor (Table S2). Liver fibrosis stages were pathologically determined from the peritumoral liver parenchyma of patients with liver tumors or from percutaneous liver biopsy samples in patients with hepatitis-related fibrosis (Table S3). In our study, five separate models were built to perform the following binary classification of liver stages: N *vs.* F0–4, N–F0 *vs.* F1–4, N–F1 *vs.* F2–4, N–F2 *vs.* F3–4, and N–F3 *vs.* F4. The detailed information is shown in Table S1.

Input data: *MRI technique*

Dynamic contrast-enhanced MR images were obtained with MRI units from four manufacturers (GE HealthCare, Chicago, IL, USA; Siemens Healthineers, Erlangen, Germany; Philips Healthcare, Best, the Netherlands; United Imaging Healthcare, Shanghai, China). Gadopentetate dimeglumine (Gd-DTPA) was injected at a dose of 0.1 mmol of gadolinium per kilogram of body weight. The delayed phase of dynamic contrast-enhanced MR images was selected for our study since this phase is more associated with the bright-appearing reticulations suggestive of hepatic fibrosis as compared with other phases (23). The time interval between injection and acquisition of delayed phase was 3 minutes. The parameters used to acquire the images are summarized in *Table 2*.

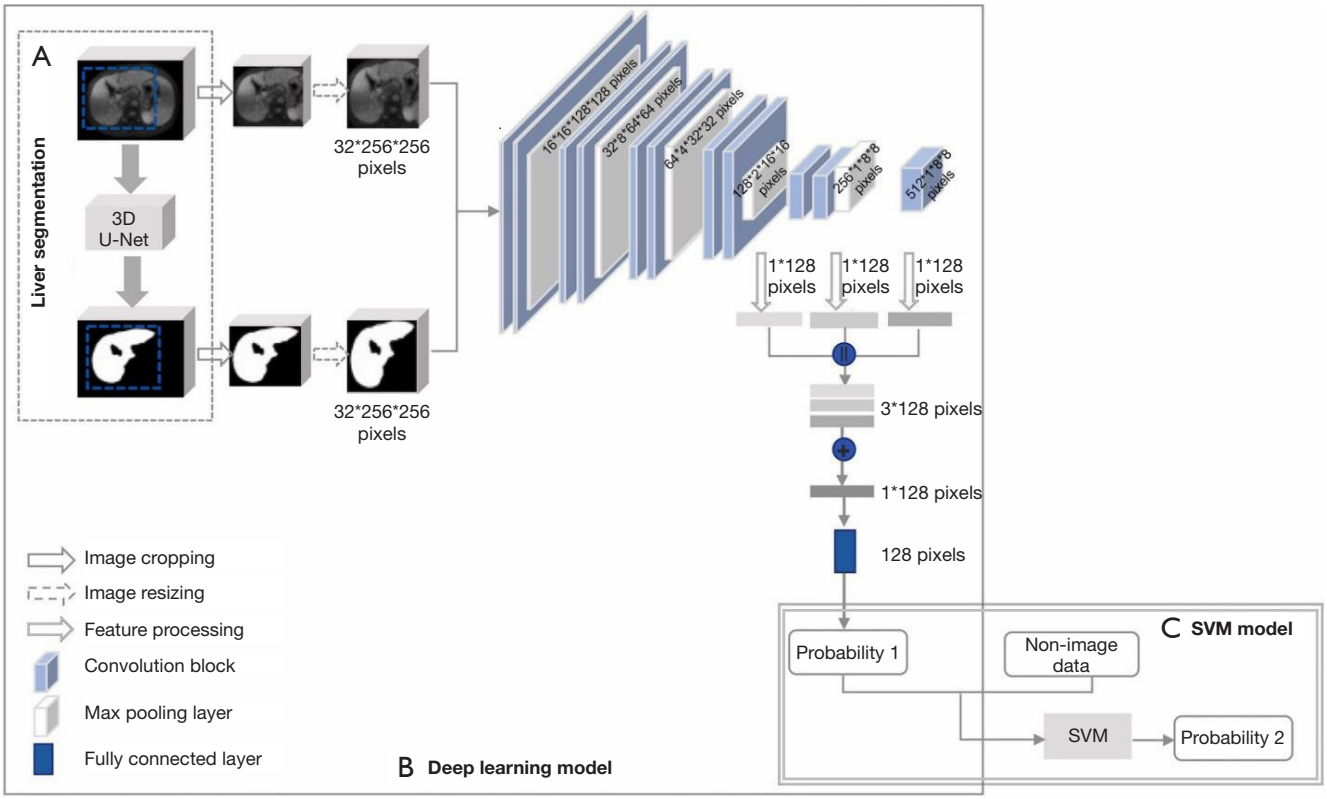

Figure 2 The schematics of two different models for staging liver fibrosis. (A) The input MR images were first processed with the self-developed 3D U-Net algorithm for liver segmentation (dashed border box). (B) MR images were cropped within the segmented liver region and then fed into the deep learning model (solid border box). (C) The establishment procedure of SVM model using the output of DL model and nonimage data as the input (double border box). 3D, three-dimensional; DL, deep learning; MR, magnetic resonance; SVM, support vector machine; U-Net, convolutional networks for biomedical image segmentation.

Input data: nonimage information

Serological examinations were performed within 2 weeks of MR examination. Platelet count, aspartate aminotransferase (AST) level, and alanine aminotransferase (ALT) level were recorded. APRI was calculated as follows: APRI = [(AST/upper limit of normal AST) ×100]/platelet count (10^9/L). Meanwhile, FIB-4 was calculated as follows: FIB-4 = [age (years)] × [AST (U/L)]/[platelet count (10^9/L)] × [ALT (U/L)$^{1/2}$] (24,25).

The following parameters were also collected for our analysis: (I) MR information including manufactures and static magnetic field (0=1.5 T and 1=3.0 T); (II) hepatitis B virus (HBV) status (0= absent and 1= present) and hepatitis C virus (HCV) status (0= absent and 1= present) judged according to the positivity of HBV surface antigen and HCV antibody, respectively. These parameters were included because MR information affects the quality of the MR images (26), and the presence of HBV and HCV infections are associated with certain subtypes of liver cirrhosis (27). The statistical details of these parameters are displayed in *Tables 1*,*2*.

Model construction and evaluation

Data preprocessing

In this study, an investigator with extensive experience in DL who was blinded to the pathological fibrosis staging and clinical information employed a self-developed 3D U-Net model to segment the liver surface (*Figure 2A*). Subsequently, she performed region of interest (ROI) extractions on the delayed MR image and liver surface for all participants, including both the control and liver fibrosis groups (*Figure 2B*). The extracted ROIs served as inputs of the 3D FADL model (DL model) (Appendix 2).

Model construction and evaluation via MR delayed images and 3D FADL

Separate 3D FADL models were built to perform the five binary classification tasks for the liver fibrosis stages. Each of the classification models was subjected to data preprocessing procedures (*Figure 2A*), augmentation, model training (*Figure 2B*), and model evaluation (Appendix 2).

Model construction and evaluation from the outputs of DL model and nonimage data with support vector machine (SVM) algorithm

In addition to DL model, we also used the SVM algorithm to build models by combining the outputs of DL model with nonimage data (*Figure 2C*). The combined models included (I) model A—DL model combined with APRI and FIB-4); (II) model B—DL model combined with virus status (HBV and HCV status); (III) model C—DL model combined with MR information (MR manufacturers and static field strength); (IV) model D—DL model combined with biomarkers and virus status; (V) model E—DL model combined with biomarkers and MR information; (VI) model F—DL model combined with virus status and MR information; and (VII) full model—DL model combined with biomarkers, virus status, and MR information.

The models were built for five separate binary classification tasks: N *vs.* F0–4, N–F0 *vs.* F1–4, N–F1 *vs.* F2–4, N–F2 *vs.* F3–4, and N–F3 *vs.* F4. The combined models were trained from the training data and fivefold validation strategy and were finally evaluated on the test data.

Radiologists' visual grading of hepatic fibrosis

For test data, two abdominal radiologists (A and B), each with 14 and 13 years of relevant experience and blinded to the pathological fibrosis staging and clinical information, independently reviewed the anonymized MR delayed images using a Digital Imaging and Communications in Medicine (DICOM) viewer. The readers graded the degree of liver fibrosis using a six-point scale (Appendix 3).

Statistical analysis

Descriptive statistics are summarized as the median and IQR. In the test cohort, the area under the receiver operating characteristic curve (AUROC) was calculated for each model's ability to diagnose fibrosis stages F0–4, F1–4, F2–4, F3–4, and F4, and AUROC values and accuracy were compared via the DeLong test (28) and McNemar test. Accuracy, sensitivity, and specificity were calculated, and the modified Wilson method was applied to determine the 95% confidence intervals (CIs) (29). The inter-reader agreement between the two radiologist and the consistency between the radiologists, DL model, full model, and the gold standard in liver fibrosis staging tasks were assessed with the Cohen kappa coefficient (30).

Statistical analyses were conducted with R software version 3.3.3 (The R Foundation for Statistical Computing), and graphs were drawn with GraphPad Prism version 9.0.0 (GraphPad Software, Siemens Healthineers, Erlangen, Germany; www.graphpad.com) for MacOS. A P value <0.05 indicated a statistical significance.

In addition, the t-distributed stochastic neighbor embedding (t-SNE) technique was used to reduce the dimensionality and facilitate the visualization of different fibrosis stages (Appendix 4) (31). The t-SNE visualization was performed with the Python 3.7 "sklearn" package.

Results

Baseline characteristics

The characteristics of the study population are summarized in *Table 1*. There were no significant differences in age, gender, body mass index, the presence of HBV and HCV, or biomarkers between the training and test cohorts.

Overall diagnostic performance of the DL model and combined models (A–F model and full model)

On the test cohort, full model offered significantly higher AUROCs for fibrosis stages F0–4 and F1–4, with AUROCs of 0.99 and 0.98, respectively, as compared with the DL model. There were no significant differences in the AUROCs for fibrosis stages F2–4, F3–4, or F4 between the full model and DL model. The detailed results of the diagnostic performance of these models, including the DL model, full model, and the A–F model [representing the collective group of combined models A through F described in section "Model construction and evaluation from the outputs of DL model and nonimage data with support vector machine (SVM) algorithm"], are displayed in *Table 3*. The visualization of their diagnostic performance is illustrated in *Figure 3*.

t-SNE visualization results

The internal features learned by the DL model were

Table 3 Diagnostic performance of the different models for staging liver fibrosis on the test set

Model and parameter	Accuracy (%)	AUROC	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	LR+	LR–	P value
N *vs.* F0–4									
DL model	88.5 (80.9, 93.3)	0.92 (0.85, 0.96)	88.8 (80.0, 94.0)	87.5 (69.0, 95.7)	95.9 (88.7, 98.9)	70.0 (52.1, 83.3)	7.1 (2.46, 20.52)	0.13 (0.08, 0.2)	0.003**
Model A	93.3 (86.8, 96.7)	0.95 (0.89, 0.98)	95.0 (87.8, 98.0)	87.5 (69.0, 95.7)	96.2 (89.4, 99.0)	84.0 (65.3, 93.6)	7.6 (2.63, 21.93)	0.06 (0.03, 0.1)	0.035*
Model B	91.3 (84.4, 95.4)	0.96 (0.9, 0.98)	91.2 (83.0, 95.7)	91.7 (74.2, 98.5)	97.3 (90.8, 99.5)	75.9 (57.9, 87.8)	10.95 (2.9, 41.35)	0.1 (0.06, 0.15)	0.034*
Model C	90.4 (83.2, 94.7)	0.92 (0.85, 0.96)	91.2 (83.0, 95.7)	87.5 (69.0, 95.7)	96.1 (89.0, 98.9)	75.0 (56.6, 87.3)	7.3 (2.53, 21.09)	0.1 (0.06, 0.16)	0.012*
Model D	97.1 (91.9, 99.2)	0.98 (0.93, 0.99)	100.0 (95.4, 100.0)	87.5 (69.0, 95.7)	96.4 (89.9, 99.0)	100.0 (84.5, 100.0)	8.0 (2.78, 23.06)	0 (0, 0)	0.45
Model E	93.3 (86.8, 96.7)	0.95 (0.89, 0.98)	95.0 (87.8, 98.0)	87.5 (69.0, 95.7)	96.2 (89.4, 99.0)	84.0 (65.3, 93.6)	7.6 (2.63, 21.93)	0.06 (0.03, 0.1)	0.041*
Model F	94.2 (88.0, 97.3)	0.96 (0.9, 0.98)	96.2 (89.5, 99.0)	87.5 (69.0, 95.7)	96.2 (89.5, 99.0)	87.5 (69.0, 95.7)	7.7 (2.67, 22.21)	0.04 (0.02, 0.08)	0.02*
Full model	97.1 (91.9, 99.2)	0.99 (0.94, 1.0)	100.0 (95.4, 100.0)	87.5 (69.0, 95.7)	96.4 (89.9, 99.0)	100.0 (84.5, 100.0)	8.0 (2.78, 23.06)	0 (0, 0)	
N–F0 *vs.* F1–4									
DL model	91.3 (84.4, 95.4)	0.91 (0.85, 0.95)	92.4 (84.4, 96.5)	88.0 (70.0, 95.8)	96.1 (89.0, 98.9)	78.6 (60.5, 89.8)	7.7 (2.66, 22.3)	0.09 (0.05, 0.14)	0.021*
Model A	94.2 (88.0, 97.3)	0.96 (0.91, 0.99)	96.2 (89.4, 99.0)	88.0 (70.0, 95.8)	96.2 (89.4, 99.0)	88.0 (70.0, 95.8)	8.02 (2.77, 23.2)	0.04 (0.02, 0.08)	0.316
Model B	94.2 (88.0, 97.3)	0.96 (0.9, 0.98)	96.2 (89.4, 99.0)	88.0 (70.0, 95.8)	96.2 (89.4, 99.0)	88.0 (70.0, 95.8)	8.02 (2.77, 23.2)	0.04 (0.02, 0.08)	0.109
Model C	91.3 (84.4, 95.4)	0.94 (0.87, 0.97)	92.4 (84.4, 96.5)	88.0 (70.0, 95.8)	96.1 (89.0, 98.9)	78.6 (60.5, 89.8)	7.7 (2.66, 22.3)	0.09 (0.05, 0.14)	0.011*
Model D	97.1 (91.9, 99.2)	0.98 (0.93, 1.0)	100.0 (95.4, 100.0)	88.0 (70.0, 95.8)	96.3 (89.8, 99.0)	100.0 (85.1, 100.0)	8.33 (2.88, 24.09)	0 (0, 0)	0.165
Model E	94.2 (88.0, 97.3)	0.96 (0.91, 0.99)	96.2 (89.4, 99.0)	88.0 (70.0, 95.8)	96.2 (89.4, 99.0)	88.0 (70.0, 95.8)	8.02 (2.77, 23.2)	0.04 (0.02, 0.08)	0.205
Model F	93.3 (86.8, 96.7)	0.97 (0.91, 0.99)	94.9 (87.7, 98.0)	88.0 (70.0, 95.8)	96.2 (89.3, 99.0)	84.6 (66.5, 93.8)	7.91 (2.73, 22.9)	0.06 (0.03, 0.1)	0.169
Full model	97.1 (91.9, 99.2)	0.98 (0.93, 0.99)	100.0 (95.4, 100.0)	88.0 (70.0, 95.8)	96.3 (89.8, 99.0)	100.0 (85.1, 100.0)	8.33 (2.88, 24.09)	0 (0, 0)	
N–F1 *vs.* F2–4									
DL model	80.8 (72.2, 87.2)	0.86 (0.78, 0.91)	81.7 (71.2, 89.0)	78.8 (62.2, 89.3)	89.2 (79.4, 94.7)	66.7 (51.0, 79.4)	3.85 (1.98, 7.5)	0.23 (0.15, 0.36)	0.061
Model A	81.7 (73.2, 88.0)	0.86 (0.78, 0.91)	83.1 (72.7, 90.1)	78.8 (62.2, 89.3)	89.4 (79.7, 94.8)	68.4 (52.5, 80.9)	3.92 (2.01, 7.62)	0.21 (0.14, 0.33)	0.029*
Model B	81.7 (73.2, 88.0)	0.87 (0.79, 0.92)	83.1 (72.7, 90.1)	78.8 (62.2, 89.3)	89.4 (79.7, 94.8)	68.4 (52.5, 80.9)	3.92 (2.01, 7.62)	0.21 (0.14, 0.33)	0.056
Model C	81.7 (73.2, 88.0)	0.87 (0.79, 0.92)	83.1 (72.7, 90.1)	78.8 (62.2, 89.3)	89.4 (79.7, 94.8)	68.4 (52.5, 80.9)	3.92 (2.01, 7.62)	0.21 (0.14, 0.33)	0.115
Model D	82.7 (74.3, 88.8)	0.89 (0.82, 0.94)	83.1 (72.7, 90.1)	81.8 (65.6, 91.4)	90.8 (81.3, 95.7)	69.2 (53.6, 81.4)	4.57 (2.2, 9.5)	0.21 (0.14, 0.31)	0.4
Model E	83.7 (75.4, 89.5)	0.87 (0.79, 0.92)	85.9 (76.0, 92.2)	78.8 (62.2, 89.3)	89.7 (80.2, 94.9)	72.2 (56.0, 84.2)	4.05 (2.08, 7.87)	0.18 (0.11, 0.28)	0.021*

Table 3 (*continued*)

Table 3 (*continued*)

Model and parameter	Accuracy (%)	AUROC	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	LR+	LR–	P value
Model F	81.7 (73.2, 88.0)	0.88 (0.8, 0.93)	83.1 (72.7, 90.1)	78.8 (62.2, 89.3)	89.4 (79.7, 94.8)	68.4 (52.5, 80.9)	3.92 (2.01, 7.62)	0.21 (0.14, 0.33)	0.147
Full model	84.6 (76.5, 90.3)	0.9 (0.83, 0.95)	85.9 (76.0, 92.2)	81.8 (65.6, 91.4)	91.0 (81.8, 95.8)	73.0 (57.0, 84.6)	4.73 (2.28, 9.8)	0.17 (0.11, 0.27)	
N–F2 *vs.* F3–4									
DL model	76.0 (66.9, 83.2)	0.8 (0.71, 0.86)	78.0 (64.8, 87.2)	74.1 (61.1, 83.9)	73.6 (60.4, 83.6)	78.4 (65.4, 87.5)	3.01 (1.87, 4.83)	0.3 (0.2, 0.45)	0.165
Model A	76.0 (66.9, 83.2)	0.78 (0.7, 0.85)	76.0 (62.6, 85.7)	75.9 (63.1, 85.4)	74.5 (61.1, 84.5)	77.4 (64.5, 86.5)	3.16 (1.92, 5.2)	0.32 (0.22, 0.46)	0.044*
Model B	76.0 (66.9, 83.2)	0.79 (0.71, 0.86)	76.0 (62.6, 85.7)	75.9 (63.1, 85.4)	74.5 (61.1, 84.5)	77.4 (64.5, 86.5)	3.16 (1.92, 5.2)	0.32 (0.22, 0.46)	0.117
Model C	76.0 (66.9, 83.2)	0.8 (0.71, 0.86)	76.0 (62.6, 85.7)	75.9 (63.1, 85.4)	74.5 (61.1, 84.5)	77.4 (64.5, 86.5)	3.16 (1.92, 5.2)	0.32 (0.22, 0.46)	0.278
Model D	76.0 (66.9, 83.2)	0.8 (0.71, 0.86)	76.0 (62.6, 85.7)	75.9 (63.1, 85.4)	74.5 (61.1, 84.5)	77.4 (64.5, 86.5)	3.16 (1.92, 5.2)	0.32 (0.22, 0.46)	0.168
Model E	76.9 (68.0, 84.0)	0.79 (0.7, 0.86)	76.0 (62.6, 85.7)	77.8 (65.1, 86.8)	76.0 (62.6, 85.7)	77.8 (65.1, 86.8)	3.42 (2.03, 5.77)	0.31 (0.21, 0.45)	0.129
Model F	76.9 (68.0, 84.0)	0.81 (0.72, 0.87)	78.0 (64.8, 87.2)	75.9 (63.1, 85.4)	75.0 (61.8, 84.8)	78.8 (66.0, 87.8)	3.24 (1.97, 5.32)	0.29 (0.19, 0.43)	0.855
Full model	76.0 (66.9, 83.2)	0.81 (0.73, 0.88)	78.0 (64.8, 87.2)	74.1 (61.1, 83.9)	73.6 (60.4, 83.6)	78.4 (65.4, 87.5)	3.01 (1.87, 4.83)	0.3 (0.2, 0.45)	
N–F3 *vs.* F4									
DL model	79.8 (71.1, 86.4)	0.8 (0.72, 0.87)	82.8 (65.5, 92.4)	78.7 (68.1, 86.4)	60.0 (44.6, 73.7)	92.2 (83.0, 96.6)	3.88 (2.44, 6.18)	0.22 (0.14, 0.35)	0.166
Model A	79.8 (71.1, 86.4)	0.78 (0.69, 0.85)	82.8 (65.5, 92.4)	78.7 (68.1, 86.4)	60.0 (44.6, 73.7)	92.2 (83.0, 96.6)	3.88 (2.44, 6.18)	0.22 (0.14, 0.35)	0.033*
Model B	79.8 (71.1, 86.4)	0.83 (0.74, 0.89)	79.3 (61.6, 90.2)	80.0 (69.6, 87.5)	60.5 (44.7, 74.4)	90.9 (81.6, 95.8)	3.97 (2.43, 6.47)	0.26 (0.17, 0.39)	0.621
Model C	78.8 (70.0, 85.6)	0.8 (0.71, 0.87)	79.3 (61.6, 90.2)	78.7 (68.1, 86.4)	59.0 (43.4, 72.9)	90.8 (81.3, 95.7)	3.72 (2.32, 5.96)	0.26 (0.17, 0.41)	0.18
Model D	81.7 (73.2, 88.0)	0.83 (0.75, 0.89)	79.3 (61.6, 90.2)	82.7 (72.6, 89.6)	63.9 (47.6, 77.5)	91.2 (82.1, 95.9)	4.58 (2.7, 7.76)	0.25 (0.17, 0.37)	0.315
Model E	79.8 (71.1, 86.4)	0.8 (0.71, 0.87)	86.2 (69.4, 94.5)	77.3 (66.7, 85.3)	59.5 (44.5, 73.0)	93.5 (84.6, 97.5)	3.8 (2.44, 5.92)	0.18 (0.11, 0.3)	0.07
Model F	79.8 (71.1, 86.4)	0.82 (0.73, 0.88)	79.3 (61.6, 90.2)	80.0 (69.6, 87.5)	60.5 (44.7, 74.4)	90.9 (81.6, 95.8)	3.97 (2.43, 6.47)	0.26 (0.17, 0.39)	0.469
Full model	81.7 (73.2, 88.0)	0.84 (0.76, 0.9)	82.8 (65.5, 92.4)	81.3 (71.1, 88.5)	63.2 (47.3, 76.6)	92.4 (83.5, 96.7)	4.43 (2.69, 7.32)	0.21 (0.14, 0.33)	

Statistical values are expressed as the 95% CI, when applicable. AUROC values were compared between the full model and deep learning model/models A–F via the Delong test (*, P<0.05; **, P<0.01). Deep learning model: the deep learning-based model using delayed phase magnetic resonance images as input; model A: deep learning model combined with biomarkers (APRI and FIB-4); model B: deep learning model combined with virus status (hepatitis B virus and hepatitis C virus status); model C: deep learning model combined with magnetic resonance information (magnetic resonance manufacturers and static field strength); model D: deep learning model combined with biomarkers and virus status; model E: deep learning model combined with biomarkers and magnetic resonance information; model F: deep learning model combined with virus status and magnetic resonance information; full model: deep learning model combined with biomarkers, virus status, and magnetic resonance information. APRI, aminotransferase-to-platelet ratio index; AUROC, area under the receiver operating characteristic curve; CI, confidence interval; DL, deep learning; FIB-4, fibrosis index based on four factors; LR+, positive diagnostic likelihood ratio; LR–, negative diagnostic likelihood ratio; N, normal; NPV, negative predictive value; PPV, positive predictive value.

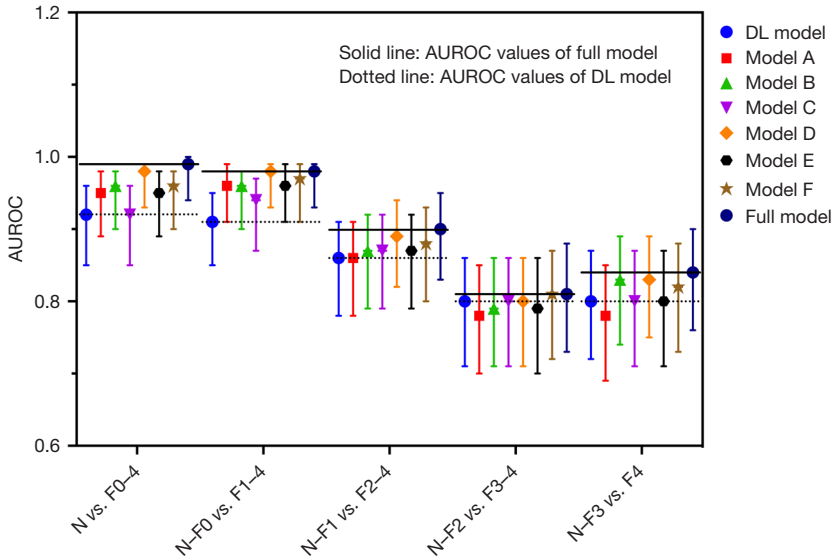

Figure 3 Comparison of the diagnostic performance between the different models. Model definitions: DL model, the deep learning-based model using delayed phase magnetic resonance images as input; model A, deep learning model combined with biomarkers (APRI and FIB-4); model B, deep learning model combined with virus status (HBV and HCV status); model C, deep learning model combined with magnetic resonance information (magnetic resonance manufacturers and static field strength); model D, deep learning model combined with biomarkers and virus status; model E, deep learning model combined with biomarkers and magnetic resonance information; model F, deep learning model combined with virus status and magnetic resonance information; full model, deep learning model combined with biomarkers, virus status, and magnetic resonance information. APRI, aminotransferase-to-platelet ratio index; AUROC, area under the receiver operating characteristic curve; DL, deep learning; FIB-4, fibrosis index based on four factors; HBV, hepatitis B virus; HCV, hepatitis C virus.

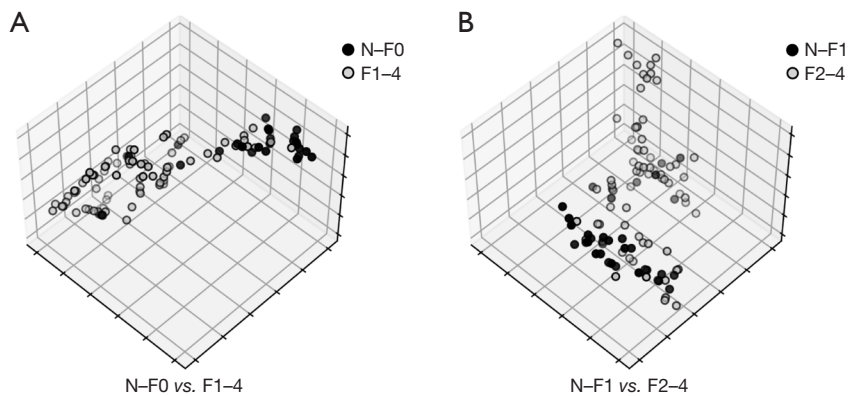

Figure 4 Illustration of the two classifiers learned by the deep learning model projected to three dimensions for visualization via the t-SNE algorithm using values of the last fully connected layer in the CNN framework on the test set. (A) Scatterplot visualizing the clustering of N–F0 (black circles) *vs.* F1–4 (white circles). (B) Scatterplot visualizing the clustering of N–F1 (black circles) *vs.* F2–4 (white circles). Scatterplots show how the algorithm clusters, with each point representing an individual and different colors representing different fibrosis stages. CNN, convolution neural network; t-SNE, t-distributed stochastic neighbor embedding.

examined with t-SNE for two representative classification tasks (*Figure 4*). For the N–F0 *vs.* F1–4 classification task (*Figure 4A*), the two groups formed distinct clusters, indicating good discriminative ability of the model. Compared with the F0 *vs.* F1–4 task, the points of the same group did not cluster together very well and only vaguely

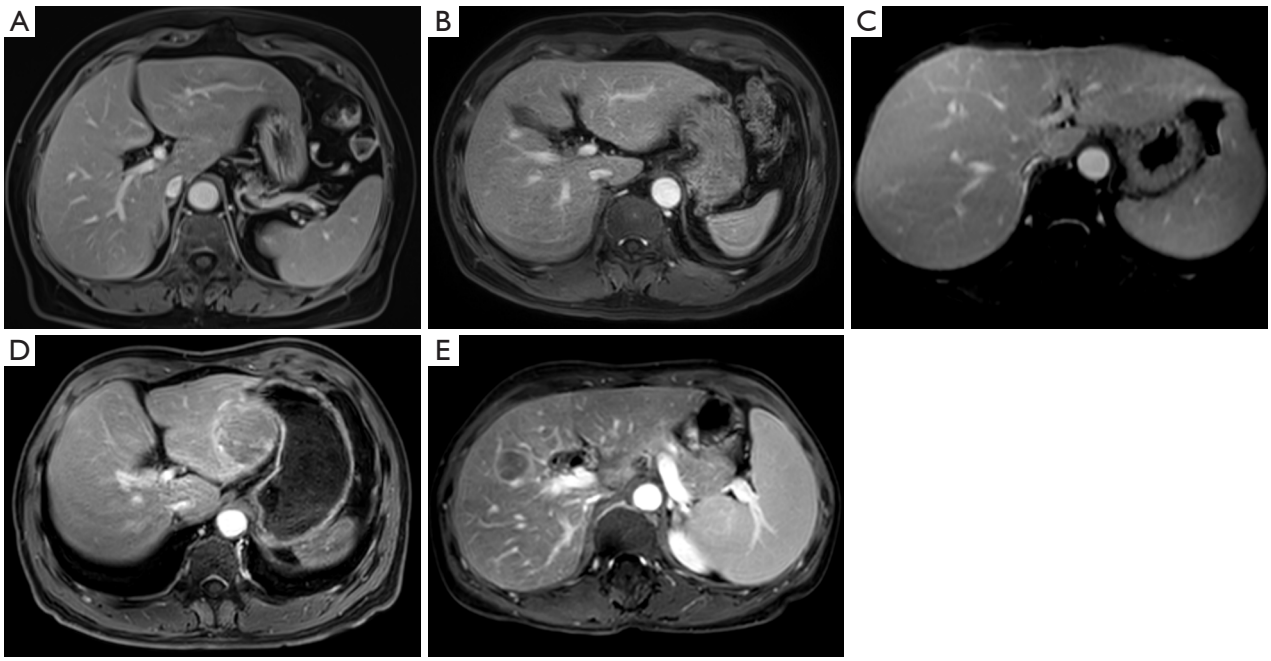

Figure 5 Representative magnetic resonance images showing cases with correctly and incorrectly classified liver fibrosis stages. (A) A correctly classified case by the DL model, full model, and two radiologists of a 68-year-old female with surgery-confirmed fibrosis stage N–F2 *vs.* F3–4 (F=2). (B) A correctly classified case by both the DL model and full model of a 60-year-old male with hepatitis B and surgery-confirmed fibrosis stage N–F3 *vs.* F4 (F=4). (C) A correctly classified case by the full model of a 54-year-old female with biopsy-confirmed fibrosis stage N–F2 *vs.* F3–4 (F=4). (D) A correctly classified case by the DL model of a 64-year-old male with surgery-confirmed fibrosis stage N–F3 *vs.* F4 (F=4). (E) A correctly classified case by two radiologists of a 55-year-old female with hepatitis B and surgery-confirmed fibrosis stage F2 *vs.* F3–4 (F=3). DL, deep learning.

separated from the points of the other group, which might be attributable to the imaging features in F1 patients being similar to those of F2 patients (*Figure 4B*).

Comparison of diagnostic performance between the DL model, full model, radiologists, and biomarkers

Compared with the diagnostic performance of radiologists, both the DL model and full model showed more balanced sensitivity and specificity. Additionally, both the DL model and full model showed better specificity, slightly lower to equivalent sensitivity for fibrosis stages F0–4, and better sensitivity for all four other classification tasks; however, the DL model had a slightly lower sensitivity in diagnosing the fibrosis stage F14. A representative classification for the DL model, full model, and two radiologists using MR images are shown in *Figure 5*. *Figure 5B,5D* indicate that breathing artifact, to a certain extent, had no effect on the diagnostic performance of the DL model or full model, which was another advantage of convolution neural network (CNN)-based image recognition. As can be seen in *Figure 5C*, the full model had a superior diagnostic performance to that of the DL model and the two radiologists in some cases. This is because the full model included extra nonimage information and could potentially obtain more detailed liver information than could human eyes. However, there might have been certain limitations regarding evaluating the details beyond the liver (e.g., splenomegaly and varices) compared with the evaluation of liver parenchyma (*Figure 5E*). This might have been because this information was not included in the training set due to the fact that the segmentation only focused on the liver in our model, which should be addressed in a further study.

Compared with biomarkers (APRI and FIB-4), both the DL model and full model showed significantly higher specificity in fibrosis stage F4, while they only showed higher specificity in fibrosis stage F3–4 as compared to FIB-4. In addition, the full model demonstrated significantly higher sensitivity in fibrosis stages F0–4 and F1–4 as compared to both biomarkers and demonstrated a higher sensitivity in

fibrosis stage F2–4 as compared to APRI. However, for the other classification tasks, there was no significant difference in either the specificity or sensitivity between the DL model, full model, and biomarkers (*Table 4*).

Consistency evaluation

The agreement of two radiologists was measured via the Cohen kappa coefficient (Table S4), which indicated moderate agreement for fibrosis stages F0–4 and F2–4 ($\kappa>0.4$) and substantial consistency for fibrosis stages F1–4, F3–4, and F4 ($\kappa>0.6$) (30).

We also compared the agreement between the radiologists, the DL model, and full model with the pathological fibrosis stage (*Table 5*). For fibrosis stage F2–4 and F3–4, both the DL model and radiologists showed moderate consistency ($0.4<\kappa<0.6$). For fibrosis stage F0–4, radiologist 1 and the DL model showed substantial agreement ($0.6<\kappa<0.8$), while radiologist 2 showed moderate agreement. For fibrosis stage F4, radiologist 2 displayed substantial agreement, while radiologist 1 and DL model showed moderate agreement. Thus, it was noted that DL model outperformed radiologist 2 in staging early-stage fibrosis and exhibited a comparative diagnostic performance with that of radiologist 1. Furthermore, approximately only 10 seconds were required for the DL model to stage liver fibrosis based on MR delayed images (liver segmentation required 7.5 seconds and hepatic fibrosis 2.7 seconds), which was shorter than the time required by human readers.

Discussion

As morphological changes (32) and vasculature distortion of liver are predictors of liver fibrosis (18,33-35), our study used 3D FADL technology to extract the above-mentioned information of the whole liver. This distinguished it from previous studies that only used the two-dimensional (2D) liver surface to capture texture information of the liver parenchyma for fibrosis prediction (36,37). With this 3D FADL technology, our study achieved a comparatively high diagnostic performance (with AUROCs of 0.99, 0.98, 0.90, 0.81, and 0.84 for the full model in diagnosing liver fibrosis stages F0–4, F1–4, F2–4, F3–4, and F4, respectively) using a relatively smaller amount of data compared as compared with that in the study by Choi *et al.* (18). There has been a degree of success achieved in predicting significant fibrosis (≥ F2), advanced fibrosis (≥ F3), and cirrhosis (≥ F4), with AUROCs of 0.74–0.96, 0.76–0.98, and 0.73–0.97, respectively, based on US, shear wave elastography (SWE), CT, and gadoxetic acid-enhanced MR (15-19,28-35,37). Although our model did not show the highest diagnostic performance for the classification task among these studies, our was the first study to screen out patients with CLD from the healthy population (N *vs.* F0–4), with high AUROC values of 0.92 for the DL model and 0.99 for the full model. Furthermore, our study employed delayed phase MR images without any requirements related to operator's technique, radiation (as with CT), or specific contrast agents such as gadoxetic acid-enhanced MRI. Given these advantages, our model could be more widely applied in routine clinical practice, especially for screening out patients with CLD from the healthy population or for patients who require regular follow-up for monitoring the efficacy of treatment.

In addition, compared with radiologists and biomarkers, our models demonstrated more balanced sensitivity and specificity values. Moreover, both the DL model and full model showed high specificity for fibrosis stage F0–4 (87.5% and 87.5%, respectively) and for fibrosis stage F2–4 (78.8% and 81.8%, respectively), as well as high sensitivity for fibrosis stage F0–4 (88.8% and 100%, respectively) and for fibrosis stage F2–4 (81.7% and 85.9%, respectively). This facilitated the identification of patients with CLD from healthy individuals and the determination of which patients needed treatment. Both the DL model and full model showed higher specificity (78.7% and 81.3%, respectively) and sensitivity (82.8% and 82.8%, respectively) in fibrosis stage F4 and thus may help in preventing patients from progressing to decompensated cirrhosis via a more attentive monitoring of patients with compensated cirrhosis (F4).

Several limitations to this study should be acknowledged. First, the sample size was relatively small, and the distribution across the different pathology fibrosis stages was unbalanced. The DL-based model could be better trained and provide a better performance if more patients were included and a greater balance of patients across all fibrosis stages. Second, our developmental data included data from patients with liver tumors (some tumors were larger than 5 cm) or with relatively obvious breathing artifacts. Despite these limitations, our model showed high accuracy in staging liver fibrosis. In addition, we believe the analysis of images with large tumors or breathing artifacts provider a more accurate reflection of real-world practice. Thus, the signal intensity of MR images was not an absolute value. Although the quantitative imaging techniques such as T2 and T1 mapping have been previously used in clinical

Table 4 Comparison of sensitivity and specificity between the DL model, full model, radiologists, and serum biomarkers

Classification task and diagnostic performance	DL model	Full model	R1	R2	APRI	FIB-4	McNemar's test P value								
							DL model *vs.* full model	DL model *vs.* R1	Full model *vs.* R1	DL model *vs.* R2	Full model *vs.* R2	DL model *vs.* APRI	Full model *vs.* APRI	DL model *vs.* FIB-4	Full model *vs.* FIB-4
N *vs.* F0–4															
Sensitivity (%)	88.8 (80.0, 94.0)	100.0 (95.4, 100.0)	100.0 (95.4, 100.0)	100.0 (95.4, 100.0)	81.2 (71.3, 88.3)	88.8 (80.0, 94.0)	0.008**	0.008**	NA	0.008**	NA	0.264	<0.001***	1.000	0.008**
Specificity (%)	87.5 (69.0, 95.7)	87.5 (69.0, 95.7)	70.8 (50.8, 85.1)	37.5 (21.2, 57.3)	95.8 (79.8, 99.8)	83.3 (64.1, 93.3)	NA	0.343	0.343	0.003**	0.003**	0.617	0.617	1.000	1.000
N–F0 *vs.* F1–4															
Sensitivity (%)	92.4 (84.4, 96.5)	100.0 (95.4, 100.0)	97.5 (91.2, 99.6)	86.1 (76.8, 92.0)	81.0(71.0, 88.1)	88.6 (79.7, 93.9)	0.041*	0.289	0.48	0.267	0.003**	0.052	<0.001***	0.547	0.008**
Specificity (%)	88.0 (70.0, 95.8)	88.0 (70.0, 95.8)	92.0 (75.0, 98.6)	88.0(70.0, 95.8)	92.0 (75.0, 98.6)	80.0 (60.9, 91.1)	NA	1	1	1	1	1	1	0.683	0.683
N–F1 *vs.* F2–4															
Sensitivity (%)	81.7 (71.2, 89.0)	85.9 (76.0, 92.2)	73.2 (61.9, 82.1)	67.6 (56.1, 77.3)	70.4 (59.0, 79.8)	91.5 (82.8, 96.1)	0.248	0.211	0.052	0.055	0.012*	0.186	0.046	0.096	0.343
Specificity (%)	78.8 (62.2, 89.3)	81.8 (65.6, 91.4)	90.9 (76.4, 96.9)	81.8 (65.6, 91.4)	84.8 (69.1, 93.3)	69.7 (52.7, 82.6)	1	0.134	0.248	1.000	1.000	0.724	1.000	0.505	0.289
N–F2 *vs.* F3–4															
Sensitivity (%)	78.0 (64.8, 87.2)	78.0 (64.8, 87.2)	60.0 (46.2, 72.4)	70.0 (56.2, 80.9)	84.0 (71.5, 91.7)	92.0 (81.2, 96.8)	NA	0.027*	0.027*	0.343	0.343	0.628	0.628	0.096	0.096
Specificity (%)	74.1 (61.1, 83.9)	74.1 (61.1, 83.9)	88.9 (77.8, 94.8)	83.3 (71.3, 91.0)	57.4 (44.2, 69.7)	46.3 (33.7, 59.4)	1	0.043*	0.043*	0.267	0.267	0.066	0.066	0.002**	0.001**
N–F3 *vs.* F4															
Sensitivity (%)	82.8 (65.5, 92.4)	82.8 (65.5, 92.4)	58.6 (40.7, 74.5)	72.4 (54.3, 85.3)	72.4 (54.3, 85.3)	96.6 (82.8, 99.8)	1	0.096	0.070	0.547	0.505	0.547	0.505	0.221	0.134
Specificity (%)	78.7 (68.1, 86.4)	81.3 (71.1, 88.5)	90.7 (82.0, 95.4)	92.0 (83.6, 96.3)	61.3 (50.0, 71.5)	37.3 (27.3, 48.6)	0.617	0.027*	0.070	0.016	0.043*	0.012*	0.002**	<0.001***	<0.001***

*, P<0.05; **, P<0.01; ***, P<0.001. APRI, aminotransferase-to-platelet ratio index; DL, deep learning; FIB-4, fibrosis index based on four factors; N, normal; R1, radiologist 1; R2, radiologist 2.

Table 5 Consistency evaluation between two radiologists, the DL model, and the gold standard

Variables	Kappa	Z score	P value
N *vs.* F0–4			
R1	0.789	8.230	<0.001
R2	0.480	5.731	<0.001
DL model	0.701	7.232	<0.001
Full model	0.915	9.365	<0.001
N–F0 *vs.* F1–4			
R1	0.895	9.124	<0.001
R2	0.668	6.936	<0.001
DL model	0.772	7.899	<0.001
Full model	0.918	9.390	<0.001
N–F1 *vs.* F2–4			
R1	0.568	6.100	<0.001
R2	0.434	4.695	<0.001
DL model	0.577	5.929	<0.001
Full model	0.656	6.715	<0.001
N–F2 *vs.* F3–4			
R1	0.494	5.236	<0.001
R2	0.536	5.500	<0.001
DL model	0.520	5.308	<0.001
Full model	0.520	5.308	<0.001
N–F3 *vs.* F4			
R1	0.520	5.350	<0.001
R2	0.658	6.719	<0.001
DL model	0.550	5.774	<0.001
Full model	0.585	6.087	<0.001

DL model: the deep learning-based model using magnetic resonance delayed images as input; full model: deep learning model combined with biomarker, virus status, and magnetic resonance information (APRI, FIB-4, hepatitis B virus, hepatitis C virus status, magnetic resonance manufacturers, and static field strength). APRI, aminotransferase-to-platelet ratio index; FIB-4, fibrosis index based on four factors; N, normal; R1, radiologist 1; R2, radiologist 2.

application, they are not widely available (38). Thus, we included augmented images with random geometric and intensity transformations for training the model. This ensured that the model was robust for different window settings. Furthermore, we used a batch normalization function to normalize the grayscale of the image data.

Conclusions

The DL-based model and full model based on delayed phase MR images showed good diagnostic performance in the assessment of liver fibrosis. It may aid in identifying patients with CLD from healthy individuals, indicate which patients require treatment, and prevent patients from progressing to decompensated cirrhosis through enhanced monitoring of those with compensated cirrhosis (F4).

Acknowledgments

We sincerely thank the editor for the valuable comments. We also acknowledge that an earlier version of this abstract was previously presented as a conference abstract at the European Congress of Radiology (ECR), Vienna, March 2–6, 2022. We have thoroughly revised the abstract to ensure originality and compliance with copyright regulations.

Footnote

Reporting Checklist: The authors have completed the STARD reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-2024-2506/rc

Data Sharing Statement: Available at https://qims.amegroups.com/article/view/10.21037/qims-2024-2506/dss

Funding: This work was supported by the National Natural Science Foundation of China (No. 81401406) and Innovative Research Team of High-Level Local Universities in Shanghai. The funders had no role in the study design, data collection, analysis, decision to publish, or preparation of the manuscript.

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-2024-2506/coif). The authors report that this work was supported by the National Natural Science Foundation of China (No. 81401406) and the Innovative Research Team of the High-Level Local Universities in Shanghai. Y.Z. and Y.Y. are currently employees of SenseTime Research, Shanghai, China. X.D. is currently an employee of SenseBrain Technology, SenseTime, Princeton, NJ, USA. Yinan Chen is currently an employee of SenseTime Research, Shanghai, and WCH-

SenseTime Joint Lab, Chengdu, China. The authors have no other conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. This study was conducted in accordance with the Declaration of Helsinki and its subsequent amendments and was approved by the Ethics Committee of Ruijin Hospital, Shanghai Jiao Tong University School of Medicine (No. RJ2018-209), and by the Ethics Committee of the First Affiliated Hospital of Zhengzhou University (No. 2017-KY-01). The requirement for informed consent was waived due to the retrospective nature of the study.

References

1. Asrani SK, Devarbhavi H, Eaton J, Kamath PS. Burden of liver diseases in the world. J Hepatol 2019;70:151-71.
2. Tsochatzis EA, Bosch J, Burroughs AK. Liver cirrhosis. Lancet 2014;383:1749-61.
3. Castera L. Invasive and non-invasive methods for the assessment of fibrosis and disease progression in chronic liver disease. Best Pract Res Clin Gastroenterol 2011;25:291-303.
4. Castera L. Noninvasive methods to assess liver disease in patients with hepatitis B or C. Gastroenterology 2012;142:1293-1302.e4.
5. Rockey DC, Caldwell SH, Goodman ZD, Nelson RC, Smith AD; American Association for the Study of Liver Diseases. Liver biopsy. Hepatology 2009;49:1017-44.
6. Ratziu V, Charlotte F, Heurtier A, Gombert S, Giral P, Bruckert E, Grimaldi A, Capron F, Poynard T; LIDO Study Group. Sampling variability of liver biopsy in nonalcoholic fatty liver disease. Gastroenterology 2005;128:1898-906.
7. Chou R, Wasson N. Blood tests to diagnose fibrosis or cirrhosis in patients with chronic hepatitis C virus infection: a systematic review. Ann Intern Med 2013;158:807-20. Erratum in: Ann Intern Med 2013;159:308.
8. Horowitz JM, Venkatesh SK, Ehman RL, Jhaveri K, Kamath P, Ohliger MA, Samir AE, Silva AC, Taouli B, Torbenson MS, Wells ML, Yeh B, Miller FH. Evaluation of hepatic fibrosis: a review from the society of abdominal radiology disease focus panel. Abdom Radiol (NY) 2017;42:2037-53.
9. Huwart L, Sempoux C, Salameh N, Jamart J, Annet L, Sinkus R, Peeters F, ter Beek LC, Horsmans Y, Van Beers BE. Liver fibrosis: noninvasive assessment with MR elastography versus aspartate aminotransferase-to-platelet ratio index. Radiology 2007;245:458-66.
10. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44.
11. Pickhardt PJ, Malecki K, Kloke J, Lubner MG. Accuracy of Liver Surface Nodularity Quantification on MDCT as a Noninvasive Biomarker for Staging Hepatic Fibrosis. AJR Am J Roentgenol 2016;207:1194-9.
12. Smith AD, Branch CR, Zand K, Subramony C, Zhang H, Thaggard K, Hosch R, Bryan J, Vasanji A, Griswold M, Zhang X. Liver Surface Nodularity Quantification from Routine CT Images as a Biomarker for Detection and Evaluation of Cirrhosis. Radiology 2016;280:771-81.
13. Goshima S, Kanematsu M, Kobayashi T, Furukawa T, Zhang X, Fujita H, Watanabe H, Kondo H, Moriyama N, Bae KT. Staging hepatic fibrosis: computer-aided analysis of hepatic contours on gadolinium ethoxybenzyl diethylenetriaminepentaacetic acid-enhanced hepatocyte-phase magnetic resonance imaging. Hepatology 2012;55:328-9.
14. Treacher A, Beauchamp D, Quadri B, Fetzer D, Vij A, Yokoo T, Montillo A. Deep Learning Convolutional Neural Networks for the Estimation of Liver Fibrosis Severity from Ultrasound Texture. Proc SPIE Int Soc Opt Eng 2019;10950:109503E.
15. Wang K, Lu X, Zhou H, Gao Y, Zheng J, Tong M, Wu C, Liu C, Huang L, Jiang T, Meng F, Lu Y, Ai H, Xie XY, Yin LP, Liang P, Tian J, Zheng R. Deep learning Radiomics of shear wave elastography significantly improved diagnostic performance for assessing liver fibrosis in chronic hepatitis B: a prospective multicentre study. Gut 2019;68:729-41.
16. Yasaka K, Akai H, Kunimatsu A, Abe O, Kiryu S. Deep learning for staging liver fibrosis on CT: a pilot study. Eur Radiol 2018;28:4578-85.
17. Lee JH, Joo I, Kang TW, Paik YH, Sinn DH, Ha SY, Kim K, Choi C, Lee G, Yi J, Bang WC. Deep learning with

ultrasonography: automated classification of liver fibrosis using a deep convolutional neural network. Eur Radiol 2020;30:1264-73.
18. Choi KJ, Jang JK, Lee SS, Sung YS, Shim WH, Kim HS, Yun J, Choi JY, Lee Y, Kang BK, Kim JH, Kim SY, Yu ES. Development and Validation of a Deep Learning System for Staging Liver Fibrosis by Using Contrast Agent-enhanced CT Images in the Liver. Radiology 2018;289:688-97.
19. 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;287:146-55.
20. Zhuang Y, Ding H, Zhang Y, Sun H, Xu C, Wang W. Two-dimensional Shear-Wave Elastography Performance in the Noninvasive Evaluation of Liver Fibrosis in Patients with Chronic Hepatitis B: Comparison with Serum Fibrosis Indexes. Radiology 2017;283:873-82.
21. Scheuer PJ. Classification of chronic viral hepatitis: a need for reassessment. J Hepatol 1991;13:372-4.
22. Wai CT, Greenson JK, Fontana RJ, Kalbfleisch JD, Marrero JA, Conjeevaram HS, Lok AS. A simple noninvasive index can predict both significant fibrosis and cirrhosis in patients with chronic hepatitis C. Hepatology 2003;38:518-26.
23. Talwalkar JA, Yin M, Fidler JL, Sanderson SO, Kamath PS, Ehman RL. Magnetic resonance imaging of hepatic fibrosis: emerging clinical applications. Hepatology 2008;47:332-42.
24. Kleiner DE, Brunt EM, Van Natta M, Behling C, Contos MJ, Cummings OW, Ferrell LD, Liu YC, Torbenson MS, Unalp-Arida A, Yeh M, McCullough AJ, Sanyal AJ; Nonalcoholic Steatohepatitis Clinical Research Network. Design and validation of a histological scoring system for nonalcoholic fatty liver disease. Hepatology 2005;41:1313-21.
25. Sterling RK, Lissen E, Clumeck N, Sola R, Correa MC, Montaner J, S Sulkowski M, Torriani FJ, Dieterich DT, Thomas DL, Messinger D, Nelson M; APRICOT Clinical Investigators. Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. Hepatology 2006;43:1317-25.
26. Erturk SM, Alberich-Bayarri A, Herrmann KA, Marti-Bonmati L, Ros PR. Use of 3.0-T MR imaging for evaluation of the abdomen. Radiographics 2009;29:1547-63.
27. Kojiro M, Shimamatsu K, Kage M. Pathomorphologic comparison of hepatitis C virus-related and hepatitis B virus-related cirrhosis bearing hepatocellular carcinoma. Princess Takamatsu Symp 1995;25:179-84.
28. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837-45.
29. Tsai WY, Chi Y, Chen CM. Interval estimation of binomial proportion in clinical trials with a two-stage design. Stat Med 2008;27:15-35.
30. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22:276-82.
31. Van Der Maaten L. Accelerating t-SNE using tree-based algorithms. J Mach Learn Res 2014;15:3221-45.
32. Colli A, Fraquelli M, Andreoletti M, Marino B, Zuccoli E, Conte D. Severe liver fibrosis or cirrhosis: accuracy of US for detection--analysis of 300 cases. Radiology 2003;227:89-94.
33. Lubner MG, Malecki K, Kloke J, Ganeshan B, Pickhardt PJ. Texture analysis of the liver at MDCT for assessing hepatic fibrosis. Abdom Radiol (NY) 2017;42:2069-78.
34. Kim H, Park SH, Kim EK, Kim MJ, Park YN, Park HJ, Choi JY. Histogram analysis of gadoxetic acid-enhanced MRI for quantitative hepatic fibrosis measurement. PLoS One 2014;9:e114224.
35. Suk KT, Kim DJ. Staging of liver fibrosis or cirrhosis: The role of hepatic venous pressure gradient measurement. World J Hepatol 2015;7:607-15.
36. Hectors SJ, Kennedy P, Huang KH, Stocker D, Carbonell G, Greenspan H, Friedman S, Taouli B. Fully automated prediction of liver fibrosis using deep learning analysis of gadoxetic acid-enhanced MRI. Eur Radiol 2021;31:3805-14.
37. Jana A, Q uH, Rattan P, Minacapelli CD, Rustgi V, Metaxas D. Deep Learning based NAS Score and Fibrosis Stage Prediction from CT and Pathology Data. 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE); 26-28 October 2020; Cincinnati, OH, USA. IEEE; 2020.
38. Yu H, Touret AS, Li B, O'Brien M, Qureshi MM, Soto JA, Jara H, Anderson SW. Application of texture analysis on parametric T(1) and T(2) maps for detection of hepatic fibrosis. J Magn Reson Imaging 2017;45:250-9.

Cite this article as: Li W, Zhu Y, Zhao G, Chen X, Zhao X, Xu H, Che Y, Chen Y, Ye Y, Dou X, Wang H, Cheng J, Xie Q, Chen K. Deep learning-based automated assessment of hepatic fibrosis via magnetic resonance images and nonimage data. Quant Imaging Med Surg 2025;15(9):8250-8264. doi: 10.21037/qims-2024-2506

Appendix 1: Materials and methods—reference standards—pathologic assessment of liver fibrosis according to Scheuer Scoring System

Pathologic assessment of liver fibrosis was performed according to Scheuer Scoring System [stage 0, absence of fibrosis (F0); stage 1, fibrous portal expansion (F1); stage 2, periportal or rare portal-portal septa (F2); stage 3, fibrous septa with architectural distortion (F3); stage 4, cirrhosis (F4)] (21). Necroinflammatory activity was defined as follows: grade 0, no portal or periportal and lobular necroinflammatory activity (A0); grade 1, portal or periportal inflammatory and minimal occasionally spotty lobular inflammation (A1); grade 2, mild piecemeal portal or periportal necrosis and mild or focal lobular necrosis (A2); grade 3, moderate piecemeal portal or periportal necrosis and moderate or noticeable hepatocellular change inside the lobule (A3); grade 4, severe piecemeal portal or periportal necrosis and severe or diffuse hepatocellular damage inside the lobule (A4).

Appendix 2: Materials and methods—model construction and evaluation using magnetic resonance (MR) delay images and three-dimensional (3D) fully automated deep learning (FADL)

Data pre-processing

Firstly, the delayed-phase MR images were fed into a self-developed 3D U-Net model to obtain the liver surfaces. Secondly, the delayed-phase images and the matched liver surfaces were resampled to the median voxel spacing of the dataset. Finally, both delayed-phase MR images and the matched liver surfaces were cropped according to the size of the liver surface, and were then resized to 32×256×256 pixels using nearest neighbor interpolation algorithm before being fed to the 3D FADL model (model DL).

Data augmentation

In the training cohort, online data augmentation was applied to reduce the potential bias caused by the limited sample size and unbalanced data. The delayed-phase MR images and the matched liver surfaces were augmented through a number of random geometric and intensity transformations, including affine transform, rotation, scaling and lighting alteration, which increased the training data pool and decreased the overfitting of the generated deep learning-based model. The online augmentation was conducted for each mini-batch during the training procedure using Python 3.7.

Model training

To be sure that the model built based on limited sample size can perform well on unseen data, we used a re-sampling technique called stratified k-fold cross-validation during the training phase. In our study, k was selected as 5. The cross-validation method split the dataset of the training cohort into training and validation sets, the model which showed the best performance on the validation dataset was selected to infer the classification results of the test data.

The pre-processed and augmented 3D delayed-phase MR images and the matched liver surfaces in 396 patients were used as the inputs to model DL. Model DL was trained on a GeForce GTX 1080Ti (NVIDIA) graphic processing unit using Python 3.7 and Pytorch 1.4.0. Five double convolution blocks were designed in this network, and each of them was followed by a max pooling layer. And then another separate convolution block was applied. Every convolution block consisted of a convolution layer, an instance norm layer and a ReLU layer. Each of the feature representations acquired from the last two max pooling layers and the last convolution block was scaled into the same size through a feature processing block. Each feature processing block consisted of an adaptive average pooling layer, a fully connected layer and a ReLU layer. Three processed feature representations were concatenated as one to obtain more compact feature representation, and was then summed along the specific dimension as a fused feature. The fused feature was input into the last fully connected layer to calculate the probabilities of belonging to each category. Adam optimizer and the additive angular margin (Arcface) loss were used to improve the classification accuracy of the predicted mode which had demonstrated better performance for face recognition compared to other loss functions (39). Among them, adam optimizer was used to minimize the cross-entropy

 https://dx.doi.org/10.21037/qims-2024-2506

between the classification outputs and target labels (pathologic fibrosis stage) (40). Arcface loss was applied as the loss function, which maximizes class separability by obtaining highly discriminative features (41). The architecture of model DL is illustrated in *Figure 2B*.

In addition, the dataset used to develop the model was imbalanced especially that the amount of data for F0 and F1 was smaller than other groups. This imbalance could result in poor classification accuracy of the model for the minority classes. To overcome this problem, the same two positive and negative samples were fed into a mini-batch (batch size was set as 4) of the training process, and the weights of different categories in the loss function were balanced, which would facilitate convergence and stability of the training procedure. Weight decay was set to 1e–4. Adam optimizer was used with an initial learning rate of 5e–5 and was updated by CosineAnnealingLR method to prevent overfitting of the model (39).

Model evaluation

The performance of model DL was evaluated on a separate test set of 104 patients by five-fold experiments. The model showing the best performance in the validation dataset was selected from the five experiments. The five selected models were used to infer classification results on the test data. The average prediction result of the five experiments on the test data was served as the final result.

Appendix 3: Materials and methods—radiologists' visual grading of hepatic fibrosis

The instructions of visual grades were as follows: (I) normal, sharp liver margin, narrow liver fissure, smooth liver surface; (II) grade 0, for the findings between normal and grade 1; (III) grade 1, blunted liver edge, widened liver fissure, no obvious liver surface irregularity or nodularity, no findings of portal hypertension; (IV) grade 2, blunted liver edge, widened liver fissure, irregular or nodular surface of liver, no findings of portal hypertension; (V) grade 3, for the findings between grade 2 and grade 4; (VI) grade 4, irregular or nodular surface of liver, findings of portal hypertension (i.e., ascites, splenomegaly, varices), redistribution of liver segmental volume, obviously coarse hepatic texture and the presence of cirrhotic nodules (18,42,43).

Appendix 4: Statistical analysis

The values obtained from the last fully hidden connected layer were used as an input to the t-distributed stochastic neighbor embedding (t-SNE) algorithm. Different stages of liver fibrosis were shown as different colored points, to demonstrate how the algorithm clustered the diseases. each point of the scatterplots represented an input image projected from the 128-dimensional output of the convolution neural network (CNN)'s last hidden layer into three dimensions and the color represented the fibrosis stage, which showed how the algorithm clusters.

References

39. Loshschilv I, Hutter F. SGDR: Stochastic gradient descent with warm restarts. 5th International Conference on Learning Representations. 2017;1608:03983.
40. Kingma DP, Ba J. Adam: A method for stochastic optimization. International Conference on Learning Representations, ICLR 2017;1412:6980.
41. Deng J, Guo J, Yang J, Xue N, Kotsia I, Zafeiriou S. ArcFace: Additive angular margin loss for deep face recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2019:4690-9.
42. Bonekamp S, Kamel I, Solga S, Clark J. Can imaging modalities diagnose and stage hepatic fibrosis and cirrhosis accurately. J Hepatol 2009;50: 17-35.
43. Rustogi R, Horowitz J, Harmath C, Wang Y, Chalian H, Ganger DR, Chen ZE, Bolster BD Jr, Shah S, Miller FH. Accuracy of MR elastography and anatomic MR imaging features in the diagnosis of severe hepatic fibrosis and cirrhosis. J Magn Reson Imaging 2012;35:1356-64.

https://dx.doi.org/10.21037/qims-2024-2506

Table S1 The pathological information of patients

Variables	Total patients (500)	Training cohort (396)	Test cohort (104)	P value
Fibrosis stages				1.000
Normal†	116 (23.2%)	92 (23.2%)	24 (23.1%)	
F0	7 (1.4%)	6 (1.5%)	1 (1.0%)	
F1	40 (8.0%)	32 (8.1%)	8 (7.7%)	
F2	97 (18.4%)	76 (19.2%)	21 (20.2%)	
F3	99 (19.8%)	78 (19.7%)	21 (20.2%)	
F4	141 (28.2%)	112 (28.3%)	29 (27.9%)	
Inflammation grades				0.986
Normal†	116 (23.2%)	92 (23.2%)	24 (23.1%)	
A0	3 (0.6%)	3 (0.8%)	0 (0%)	
A1	37 (7.4%)	30 (7.6%)	7 (6.7%)	
A2	191 (38.2%)	149 (37.6%)	42 (40.4%)	
A3	112 (22.4%)	88 (22.2%)	24 (23.1%)	
A4	41 (8.2%)	34 (8.6%)	7 (6.7%)	

P value was calculated between the training and test cohorts. †, the normal subset of control subjects with benign liver lesions or extrahepatic tumors, pathological evaluation of liver parenchyma was not performed due to ethical considerations. Instead, these individuals were defined as healthy (normal) based on the absence of a history of hepatitis, negative viral hepatitis markers, and no evidence of fatty liver as assessed by magnetic resonance imaging-derived proton density fat fraction (MRI-PDFF <5%) (22). MRI-PDFF, MRI-derived proton density fat fraction.

Table S2 Tumor distribution in healthy controls with benign liver lesions and extrahepatic tumors

Liver tumor	Number of patients	Pancreatic or duodenal tumor	Number of patients
MT	2	STPT	38
MT combined with HH	1	STPT combined with liver HH	2
HA	1	IPMN combined with liver HH	1
HH	24	PSC	26
FNH	9	PMC	6
FNH combined with HH	1	pNET combined with liver FNH	1
AML	2	Duodenal adenoma	1
AML combined with HH	1	–	–

AML, hepatic angiomyolipoma; FNH, hepatic focal nodular hyperplasia; HA, hepatocellular adenoma; HH, hepatic hemangioma; IPMN, pancreatic intraductal papillary mucinous neoplasm; MT, liver metastatic tumor; pNET, pancreatic neuroendocrine tumor; PSC, pancreatic serous cystadenoma; PMC, pancreatic mucinous cystadenoma; STPT, solid pseudopapillary tumor of pancreas.

https://dx.doi.org/10.21037/qims-2024-2506

Table S3 The clinical data of control and liver fibrosis group

Variables	Total patients (n=500)	Training cohort (n=396)	Test cohort (n=104)
Normal[†]	116 (23.2%)	92 (23.2%)	24 (23.1%)
Liver tumor[‡]	41 (8.2%)	31 (7.8%)	10 (9.6%)
Pancreatic or duodenal tumor	75 (15%)	61 (15.4%)	14 (13.5%)
Fibrosis stages			
F0	7 (1.4%)	6 (1.5%)	1 (1%)
HH[‡]	1 (0.2%)	1 (0.3%)	0 (0%)
HCC[‡]	1 (0.2%)	0 (0%)	1 (1%)
Chronic hepatitis B[§]	4 (0.8%)	4 (1.0%)	0 (0%)
Unknown etiology[§]	1 (0.2%)	1 (0.3%)	0 (0%)
F1	40 (8.0%)	32 (8.1%)	8 (7.7%)
HCC[‡]	5 (1.0%)	5 (1.3%)	0 (0%)
Chronic hepatitis B[§]	24 (4.8%)	19 (4.8%)	6 (5.8%)
DILI[§]	2 (0.4%)	1 (0.3%)	1 (1%)
Nonalcoholic fatty liver disease[§]	1 (0.2%)	1 (0.3%)	0 (0%)
Autoimmune hepatitis[§]	2 (0.4%)	2 (0.5%)	0 (0%)
Unknown etiology[§]	5 (1.0%)	4 (1.0%)	1 (1%)
F2	97 (18.4%)	76 (19.2%)	21 (20.2%)
HCC[‡]	49 (9.8%)	36 (9.1%)	13 (12.5%)
Chronic hepatitis B[§]	27 9 (5.4%)	22 (5.6%)	5 (4.8%)
DILI[§]	3 (0.6%)	2 (0.5%)	1 (1%)
Autoimmune hepatitis[§]	1 (0.2%)	1 (0.3%)	0 (0%)
AOSD[§]	1 (0.2%)	1 (0.3%)	0 (0%)
Unknown etiology[§]	16 (3.2%)	14 (3.5%)	2 (1.9%)
F3	99 (19.8%)	78 (19.7%)	21 (20.2%)
HCC[‡]	67 (13.4%)	54	13 (12.5%)
Dual-phenotype HCC[‡]	2 (0.4%)	1 (0.3%)	1 (1%)
Combined HCC[‡]	1 (0.2%)	1 (0.3%)	0 (0%)
Chronic hepatitis B[§]	15 (3.0%)	11 (13.6%)	4 (3.8%)
Chronic hepatitis C[§]	1 (0.2%)	1 (0.3%)	0 (0%)
DILI[§]	1 (0.2%)	0 (0%)	1 (1%)
Autoimmune hepatitis[§]	5 (1.0%)	4 (1.0%)	1 (1%)
Unknown etiology[§]	7 (1.4%)	6 (1.5%)	1 (1%)

Table S3 (*continued*)

https://dx.doi.org/10.21037/qims-2024-2506

Table S3 (*continued*)

Variables	Total patients (n=500)	Training cohort (n=396)	Test cohort (n=104)
F4	141 (28.2%)	112 (28.3%)	29 (27.9%)
HCC‡	115 (23.0%)	90 (22.7%)	25
Dual-phenotype HCC‡	3 (0.6%)	3 (0.3%)	0 (0%)
Combined HCC‡	2 (0.4%)	2 (0.5%)	0 (0%)
Chronic hepatitis B§	12 (2.4%)	9 (2.3%)	3 (24.0%)
Chronic hepatitis C§	2 (0.4%)	2 (0.5%)	0 (0%)
DILI§	1 (0.2%)	1 (0.3%)	0 (0%)
Nonalcoholic fatty liver disease§	2 (0.4%)	1 (0.3%)	1 (1%)
Wilson's disease§	1 (0.2%)	1 (0.3%)	0 (0%)
Autoimmune hepatitis§	1 (0.2%)	1 (0.3%)	0 (0%)
Unknown etiology§	2 (0.4%)	2 (0.5%)	0 (0%)

†, individuals in this group had either hepatic or pancreatoduodenal tumors, but all had normal liver function. They were considered to have healthy livers based on the absence of a history of hepatitis, negative viral hepatitis markers, and no evidence of hepatic steatosis as assessed by magnetic resonance imaging-derived proton density fat fraction (MRI-PDFF <5%). ‡, specimens were obtained from liver tumor resections in patients with various liver tumor, liver fibrosis stages were pathologically determined from the peri-tumoral liver parenchyma. §, specimens were obtained from percutaneous liver biopsies in patients with hepatic dysfunction of various etiologies but without liver tumors; liver fibrosis stages were pathologically determined from the biopsy samples. AOSD, adult-onset Still's disease; DILI, drug-induced liver injury; HCC, hepatocellular carcinoma; HH, hepatic hemangioma; MRI-PDFF, MRI-derived proton density fat fraction.

Table S4 Consistency analysis of two radiologists

Fibrosis stage	Kappa	Z score	P value
N *vs.* F0–4	0.480	5.215	<0.001
N–F0 *vs.* F1–4	0.668	6.926	<0.001
N–F1 *vs.* F2–4	0.595	6.072	<0.001
N–F2 *vs.* F3–4	0.677	6.996	<0.001
N–F3 *vs.* F4	0.663	6.779	<0.001

https://dx.doi.org/10.21037/qims-2024-2506