Research Article
Innovative Diagnostics, Modelling and Digital Hepatology

JOURNAL
OF HEPATOLOGY

Contrast-enhanced ultrasound-based AI model for multi-classification of focal liver lesions

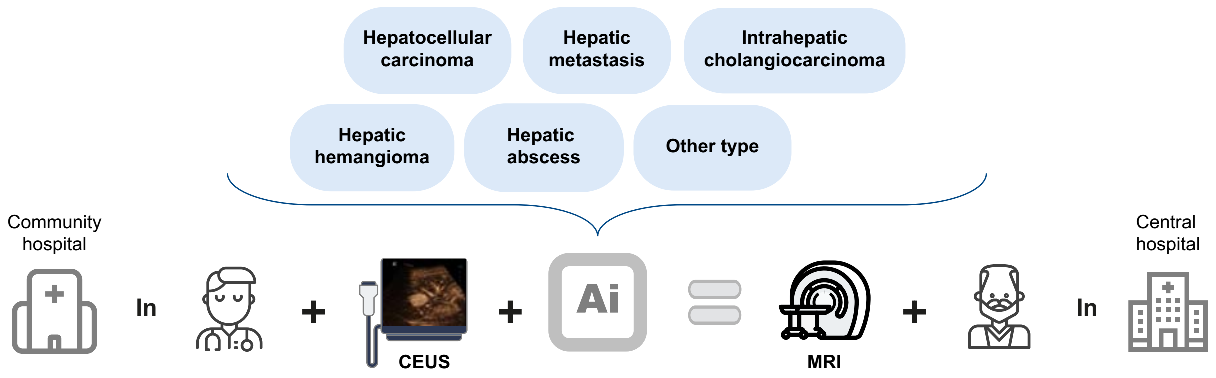
Authors

Wenzhen Ding, Yaqing Meng, Jun Ma, …, Jie Yu, Ping Liang, Kun Wang

Correspondence

jie.tian@ia.ac.cn (J. Tian), kun.wang@ia.ac.cn (K. Wang), jiemi301@163.com (J. Yu), liangping301@hotmail.com (P. Liang).

Graphical abstract



Highlights:

- Optical-flow and deep learning methods were applied to automatically analyze CEUS videos.
- Multi-type biomarker information was utilized for the non-invasive diagnostic strategy of ‘from CEUS to Biomarker to Disease’.
- By integrating CEUS, biomarker and clinical information, we built a model to classify six types of focal liver lesions.
- Model performance is better than junior CEUS radiologists, and comparable to senior CEUS/MRI radiologists.
- With model assistance, the performance of junior CEUS radiologists can be improved to the level of senior radiologists.

Impact and implications:

Ultrasound is the most common imaging examination for screening focal liver lesions (FLLs), but it lacks accuracy for multi-classification, which is a prerequisite for appropriate clinical management. Contrast-enhanced ultrasound (CEUS) offers better diagnostic performance but relies on the experience of radiologists. We developed a CEUS-based model (Model^{-DCB}) that can help junior CEUS radiologists to achieve comparable diagnostic ability as senior CEUS radiologists and senior MRI radiologists. The combination of an ultrasound device, CEUS examination and Model^{-DCB} means that even patients in remote areas can be accurately diagnosed through examination by junior radiologists.

https://doi.org/10.1016/j.jhep.2025.01.011

J. Hepatol. 2025, 83, 426–439

Research Article
Innovative Diagnostics, Modelling and Digital Hepatology

JOURNAL OF HEPATOLOGY

Contrast-enhanced ultrasound-based AI model for multi-classification of focal liver lesions

Wenzhen Ding[1,†], **Yaqing Meng**[2,3,†], **Jun Ma**[1,†], **Chuan Pang**[1,†], Jiapeng Wu[1], Jie Tian[2,3,4,*], Jie Yu[1,*], Ping Liang[1,*], Kun Wang[2,3,*]

Journal of Hepatology **2025**. vol. 83 | 426–439

Background & Aims: Accurate multi-classification is a prerequisite for appropriate management of focal liver lesions (FLLs). Ultrasound is the most common imaging examination but lacks accuracy. Contrast-enhanced ultrasound (CEUS) offers better performance but is highly dependent on operator experience. Therefore, we aimed to develop a CEUS-based artificial intelligence (AI) model for FLL multi-classification and evaluate its performance in multicenter clinical tests.

Methods: Since January 2017 to December 2023, CEUS videos, immunohistochemical biomarkers and clinical information on solid FLLs >1 cm in adults were collected from 52 centers to build and test the model. The model was developed to classify FLLs into six types: hepatocellular carcinoma, hepatic metastasis, intrahepatic cholangiocarcinoma, hepatic hemangioma, hepatic abscess and others. First, Module-Disease, Module-Biomarker and Module-Clinic were built in training set A and a validation set. Then, three modules were aggregated as Model$^{\text{-DCB}}$ in training set B and an internal test set. Model$^{\text{-DCB}}$ performance was compared with CEUS and MRI radiologists in three external test sets.

Results: In total 3,725 FLLs from 52 centers were divided into training set A (n = 2,088), the validation set (n = 592), training set B (n = 234), the internal test set (n = 110), and external test sets A (n = 113), B (n = 276) and C (n = 312). In external test sets A, B and C, Model$^{\text{-DCB}}$ achieved significantly better performance (accuracy from 0.85 to 0.86) than junior CEUS radiologists (0.59-0.73), and comparable performance to senior CEUS radiologists (0.79-0.85) and senior MRI radiologists (0.82-0.86). In multiple subgroup analyses on demographic characteristics, tumor characteristics and ultrasound devices, its accuracy ranged from 0.79 to 0.92.

Conclusions: CEUS-based Model$^{\text{-DCB}}$ provides accurate multi-classification of FLLs. It holds promise for a wide range of populations, especially those in remote areas who have difficulty accessing MRI.

Clinical trial: NCT04682886.

Introduction

The liver is one of the organs most prone to lesions in the human body.[1] Common types of focal liver lesions (FLLs) include hepatocellular carcinoma (HCC), hepatic metastatic carcinoma (HM), intrahepatic cholangiocarcinoma (ICC), hepatic hemangioma (HH) and hepatic abscess (HA),[2] while rare types include hepatic adenomas, hepatic lymphomas *etc.*[3] Treatment methods recommended by the guidelines vary greatly for different FLL types.[4,5] For example, even though HCC and HM are both malignant FLLs, their treatments are completely different.[6,7] Simply distinguishing between benign and malignant or only capable of identifying one or two types of FLL is not enough to provide sufficient information to make correct treatment decisions for the vast majority of patients with FLLs. Therefore, FLL diagnosis should shift to precise multi-classification.

Ultrasound is the most common and easily deployed imaging examination for liver disease. It is widely used in the screening of FLLs due to its convenience, low cost and real-time results, but its diagnostic capability for FLL is not satisfactory.[8] MRI is widely regarded as the best imaging examination for the diagnosis of FLLs, offering diagnostic performance second only to pathology. However, its high cost and logistical limitations affect its applicability to some extent. Contrast-enhanced ultrasound (CEUS) can transform conventional ultrasound examination from a "screening" to a "diagnostic" approach, and it plays an important role in FLL diagnosis.[9] It has high temporal resolution, which can coherently record the vascular perfusion, hemodynamic

[*] Corresponding authors. Addresses: CAS Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China. (K. Wang), or Department of Interventional Ultrasound, Chinese PLA General Hospital, 28 Fuxing Road, Beijing, 100853, China; (J. Yu), or (P. Liang).
E-mail addresses: jie.tian@ia.ac.cn (J. Tian), kun.wang@ia.ac.cn (K. Wang), jiemi301@163.com (J. Yu), liangping301@hotmail.com (P. Liang).

[†] Co–first authors.

https://doi.org/10.1016/j.jhep.2025.01.011

characteristics, and vascular distribution pattern of FLLs, and provides critical information for multi-classification.[10] However, CEUS still inevitably share some common dilemmas, including high reliance on experience and reduced diagnostic ability due to doctor fatigue.[11] Additionally, CEUS faces unique challenges due to its high temporal resolution. A typical CEUS video of FLL contains thousands of frames, far more than other liver imaging examinations. The rich information gives CEUS great potential in terms of diagnostic capabilities,[12] but how to accurately and comprehensively capture key dynamic features related to multi-type FLL classifications becomes the dilemma beyond the capabilities of ultrasound radiologists.[2]

To overcome these challenges, deep learning (DL)-based artificial intelligence (AI) is considered a promising solution.[13] AI can rapidly analyze and quantify large amounts of information, accurately discover and learn hidden features that doctors cannot identify, and objectively perform high-throughput diagnoses without feeling tired.[14] Various efforts have been made to develop AI models for FLL classification and have achieved benchmark-level progress. Kuang *et al.* constructed a CEUS-model with an AUC of 0.934 for identifying malignant FLLs in 211 patients.[15] Yang *et al.* constructed an ultrasound-model with an AUC of 0.913 for identifying hepatic echinococcosis in 548 patients.[16] However, AI models in these studies could only analyze static ultrasound/CEUS images in isolation and did not have the ability to coherently analyze dynamic videos. Moreover, these models only roughly differentiated certain common types of FLLs, lacking the ability to diagnose rare types.

Recent studies have revealed that the expression of biomarkers, such as hepatocyte antigen (Hep), glypican-3 (GPC3), cytokeratin (CK)7, and CK19, are important references for pathologists to diagnose FLL types.[17,18] However, such information can only be obtained invasively through liver biopsy or postoperative immunohistochemistry. We hypothesize that the biomarker expression at the microscopic level is reflected in the hidden spatiotemporal features of macroscopic CEUS videos, and they can be effectively recognized and learned by sophisticated DL models. Therefore, by integrating CEUS-based DL models trained to predict biomarker expression and DL models trained to directly classify FLLs, we should be able to further enhance the non-invasive multi-classification of patients with FLL during their CEUS examinations. However, this hypothesis has not been confirmed by relevant investigations yet.

In this study, we combined the conventional AI strategy of 'from image to FLL', the new AI strategy of 'from image to biomarker', and clinical characteristics to develop multiple CEUS-based DL models for accurately classifying HCC, HM, ICC, HH, HA, and other types (OT, including hepatic adenomas, hepatic lymphomas, focal nodular hyperplasia, neuroendocrine neoplasm, hepatic sarcoma and hepatic lipoma). Their performances were validated and compared in multiple large-sample, multicenter, prospective patient cohorts. Then, the best one was compared with CEUS and MRI radiologists, respectively. Furthermore, their ability to assist CEUS radiologists was also explored.

Patients and methods

This study was launched on January, 2017 (NCT04682886) and approved by the ethics committee of the Chinese PLA General Hospital (S2017-046-03). The registration site on ClinicalTrials.gov is the lead institution of this study, while the other 51 institutions were registered and approved for ethics at their respective hospitals. The inclusion criteria were: 1) distinct solid nodule larger than 1 cm with CEUS video; 2) diagnosed malignant nodules with pathological confirmation; 3) diagnosed benign nodules with clinical confirmation, MRI and follow-up. Detailed diagnostic criteria were shown in the Supplementary Material. The exclusion criteria were: 1) age less than 18 years; 2) poor CEUS image quality; 3) missing clinical information.

FLLs were classified into six types: HCC, HM, ICC, HH, HA, and OT. Clinical information of all FLLs were collected, including age, sex, tumor size, disease history, and serological index; Biomarker information of partial FLLs were collected, including Hep, GPC3, CK7 and CK19 (Supplementary Material).

CEUSs from 52 centers were collected to establish the AI model. From January 2017 to December 2022, cases from centers 1-36 were randomly divided into training set A and a validation set at a ratio of 4:1. Cases from centers 37-49 were all assigned to training set B. Cases from centers 50, 51 and 52 were assigned to the internal test set, and external test sets A and B, respectively. Since January 2023 to December 2023, cases from all 52 centers constituted the prospective external test set C (Fig. 1).

CEUS collection protocol

A bolus injection of 2 ml of SonoVue (Bracco SpA, Milan, Italy) was injected via the antecubital vein, followed by a 5 ml saline flush. CEUS videos were recorded continuously for arterial, portal, and delayed phases. In total 226 US devices from nine major manufacturers were used in different centers (Table S1).

CEUS processing

The annotation of FLLs was performed using a combination of manual initiation and automatic segmentation. A radiologist (eight-year CEUS experience) was invited to outline the lesion border on the frame with the largest tumor size in arterial phase, which was the only manual initiation. All subsequent steps of the process were automated. A rectangular box minimizing the coverage of this manually outlined region was then generated and expanded 20 pixels outward as the region of interest (ROI). After that, similar ROIs were generated on all frames of the CEUS video, so that the time-intensity curve (TIC) based on these ROIs was obtained (Fig. S1).

For each video, 80 frames were collected evenly from the beginning to the peak of TIC curve, and 20 frames would be collected evenly from the peak to the end (Fig. S1). Then, a specially designed two-stream model was used to extracted spatial and temporal information from these 100 ROIs (Supplementary Material and Fig. S2).[19] Each model included the spatial branch and the temporal branch (Fig. 2A). The spatial branch extracted spatial features through ResNet34.[20] The temporal branch combined the pixel displacement between two adjacent frames into optical flow maps,[21] and then extracted temporal features through ResNet18.[20] We found that if we downsampled each CEUS video into 100 frames, the corresponding optical flow changes in adjacent frames were too small, which misled model training, consumed excessive computing power, and sometimes caused our computing

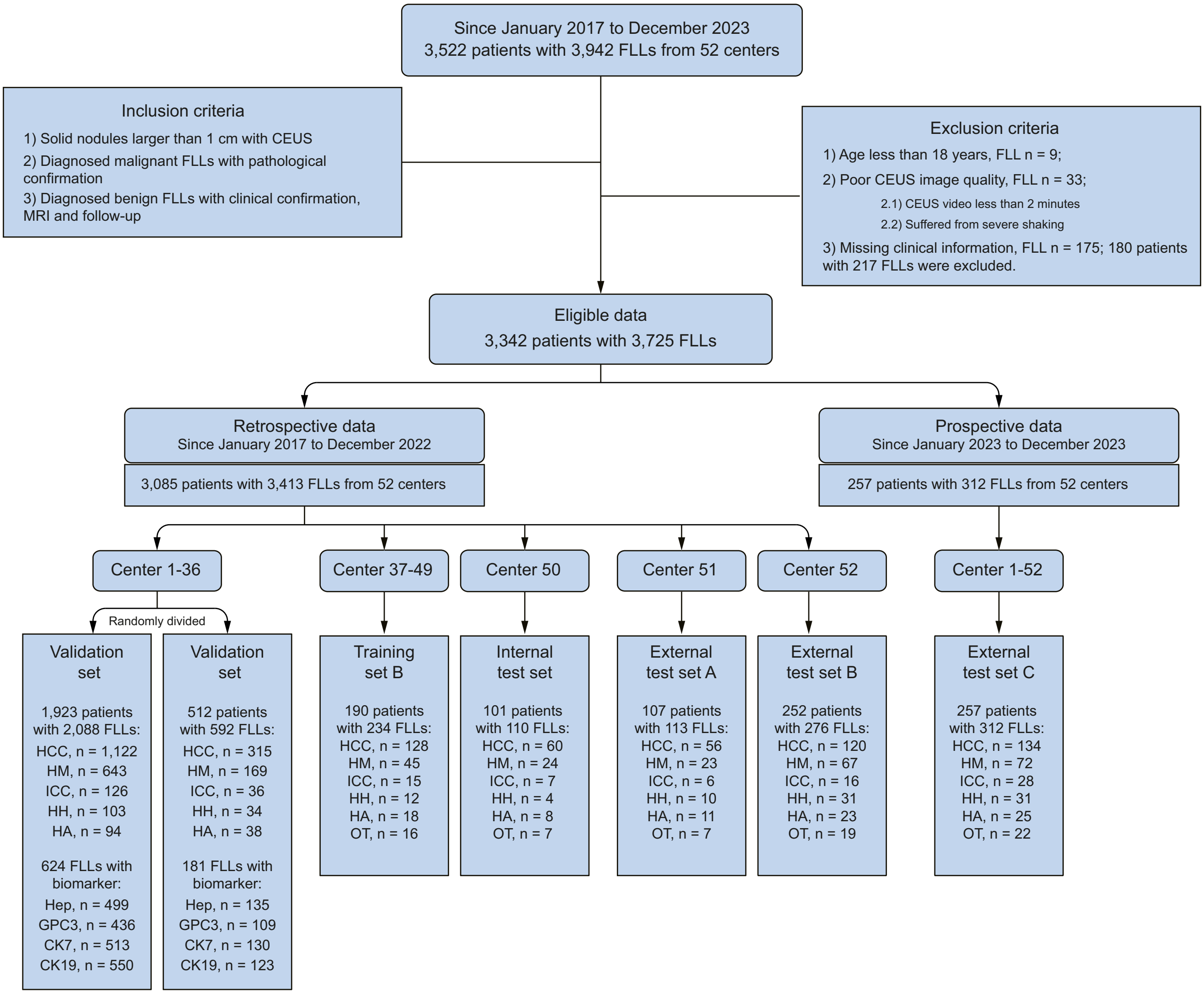

Since January 2017 to December 2023
3,522 patients with 3,942 FLLs from 52 centers
Inclusion criteria
1) Solid nodules larger than 1 cm with CEUS
2) Diagnosed malignant FLLs with pathological confirmation
3) Diagnosed benign FLLs with clinical confirmation, MRI and follow-up
Exclusion criteria
1) Age less than 18 years, FLL n = 9;
2) Poor CEUS image quality, FLL n = 33;
2.1) CEUS video less than 2 minutes
2.2) Suffered from severe shaking
3) Missing clinical information, FLL n = 175; 180 patients with 217 FLLs were excluded.
Eligible data
3,342 patients with 3,725 FLLs
Retrospective data
Since January 2017 to December 2022
3,085 patients with 3,413 FLLs from 52 centers
Prospective data
Since January 2023 to December 2023
257 patients with 312 FLLs from 52 centers
Center 1-36
Center 37-49
Center 50
Center 51
Center 52
Center 1-52
Randomly divided
Validation set
1,923 patients with 2,088 FLLs:
HCC, n = 1,122
HM, n = 643
ICC, n = 126
HH, n = 103
HA, n = 94
624 FLLs with biomarker:
Hep, n = 499
GPC3, n = 436
CK7, n = 513
CK19, n = 550
Validation set
512 patients with 592 FLLs:
HCC, n = 315
HM, n = 169
ICC, n = 36
HH, n = 34
HA, n = 38
181 FLLs with biomarker:
Hep, n = 135
GPC3, n = 109
CK7, n = 130
CK19, n = 123
Training set B
190 patients with 234 FLLs:
HCC, n = 128
HM, n = 45
ICC, n = 15
HH, n = 12
HA, n = 18
OT, n = 16
Internal test set
101 patients with 110 FLLs:
HCC, n = 60
HM, n = 24
ICC, n = 7
HH, n = 4
HA, n = 8
OT, n = 7
External test set A
107 patients with 113 FLLs:
HCC, n = 56
HM, n = 23
ICC, n = 6
HH, n = 10
HA, n = 11
OT, n = 7
External test set B
252 patients with 276 FLLs:
HCC, n = 120
HM, n = 67
ICC, n = 16
HH, n = 31
HA, n = 23
OT, n = 19
External test set C
257 patients with 312 FLLs:
HCC, n = 134
HM, n = 72
ICC, n = 28
HH, n = 31
HA, n = 25
OT, n = 22

Fig. 1. Flowchart of data collection and seven cohort division. Since January 2017 to December 2022, cases from centers 1-36 were randomly divided into training set A and a validation set. Cases from centers 37-49 were all divided into training set B. Cases from center 50, 51, and 52 were respectively divided into an internal test set, and external test sets A and B. Since January 2023 to December 2023, cases from all 52 centers constituted the prospective external test set C. CEUS, contrast-enhanced ultrasound; FLLs, focal liver lesions; HA, hepatic abscess; HCC, hepatocellular carcinoma; HH, hepatic hemangioma; HM, hepatic metastasis; ICC, intrahepatic cholangiocarcinoma; OT, other type.

system to crash (two RTX 4090 Ti graphics cards (NVIDIA, USA) for experiments). Therefore, we compared the performances of downsampling the optical flow map to 20, 15, 10, and 5 frames per video, and finally selected 10 frames per video as the model input to ensure a balance between inference efficiency, classification accuracy, and system stability (Fig. S3).

Two-stage model construction

In the first stage, three modules were built, namely Module-Disease, Module-Biomarker, and Module-Clinic (Fig. 2B). LSTM (long short-term memory) was used to integrate the extracted features from the spatial and temporal branches to develop Module-Disease and Module-Biomarker.[22] Prediction probability was the output after adaptive weighting. We adopted the weighted sum of spatial branch classification loss, temporal branch classification loss, and final classification loss as loss functions. Minority class weighting was used to overcome imbalances caused by class distributions. Detailed description, function definitions and related formulas are provided in the Supplementary Material.

In Module-Disease, we compared two model construction strategies. Strategy A was to directly build the multi-classification model. Strategy B adopted the distributed training approach and aggregated multiple binary-classification sub-models to build the multi-classification model. Five sub-models were trained, including: HCC *vs.* non-HCC, HM *vs.* non-HM, ICC *vs.* non-ICC, HH *vs.* non-HH, and HA *vs.* non-HA. In Module-Biomarker, four binary-classification sub-models were developed, including: Hep-positive *vs.* Hep-negative,

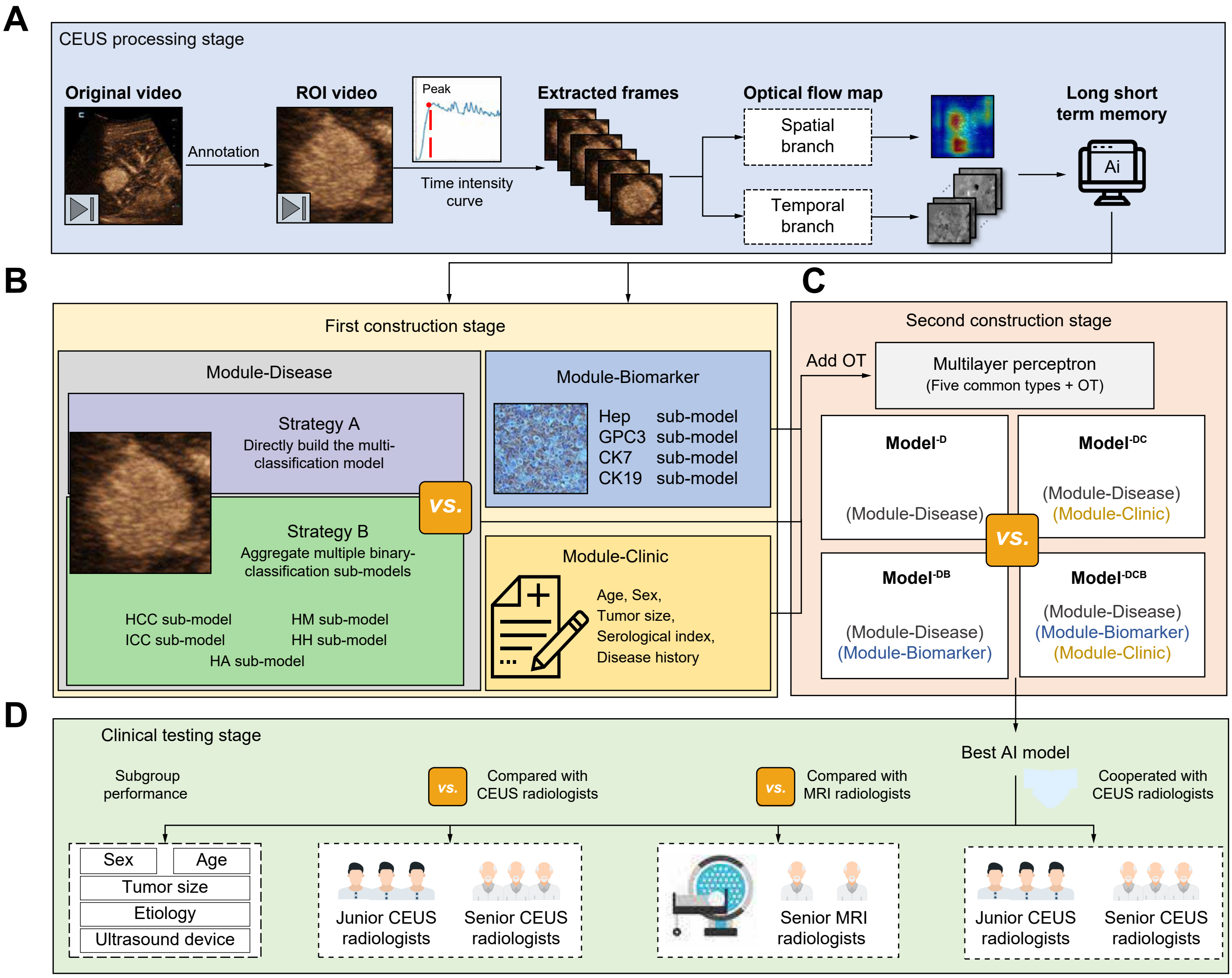


Fig. 2. Experimental design flowchart. (A) In the CEUS processing stage, we extracted downsampled key-frames from a CEUS video based on its TIC, generated the optical flow map from this keyframe sequence, and integrated hidden features from both spatial and temporal branches by LSTM. This strategy was applied to construct the Module-Disease and Module-Biomarker models, respectively. (B) In the first construction stage, three modules were developed. The Module-Disease were obtained by comparing direct classification of the five FLL types (Strategy A) with aggregating multiple binary-classification sub-models (Strategy B). The Module-Biomarker contained four trained binary-classification sub-models for four biomarkers. The Module-Clinic were built by selecting key clinical information with high correlation to FLLs. (C) In the second construction stage, four multi-classification models, namely Model^{-D}, Model^{-DC}, Model^{-DB}, and Model^{-DCB}, were built, and each one was composed of corresponding modules through multilayer perceptron. Their performances were compared, and the best one was selected for the nest stage. (D) In the clinical testing stage, we compared the AI model performance with six CEUS radiologists (three junior radiologists and three senior radiologists) and two MRI senior radiologists. The cooperation between CEUS radiologists and the model was explored. The performances of the model in multiple subgroups were also evaluated. CEUS, contrast-enhanced ultrasound; LSTM, long short-term memory; FLLs, focal liver lesions; HA, hepatic abscess; HCC, hepatocellular carcinoma; HH, hepatic hemangioma; HM, hepatic metastasis; ICC, intrahepatic cholangiocarcinoma; OT, other type; TIC, time-intensity curve. (This figure appears in color on the web.)

GPC3-positive *vs.* GPC3-negative, CK7-positive *vs.* CK7-negative, and CK19-positive *vs.* CK19-negative. In Module-Clinic, all collected clinical information was filtered by the correlation coefficient of FLL type (Supplementary Material).

In the second stage, the corresponding modules were combined to build four multi-classification models through multilayer perceptron (MLP),[23] namely Model^{-D}, Model^{-DC}, Model^{-DB}, and Model^{-DCB}, for performance comparison (Fig. 2C). Moreover, we added OT cases as a complementary classification in this stage, so that six types of FLL were classified (five common types + OT). This was achieved by leveraging MLP without training an OT *vs.* non-OT sub-model, but setting OT as an extra classification outside the original five common types in MLP (Fig. S4). The basic rule was that if a case was predicted to have low probabilities for all five FLL types, it would be classified as OT. This was because the number of OT cases was bound to be too small to support the CEUS video-based sub-model training.

Model evaluation

Model performance was evaluated in four test sets through multi-classification index, subgroup analysis (sex, age, tumor size, etiology, cirrhosis, fatty liver, and ultrasound device manufacturer), and six-type radar chart analysis. We also conducted subgroup analyses on cases with pathological diagnosis of FLLs, pathological diagnosis of cirrhosis, and pathological diagnosis of fatty liver. In the external test sets A, B, and C, we compared model performance with six CEUS radiologists (three junior and three senior radiologists) and two senior MRI radiologists (Table S2).[24] All radiologists were blinded to the FLL diagnoses. The illustration of CEUS imaging findings used for FLL diagnosis by CEUS radiologists were shown in Fig. S5. After a 1-month washout period, we evaluated the performance of these CEUS radiologists with model assistance and compared it with their performance 1 month earlier (Fig. 2D). We also recorded the changes in diagnosis time and diagnostic confidence of CEUS radiologists before and after model assistance.

Model visualization

To generate a visual explanation of the model diagnostic process, we converted feature maps into pseudo-colored maps using the OpenCV method through Grad-CAM (Gradient Weighted Class Activation Mapping), which displays the pixels in the ROIs that provide the greatest contribution to the classification output.[25] In addition, we also displayed the disease classification probabilities from Module-Disease and biomarker probabilities from Module-Biomarker.

Model generalization and robustness experiment

Since the study population was patients with FLL who underwent CEUS, the FLL distribution did not align with the natural FLL distribution. We performed a generalizability experiment for the model by simulating three clinical scenarios. First, health check-up center (a lower proportion of malignant tumors and a higher proportion of benign tumors, 1:9); second, hospital outpatient clinic (an equal proportion of malignant and benign tumors, 5:5); three, hospital inpatient ward (a higher proportion of malignant tumors and a lower proportion of benign tumors, 9:1). We randomly selected cases from three external test sets according to the above ratios to form the following cohorts: health check-up center cohort (n = 100), hospital outpatient cohort (n = 100), and hospital inpatient cohort (n = 100). We used the model to perform diagnoses in these three cohorts and observed the impact of varying FLL prevalence on model performance. This experiment was repeated 100 times (Fig. S6A).

Since we used a combination of manual and automated ROI annotation, we performed a robustness test to assess the impact of variability in manual annotations on the model's performance. By expanding, shrinking, moving, or combining these methods, we randomly adjusted the manually annotated ROI by the radiologists to simulate the impact of variability in annotation on model performance. This experiment was repeated 100 times (Fig. S6B).

Statistical analysis

Continuous variables were summarized as means ± SDs, and categorical variables were categorized as numbers and percentages. Performance of the binary-classification model was visualized by a ROC (receiver-operating characteristic) curve and evaluated by AUC. Performance of the multi-classification model was displayed by confusion matrix and Macro-ROC, and evaluated by Accuracy, Macro-AUC, Macro-Specificity, Macro-Recall (same as sensitivity), Macro-Precision (same as positive predictive value), Macro-NPV (negative predictive value), Macro-F1, and six-type radar chart analysis (Supplementary Material). 95% CIs were evaluated by bootstrapping with 1,000 resamples. In addition, for multi-class diagnosis research, Accuracy could better evaluate model performance than Macro-AUC. Therefore, for comparisons between AI models and comparisons between AI and radiologists, we conducted significance analysis on Accuracy by McNemar test.[26] p <0.05 indicated significant difference. Statistical analysis was performed using R (Version 4.0.0).

Results

Clinical characteristics

Since January 2017 to December 2022, 2,914 FLLs from 49 centers were retrospectively collected and divided into training set A (n = 2,088), the validation set (n = 592), and training set B (n = 234). FLLs from the other three independent centers were divided into the internal test set (n = 110), and external test sets A (n = 113) and B (n = 276). Since January 2023 to December 2023, 312 FLLs from 52 centers were prospectively collected as the prospective external test set C (Fig. 1, Table 1). As a result, a total of 3,725 CEUS videos (corresponding to 3,725 FLLs) from 3,342 patients were collected in this study (Table S3). Each CEUS video continuously recorded the arterial phase, portal phase, and delayed phase of the FLL, containing more than 1,000 frames. In other words, more than 9,000 min (nearly 4,000,000 frames) were collected to develop and validate the CEUS-based FLL multi-classification AI model.

Patient baseline characteristics (age, sex, viral hepatitis history, malignancy history, tumor size, and FLL type) showed no significant differences between training set A and the validation set, between training set B and the internal test set, and between external test sets A, B, and C (all p >0.05, Table 1). Additionally, training set A had 624 FLLs with various biomarker information, including Hep (n = 499), GPC3 (n = 436), CK7 (n = 513), and CK19 (n = 550). The validation set had 181 FLLs with various biomarker information, including Hep (n = 135), GPC3 (n = 109), CK7 (n = 130), and CK19 (n = 123) (Fig. 1). There were no significant differences in the biomarker information between these two datasets (all p >0.20, Table S4).

Performance of Module-Disease

Compared with strategy A, strategy B had a generally better performance in training set A and the validation set (Table S5). In the validation set, Strategy B showed more correct classifications in the confusion matrix, and it was significantly better than strategy A in terms of Accuracy (0.83, 95% CI 0.80-0.86 *vs.* 0.77, 95% CI 0.74-0.79, p <0.001), while also exhibiting better Macro-AUC (0.86, 95% CI 0.83-0.89 *vs.* 0.81, 95% CI 0.79-0.83), Macro-Specificity (0.94, 95% CI 0.89-0.97 *vs.* 0.92, 95% CI 0.88-0.95), Macro-Recall (0.79, 95% CI 0.76-0.82 *vs.* 0.71, 95% CI 0.67-0.75), Macro-Precision (0.85, 95% CI 0.82-0.88 *vs.* 0.75, 95% CI 0.72-0.77), Macro-NPV (0.95, 95% CI 0.89-0.98 *vs*. 0.93, 95% CI 0.87-0.95) and Macro-F1 (0.81, 95% CI 0.78-0.83 *vs.* 0.72, 95% CI 0.68-0.76) (Fig. 3A). Therefore, strategy B was chosen to build Module-Disease, and AUCs of five binary-classification sub-models in strategy B are listed in Table S6.

Table 1. Baseline characteristics of training, validation and test sets.

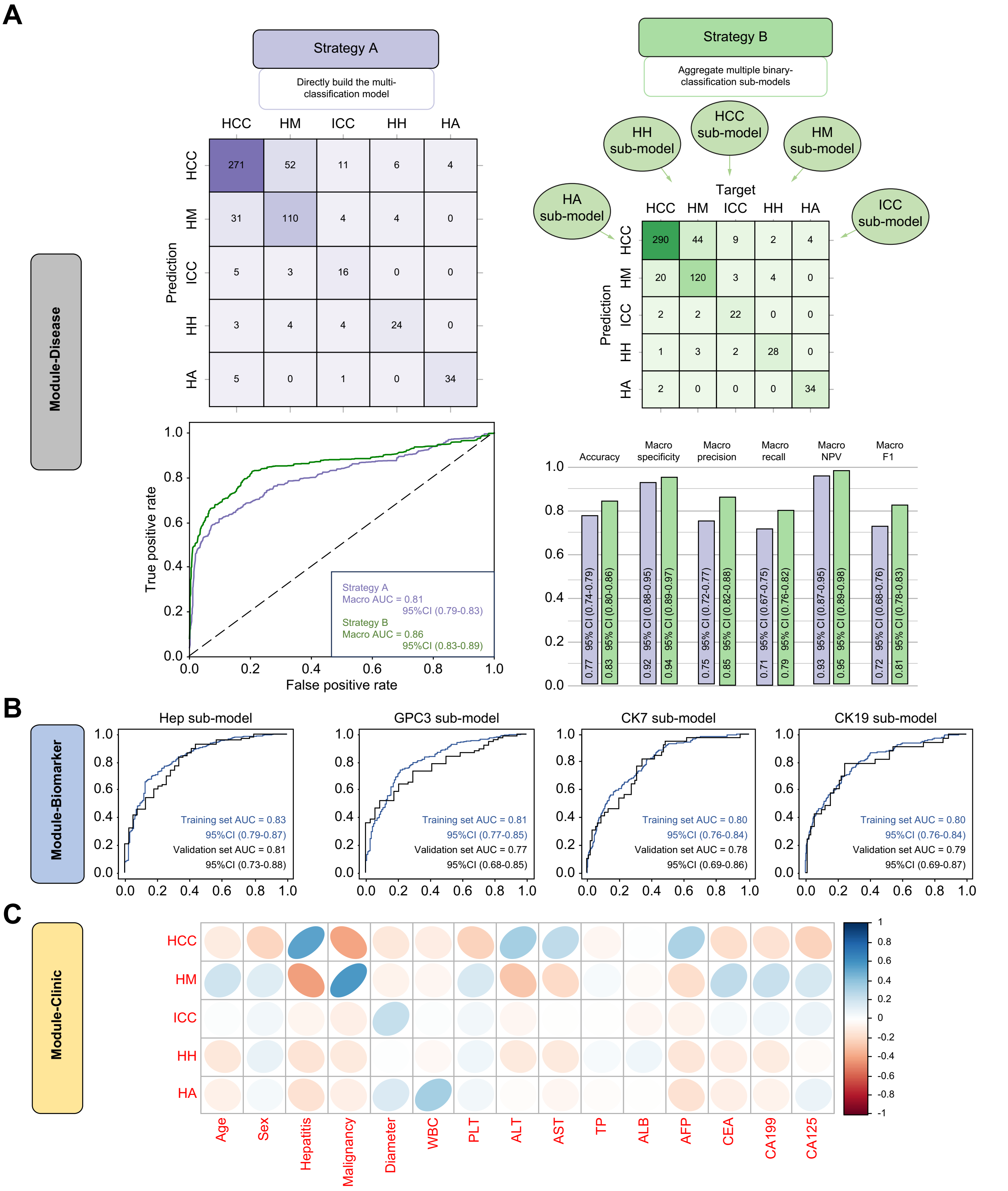
	Training set A	Validation set		Training set B	Internal test set		External test set A	External test set B	External test set C	
	n = 2,088	n = 592	*p*1	n = 234	n = 110	*p*2	n = 113	n = 276	n = 312	*p*3
Age (%)			0.362			0.938				0.745
<50 years	590 (28.3%)	150 (25.3%)		34 (14.5%)	17 (15.5%)		27 (23.9%)	57 (20.7%)	66 (21.2%)	
50-60 years	732 (35.1%)	219 (37.0%)		83 (35.5%)	37 (33.6%)		35 (31.0%)	98 (35.5%)	118 (37.8%)	
>60 years	766 (36.7%)	223 (37.7%)		117 (50.0%)	56 (50.9%)		51 (45.1%)	121 (43.8%)	128 (41.0%)	
Sex (%)			0.127			0.678				0.794
Male	1,515 (72.6%)	410 (69.3%)		157 (67.1%)	77 (70.0%)		80 (70.8%)	192 (69.6%)	225 (72.1%)	
Female	573 (27.4%)	182 (30.7%)		77 (32.9%)	33 (30.0%)		33 (29.2%)	84 (30.4%)	87 (27.9%)	
Viral hepatitis history (%)			0.139			0.634				0.822
No	952 (45.6%)	249 (42.1%)		109 (46.6%)	55 (50.0%)		53 (46.9%)	139 (50.4%)	155 (49.7%)	
Yes	1,136 (54.4%)	343 (57.9%)		125 (53.4%)	55 (50.0%)		60 (53.1%)	137 (49.6%)	157 (50.3%)	
Malignancy history (%)			0.335			0.975				0.902
No	1,232 (59.0%)	363 (61.3%)		153 (65.4%)	71 (64.5%)		76 (67.3%)	179 (64.9%)	205 (65.7%)	
Yes	856 (41.0%)	229 (38.7%)		81 (34.6%)	39 (35.5%)		37 (32.7%)	97 (35.1%)	107 (34.3%)	
Tumor size (%)			0.08			0.11				0.821
<3 cm	457 (21.9%)	127 (21.5%)		46 (19.7%)	32 (29.1%)		28 (24.8%)	53 (19.2%)	72 (23.1%)	
3-5 cm	835 (40.0%)	269 (45.4%)		120 (51.3%)	42 (38.2%)		39 (34.5%)	100 (36.2%)	99 (31.7%)	
5-10 cm	624 (29.9%)	157 (26.5%)		52 (22.2%)	27 (24.5%)		33 (29.2%)	88 (31.9%)	103 (33.0%)	
>10 cm	172 (8.2%)	39 (6.6%)		16 (6.8%)	9 (8.2%)		13 (11.5%)	35 (12.7%)	38 (12.2%)	
FLL type (%)			0.305			0.986				0.891
HCC	1,122 (53.7%)	315 (53.2%)		128 (54.7%)	60 (54.5%)		56 (49.6%)	120 (43.5%)	134 (42.9%)	
HM	643 (30.8%)	169 (28.5%)		45 (19.2%)	24 (21.8%)		23 (20.4%)	67 (24.3%)	72 (23.1%)	
ICC	126 (6.0%)	36 (6.1%)		15 (6.4%)	7 (6.4%)		6 (5.3%)	16 (5.8%)	28 (9.0%)	
HH	103 (4.9%)	34 (5.7%)		12 (5.1%)	4 (3.6%)		10 (8.8%)	31 (11.2%)	31 (9.9%)	
HA	94 (4.5%)	38 (6.4%)		18 (7.7%)	8 (7.3%)		11 (9.7%)	23 (8.3%)	25 (8.0%)	
OT	0 (0.0%)	0 (0.0%)		16 (6.8%)	7 (6.4%)		7 (6.2%)	19 (6.9%)	22 (7.1%)	

FLL, focal liver lesion; HA, hepatic abscess; HCC, hepatocellular carcinoma; HH, hepatic hemangioma; HM, hepatic metastasis; ICC, intrahepatic cholangiocarcinoma; OT, other type.

*p*1 showed the difference between training set A and the validation set.

*p*2 showed the difference between training set B and the internal test set.

*p*3 showed the difference among external test sets A, B and C.



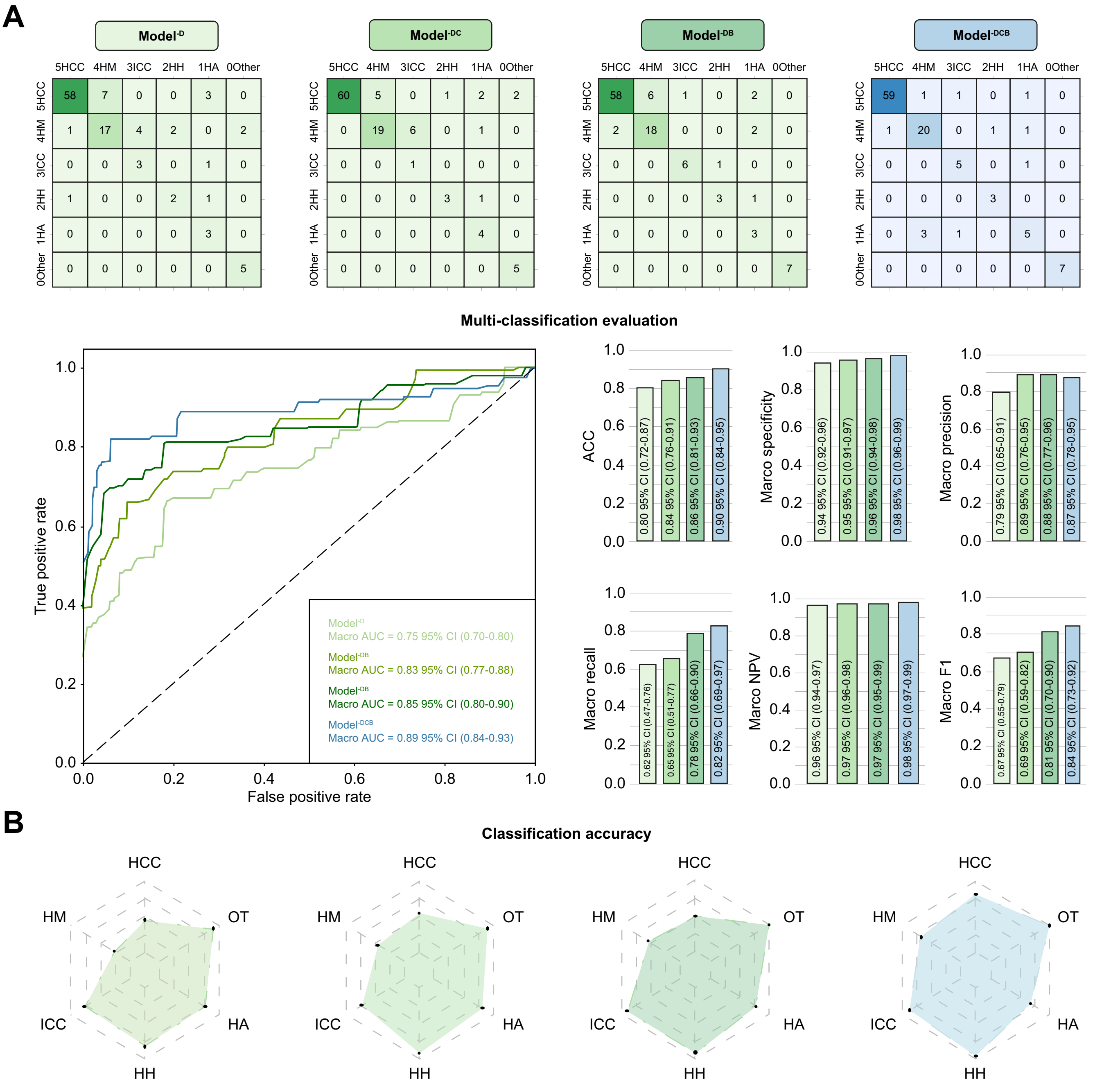
Strategy A

Prediction	HCC	HM	ICC	HH	HA
HCC	271	52	11	6	4
HM	31	110	4	4	0
ICC	5	3	16	0	0
HH	3	4	4	24	0
HA	5	0	1	0	34

Strategy B

Prediction	HCC	HM	ICC	HH	HA
HCC	290	44	9	2	4
HM	20	120	3	4	0
ICC	2	2	22	0	0
HH	1	3	2	28	0
HA	2	0	0	0	34

Fig. 3. Performance evaluation of Module-Disease, Module-Biomarker, and Module-Clinic in the first construction stage. (A) In Module-Disease, for the five-type FLL classification in the validation set (n = 592), strategy B (green) using distributed training generally outperformed strategy A (purple) using direct training, in terms of confusion matrix, Macro-AUC, Accuracy, Macro-Precision, Macro-Recall, and Macro-F1. (B) In Module-Biomarker, receiver-operating characteristic curves of the four biomarker sub-models are shown for the training set A (blue) and validation set (black). (C) In Module-Clinic, 17 types of clinical information were selected, because each of them had a correlation coefficient of more than the moderate degree (absolute value >0.2) with at least one FLL type. Ellipse direction represented the



Model^{-D}

	5HCC	4HM	3ICC	2HH	1HA	0Other
5HCC	58	7	0	0	3	0
4HM	1	17	4	2	0	2
3ICC	0	0	3	0	1	0
2HH	1	0	0	2	1	0
1HA	0	0	0	0	3	0
0Other	0	0	0	0	0	5

Model^{-DC}

	5HCC	4HM	3ICC	2HH	1HA	0Other
5HCC	60	5	0	1	2	2
4HM	0	19	6	0	1	0
3ICC	0	0	1	0	0	0
2HH	0	0	0	3	1	0
1HA	0	0	0	0	4	0
0Other	0	0	0	0	0	5

Model^{-DB}

	5HCC	4HM	3ICC	2HH	1HA	0Other
5HCC	58	6	1	0	2	0
4HM	2	18	0	0	2	0
3ICC	0	0	6	1	0	0
2HH	0	0	0	3	1	0
1HA	0	0	0	0	3	0
0Other	0	0	0	0	0	7

Model^{-DCB}

	5HCC	4HM	3ICC	2HH	1HA	0Other
5HCC	59	1	1	0	1	0
4HM	1	20	0	1	1	0
3ICC	0	0	5	0	1	0
2HH	0	0	0	3	0	0
1HA	0	3	1	0	5	0
0Other	0	0	0	0	0	7

Fig. 4. Performance evaluation of Model^{-D}, Model^{-DC}, Model^{-DB} and Model^{-DCB} in the second construction stage. (A) In the internal test set (n = 110), Model^{-DCB} (light blue) outperformed Model-D (pale green), Model^{-DC} (light green), and Model^{-DB} (dark green) in terms of confusion matrix, Macro-AUC, ACC, Macro-Specificity, Macro-Recall, Macro-NPV, and Macro F1. (B) Radar charts demonstrate the different performances between these four models for the six-type classification of FLLs. ACC, Accuracy; FLL, focal liver lesion; HA, hepatic abscess; HCC, hepatocellular carcinoma; HH, hepatic hemangioma; HM, hepatic metastasis; ICC, intrahepatic cholangiocarcinoma; OT, other type. (This figure appears in color on the web.)

positive or negative correlation between factor and FLL type. Ellipse color and shape represented the correlation degree. AFP, alpha-fetoprotein; ALB, albumin; ALT, alanine aminotransferase; AST, aspartate aminotransferase; CA19-9, cancer antigen 19-9; CA125, cancer antigen 125; CEA, carcinoembryonic antigen; CEUS, contrast-enhanced ultrasound; CK, cytokeratin; LSTM, long short-term memory; FLLs, focal liver lesions; GPC3, glypican-3; HA, hepatic abscess; HCC, hepatocellular carcinoma; HH, hepatic hemangioma; HM, hepatic metastasis; ICC, intrahepatic cholangiocarcinoma; OT, other type; PLT, platelet count; TP, total protein; WBC, white blood cell count. (This figure appears in color on the web.)

Performance of Module-Biomarker

Four biomarker sub-models were also developed and validated in training set A and the validation set, and their AUCs were: 0.83, 95% CI 0.79-0.87 and 0.81, 95% CI 0.73-0.89 for the Hep sub-model; 0.81, 95% CI 0.77-0.85 and 0.77, 95% CI 0.68-0.85 for the GPC3 sub-model; 0.80, 95% CI 0.76-0.84 and 0.78, 95% CI 0.69-0.86 for the CK7 sub-model; and 0.80, 95% CI 0.76-0.84 and 0.79, 95% CI 0.69-0.87 for the CK19 sub-model (Fig. 3B, Table S6). These results revealed a good biomarker prediction capability of our CEUS-based AI model.

Selected clinical information in Module-Clinic

Seventeen types of clinical information were chosen in Module-Clinic, including age, sex, alpha-fetoprotein, carcinoembryonic antigen, cancer antigen 19-9 *etc* (Fig. 3C). Each of them had a correlation coefficient of more than moderate degree (absolute value >0.2) with at least one FLL type (Table S7).

Performances of four multi-classification models

Model^{-D}, Model^{-DC}, Model^{-DB}, and Model^{-DCB} were trained in training set B and evaluated in the internal test set for the six-

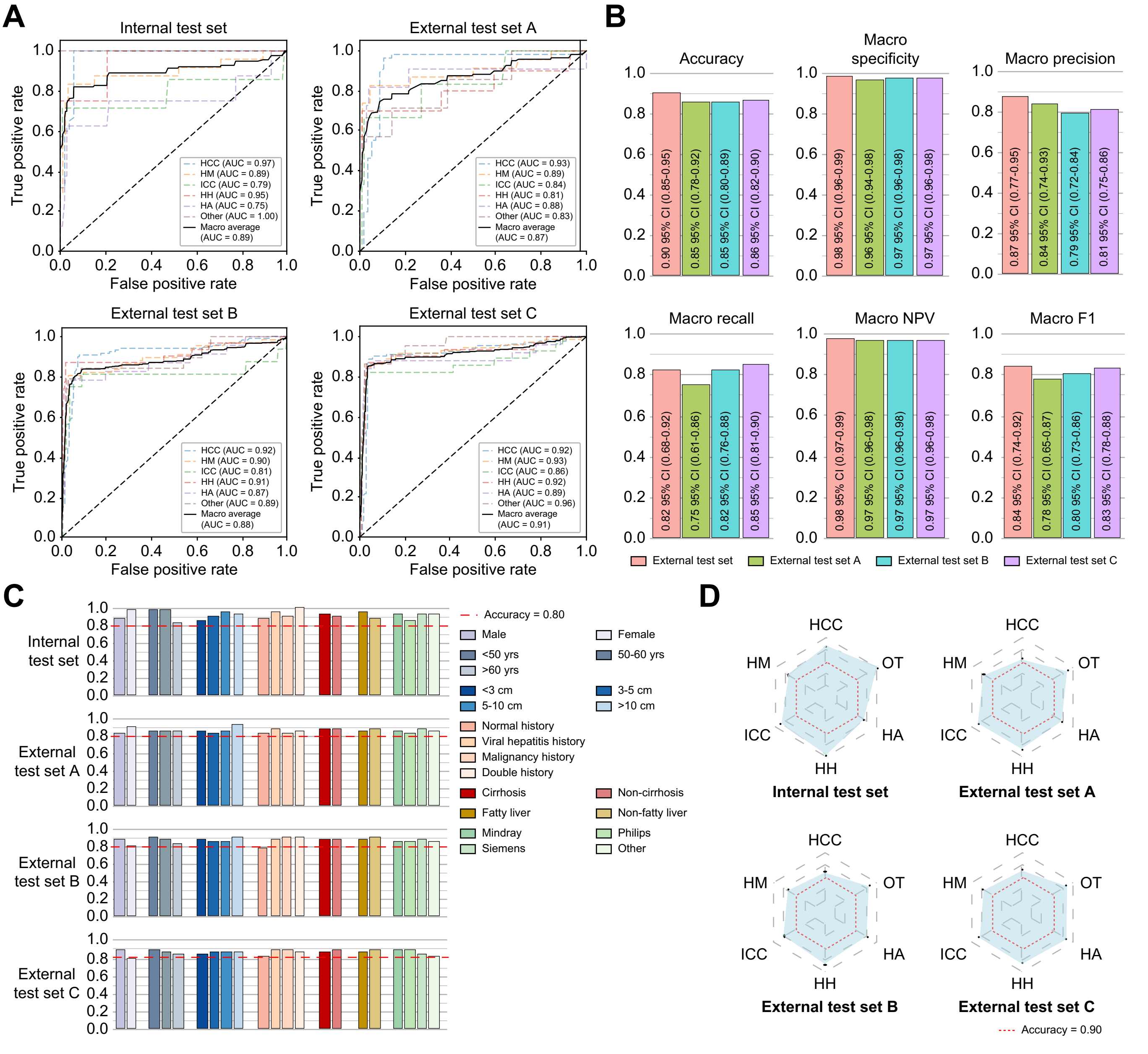


Fig. 5. Performances of Model^{-DCB} in the internal test set and external test sets A, B, and C. (A) Receiver-operating characteristic curves and Macro-AUCs of Model^{-DCB} in four test sets. (B) Accuracy, Macro-Specificity, Macro-Precision, Macro-Recall, Macro-NPV, and Macro-F1 of Model-DCB in four test sets. (C) Most accuracies of Model-DCB exceeded 0.80 (red dotted lines) in subgroup analyzes, including gender, age, tumor size, etiology, cirrhosis, fatty liver, and ultrasound device manufacturer; (D) All accuracies for the six-type classification of Model^{-DCB} exceeded 0.90 (red dotted hexagons) in four test sets. HA, hepatic abscess; HCC, hepatocellular carcinoma; HH, hepatic hemangioma; HM, hepatic metastasis; ICC, intrahepatic cholangiocarcinoma; OT, other type. (This figure appears in color on the web.)

type classification of FLLs. Our results showed that Model^{-DCB} outperformed the other three models in both datasets (Table S8). In the internal test set, Model^{-DCB} was clearly the best in the confusion matrix, and provided the highest Accuracy (0.90, 95% CI 0.84-0.95), Macro-AUC (0.89, 95% CI 0.84-0.93), Macro-Specificity (0.98, 95% CI 0.97-0.99), Macro-Recall (0.82, 95% CI 0.69-0.97), Macro-NPV (0.98, 95% CI 0.97-0.99), and Macro-F1 (0.84, 95% CI 0.73-0.92) among the four models (Fig. 4A). Therefore, it was chosen for further investigations in the next clinical testing stage.

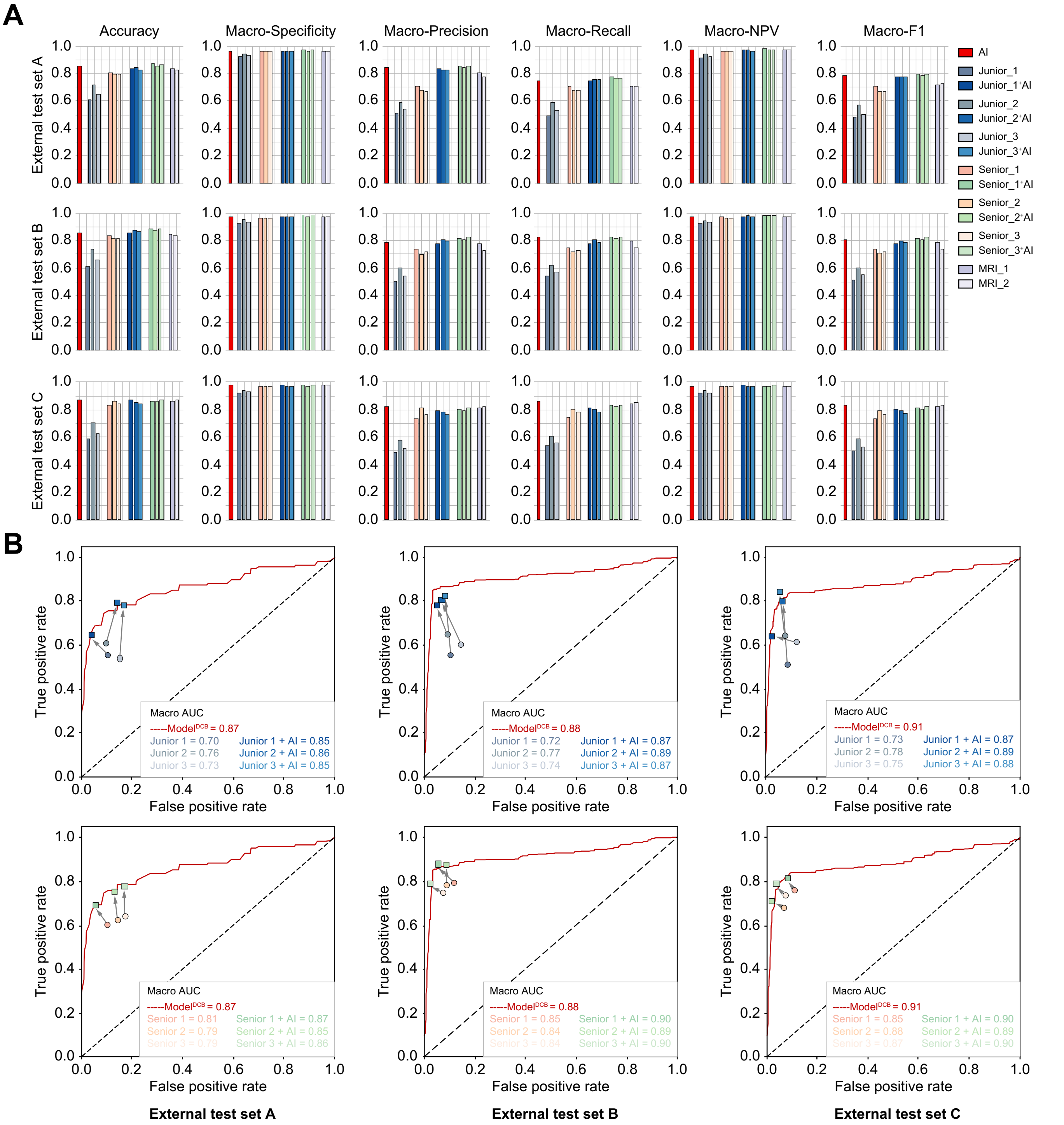


Fig. 6. Comparison and cooperation between Model^{-DCB} and radiologists in external test sets A, B, and C. (A) The Accuracy, Macro-Specificity, Macro-Precision, Macro-Recall, Macro-NPV, and Macro-F1 of Model^{-DCB}, junior CEUS radiologists (without and with AI assistance), senior CEUS radiologists (without and with AI assistance), and senior MRI radiologists were plotted for comparisons. (B) With the Model^{-DCB} assistance, junior CEUS radiologists achieved much better Macro-AUC than before. (C) With the Model^{-DCB} assistance, senior CEUS radiologists achieved slightly better Macro-AUC than before. (This figure appears in color on the web.)

In the internal test set, compared with Model^{-D}, Model^{-DC} had a 5.0% improvement in Accuracy (0.80, 95% CI 0.72-0.87 *vs.* 0.84, 95% CI 0.76-0.91, p = 0.28), but Model^{-DB} achieved 7.5% improvement (0.80, 95% CI 0.72-0.87 *vs.* 0.86, 95% CI 0.81-0.93, p = 0.19), which revealed that Module-Biomarker had a greater effect on improving classification accuracy than Module-Clinic. For Model^{-D}, Model^{-DC}, and Model^{-DB}, their accuracy in diagnosing HCC and HM was always worse than that in diagnosing the other four types of FLLs (Fig. 4B). However, by integrating all three modules through MLP, Model^{-DCB} achieved similar levels of diagnostic accuracy for all six types of FLLs, reaching or exceeding 0.94 (Table S8).

Performances of Model^{-DCB} in retrospective and prospective test sets

Model^{-DCB} had accuracies (95% CI) of 0.90 (0.85-0.95), 0.85 (0.78-0.92), 0.85 (0.80-0.89), and 0.86 (0.82-0.90) in the internal test set, and external test sets A, B, and C, respectively (Fig. 5A). Macro-AUC, Macro-Specificity, Macro-Recall, Macro-Precision, Macro-NPV, and Macro-F1 of Model^{-DCB} ranged from 0.87 to 0.91, from 0.96 to 0.98, from 0.75 to 0.85, from 0.79 to 0.87, from 0.97 to 0.98, and from 0.78 to 0.84 in the four test sets (Fig. 5B, Table S9). Its accuracies in all subgroups ranged from 0.77 to 1.00 (Fig. 5C and Fig. S7, Table S9). Accuracies of the six-type classification of FLLs were all above 0.90 (HCC 0.91-0.96; HM 0.93-0.94; ICC 0.95-0.97; HH 0.96-0.99, HA 0.94-0.97 and OT 0.97-1.00, Fig. 5D, Table S9). These results indicated that Model^{-DCB} was accurate, stable and reliable in the multi-type FLL classifications.

Comparison with CEUS and MRI radiologists

After comparing Model^{-DCB} with three junior CEUS radiologists, three senior CEUS radiologists, and two Senior MRI radiologists in external test sets A, B, and C, we found it was significantly better than junior CEUS radiologists in terms of accuracy (all p <0.05), and comparable to senior CEUS radiologists and MRI radiologists (all p >0.05, Fig. 6A). Detailed performance comparison data are listed in Tables 2 and Fig. S10. We also performed subgroup analyses between Model^{-DCB} and two Senior MRI radiologists in these three datasets, and no significant differences emerged (Table S11).

Cooperation with CEUS radiologists

One month later, all six CEUS radiologists performed the diagnosis again with Model^{-DCB} assistance in three external validation sets. The information provided by Model^{-DCB} to CEUS radiologists was shown in Fig. S8. Accuracy significantly improved for all junior CEUS radiologists (from 0.59-0.73 to 0.82-0.87, all p <0.05) (Fig. 6B), as did Macro-Recall (from 0.49-0.60 to 0.75-0.80), Macro-Precision (from 0.49-0.60 to 0.75-0.83) and Macro-F1 (from 0.47-0.58 to 0.76-0.80). Similar diagnostic enhancement was also observed for senior CEUS radiologists, but the improvements were not statistically significant (all p >0.05). Detailed comparison data are listed in Table S10. The addition of model assistance did not significantly affect the diagnosis time of radiologists, but slightly improved the diagnostic confidence of junior radiologists (Tables S12 and S13).

Model visualization

The optical flow map, heat map, Module-Disease probability scores of its five sub-models, and Module-Biomarker probability scores of its four sub-models are visualized for three different cases in Fig. S8 as examples. For the patients with HCC, the optical flow maps and heat maps of CEUS frames displayed a rapid change pattern (Fig. S8A and B), which probably corresponded to a typical CEUS dynamic feature in HCC: “rapid hyperenhancement in arterial phase and washout in portal phase” (Fig. S5). This pattern was rarely seen in other FLL types (Fig. S8C).

Model generalization and robustness experiment

In the generalization experiment, there was no significant difference in the accuracy of Model^{-DCB} in three simulated clinical scenarios, which proved that the difference of FLL prevalence did not affect the performance of Model^{-DCB} (Fig. S9A).

In the robustness experiment, we found that randomly adjusting the ROI did not have a significant impact on the

Table 2. Clinical testing stage.

	External test set C (prospective multicenter test set)						
	Accuracy	*p*	Macro-Specificity	Macro-Recall	Macro-Precision	Macro-NPV	Macro-F1
AI	0.86 (0.82–0.90)	reference	0.97 (0.96–0.98)	0.85 (0.81–0.90)	0.81 (0.75–0.86)	0.97 (0.96–0.98)	0.83 (0.78–0.88)
Junior1	0.59 (0.53–0.64)	<0.01	0.92 (0.90–0.93)	0.53 (0.47–0.60)	0.49 (0.43–0.54)	0.91 (0.90–0.92)	0.50 (0.44–0.56)
Junior1+AI	0.83 (0.80–0.88)	0.31	0.97 (0.96–0.97)	0.77 (0.71–0.83)	0.75 (0.69–0.81)	0.96 (0.96–0.97)	0.76 (0.69–0.82)
Junior2	0.70 (0.65–0.75)	<0.01	0.94 (0.93–0.95)	0.60 (0.53–0.66)	0.57 (0.52–0.63)	0.94 (0.93–0.95)	0.58 (0.52–0.64)
Junior2+AI	0.85 (0.81–0.89)	0.64	0.97 (0.96–0.98)	0.80 (0.74–0.86)	0.78 (0.71–0.84)	0.97 (0.96–0.98)	0.79 (0.73–0.84)
Junior3	0.63 (0.57–0.68)	<0.01	0.92 (0.91–0.93)	0.55 (0.49–0.62)	0.52 (0.46–0.58)	0.92 (0.91–0.93)	0.53 (0.46–0.59)
Junior3+AI	0.84 (0.80–0.88)	0.42	0.97 (0.96–0.98)	0.78 (0.72–0.83)	0.76 (0.70–0.82)	0.97 (0.96–0.97)	0.77 (0.71–0.82)
Senior1	0.83 (0.80–0.85)	0.25	0.97 (0.95–0.98)	0.74 (0.71–0.78)	0.73 (0.69–0.80)	0.96 (0.95–0.97)	0.73 (0.70–0.78)
Senior1+AI	0.86 (0.82–0.90)	0.91	0.97 (0.96–0.98)	0.83 (0.77–0.88)	0.80 (0.75–0.86)	0.97 (0.96–0.98)	0.81 (0.75–0.86)
Senior2	0.85 (0.83–0.87)	0.71	0.97 (0.95–0.98)	0.8 (0.76–0.85)	0.81 (0.78–0.86)	0.97 (0.96–0.98)	0.79 (0.76–0.84)
Senior2+AI	0.85 (0.81–0.89)	0.73	0.97 (0.96–0.98)	0.82 (0.76–0.87)	0.79 (0.73–0.85)	0.97 (0.96–0.98)	0.80 (0.74–0.85)
Senior3	0.83 (0.79–0.88)	0.29	0.97 (0.95–0.98)	0.77 (0.73–0.82)	0.76 (0.72–0.82)	0.96 (0.96–0.97)	0.76 (0.72–0.81)
Senior3+AI	0.86 (0.82–0.90)	0.99	0.97 (0.96–0.98)	0.83 (0.77–0.87)	0.80 (0.75–0.86)	0.97 (0.96–0.98)	0.81 (0.76–0.86)
MRI1	0.86 (0.82–0.89)	0.16	0.97 (0.96–0.98)	0.84 (0.79–0.89)	0.81 (0.75–0.86)	0.97 (0.96–0.98)	0.82 (0.77–0.87)
MRI2	0.86 (0.82–0.90)	0.99	0.97 (0.96–0.98)	0.84 (0.79–0.90)	0.81 (0.75–0.87)	0.97 (0.96–0.98)	0.83 (0.77–0.88)

AI, artificial intelligence; NPV, negative predictive value.

Comparison of Accuracy between AI and radiologists was performed by McNemar test.

performance of the model on the three external test sets (Fig. S9B).

Analysis of misdiagnosis cases

Fig. S10 shows the cases misdiagnosed by Model^{-DCB} in three external test sets (n = 701). The largest and the most frequent misdiagnosis was between HCC and HM (n = 27, 5.7%, 27/472). They are also generally considered to be very difficult to distinguish, particularly due to the wide variety of origins of HM, which leads to significant variability in CEUS features. The most critical misdiagnosis was malignant FLLs (HCC, HM, ICC) misdiagnosed as benign FLLs (HH, HA) (n = 21, 4.0%, 21/522).

It is sometimes difficult to distinguish necrotic metastatic lesions and HA on CEUS videos, but our results showed the misdiagnosis rate between HM and HA was relatively low (n = 7, 3.2%, 7/221). This was because their clinical and laboratory findings (medical history and white blood cell count) were very different, especially when HA progressed to liquefaction. This also revealed the importance of Module-Clinic (Fig. S8C).

Discussion

In this study, we established a massive FLL database, including 3,725 CEUS videos and 805 biomarker results from 52 centers, to build and validate our AI model (Model^{-DCB}) for six-type FLL classifications. Three independent test sets, including two single-center retrospective sets and one multicenter prospective set, were employed to evaluate the diagnostic performance of Model^{-DCB}, which achieved accuracies of 0.85, 0.85, and 0.86, respectively. In all test sets, Model^{-DCB} showed significantly better performance than junior CEUS radiologists (all p <0.05) and equivalent performance to senior CEUS radiologists and senior MRI radiologists (all p >0.05). With its assistance, junior CEUS radiologists significantly improved their classification accuracy (p <0.05), reaching the level of senior radiologists.

Accurate classification of FLLs is of great significance for treatment selection, prognosis prediction, and appropriate disease management.[27,28] In recent years, studies on diagnosing FLL using AI technology have been continually emerging. We have summarized 15 outstanding studies from 2018 to now and listed them in Table S14. Compared with previous studies based on ultrasound/CEUS, our study was the first to use an AI method specifically designed for dynamic video analysis, with the largest sample size and multicenter prospective evaluation. Therefore, our results are more solid and reliable. Only one study had a larger data size than ours,[16] but it was mainly for hepatic echinococcosis diagnosis rather than multi-type FLL classification. The robust performance of Model^{-DCB} across 84 subgroups, including sex, age, tumor size, etiology, cirrhosis, fatty liver and ultrasound device manufacturer, has not been seen in other studies, suggesting that our model is likely to be more generalizable.

Previous studies on the use of DL to diagnose FLLs had given us a lot of inspiration. Hamm CA *et al.*'s study using multiphasic MRI had a good reference value for CEUS in terms of processing temporal information.[29,30] Compared with previous studies based on contrast-enhanced CT (CECT) or MRI, our study still had advantages in terms of data size, classification types, and prospective validation. Ying *et al.*'s study using CECT was the only one with a larger data size than ours.[31] Their AI model (LiAIDS) was excellent and can classify HCC, HM, ICC, HH, HA, which was the same with our Model^{-DCB}. However, unlike our model, instead of classifying extremely challenging rare types of FLL, such as hepatic adenomas, hepatic lymphomas, *etc*., LiAIDS was trained to classify hepatic cysts, which can be easily diagnosed by most doctors in clinical practice, without using AI. In addition, compared with CECT and MRI, CEUS has inherent advantages (real-time imaging, no radiation, low cost, and shorter examination times).

Unlike previous studies that relied on a 'from image to disease' strategy (Module-Disease), we also made full use of the biomarker information of FLL and added the 'from image to biomarker to disease' strategy (Module-Biomarker). The relevant results proved that our hypothesis was correct. Whether from Model^{-D} to Model^{-DB} (Accuracy from 0.80 to 0.86) or from Model^{-DC} to Model^{-DCB} (Accuracy from 0.84 to 0.90), adding Module-Biomarker improved the multi-classification performance for FLLs (Fig.4B). Furthermore, the way in which Model^{-DCB} intelligently applied biomarker information was consistent with previous biomarker studies. For example, Hep was a well-established biomarker for HCC but is rarely found in other FLL types [17, 18]. Therefore, the Hep sub-model synergized with the HCC sub-model to enhance the identification of HCC (Fig. S8A) and played a decisive role when the predicted probability of the HCC sub-model was close to that of another FLL sub-model (Fig. S8B). The predictive ability of biomarkers also gave Model^{-DCB} the potential to quantitatively evaluate the aggressiveness of malignant FLL and predict prognosis, which may help achieve better treatment decisions before pathological analysis.

In three external test sets, we found that the assistance of Model^{-DCB} can effectively improve the diagnostic performance of CEUS radiologists at all levels, especially for junior CEUS radiologists. With Model^{-DCB} assistance, both junior and senior CEUS radiologists could achieve diagnostic performances comparable to or even better than those of senior MRI radiologists, which were generally considered the silver standard for FLL diagnosis, second only to biopsy.[24,32] Such AI-assisted improvement may have great significance for real-world clinical practice. Globally, the number of ultrasound devices and their application regions far exceed those of MRI, and ultrasound is also a cheaper, more convenient, and faster imaging method. CEUS technology enhances the diagnostic capabilities of ultrasound, and Model-^{DCB} further reduces the experience requirements of CEUS radiologists, allowing even junior or inexperienced CEUS radiologists in remote or underdeveloped regions to provide patients with FLL with diagnostic services comparable to those of senior MRI radiologists. Model-^{DCB} also may greatly simplify the process from diagnosis to treatment for patients with FLL. In the future, some of them may not need to undergo days or weeks of exhausting back-to-back examinations in ultrasound, MRI, and pathology departments. For patients with small malignant FLLs, ultrasound screening, diagnosis, and interventional ablation may be completed within 1 day solely in the ultrasound department, saving the precious time of doctors and reducing the medical burden for patients. As the incidence of FLL continues to increase worldwide,[33] promoting AI models that can assist CEUS radiologists in the accurate diagnosis and management of FLL would be more meaningful.[31,34]

Although ultrasound is the most widely used tool for liver imaging examinations, it still has great imaging variability due to numerous manufacturers and different system parameter settings. To overcome such variability, we used the two-stream model strategy. Because the optical flow was generated by calculating the difference between two adjacent CEUS frames, this method paid more attention to the changing trend of the video rather than the video itself, thus circumventing the imaging variability in CEUS. 226 US devices from nine major manufacturers participated in our study, yet Model^{-DCB} still achieved high stability in all test sets (Fig. 5C). Fig. S5 shows two examples. CEUS images acquired from Mindray and Siemens devices showed obvious differences in color saturation, brightness, and contrast, but they were well normalized by converting to optical flow maps.

Another unique advantage of Model^{-DCB} was that it achieved high-accuracy classification of OT without training an OT sub-model. OT included a variety of FLL types with extremely low incidence and lacked unified imaging characteristics, so it was almost impossible to train an efficacious OT sub-model (Fig. S11). However, a distributed training strategy can improve the diagnostic performance for OT by improving the diagnostic performance of other sub-models. The diagnostic capability for OT is indispensable for the translation of AI models from experiment to clinical, because it can prevent AI from making misleading diagnoses uncontrollably when faced with unknown diseases.

Last but not least, Model^{-DCB} only required very limited computing power (two RTX 4090 Ti graphics cards) for training, which we deliberately set up so that any ultrasound manufacturer can easily make their device independently support the application of the model, without greatly affecting the size, portability and cost of the device. Moreover, Model^{-DCB} required little manual work for ROI definition. These two characteristics are especially critical for ultrasound examinations, because ultrasound radiologists need to acquire images and make diagnoses at the same time, and ultrasound devices are frequently moved between beds. A huge computing hardware that cannot be stuffed into the ultrasound device and heavy manual annotation burden will make it difficult for AI models to achieve clinical translation in real ultrasound scenarios.

Our study has some limitations. First, CEUSs in this study were all collected in Chinese centers, and international verification will be needed in the future. Second, the 1-year prospective test set might not capture longitudinal variations or future practices, and we will continue to collect data and validate the model in the future. Third, the distribution of FLL types in this study did not exactly match the natural distribution, but was representative of the distribution of patients undergoing CEUS. Fourth, the diagnosis of some benign FLLs was not based on pathological diagnosis but on clinical findings, laboratory results, and imaging examination.

In conclusion, by effectively integrating CEUS videos, biomarker information, and clinical information, Model^{-DCB} achieved accurate multi-type FLL classification in multiple clinical tests, including independent retrospective tests and a prospective multicenter test. Its assistance improved the diagnostic performance of all participating CEUS radiologists, especially junior radiologists. We believe that Model^{-DCB}'s multi-classification capability, cross-manufacturer stability, low computing power requirement, and inexpensive and easy-to-use features give it great potential for large-scale clinical applications that will benefit a wide range of populations, especially patients with FLL in remote, suburban or underdeveloped areas who have difficulty accessing MRI.

Affiliations

[1]*Department of Interventional Ultrasound, Chinese PLA General Hospital, Beijing 100853, China;* [2]*CAS Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing, China;* [3]*School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China;* [4]*Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, School of Medicine and Engineering, Beihang University, Beijing 100191, China*

Abbreviations

AI, artificial intelligence; CECT, contrast-enhanced CT; CEUS, contrast-enhanced ultrasound; CK, cytokeratin; DL, deep learning; FLLs, focal liver lesions; GPC3, glypican-3; HA, hepatic abscess; HCC, hepatocellular carcinoma; Hep, Hepatocyte antigen; HH, hepatic hemangioma; HM, hepatic metastatic carcinoma; ICC, intrahepatic cholangiocarcinoma; MLP, multilayer perceptron; OT, other types; TIC, time-intensity curve.

Financial support

This work is supported by the National Key Research and Development Program of China (No. 2023YFF1204600 and 2022YFC2407405), the National Natural Science Foundation of China (No. 82441010, 82441011, 92159305, 82030047, 81227901, 81930053, 82272029, 82227804, 82402280 and 92259303), the Beijing Science Fund for Distinguished Young Scholars (No. JQ22013), the R&D project of Pazhou Lab (Huangpu) (2023K0607). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflicts of interest

The authors of this study declare that they do not have any conflict of interest.

Please refer to the accompanying ICMJE disclosure forms for further details.

Authors' contributions

DWZ, MYQ, PC, JT, YJ, LP and WK conceived and designed the study. MJ, PC, WJP, YJ, and LP acquired the data. DWZ and MYQ did the statistical analyses. MYQ, TJ and WK developed, trained, and applied the artificial neural network. YJ and WK implemented quality control of data and the algorithms. YJ and LP verified the underlying raw data. All authors had access to the data presented in the manuscript. All authors analyzed and interpreted the data. DWZ, and MYQ prepared the first draft of the manuscript. YJ, LP and WK revised the manuscript. All authors contributed to manuscript preparation. All authors were responsible for the decision to submit the manuscript for publication.

Data availability statement

Authors will share deidentified individual participant imaging data on request with researchers who provide a methodologically viable proposal and do analyses that achieve the aims of the proposal. Data sharing requests can be directed to DWZ by email. To gain access data requestors will need to sign a data access agreement.

Acknowledgements

We thank the organizations and individuals who participated in this study: Baoji Central Hospital, DGL; Beijing Aerospace Center Hospital, GJ; Chengdu Fifth People's Hospital, LM; China-Japan Union Hospital of Jilin University, WH; First Hospital of Shanxi Medical University, LLP; Fujian Cancer Hospital, TLN; Gansu Provincial Hospital, ZZB; General Hospital of Northern Theater Command, CJY; General Hospital of Western Theater Command, DD; Harbin First Hospital, WJY; Harbin Medical University cancer hospital, CW; Hubei Cancer Hospital, HJG; Hunan Provincial Hospital, WA; Jiangsu Provincial Hospital, YXH; Jilin Central General Hospital, GB; Jingzhou First People's Hospital, XL; Jining First People Hospital, WSL; Maanshan People Hospital, ZHP; Panzhihua Central Hospital, YF; Peking University Cancer Hospital, YK; Peking University Shenzhen Hospital, SDS; Second Hospital of Shanxi Medical University, WXH; Shandong Cancer

Hospital, XGQ; Shijiazhuang Fifth Hospital, DRQ; Sichuan Cancer Hospital, LM; Sun Yat-sen University Cancer Center, ZJH; Tangdu Hospital, YYL; The 2nd Affiliated Hospital of Harbin Medical University, ZXL; The Affiliated Hospital of Qingdao University, ZC; The First Affiliated Hospital of Dalian Medical University, WH; The First Affiliated Hospital of Guangxi Medical University, YH; The First Affiliated Hospital of Zhengzhou University, QCC; The First Affiliated Hospital, Zhejiang University School of Medicine, JTA; The first hospital of Jilin University, ZDZ; The Fourth Hospital of Hebei Medical University, JXH; The People Hospital of Qiannan, LW; The Second Affiliated Hospital of Kunming Medical University, BR; The Second Affiliated Hospital, Sun Yat-sen University, LBM; The Seventh Affiliated Hospital, Sun Yat-sen University, XZF; The Sixth Affiliated Hospital, Sun Yat-sen University, LGJ; The Third Affiliated Hospital, Sun Yat-sen University, ZRQ; The Third People's Hospital of Shenzhen, QYY; Tianjing Third Central Hospital, JX; Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, XMX; Wenzhou Central Hospital, CLM; Xiangya Hospital Central South University, LJT; Xianyang Central Hospital, ZWA; Xijing Hospital, HGB; Yantai Qishan Hospital, WS; Yunnan Cancer Hospital, LXM.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhep.2025.01.011.

References

Author names in bold designate shared co-first authorship

[1] Rebecca LS, Angela NG, Ahmedin J. Cancer statistics, 2024. CA Cancer J Clin 2024;74.

[2] Rumgay H, Arnold M, Ferlay J, et al. Global burden of primary liver cancer in 2020 and predictions to 2040. J Hepatol 2022;77:1598–1606.

[3] **Asrani SK**, **Devarbhavi H**, Eaton J, et al. Burden of liver diseases in the world. J Hepatol 2019;70:151–171.

[4] Reig M, Forner A, Rimola J, et al. BCLC strategy for prognosis prediction and treatment recommendation: the 2022 update. J Hepatol 2022;76 :681–693.

[5] **Vitale A**, **Trevisani F**, Farinati F, et al. Treatment of hepatocellular carcinoma in the precision medicine era: from treatment stage migration to therapeutic hierarchy. Hepatology 2020;72:2206–2218.

[6] Kloeckner R, Galle PR, Bruix J. Local and regional therapies for hepatocellular carcinoma. Hepatology 2021;73(Suppl 1):137–149.

[7] Gunasekaran G, Bekki Y, Lourdusamy V, et al. Surgical treatments of hepatobiliary cancers. Hepatology 2021;73(Suppl 1):128–136.

[8] Lencioni R, Piscaglia F, Bolondi L. Contrast-enhanced ultrasound in the diagnosis of hepatocellular carcinoma. J Hepatol 2008;48:848–857.

[9] Kono Y, Lyshchik A, Cosgrove D, et al. Contrast enhanced ultrasound (CEUS) liver imaging reporting and data system (LI-RADS(R)): the official version by the American college of radiology (ACR). Ultraschall Med 2017;38:85–86.

[10] **Dietrich CF**, **Nolsøe CP**, Barr RG, et al. Guidelines and good clinical practice recommendations for contrast enhanced ultrasound (CEUS) in the liver – update 2020 – WFUMB in cooperation with EFSUMB, AFSUMB, AIUM, and FLAUS. Ultraschall der Medizin - Eur J Ultrasound 2020;41:562–585.

[11] Chernyak V, Fowler KJ, Do RKG, et al. LI-RADS: looking back, looking forward. Radiology 2023;307.

[12] **Zheng W**, **Li Q**, Zou XB, et al. Evaluation of contrast-enhanced US LI-RADS version 2017: application on 2020 liver nodules in patients with hepatitis B infection. Radiology 2020;294:299–307.

[13] Pranav R, Matthew PL. The current and future state of AI interpretation of medical images. N Engl J Med 2023;388.

[14] Pranav R, Emma C, Oishi B, et al. AI in health and medicine. Nat Med 2022;28.

[15] Hu HT, Wang W, Chen LD, et al. Artificial intelligence assists identifying malignant versus benign liver lesions using contrast-enhanced ultrasound. J Gastroenterol Hepatol 2021;36(10):2875–2883.

[16] **Yang Y**, **Cairang Y**, Jiang Ta, et al. Ultrasound identification of hepatic echinococcosis using a deep convolutional neural network model in China: a retrospective, large-scale, multicentre, diagnostic accuracy study. The Lancet Digital Health 2023;5:e503–e514.

[17] Ihab Shafek A. Efficacy of expressions of Arg-1, Hep Par-1, and CK19 in the diagnosis of the primary hepatocellular carcinoma subtypes and exclusion of the metastases. Histol Histopathol 2021;36.

[18] Kenta M, Takamichi I, Haruhiko T, et al. Integrated analyses of the genetic and clinicopathological features of cholangiolocarcinoma: cholangiolocarcinoma may be characterized by mismatch-repair deficiency. J Pathol 2024;263.

[19] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. Adv Neural Inf Process Syst 2014;1.

[20] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. IEEE; 2016.

[21] Baker S, Scharstein D, Lewis JP, et al. A database and evaluation methodology for optical flow. Int J Computer Vis 2010;92:1–31.

[22] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9:1735–1780.

[23] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–444.

[24] van der Pol CB, McInnes MDF, Salameh J-P, et al. CT/MRI and CEUS LI-RADS major features association with hepatocellular carcinoma: individual patient data meta-analysis. Radiology 2022;302:326–335.

[25] Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: IEEE international conference on computer vision; 2017.

[26] M E, A D. Application of the McNemar test to non-independent matched pair data. Stat Med 1991;10.

[27] Choi GH, Yun J, Choi J, et al. Development of machine learning-based clinical decision support system for hepatocellular carcinoma. Sci Rep 2020;10:14855.

[28] Zhao J, Sun Z, Yu Y, et al. Radiomic and clinical data integration using machine learning predict the efficacy of anti-PD-1 antibodies-based combinational treatment in advanced breast cancer: a multicentered study. J Immunother Cancer 2023;11.

[29] Hamm CA, Wang CJ, Savic LJ, et al. Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multiphasic MRI. Eur Radiol 2019;29:3338–3347.

[30] Wang CJ, Hamm CA, Savic LJ, et al. Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. Eur Radiol 2019;29:3348–3357.

[31] Ying H, Liu X, Zhang M, et al. A multicenter clinical AI system study for detection and diagnosis of focal liver lesions. Nat Commun 2024;15.

[32] Marrero JA, Kulik LM, Sirlin CB, et al. Diagnosis, staging, and management of hepatocellular carcinoma: 2018 practice guidance by the American association for the study of liver diseases. Hepatology 2018;68:723–750.

[33] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin 2020;70:7–30.

[34] Su-E C, Lin-Qi Z, Si-Chi K, et al. Multiphase convolutional dense network for the classification of focal liver lesions on dynamic contrast-enhanced computed tomography. World J Gastroenterol 2020;26.

Keywords: Focal liver lesion multi-classification; Deep learning; Contrast-enhanced ultrasound; immunohistochemical biomarker; multi-center retrospective and prospective study.

Received 19 September 2024; received in revised form 26 December 2024; accepted 8 January 2025; available online 21 January 2025