Appendix

Magnetic resonance imaging acquisition

Patients from Center 1 were examined using a 1.5 Tesla magnetic resonance imaging (MRI) scanner (MAGNETOM Aera, Siemens Healthineers) (n = 337). Patients from Center 2 were examined using a 3.0 Tesla MRI scanner (Verio, Siemens Healthineers; Signa HDxt, GE HealthCare; Ingenia, Philips Healthcare; MAGNETOM Skyra, Siemens Healthineers) (n = 38; n = 31; n = 4; n = 3). Patients from Center 3 were examined using 3.0 Tesla MRI scanners (Ingenia, Philips Healthcare and MAGNETOM Skyra, Siemens Healthineers) (n = 75 and n = 22).

The patients were required to fast for 4-6 h before the examination, and underwent breathing and breath-holding training to achieve enhanced image quality. For Center 1 and 2, the dynamic contrast-enhanced MRI scanning was performed with a T1- weighted fat-suppressed gradient-echo sequence before and after the injection of the contrast agent. For Center 3, Ingenia 3.0T scanning was performed with a T1 high-resolution isotropic volume excitation sequence and MAGNETOM Skyra was performed with a volumetric interpolated body examination sequence. Besides dynamic contrast-enhanced MRI, routine sequences were performed, including T2-weighted and diffusion-weighted imaging.

The scanning area covered the entire liver, with in-plane spatial resolution ranging from 0.32 × 0.32 mm2 to 1.39 × 1.39 mm2, slice thickness of 3-7 mm, slice spacing of 0-3 mm, number of slices ranging from 28 to 144, field of view ranging from 297 × 380 mm2 to 400 × 420 mm2, repetition time 2.86-6.89, echo time 1.23-2.39, and flip angle 7.8°-18°.

Feature categorization by radiologists

The annotation of nonrim arterial phase hyper-enhancement (APHE) was based on the images of the pre-contrast phase and arterial phase. The annotation of washout was based on portal venous phase and arterial phase. The annotation of enhancing capsule was based on portal venous phase and transitional phase. Capsule was positive if it existed in either portal venous phase or transitional phase.

Nonrim APHE classifier

First, the para-lesion liver parenchyma was defined through operations including lesion region dilation (using disk kernel with size as one quarter of lesion radius) and exclusive or to obtain the para-lesion region, as well as intersection with the liver region to obtain para-lesion liver parenchyma.

Considering that the APHE definition in Liver Imaging Reporting and Data System (LI-RADS) guidelines contains two conditions, including enhancement in arterial phase greater in whole or in part than in liver, the signal intensity of the entire lesion region could not effectively recognize the enhancement for part enhancement. Therefore, the lesion region was segmented into enhanced and non-enhanced regions by Otsu's method[1], which was used to perform automatic image thresholding. This threshold was determined by maximizing inter-class intensity variance[1,2]. The pixels with signal intensity larger than the threshold were regarded as enhanced lesion regions. Compared with the mean signal intensity, the 75th percentile of the signal intensity within the enhanced lesion region could effectively represent enhancement, which was confirmed by the experiment. Therefore, the 75th percentile of signal intensity within the enhanced lesion region was determined for the APHE score calculation.

Considering the intensity variation of liver parenchyma between various phases, if the 75th percentile of the signal intensity within the enhanced lesion

region in pre-contrast phase (Pre) was larger than that of para-lesion liver parenchyma in Pre, the APHE score for each lesion slice was defined as expressed in Eq. (1). Otherwise, the APHE score was defined as $Score\ AP$ based only on the arterial phase.

$$APHE\ score = \frac{Score\ AP}{Score\ Pre} \quad (1)$$

where $Score\ AP$ and $Score\ Pre$ are signal intensity ratios of lesion to liver parenchyma in arterial phase (AP) and Pre, respectively, which were calculated as expressed in Eqs. (2) and (3).

$$Score\ AP = \frac{I_{tAP}}{I_{lAP}} \quad (2)$$

$$Score\ Pre = \frac{I_{tPre}}{I_{lPre}} \quad (3)$$

where I_{tAP}, I_{lAP}, I_{tPre}, and I_{lPre} are the 75th percentiles of the signal intensity within the enhanced lesion region in AP, mean signal intensity of para-lesion liver parenchyma in AP, 75th percentile of the signal intensity within enhanced lesion region in Pre, and mean signal intensity of para-lesion liver parenchyma in Pre, respectively.

After determining the APHE score for all lesion slices, the APHE final score for the patient was calculated as the maximum of all the APHE scores.

Nonperipheral washout classifier

Portal venous phase (PVP) and AP were considered together to judge whether washout existed. One pre-condition of washout was that the signal intensity ratio of the lesion to liver parenchyma in PVP was smaller than that in AP, as expressed in Eq. (4). When this condition was met, the washout score for each lesion slice was determined using Eq. (5), where I_{tPVP} and I_{lPVP} are the mean signal intensities of the whole lesion or the dark region within the lesion (if extremely strong enhancement in AP) in PVP and para-lesion liver parenchyma in PVP. I_{tAP} and I_{lAP} are mean signal intensities of the whole lesion in AP and

para-lesion liver parenchyma in AP, respectively. Unlike APHE score, washout score is the signal intensity ratio of liver parenchyma to the lesion in PVP, not the signal intensity ratio of the lesion to liver parenchyma in PVP, expressed as:

$$\frac{I_{tPVP}}{I_{lPVP}} < \frac{I_{tAP}}{I_{lAP}} \quad (4)$$

$$Washout\ score = \frac{I_{lPVP}}{I_{tPVP}} \quad (5)$$

After calculating the washout score for all lesion slices, the washout final score for the patient was calculated as the maximum of all the washout scores.

Enhancing capsule classifier

The workflow of capsule detection is as follow. First, the lesion and its surrounding region in the raw image slice were obtained through operations including lesion region dilation (using disk kernel with size as the lesion radius), intersection with raw image slice, and intersection with liver mask. Then, the Frangi filter[3,4], which leveraged the eigenvalues of the Hessian matrix to enhance tubular structures, was used to enhance the image. After Frangi enhancement, adaptive thresholding was used to obtain regions with high intensity value. After adaptive thresholding, fragmented small regions with sizes smaller than one tenth of the radius were removed. The candidate capsule regions were obtained after removing enhanced regions far from the lesion boundary using a lesion mask to limit the capsule position range surrounding the lesion contour.

Next, capsule post-processing was conducted to remove small points and confirm capsule regions by comparing the signal intensity of the candidate capsule region with that of the regions on both sides of it. Each candidate capsule region in the raw image slice was classified as a true capsule if its signal intensity exceeded that of the regions on either side of it. The 75th percentile of the signal intensity for the candidate capsule region, along with the mean signal intensities of the surrounding regions on the upper and lower sides were calculated to

compare the signal intensity. Based on this, the signal intensity ratio of the candidate capsule region to that of either side surrounding the region was calculated, as expressed in Eqs. (6)-(9) for both PVP and transitional phase (TP).

$$R_{1PVPi} = \frac{I_{tcrPVPi}}{I_{sr1PVPi}} \tag{6}$$

$$R_{2PVPi} = \frac{I_{tcrPVPi}}{I_{sr2PVPi}} \tag{7}$$

$$R_{1TPi} = \frac{I_{tcrTPi}}{I_{sr1TPi}} \tag{8}$$

$$R_{2TPi} = \frac{I_{tcrTPi}}{I_{sr2TPi}} \tag{9}$$

where $I_{tcrPVPi}$ and I_{tcrTPi} were the 75th percentile of signal intensities within the candidate capsule region i in PVP and TP, respectively. $I_{sr1PVPi}$, $I_{sr2PVPi}$, I_{sr1TPi}, and I_{sr2TPi} were the mean signal intensities of two regions on either side of the capsule region i in PVP and TP, respectively. For the candidate capsule region, the 75th percentile of signal intensity rather than the mean signal intensity could effectively represent enhancing capsule, which was confirmed by the experiment.

Only when both signal intensity ratios were larger than the capsule enhancement threshold, that is, conditions (10) and (11) were met at the same time, the candidate capsule region could be regarded as the true capsule region.

$$R_{1PVPi} > capsule\_enhancement\_threshold \tag{10}$$

$$R_{2PVPi} > capsule\_enhancement\_threshold \tag{11}$$

After confirming all candidate capsule regions, all the true capsule regions were skeletonized and the skeleton length was calculated for each true capsule region. Finally, the capsule score for each lesion slice could be calculated as expressed in Eqs. (12)-(14):

$$Capsule\ score = \max\left(Capsule\ score\ PVP, Capsule\ score\ TP\right) \tag{12}$$

$$Capsule\ score\ PVP = \frac{\sum_1^{nPVP} l_{iPVP}}{L} \tag{13}$$

$$Capsule\ score\ TP = \frac{\sum_1^{nTP} l_{iTP}}{L} \tag{14}$$

where $nPVP$ and nTP are the numbers of true capsule regions in PVP and TP, respectively; l_{iPVP} and l_{iTP} are the skeleton lengths of the true capsule region i in PVP and TP, respectively; and L is the lesion perimeter.

After calculating the capsule score for all lesion slices, the capsule final score for the patient was calculated as the maximum of all the capsule scores.

Three comparison methods for Liver Imaging Reporting and Data System categorization

The effectiveness of the proposed method in terms of accuracy and generalizability was verified by comparing it with three other LI-ARDS categorization methods, including a random forest classifier based on radiomics features (Radiomics)[5], AlexNet[6], and VGG16[7]. These models trained end-to-end models to determine LI-RADS categories, without using the three major features of LI-RADS.

(1) Radiomics

After registration, segmentation, and lesion mask check and modification, if necessary, the "pyradiomics" toolbox was used to extract radiomics data from dynamic contrast-enhanced MRI four-phase images. For each phase, 107 features were extracted, including 14 shape features (volume, surface area, maximum/minimum diameters, sphericity, compactness, elongation, and so forth), 18 first-order statistics features (mean, median, minimum, maximum, standard deviation, percentiles, skewness, kurtosis, entropy, and so forth), and 75 texture features (including gray level co-occurrence matrix, gray level run length matrix, gray level size zone matrix, neighborhood gray tone difference matrix, and gray level dependence matrix). In total, 428 features were used to train the random forest classifier to determine LI-RADS categories, including LI-RADS grade 3 (LR-3), 4 (LR-4), and 5 (LR-5). During the training process, grid

search was implemented to search the most suitable parameters (n_estimators, max_depth, min_samples_leaf, and max_features) for the random forest classifier by maximizing the validation score of the classifier. The experiments included five repetitions of five-fold cross-validation. For each five-fold cross-validation, the output of the model was ensembled by average.

(2) AlexNet

1) Pre-processing

The AlexNet model was built and trained referring to the study by Wu et al.[6]. Wu et al. selected center image slices of the pre-contrast phase, arterial phase and washout phase (late PVP or TP) to adapt to the 2D AlexNet with each image slice in various channels, which was originally designed for red, green, and blue channels of the nature images. We also applied the 2D AlexNet pre-trained by ImageNet, but we considered all lesion slices in the following manner. After registration, segmentation, and lesion mask check and modification, if necessary, the lesion slices were pre-processed as performed for the proposed method. The region of interest images of each lesion slice were cropped based on the mask bounding box (with a fixed margin 20 pixels), and the volume of interest with lesion was resampled to 16 × 56 × 56, with 16 slices and each slice with a pixel size 56 × 56. Then, the 16 layers were arranged by 4 × 4 to form a larger image with pixel size 224 × 224, which was the input size of AlexNet. Instead of using the pre-contrast phase, arterial phase and washout phase (late PVP or TP) in the study of Wu et al., we chose AP, PVP, and TP and placed these three-phase images in the three separate channels. The model input was of size 224 × 224 × 3.

Image augmentation was implemented referring to the study by Wu et al., rotating by degrees of {−60, −30, 30, 60}, and flipping both horizontally and vertically. As a result, seven images for each lesion were used as inputs to AlexNet. After image augmentation, the signal intensity was scaled to the range

between 0 and 1.

2) Model training

For AlexNet, the model was trained using the Adam optimizer, with a batch size of 128, the initial learning rate set to 1e-4, and it was reduced on the plateau with a factor of 0.2, the patience of 5, and the minimum learning rate of 1e-6. The epoch was 250, and early stopping was set with patience set as 20. The model was implemented in Tensorflow (version 2.7.0, Google). Five repetitions of five-fold cross-validation were performed to train the model. We selected the most appropriate model in each fold and ensembled their outputs for final predictions. The model was trained on a graphics processing unit with 16 GB memory (Tesla P100, NVIDIA).

3) Loss and evaluation metric

The loss was categorical cross-entropy loss, and the metric was categorical accuracy. The model was trained by maximizing the validation accuracy through transfer learning from the weights trained by ImageNet.

(3) VGG16

1) Pre-processing

Yamashita et al.[7] demonstrated that VGG16 was pre-trained also by ImageNet and retrained by transfer learning with triple-phase images as the input (pre-contrast, late arterial and delayed phase). Unlike the AlextNet method, the VGG16 method also considered lesion diameter as the input. Besides the pre-processing steps and image augmentation for AlexNet, for VGG16 method, the lesion diameters were normalized with mean as 0 and the standard deviation as 1. We also chose AP, PVP, and TP for VGG16 and placed these three-phase images in three separate channels, exactly similar to that for AlexNet.

2) Model training

For VGG16, the model was trained using the Adam optimizer, with a batch size

of 128, the initial learning rate set to 1e-4, and it was reduced on the plateau with a factor of 0.2, patience of 5, and minimum learning rate of 1e-6. The epoch was 250, and early stopping was set with patience as 15. The model was implemented in Tensorflow (version 2.7.0, Google). The experiments included five repetitions of five-fold cross-validation. For each five-fold cross-validation, the output of the model was ensembled by average. The model was trained on a graphics processing unit with 16 GB memory (Tesla P100, NVIDIA).

3) Loss and evaluation metric

The loss was categorical cross-entropy loss, and the metric was categorical accuracy. The model was trained by maximizing the validation accuracy through transfer learning from the weights trained by ImageNet.

Statistics

Kappa coefficient

The kappa statistic (κ) was divided into several range groups to represent various strengths of agreement[8,9].

$\kappa < 0$: poor agreement; $0.01 \leq \kappa \leq 0.20$: slight agreement; $0.21 \leq \kappa \leq 0.40$: fair agreement; $0.41 \leq \kappa \leq 0.60$: moderate agreement; $0.61 \leq \kappa \leq 0.80$: substantial agreement; and $0.81 \leq \kappa \leq 1.0$: almost perfect agreement.

Mixed-effects analyses

The notably lower cirrhosis rate in Center 2 could, in principle, induce intra-center correlation and potentially affect the generalizability of validation results. To directly address this concern while respecting the ordered nature of LI-RADS (LR) grades (LR-3/4/5), we performed mixed-effects and sensitivity analyses on multiple definitions of prediction error. For binary prediction errors (overcall among true LR $\in$ {3,4}; undercall among true LR $\in$ {4,5}; and any error in the

full cohort), we fitted generalized linear mixed models with a center random intercept, adjusting for true LR grade, and quantified clustering using the center-level variance and intra-center correlation. Error magnitude was defined as |predicted LR−true LR| (0/1/2) and analyzed using an ordinal mixed-effects model with a center random intercept; to accommodate potential grade-dependent variability, a heteroskedastic ordinal model with grade-specific scale parameters was additionally fitted to test center and cirrhosis effects.

Center × etiology–stratified resampling analysis

To evaluate model robustness under balanced subgroup representation across centers and etiologies, we performed a center × etiology–stratified resampling analysis in the validation cohort restricted to LR-3/4/5. Etiology was operationalized as HBV status (HBV+/HBV−). We conducted 2,000 stratified bootstrap iterations, sampling with replacement within each Center × HBV stratum while preserving the original stratum size. Performance was quantified using overall accuracy, mean absolute error of ordinal grading defined as |predicted LR−true LR|, within-one-grade error (|predicted LR – true LR| ≤ 1), and quadratic-weighted kappa. To provide a balanced evaluation in which each subgroup contributes equally, we additionally computed macro-averaged metrics by estimating each metric within each stratum and then taking an unweighted average across strata in each iteration. As a sensitivity analysis for HBV effects while controlling for center and true LR category, we fitted proportional-odds ordinal regression models for (i) absolute error (0/1/2) and (ii) signed error direction (−2/−1/0/+1/+2), treating center as a fixed effect due to near-zero random-effect variance with only three centers.

Supplemental results

Mixed-effects analyses

Across generalized linear mixed models for overcall, undercall, and any error analysis, the center random-intercept variance was 0 (boundary/singular fits), yielding intra-center correlation ≈ 0 and indicating negligible within-center clustering of prediction errors. Consistently, the ordinal mixed-effects model for error magnitude showed near-zero center variance (≈3.94×10−9; intra-center correlation ≈ 0). In the heteroskedastic ordinal model, center was not associated with greater error magnitude (Center 2 vs 1: $P = 0.63$; Center 3 vs 1: $P = 0.65$), and cirrhosis was not associated with error magnitude ($P = 0.63$). Error magnitude was higher for true LR-4 than LR-3 lesions (OR ≈ 3.31, $P < 0.001$), while cirrhosis showed a non-significant trend toward fewer errors (any error: $P = 0.12$; error magnitude: $P = 0.63$). Collectively, these findings suggest that the lower cirrhosis rate in Center 2 did not induce meaningful clustering or systematic center effects that would bias validation performance or compromise generalizability.

Center × etiology–stratified resampling analysis

In 2,000 Center × HBV–stratified bootstrap evaluations preserving subgroup sizes, performance was stable (median [2.5th–97.5th percentiles]): overall accuracy 0.773 [0.721–0.821], mean absolute grading error (MAE) 0.271 [0.211–0.335], within-one grade error 0.956 [0.928–0.980], and quadratic-weighted kappa (QWK) 0.723 [0.631–0.796]. Under macro-averaged (equal-weight) subgroup evaluation, results remained consistent for adjacent-category performance (within-one grade error 0.956 [0.914–0.985]; MAE 0.261 [0.186–0.356]), with a lower QWK 0.643 [0.525–0.781], indicating heterogeneity in strict ordinal agreement when each subgroup is weighted equally. Stratum-specific analyses showed the smallest subgroup (Center 3 / HBV−, $n = 11$) had higher ordinal error

(MAE 0.364) and lower agreement (within-one grade error 0.909; QWK 0.267), accounting for the reduced macro-averaged QWK. In ordinal regression sensitivity analyses adjusting for true LR grade and center, HBV was not associated with greater absolute error (OR 1.32; 95% confidence intervals (CI) 0.59–2.94; P = 0.494), while it showed a marginal shift toward overcalling in signed error direction (OR 2.05; 95%CI 0.98–4.30; P = 0.058). Overall, these analyses support robust ordinal grading performance under subgroup-balanced evaluation, without evidence that HBV materially increases the magnitude of grading errors after accounting for center and the true LI-RADS grade.

References

1 **Otsu N**. A Threshold Selection Method from Gray-Level Histograms. *IEEE Trans. Syst., Man, Cybern.* 1979; **9**: 62–66 [DOI: 10.1109/TSMC.1979.4310076]

2 **Sankur B**. Survey over image thresholding techniques and quantitative performance evaluation. *J Electron Imaging* 2004; **13**: 146 [DOI: 10.1117/1.1631315]

3 **Frangi AF**, Niessen WJ, Vincken KL, Viergever MA. Multiscale vessel enhancement filtering. *Lecture Notes in Computer Science* 1998 [DOI: 10.1007/bfb0056195]

4 **Longo A**, Morscher S, Najafababdi JM, Jüstel D, Zakian C, Ntziachristos V. Assessment of hessian-based Frangi vesselness filter in optoacoustic imaging. *Photoacoustics* 2020; **20**: 100200 [PMID: 32714832 DOI: 10.1016/j.pacs.2020.100200]

5 **Alksas A**, Shehata M, Saleh GA, Shaffie A, Soliman A, Ghazal M, Khalifeh HA, Razek AA, El-Baz A. A Novel Computer-Aided Diagnostic System for Early Assessment of Hepatocellular Carcinoma. *2020 25th International Conference on Pattern Recognition (ICPR)* 2021 [DOI: 10.1109/icpr48806.2021.9413044]

6 **Wu Y**, White GM, Cornelius T, Gowdar I, Ansari MH, Supanich MP, Deng J.

Deep learning LI-RADS grading system based on contrast enhanced multiphase MRI for differentiation between LR-3 and LR-4/LR-5 liver tumors. *Ann Transl Med* 2020; **8**: 701 [PMID: 32617321 DOI: 10.21037/atm.2019.12.151]

7 **Yamashita R**, Mittendorf A, Zhu Z, Fowler KJ, Santillan CS, Sirlin CB, Bashir MR, Do RKG. Deep convolutional neural network applied to the liver imaging reporting and data system (LI-RADS) version 2014 category classification: a pilot study. *Abdom Radiol (NY)* 2020; **45**: 24–35 [PMID: 31696269 DOI: 10.1007/s00261-019-02306-7]

8 **Landis JR**, Koch GG. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 1977; **33**: 159-174 [PMID: 843571 DOI: 10.2307/2529310]

9 **McHugh ML**. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)* 2012; **22**: 276-282 [PMID: 23092060 DOI: 10.11613/BM.2012.031]

Supplemental Material

Tables

Table S1 Magnetic resonance imaging parameters for the three centers

Scanning parameters	Center 1	Center 2		Center 3			
Scanners	Siemens Aera 1.5T (n = 337)	Siemens Verio 3.0T (n = 38)	GE Signa HDxt 3.0T (n = 31)	Philips Ingenia 3.0T (n = 4)	Siemens Skyra 3.0T (n = 3)	Philips Ingenia 3.0T (n = 75)	Siemens Skyra 3.0T (n = 22)
Repetition time (msec)	3.47-6.89	3.90-4.19	2.86-4.60	3.72-3.74	4.11-4.15	3.04-3.69	3.85-3.97
Echo time (msec)	1.31-2.39	1.89-1.93	1.35-1.73	1.32-1.38	1.24-1.96	1.32-1.38	1.23-1.29
Flip angle (°)	10-18	7.8-15	12-15	10	9-12	10	9-10
Matrix	240*320	260*320	256*256	266*268	260*320	266*268	260*320
Slice thickness (mm)	3-4	3-5	5-6	6	3-3.6	6	3-7
Slice spacing (mm)	0	0	0	3	3	3	3
Field of view (mm^2)	297*380	308*380	380*380	400*400	341*420	400*400	342*420

Table S2 Distribution of Liver Imaging Reporting and Data System features and Liver Imaging Reporting and Data System categories across datasets from three centers

LI-RADS feature or LI-RADS category	Internal datasets (*n* = 337)	Training and validation set (*n* = 275)	Internal testing set (*n* = 62)	External testing set 1 (*n* = 85)	External testing set 2 (*n* = 104)
APHE					
Positive	288 (85.5)	234 (85.1)	54 (87.1)	73 (85.9)	95 (91.3)
Negative	49 (14.5)	41 (14.9)	8 (12.9)	12 (14.1)	9 (8.7)
Washout					
Positive	221 (65.6)	181 (65.8)	40 (64.5)	44 (51.8)	80 (76.9)
Negative	116 (34.4)	94 (34.2)	22 (35.5)	41 (48.2)	24 (23.1)
Capsule					
Positive	182 (54.0)	148 (53.8)	34 (54.8)	44 (51.8)	56 (53.8)
Negative	155 (46.0)	127 (46.2)	28 (45.2)	41 (48.2)	48 (46.2)
LI-RADS category					
LR-3	74 (22.0)	60 (21.8)	14 (22.6)	21 (24.7)	15 (14.4)
LR-4	69 (20.5)	61 (22.2)	8 (12.9)	13 (15.3)	18 (17.3)
LR-5	194 (57.6)	154 (56.0)	40 (64.5)	51 (60.0)	71 (68.3)

Note: Data are numbers of lesions, with percentages in parentheses. APHE: Nonrim arterial phase hyper-enhancement; LI-RADS: Liver Imaging Reporting and Data System.

Table S3 Results of inter-observer agreement for feature classification and Liver Imaging Reporting and Data System categorization

	Center 1	Center 2	Center 3
Feature classification			
APHE	0.833	0.874	0.874
Washout	0.868	0.890	0.842
Capsule	0.792	0.874	0.923
LI-RADS categorization	0.865	0.947	0.929

APHE: Nonrim arterial phase hyper-enhancement; LI-RADS: Liver Imaging Reporting and Data System.

Table S4 Classification results of mapped lesion diameter intervals

Datasets	Accuracy	Quadratic weighted Cohen's Kappa coefficient
Training set and validation set from Center 1	80.0% (220/275)	0.773
Internal testing set from Center 1	85.5% (53/62)	0.774
External testing set 1 from Center 2	92.9% (79/85)	0.768
External testing set 2 from Center 3	94.2% (98/104)	0.894

Figures

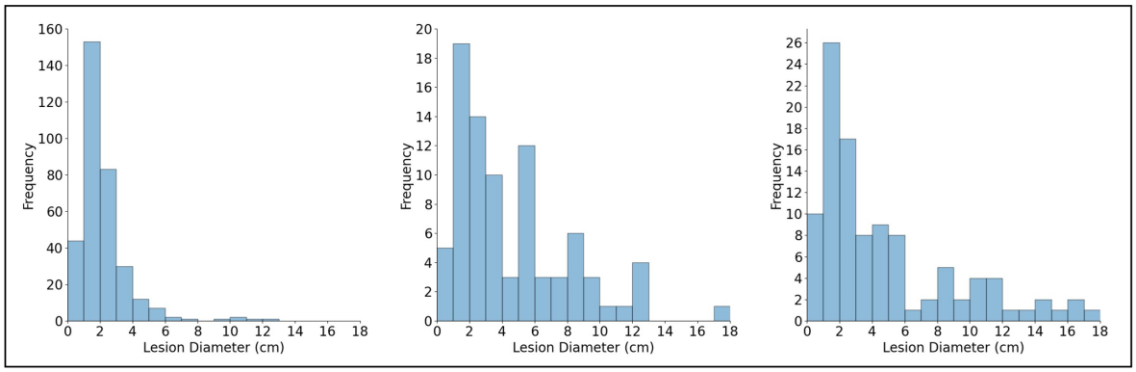

Figure S1 Distribution of lesion diameters for Center 1 (*left*), Center 2 (*middle*), and Center 3 (*right*).

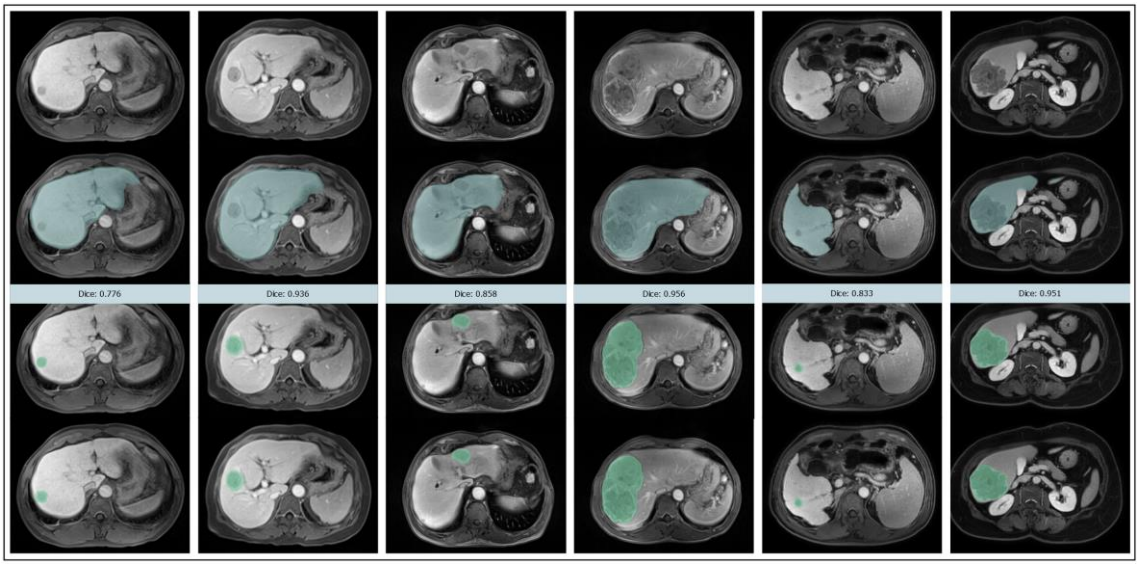

Figure S2 Representative segmentation results along with the manual lesion annotations by the radiologist. Six columns represent six patients. The first row demonstrates axial images from the transitional phase of dynamic contrast-enhanced MRI scans. The second row displays liver segmentation results (olive green indicating the segmented liver region) by 3D U-Net. The third row presents lesion segmentation results (light green indicating the segmented lesion region) by nnU-Net. The fourth row displays manual lesion annotations (light green indicating the annotated lesion region) by the radiologist. The lesion sizes for the patients shown here range from 13 to 128 mm. Gadoxetic acid contrast agent was injected at a dose of 0.025 mmol/kg and a flow rate of 2 mL/s.

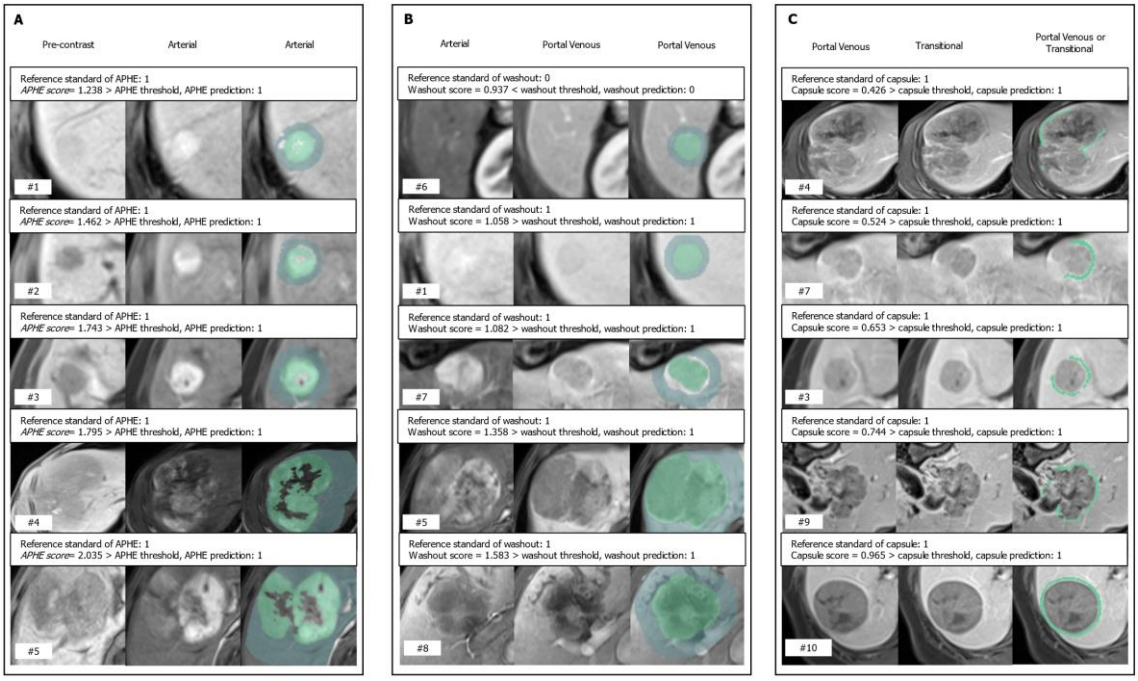

Figure S3 Representative results for feature characterization. A: Nonrim arterial phase hyper-enhancement (APHE) characterization; B: Washout characterization; C: Capsule characterization. Feature scores represent the degree of enhancement, the degree of washout, and capsule relative length, respectively. For all features, higher scores indicate a greater probability of feature presence. Patient identity numbers are displayed at the bottom left of each image. For each feature characterization, the results for five patients are demonstrated. For each group of the five patients, as the degree of enhancement, the degree of washout, or capsule relative length increases, the feature scores output by the model also increase, which demonstrates the effectiveness of the algorithms. The visualization results display the evidence for the judgment. The color definitions for various regions are as follows: A and B, olive green indicates the para-lesion liver parenchyma; A: light green indicates the enhanced lesion region; B: light green indicates the entire lesion region or dark regions within the lesion; C: light green indicates the detected capsule region. All these regions are model outputs. Gadoxetic acid contrast agent was injected at a dose of 0.025 mmol/kg and a flow rate of 2 mL/s. $APHE\ score = SEL_e / SL_i$, $Washout\ score = SL_i/SL_e$, $Capsule\ score = Capsule\ length/\ Lesion\ perimeter$. SEL_e: signal intensity of enhanced lesion region; SL_e: signal intensity of lesion region; SL_i: signal intensity of para-lesion liver parenchyma.

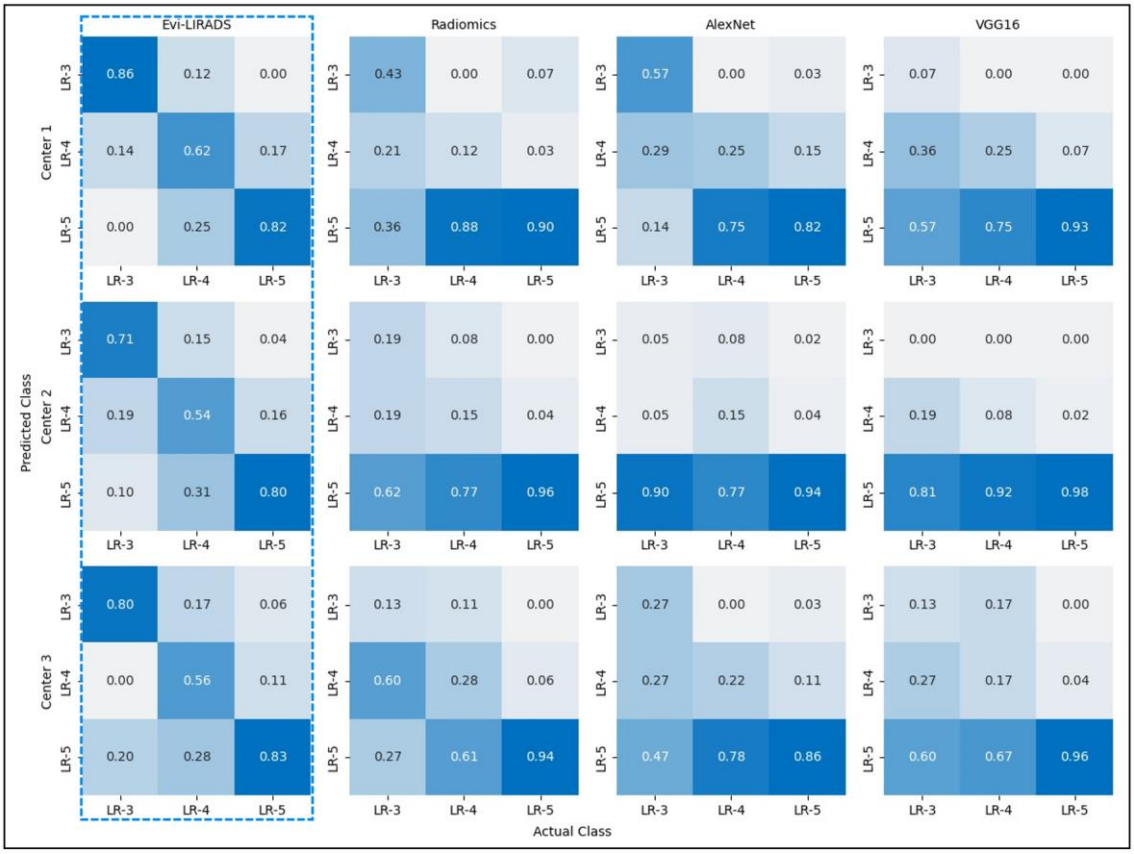

Figure S4 Confusion matrix for categorization of Liver Imaging Reporting and Data System (LI-RADS) grade 3, 4, and 5 of each method, with three rows representing the three centers and four columns for the proposed evidence-based radiologist-supervised automated LI-RADS (Evi-LIRADS) and the three comparison methods.