Multi-modal AI for Opportunistic Screening, Staging and Progression Risk Stratification of Steatotic Liver Disease

Supplementary Material

No Institute Given

Section 1. Data procssing.

Liver-spleen 3D segmentation.

To segment the liver and spleen from CTs, we utilized a framework described in [1]. Briefly,region of interest of the abdominal region was extracted at first, which is a key step towards fine segmentation of liver and spleen, given the diversity of CT scans exhibiting large variations in the field of view (FoV). This framework enables highly efficient detection and segmentation of the organs of interest. We refined and tailored it using a pre-trained model on FLARE dataset, expanded the training dataset by incorporating a large number of unenhanced CTs, and performed z-direction sub-volume cropping as a data augmentation to simulate the large variations in FoV. As a result, we had a highly efficient and accurate segmentation model that achieves a dice score of 98.0% and 97.2% for liver and spleen segmentation on the FLARE2022[2] validation.

Liver image preprocessing

We cropped liver in 3D based on the segmentation and truncated voxels to $[-175, 275]$ HU and normalized to $[0, 1]$ and then resampled to the input size (384, 256, 64) for training and inference. We also applied data augmentation during training, by randomly selecting from a pool of operations, including: 1) randomly shifting the volume within 5 voxels in x, y, z direction; 2) randomly rotating across the slice within $\pm 25^{\circ}$ in x, y direction; 3) randomly flipping the slice horizontally in x, y direction; 4) randomly adding gaussian noise to the image.

Section 2. Training and implementation details.

Training details.

We developed our models using the gold-standard **GS** dataset and ablation studies were conducted via five-fold cross-validation. We held 20% data for validation and 80% data for training at each fold. For training the MAOSS, we conducted two stages of training. In the first stage, we pretrained the backbone on the

modality-complete subset i.e. paired image and non-image data and only the ordinal regression head was optimised during the first training stage with gold standards. Then, we kept all the multi-head self-attention (MSA) blocks frozen in the second stage, only optimizing the plugged Missing-aware Multi-Modality Alignment (MAMA) blocks by training on both the modality complete subset and the modality-incomplete i.e. with missing non-image features of the GS dataset. At the same time, we also incoprated the non-gold standard (NGS) dataset for learning the curated distllation head.

Implementation details.

The image feature encoder was implemented with a 3D ResNet-34 architecture. The number of code words K was set to 8 for texture encoding. The length of prompt vectors in MAMA was set to 16. C-the feature dimensionality was set to 512 for both image and texture embedding. All models were trained using Adam with an initial learning rate of 10^{-4} and batch size of 60, using four NVIDIA Tesla V100 GPUs. For curated distillation training, batch size for the GS and NGS samples denoted as B_{GS} and B_{NGS} were both 30. λ in the loss function was set to 0.5.

Section 3. Numerical features used in model development

Demographics&physical data

Age, Gender, Body mass index (BMI), Height, Weight;

Laboratory results

Laboraty results used in the study includes hematology, liver function tests, lipid profile, cardiac markers etc: White blood cells (WBC), Red blood cells (RBC), Hemoglobin (Hgb), Hematocrit (HCT), Platelet (PLT), Total protein (TP), Albumin (ALB), Albumin/Globulin (A/G), Aspartate aminotransferase (AST), Alanine transaminase (ALT), Gamma-glutamyl transferase (GGT), Alkaline phosphatase (ALP), Prealbumin (PALB), Cholinesterase (CHE), Total bilirubin (TBIL), Direct bilirubin (TBIL), Indirect bilirubin (IDBIL), Total bile acid (TBA), Creatine kinase (CK), CKMB (Creatine Kinase-MB), Lactate dehydrogenase (LDH), Hydroxybutyrate dehydrogenase (HBDH), Blood urea nitrogen (BUN), Uric acid (UA), Cystatin C (CysC), Blood cholesterol (CHOL), Triglycerides (TG), High-density lipoprotein (HDL-C), Low-density lipoprotein (LDL-C), Apolipoprotein A-1 (Apoa1), Apolipoprotein B-100 (apoB), Blood glucose (GLUO), Lipase (LiPA), Creatinine (CREA), Prothrombin time (PT), plasma thromboplastin antecedent (PTA), International normalised ratio (INR), Prothrombin time Test and INR (PT/INR), activated partial thromboplastin time (APTT), Fibrinogen (FIB), Thrombin time (TT), D-dimer (DD), Procollagen type III amino terminal peptide (PIIIP), Hyaluronic acid (HA), Laminin (LM).

CT biomarkers

We conducted a series of conventional CT-based biomarkers in 3D and 2D over both the liver and spleen. We computed the histogram of CT values and derived the mean attenuation of the liver and spleen. We also compute the liver-spleen attenuation ratio (LSR) and the liver-spleen attenuation difference (LSD). For 2D biometry, we automatically assess regional attenuation in the liver[3]. Three liver peripheral ROIs were generated using morphological operations. First, the liver mask was eroded to locate the central slice, and then dilation was performed from the center in three directions - laterally, anteriorposterior- to define a central ROI where major vessel structures are located. Thereafter, three rounded ROIs were placed between the central ROI and the original liver mask.

Section 4. Ablation study

We investigate the effectiveness of the main components on the validation set by excluding one of them from the full setting of MAOSS. Table S2 shows that boosted distillation significantly improves the model performance in each group and overall it brings about 2% and 1.4% increments in mean-BACC and mean-AUC. Random mask of image tokens only gives a moderate increase in mean-AUC but the increase is barely observed in mean-ACC. By removing numerical features, the single modal setting learned with images achieves an mean-BACC and mean-AUC of 85.3% and 88.9% which shows about 0.8% and 1% drop in mean-BACC and mean-AUC respectively, compared with multi-modal learning. Lastly, we found that texture encoding is especially important for improving the detection in the early stage of steatosis i.e. mild-moderate. With texture encoding, it increases the BACC by approx. 3% and 1.8% in identifying mild and moderate steatosis, respectively.

To interpret the learned models, we utilize t-SNE [4]. Figure S3 display the full setting of MAOSS, we observe that the learned ordinal regression and curated distillation tokens converge towards different vectors, indicating they have learned distinct distribution of the data. This aligns with our expectations since they were trained with the GS and NGS datasets, respectively. By averaging the token embeddings i.e., $\frac{1}{2}\mathrm{x}_{ord} + \frac{1}{2}\mathrm{x}_{dist}$, we note the distribution of the joint representation with well-defined and -separated clusters with clearer decision boundaries compared to either of the tokens standing alone.

Section 5. Comparision with other state of the arts

We compare with three groups of methods: 1) imputing-based methods that are trained with biometrics data in a single modal setting where two classic methods are compared: Multiple Imputation by Chained Equation (MICE) [5] and Imputation by K-nearest neighbor [6]. 2) We compare state-of-the-art image-based methods trained in a single modality setting, including the classic CNN model ResNet [7] and the transformer-based model DeiT [8]. Additionally, we evaluate

the texture encoding method, DeepTen [9]. 3) multi-modal learning methods with missing modalities that are trained on image and biometrics together. MultiPrompt [10] is a strong baseline to investigate missing-aware prompt tuning on pre-trained models and we compare it with the input prompt setting in our work. In Table S3, we observe that multimodal-learning methods overall outperform the single modal learning with imputaion by a large margin, which is as expected since rich information in the image representation is leveraged by joint learning with clinical features. Our proposed method surpasses the strong multimodal baseline presented by [10] on both internal and external tests, demonstrating its superior capability in managing missing modalities and leveraging the relationship between image and clinical features. Notably, due to the unavailability of clinical features in the MRI-PDFF dataset, both our method and the baseline achieve similar AUCs across each group. This underscores the critical importance of clinical features in enhancing model performance. Without these features, both our approach and the baseline are reduced to relying solely on image data, thereby limiting their performance.

Section 6. Identifying patients at-risk of advanced liver fibrosis.

A total of 122 patients, accounting for about 10% (122/1192) of the whole screened population, were confirmed with liver biopsy who developed advanced liver fibrosis $\geq$F3. Again, we applied the AASLD and MAOSS pathway for a risk stratification to detect patients at risk of advanced fibrosis. The results were summarized in Table S16. We found the MAOSS pathway again was significantly superior to the AASLD pathway for identifying patient at risk of advanced fibrosis. The sensitivity of the MAOSS pathway was 54.1% (95% CI: 44.6-62.6%), significantly higher than that of the AASLD pathway's 23% (95% CI: 15.6-30.5%) ($p < .001$). MAOSS pathway also demonstrated a superior ($p = .002$) capability for excluding advanced fibrosis patients with a higher NPV of 94.2% (95% CI: 92.8-95.6%), compared with that of the AASLD pathway's 91.5% (95% CI: 89.9-93%).

Section6. Supplement Figures.

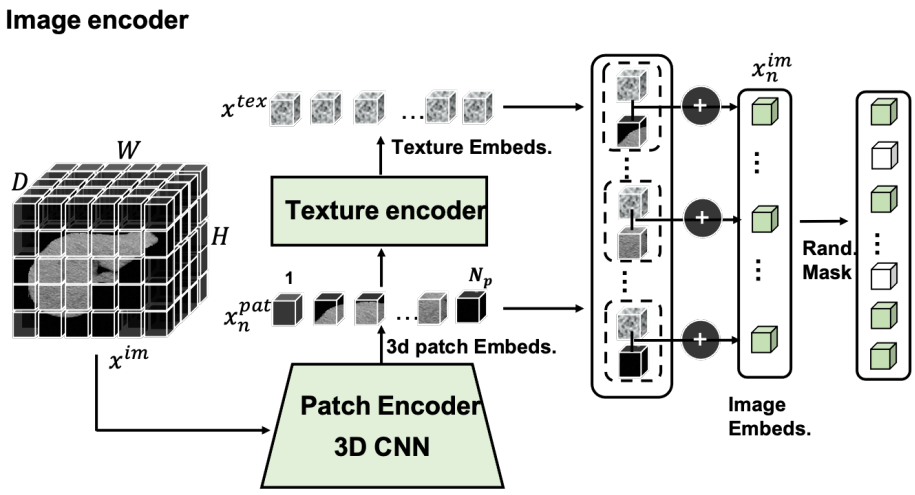

Fig. S1: Image Encoder Design detail.

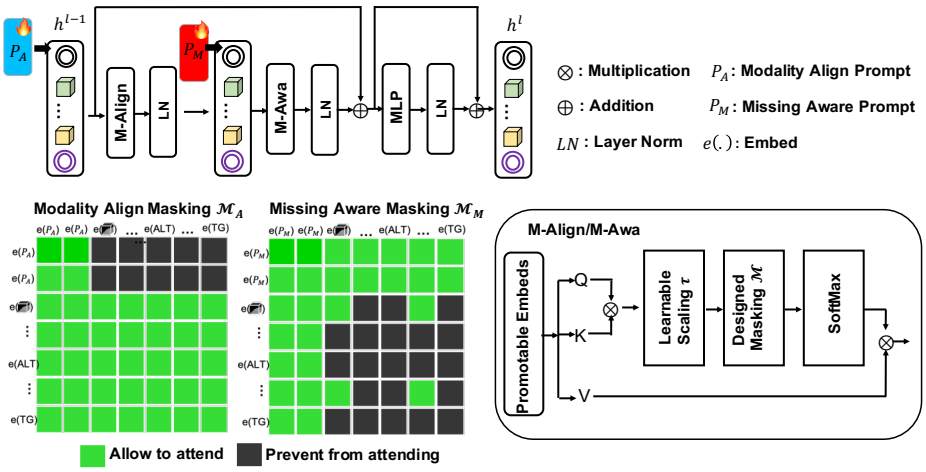

Fig. S2: Missing-Aware Modality Alignment (MAMA) module.

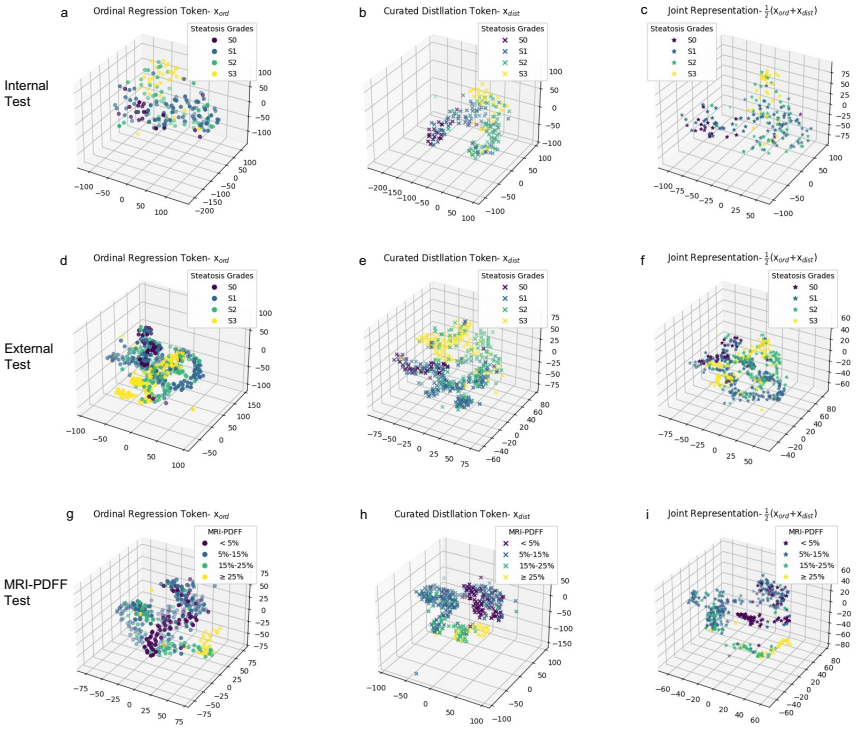

Fig. S3: Visual interpretation of latent space of MAOSS using t-SNE. a-c: on internal gold standard testset. d-f: on external gold standard testset. h-j: on MRI-PDFF testset. x_{ord}: ordinal regression token embedding; x_{dist}: curated distllation token embedding. $\frac{1}{2}(x_{ord} + x_{dist})$: joint representation of MAOSS.

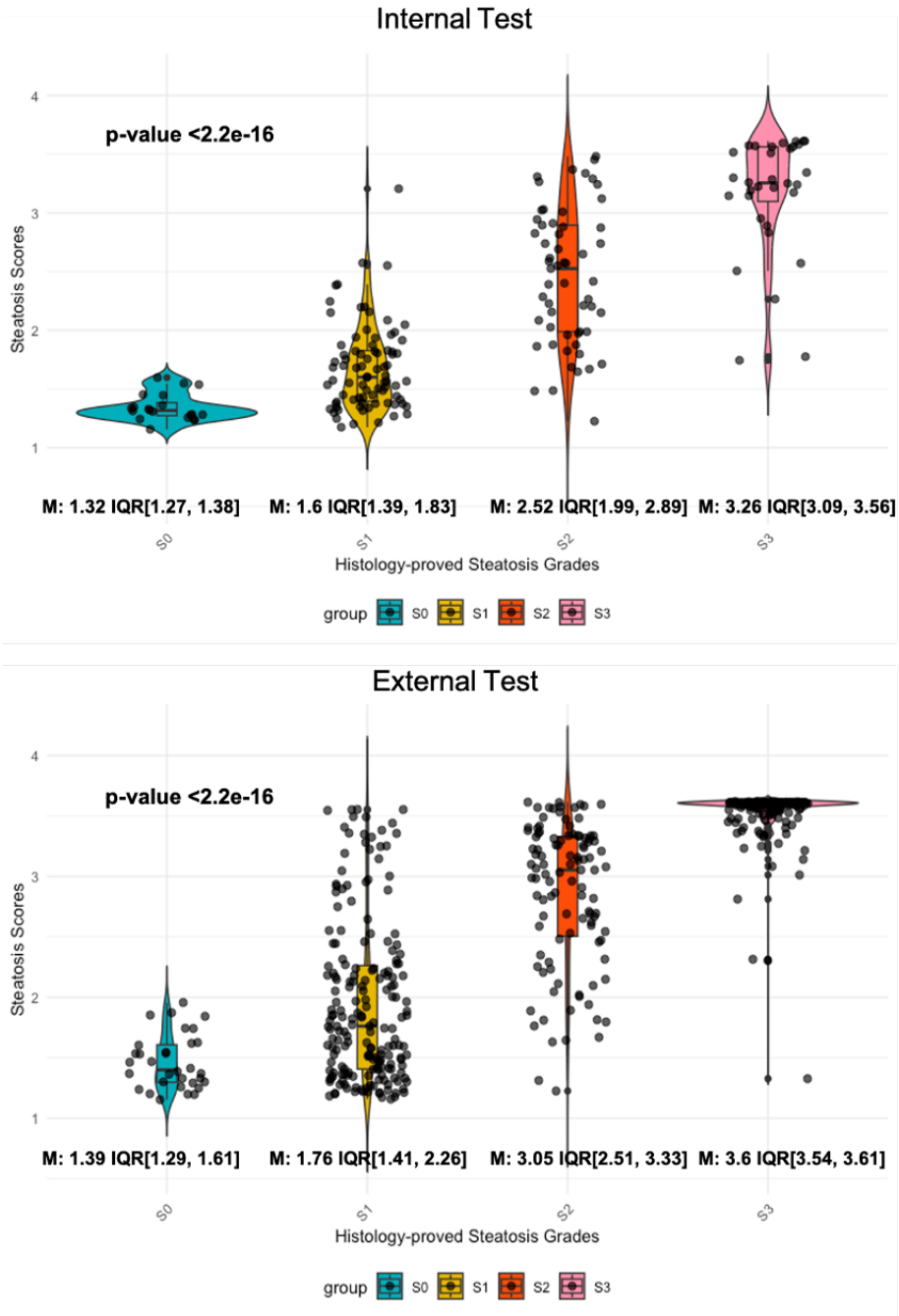

Fig. S4: MAOSS score distribution against the histologically-proved steatosis grades in the internal and external testset. The Kruskal–Wallis test was used to compare the MAOSS scores in different stage of steatosis. A two-sided p value less than 0.05 was considered statistically significant. M: Median, IQR: Interquartile range.

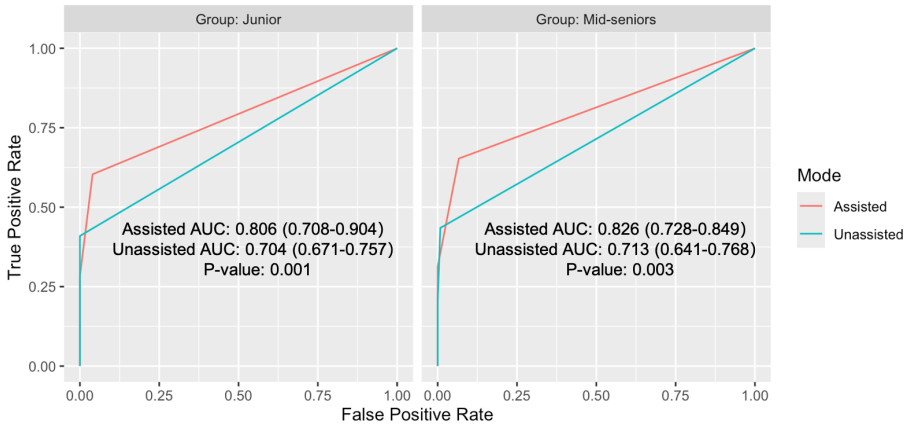

Fig. S5: Primary comparison of the area under ROC curves in the multi-reader multi-case study ROC curves evaluating the mild-severe steatosis S0 versus ≥S1 performance of junior and mid-senior radiologists assisted and unassisted with the MAOSS while interpreting the CT images. ROC=receiver operating characteristic. Numbers in parentheses are areas under curves (AUC) with 95% CI. Delong tests were performed and p value <0.05 indicates the significant difference between the compared groups.

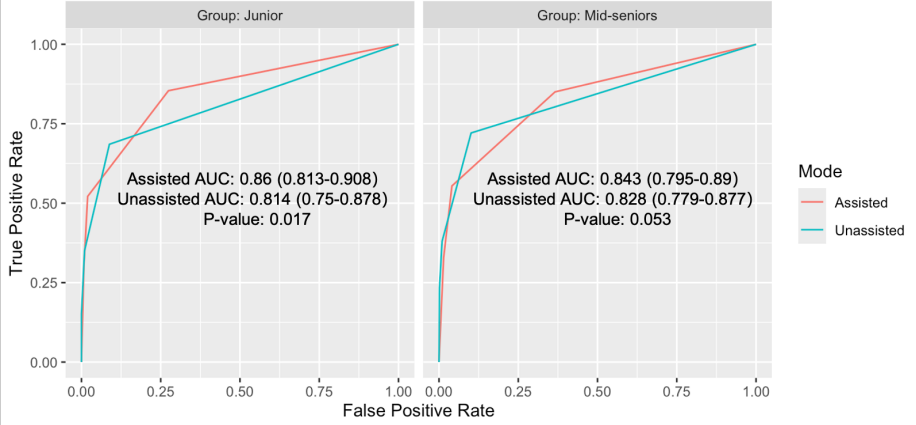

Fig. S6: Primary comparison of the area under ROC curves in the multi-reader multi-case study ROC curves evaluating the moderate-severe steatosis $\leq$S1 versus$\geq$S2 performance of junior and mid-senior radiologists assisted and unassisted with the MAOSS while interpreting the CT images. ROC=receiver operating characteristic. Numbers in parentheses are areas under curves (AUC) with 95% CI. Delong tests were performed and p value <0.05 indicates the significant difference between the compared groups.

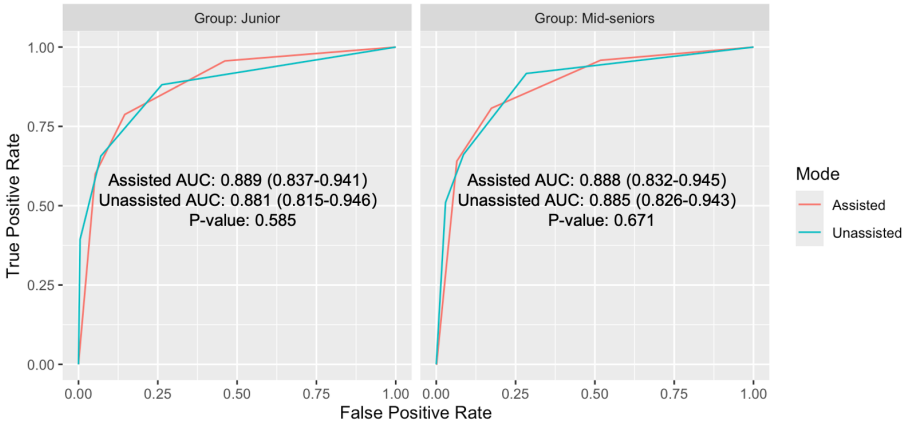

Fig. S7: Primary comparison of the area under ROC curves in the multi-reader multi-case study ROC curves evaluating the severe steatosis ≤S2 versus S3 performance of junior and mid-senior radiologists assisted and unassisted with the MAOSS while interpreting the CT images. ROC=receiver operating characteristic. Numbers in parentheses are areas under curves (AUC) with 95% CI. Delong tests were performed and p value <0.05 indicates the significant difference between the compared groups.

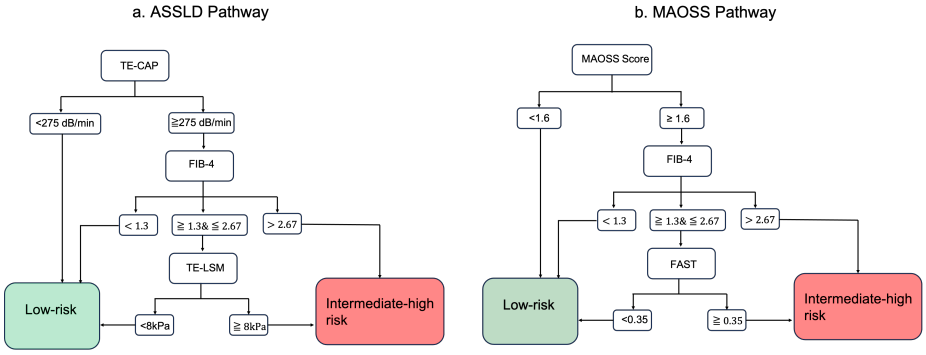

Fig. S8: ASSLD and MAOSS pathway for identifying steatosis patients at risk of steatohepatitis (NAS≥4+≥F2) or advanced liver fibrosis (≥F3).

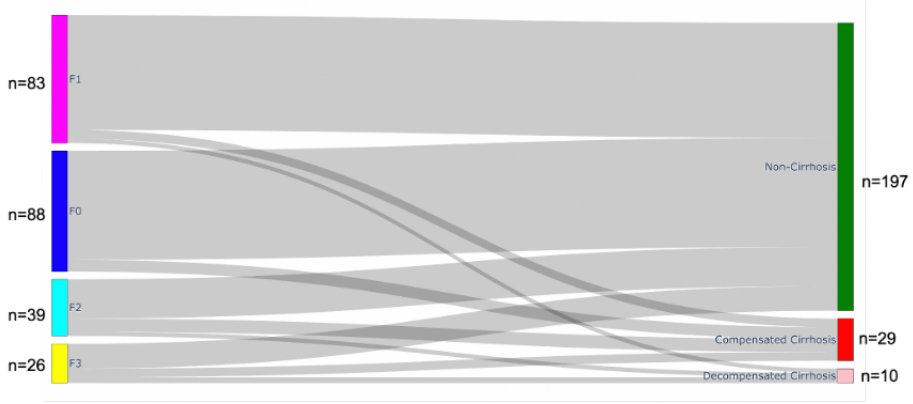

Fig. S9: Sankey diagram illustrates the progression of liver fibrosis in a subcohort of patients (n=236) with liver fibrosis stage F0-F3 in the risk-stratification cohort. 197 patients have not developed cirrhosis, 29 patients developed compensated cirrhosis and 10 patients developed decompensated cirrhosis. F0= None liver fibrosis, F1= perisinusoidal or periportal fibrosis, F2= perisinusoidal and periportal fibrosis, F3= bridging fibrosis.

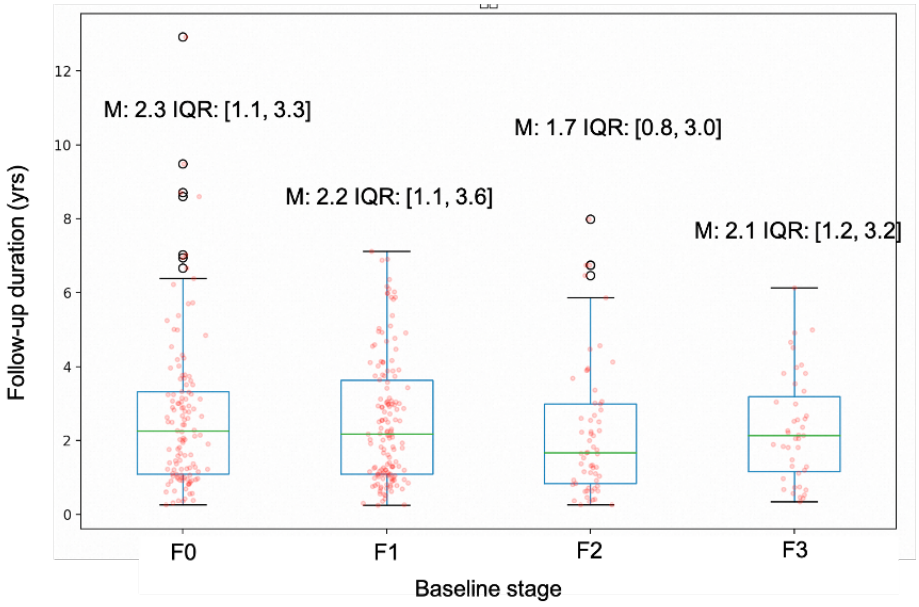

Fig. S10: Follow-up duration for MASLD patients with different stages of liver fibrosis F0-F3. F0= None liver fibrosis, F1= perisinusoidal or periportal fibrosis, F2= perisinusoidal and periportal fibrosis, F3= bridging fibrosis.

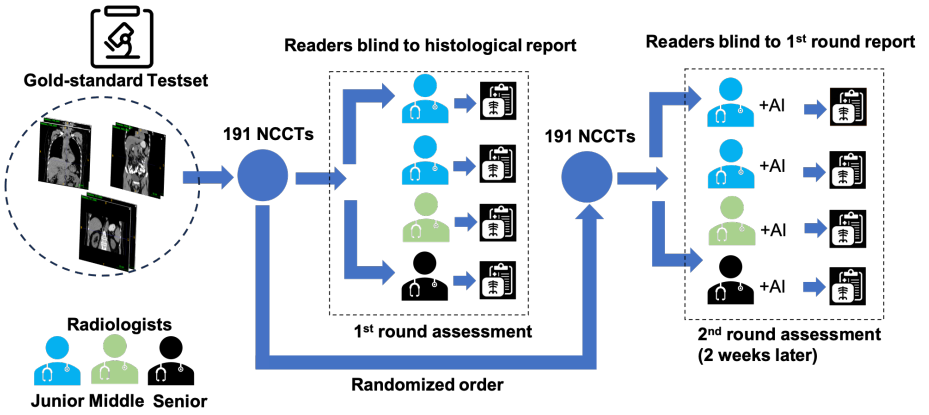

Fig. S11: Design of the Reader Study Pipeline. Radiologists ($n = 11$) with varying levels of experience participated in a two-round reader study. In the first round, each reader independently assessed 191 NCCTs. After a washout period of two weeks, the second round was conducted with the assistance of MAOSS.

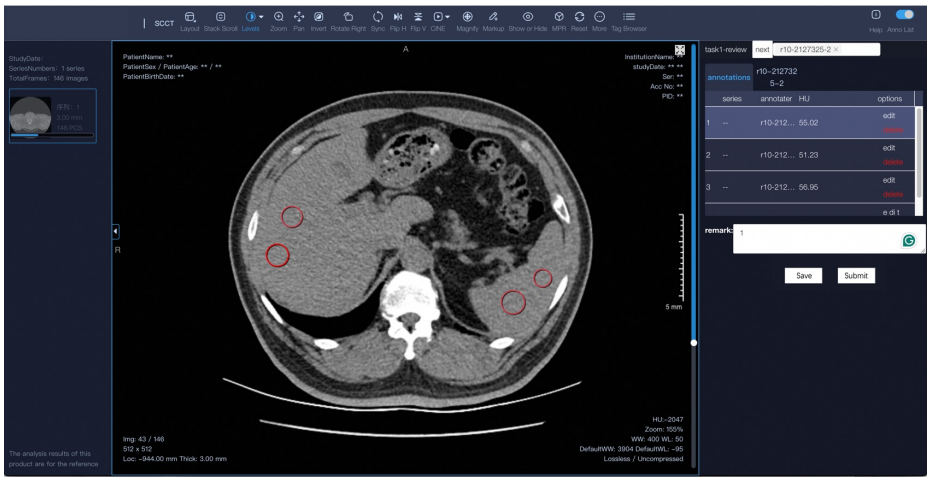

Fig. S12: In-house developed reader annotation platform for conducting diagnostic tasks. This interface illustrates how readers perform diagnoses **without** the assistance of MAOSS. Readers can freely place circular Regions of Interest (ROIs) to measure liver attenuation changes, compare liver-spleen attenuation, and assign a steatosis grade as the final result.

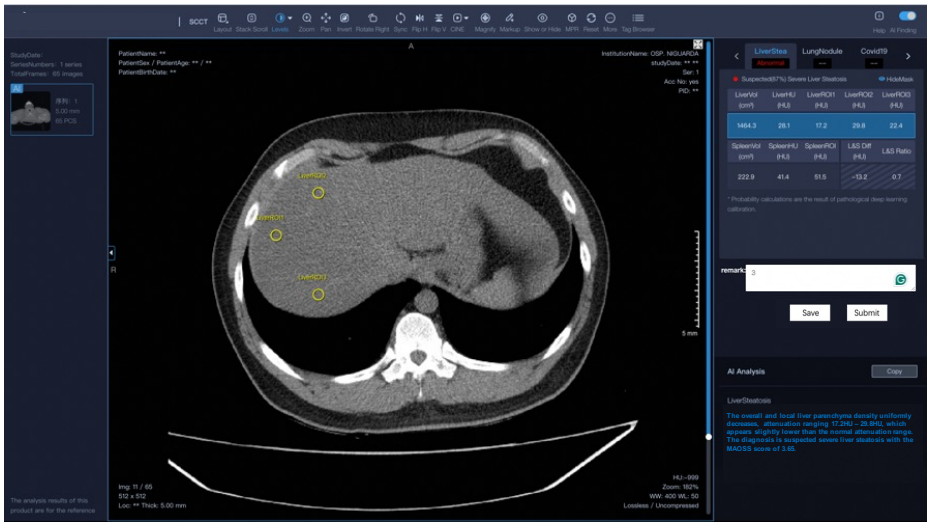

Fig. S13: In-house developed reader annotation platform assists in conducting diagnostic tasks. The interface shows how readers perform diagnoses with the assistance of MAOSS, which provides measured CT biomarkers and steatosis grades to aid readers in evaluating and grading steatosis.

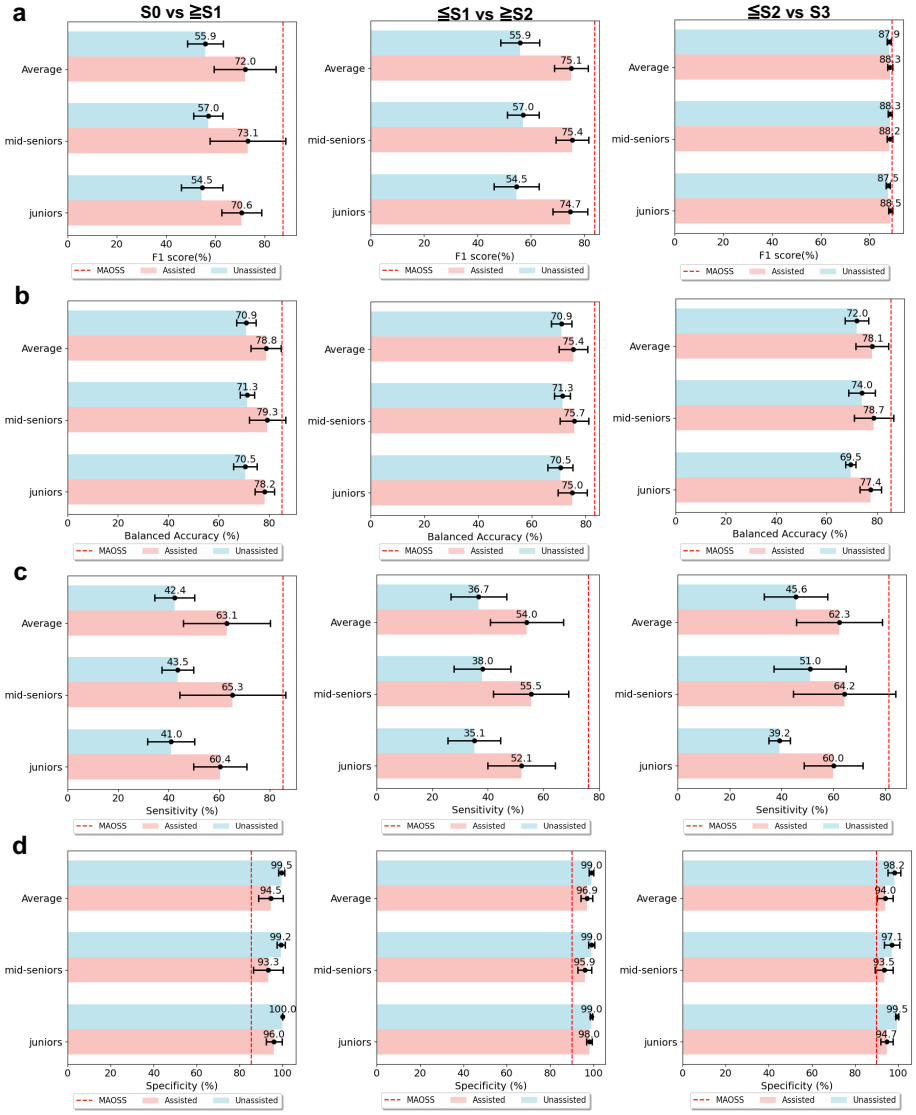

Fig. S14: Diagnostic performance of grouped radiologist with and without AI-assistance and comparison with MAOSS. We reported the measure of center and 95% confidence intervals (error bars) of the **a.** F1-score, **b.** balanced accuracy, **c.** sensitivities and **d.** specificities of readers (juniors: experience $\leq$2 yrs, mid-seniors: experience>5 yrs, average: average of all readers) and the MAOSS. Columns from left to right: S0 vs $\geq$S1, $\leq$S1 vs $\geq$S2 and $\leq$S2 vs S3.

Section7. Supplementary Tables.

Table S1: Patient Characteristics

Characteristic	Development Dataset (n=1783)		Internal Test (n=191)	External Test (n=347)	MRI-PDFF Test (n=375)	Real-World Test (n=18,504)	Risk Stratification (n=1,192)
	with gold standards (n=680)	without gold standards (n=1103)					
Age (y)	41 (33, 49)	40 (31, 54)	46 (35, 58)	37 (29, 51)	51 (40, 63)	52 (39, 64)	42 (33, 52)
Sex							
M	392 (57.6)	603 (54.6)	98 (51.3)	139 (40.4)	198 (52.8)	11202 (60.5)	579 (48.6)
F	288 (42.4)	501 (45.4)	93 (48.7)	205 (59.6)	177 (47.2)	7302 (39.5)	613 (51.4)
BMI	27.2 (24.1, 29.7)	N.A	28.1 (26.2, 30.1)	33.8 (29, 39.8)	N.A	N.A	24 (21.9, 26.7)
AST (IU/L)	34.8 (22, 49)	N.A	36.8 (29.3, 46.8)	29.1 (19.9, 41.5)	N.A	N.A	34 (20, 59)
ALT (IU/L)	54.4 (31.7, 77.5)	N.A	59.4 (46.2, 79)	38.8 (22.9, 65.1)	N.A	N.A	40 (20, 77)
AST/ALT ratio	0.7 (0.6, 0.8)	N.A	0.6 (0.5, 0.7)	0.7 (0.6, 0.8)	N.A	N.A	0.8 (0.6, 1.1)
GGT (IU/L)	54.2 (27, 69.4)	N.A	71.6 (61.6, 80.2)	33.4 (19.4, 73.4)	N.A	N.A	43 (18, 84.3)
PLT (10^9/L)	217 (183, 246.3)	N.A	204.8 (184, 239.6)	232 (198.7, 275.3)	N.A	N.A	192 (156, 235)
Glucose (mmol/L)	5.4 (4.9, 5.7)	N.A	5.6 (5.4, 5.9)	5.3 (4.9, 5.7)	N.A	N.A	5.6 (4.7, 5.8)
Triglycerides (mmol/L)	1.8 (1.1, 2.5)	N.A	2.2 (1.9, 2.5)	2.2 (1.5, 2.9)	N.A	N.A	1.1 (0.8, 1.6)
Cholesterol (mmol/L)	4.2 (3.7, 4.8)	N.A	4.2 (3.9, 4.5)	4.9 (4.4, 5.5)	N.A	N.A	4.7 (4.2, 5.3)
HDL Cholesterol (mmol/L)	1 (0.9, 1.2)	N.A	1.1 (1, 1.1)	1.2 (1, 1.6)	N.A	N.A	1.2 (0.9, 1.4)
LDL Cholesterol (mmol/L)	2.9 (2.6, 3.1)	N.A	2.9 (2.8, 3)	3.1 (2.4, 3.8)	N.A	N.A	2.4 (1.7, 3)
TBIL (mol/L)	14.2 (11.1, 15.2)	N.A	15.5 (14.3, 16.5)	13.9 (9.4, 14.7)	N.A	N.A	13.7 (10.3, 18.5)
FIB-4	0.8 (0.4, 1.2)	N.A	0.7 (0.1, 1.2)	0.7 (0.4, 0.7)	N.A	N.A	0.4 (0.2, 0.9)
APRI	0.3 (0.2, 0.5)	N.A	0.3 (0.1, 0.4)	0.7 (0.4, 0.7)	N.A	N.A	0.4 (0.2, 0.9)
CAP (dB/m)	268.6 (241.8, 290.8)	N.A	278.9 (264.5, 294.2)	312.4 (279.9, 350)	N.A	N.A	234 (215.8, 226)
LSM by VCTE (kPa)	9.6 (7.4, 11)	N.A	10.3 (6.7, 11.6)	10.9 (6.8, 12.3)	N.A	N.A	9.1 (6.5, 12.6)
NCCT scans	680	1103	191	477	375	18,577	1,192
CT Thickness							
1–4 mm	513 (75.4)	507 (46)	153 (80.1)	119 (24.9)	370 (98.7)	654 (3.5)	843 (70.7)
5–10 mm	167 (24.6)	596 (54)	38 (19.9)	358 (75.1)	5 (1.3)	17,923 (96.5)	349 (29.3)
Body part examined							
Chest	573 (84.3)	259 (23.5)	53 (27.7)	357 (74.8)	159 (42.4)	15,379 (82.8)	120 (10.1)
Abdomen	107 (15.7)	845 (76.5)	138 (72.3)	120 (25.2)	216 (57.6)	3198 (17.2)	1,072 (89.9)
Time interval between CT & biopsy (d)	6 (2, 7)	N.A	7 (4, 10)	7 (6, 8)	5 (2, 23)	N.A	5 (1, 7)
Steatosis score*							
S0-None	203 (29.9)	438 (39.7)	20 (10.5)	17 (4.9)	75 (20)	16,046 (86.7)	865 (72.6)
S1-Mild	250 (36.8)	308 (27.9)	82 (42.9)	153 (44.1)	214 (57)	2,458 (13.3)	172 (14.4)
S2-Moderate	138 (20.3)	112 (10.1)	57 (29.8)	77 (22.2)	55 (14.7)		88 (7.4)
S3-Severe	89 (13)	246 (22.3)	32 (16.8)	100 (28.8)	31 (8.3)		67 (5.6)

Note: Data are numbers of participants with percentages in parentheses n(%), median (IQR), unless otherwise specified, N.A (not available). BMI= body mass index. AST= Aspartate aminotransferase. ALT= Alanine aminotransferase. HDL= High-density lipoproteins. LDL= Low-density lipoproteins. GGT=γ-glutamyl transpeptidase. TBIL= Total bilirubin. FIB-4= fibrosis-4 index. APRI= AST to Platelet Ratio Index. LSM= Liver Stiffness Measurement. CAP= Controlled Attenuation Paramter. VCTE= Vibration-Controlled Transient Elastography. Steatosis scores*: S0 (<5% of hepatocytes involved or MRI-PDFF< 5%), S1 (5%–33% of hepatocytes involved or MRI-PDFF 5%-15%), S2 (34%–66% of hepatocytes involved or MRI-PDFF 15%-25%), S3 (>66% of hepatocytes involved or MRI-PDFF $\geq$25%).

Table S2: Ablation study of critical components of MAOSS.

Models.	S0 vs ≥S1		≤S1 vs ≥S2		≤S2 vs S3		mean-BACC	mean-AUC
	BACC	AUC	BACC	AUC	BACC	AUC		
MAOSS	**85.1**$_{\pm 2.2}$	**88.7**$_{\pm 2.1}$	**86.0**$_{\pm 1.9}$	**90.6**$_{\pm 1.4}$	**93.6**$_{\pm 1.4}$	**96.0**$_{\pm 1.4}$	**88.2**$_{\pm 1.8}$	**91.8**$_{\pm 1.6}$
w/o distill	83.2$_{\pm 3.2}$	87.5$_{\pm 2.4}$	83.7$_{\pm 3.8}$	89.4$_{\pm 1.9}$	91.6$_{\pm 1.7}$	94.4$_{\pm 1.7}$	86.2$_{\pm 2.9}$	90.4$_{\pm 2.0}$
w/o rand.mask	82.4$_{\pm 2.3}$	86.8$_{\pm 2.7}$	83.4$_{\pm 2.7}$	88.6$_{\pm 2.4}$	92.5$_{\pm 1.6}$	94.3$_{\pm 1.5}$	86.1$_{\pm 2.2}$	89.9$_{\pm 2.2}$
w/o numerical.feat	81.2$_{\pm 2.1}$	84.9$_{\pm 2.5}$	82.6$_{\pm 2.6}$	88.6$_{\pm 1.9}$	92.2$_{\pm 1.4}$	94.0$_{\pm 1.6}$	85.3$_{\pm 2.0}$	89.2$_{\pm 2.0}$
w/o texture	78.2$_{\pm 2.1}$	84.8$_{\pm 1.8}$	80.8$_{\pm 2.3}$	88.4$_{\pm 3.0}$	91.3$_{\pm 2.0}$	93.0$_{\pm 1.9}$	83.2$_{\pm 2.1}$	88.5$_{\pm 2.2}$

AUC = area under the receiver operating characteristic curv, BACC= balanced accuracy.

Table S3: Performance (AUC) comparison of different methods and modalities on internal, external and MRI-PDFF testset.

Models	Modality	S0 vs ≥ S1			≤S1 vs ≥S2			≤S2 vs S3		
		Internal	External	MRI-PDFF	Internal	External	MRI-PDFF	Internal	External (%)	MRI-PDFF
MICE[5]	UC	0.866 [0.806, 0.922]	0.864 [0.827, 0.909]	N.A	0.886 [0.835, 0.929]	0.887 [0.855, 0.915]	N.A	0.906 [0.833, 0.961]	0.872 [0.835, 0.906]	N.A
KNNImpute[6]		0.867 [0.805, 0.918]	0.886 [0.841, 0.901]	N.A	0.886 [0.837, 0.935]	0.866 [0.828, 0.903]	N.A	0.921 [0.859, 0.968]	0.822 [0.791, 0.849]	N.A
Deep-Ten[9]	UI	0.883 [0.818, 0.929]	0.846 [0.805, 0.886]	0.877 [0.841, 0.912]	0.921 [0.878, 0.957]	0.911 [0.883, 0.937]	0.930 [0.895, 0.959]	0.911 [0.842, 0.967]	0.935 [0.908, 0.958]	0.975 [0.958, 0.989]
Resnet50-3D[7]		0.879 [0.817, 0.935]	0.862 [0.822, 0.895]	0.928 [0.897, 0.953]	0.914 [0.873, 0.949]	0.920 [0.895, 0.943]	0.921 [0.882, 0.958]	0.912 [0.849, 0.964]	0.958 [0.936, 0.975]	0.977 [0.959, 0.992]
DeiT[8]		0.795 [0.721, 0.858]	0.745 [0.684, 0.801]	0.722 [0.666, 0.773]	0.879 [0.827, 0.924]	0.925 [0.9, 0.948]	0.911 [0.870, 0.951]	0.897 [0.822, 0.952]	0.929 [0.903 0.953]	0.952 [0.919, 0.978]
MultiPrompt[10]	MM	0.891 [0.833, 0.939]	0.888 [0.853, 0.921]	0.925 [0.893, 0.952]	0.917 [0.876, 0.952]	0.889 [0.855, 0.919]	0.927 [0.889, 0.958]	**0.924** [**0.867, 0.967**]	0.917 [0.888, 0.943]	0.987 [0.977, 0.996]
MAOSS (ours)		**0.918** [**0.872, 0.959**]	**0.905** [**0.867, 0.935**]	**0.929** [**0.900, 0.955**]	**0.923** [**0.884, 0.959**]	**0.934** [**0.912, 0.954**]	**0.930** [0.892, 0.965]	0.923 [0.873, 0.966]	**0.965** [**0.945, 0.980**]	**0.987** [**0.975, 0.995**]

Note: Data in brackets are 95% CIs. UC: Unimodal Clinical, UI: Unimodal Image, MM, Multimodal methods. AUC = area under the receiver operating characteristic curv.

Table S4: Diagnostic performance and accuracy of MAOSS for grading liver steatosis on internal, external and MRI-PDFF testset.

Steatosis Grading	AUC	Sensitivity (%)	Specificity (%)	Accuracy (%)	PPV (%)	NPV (%)
S0 vs. S1 or higher						
Independent test (n=20 vs. n=171)	0.918 [0.872, 0.959]	74.7 (128/171) [67.9, 81.1]	100 (20/20) [100, 100]	77.5 (148/191) [71.7, 83.3]	100 (128/128) [100, 100]	20 (20/63) [20.3, 43.1]
External test (n=32 vs. n=445)	0.903 [0.867, 0.935]	79 (352/445) [75.2, 82.7]	97 (31/32) [89.7, 100]	80.3 (383/477) [76.7, 84.1]	99.7 (352/353) [99.1, 100]	25 (31/124) [17.7, 32.8]
MRI-PDFF test (n=75 vs. n=300)	0.929 [0.90, 0.955]	75.9 (228/300) [70.7, 80.8]	90.8 (68/75) [83.7, 96.9]	78.8 (296/375) [74.4, 83.2]	97 (228/235) [94.7, 99.1]	48.5 (68/140) [40.3, 56.4]
S0 or S1 vs. S2 or S3						
Independent test (n=102 vs. n=89)	0.923 [0.884, 0.958]	84.5 (75/89) [75.9, 91.6]	85.4 (87/102) [78.2, 92.0]	84.7 (162/191) [79.1, 89.5]	83.4 (75/90) [75.8, 90.6]	86.3 (87/101) [79.1, 92.8]
External test (n=223 vs. n=254)	0.934 [0.912, 0.954]	79.1 (201/254) [73.8, 83.9]	93.3 (208/223) [89.6, 96.3]	85.8 (409/477) [82.4, 88.9]	93.2 (201/216) [89.7, 96.2]	79.6 (208/261) [74.9, 84.2]
MRI-PDFF test (n=289 vs. n=86)	0.93 [0.892, 0.965]	82.5 (71/86) [73.8, 89.9]	94.5 (273/289) [92, 96.9]	91.8 (344/375) [88.8, 94.4]	81.7 (71/87) [73.3, 88.9]	94.8 (273/288) [92.1, 97.2]
S2 or lower vs. S3						
Independent test (n=159 vs. n=32)	0.923 [0.873, 0.966]	84.4 (27/32) [70.6, 96.6]	88.6 (141/159) [83.8, 93.2]	88 (168/191) [83.2, 92.1]	60.2 (27/45) [45.2, 74.5]	86.6 (141/146) [93.3, 99.3]
External test (n=330 vs. n=147)	0.965 [0.945, 0.98]	92.4 (136/147) [88, 96.5]	92.7 (306/330) [89.8, 95.3]	92.7 (442/477) [90.4, 95]	96.6 (136/160) [79.3, 90.4]	96.5 (306/317) [86.0, 94.8]
MRI-PDFF test (n=344 vs. n=31)	0.987 [0.975, 0.995]	93.7 (29/31) [83.3, 100]	94.2 (324/344) [91.7, 96.5]	94.2 (353/375) [91.7, 96.3]	59.5 (29/49) [44.8, 74.3]	99.4 (324/326) [98.5, 100]

Note: Data in brackets are 95% CIs. Unless otherwise specified, data in parentheses are numbers of images. Hepatic steatosis grades as represented as none (grade S0), mild (S1), moderate (S2), and severe (S3). AUC = area under the receiver operating characteristic curve, NPV = negative predictive value, PPV = positive predictive value.

Table S5: Diagnostic performance of MAOSS for grading liver steatosis at different thresholds on internal testset (90% sensitivity and 90% specificity)

Hepatic Steatosis Classification	Sensitivity (%)	Specificity (%)	Accuracy (%)	PPV (%)	NPV (%)
S0 vs ≥S1 (n=20 vs. n=171)					
Optimal threshold†	74.7 (128/171) [67.9, 81.1]	100 (20/20) [100, 100]	77.1 (147/191) [71.2, 82.7]	100 (128/128) [100, 100]	20 (20/63) [20.3, 43.1]
Threshold for 90% sensitivity	91.2 (156/171) [86.7, 95.1]	75.2 (15/20) [55.6, 93.8]	89.1 (170/191) [84.8, 93.2]	96.9 (156/161) [94.2, 99.4]	50.3 (15/30) [30.4, 68]
Threshold for 90% specificity	77.9 (133/171) [71.6, 84.1]	95 (19/20) [83.3, 100]	79.2 (151/191) [73.3, 84.8]	99.3 (133/134) [97.6, 100]	33.4 (19/57) [22, 45.9]
≤S1vs ≥S2 (n=102 vs. n=89)					
Optimal threshold†	84.5 (75/89) [75.9, 91.6]	85.4 (87/102) [78.2, 92.0]	84.3 (161/191) [78.5, 89.5]	83.4 (75/90) [75.8, 90.6]	86.3 (87/101) [79.1, 92.8]
Threshold for 90% sensitivity	90.9 (81/89) [84.3, 96.6]	71.7 (73/102) [63, 80.4]	80.1 (153/191) [74.3, 85.9]	73.6 (81/110) [65.5, 81.3]	90.2 (73/81) [83.5, 96.2]
Threshold for 90% specificity	77.7 (69/89) [69.3, 85.9]	96 (98/102) [84.5, 96]	83.8 (160/191) [74.3, 85.9]	97.3 (69/71) [79.5, 94.4]	82.2 (99/112) [74.5, 89.3]
≤S2vs S3 (n=159 vs. n=32)					
Optimal threshold†	84.4 (27/32) [70.6, 96.6]	88.6 (141/159) [83.8, 93.2]	87.5 (167/191) [82.7, 92.1]	60.2 (27/45) [45.2, 74.5]	96.6 (141/146) [93.3, 99.3]
Threshold for 90% sensitivity	90.4 (29/32) [77.8, 97.7]	77.4 (123/159) [70.1, 83.5]	81.1 (155/191) [75.4, 86.4]	47.2 (29/61) [34.4, 59.3]	97.7 (127/130) [94.6, 100]
Threshold for 90% specificity	70.8 (25/32) [63.6, 91.4]	91.8 (146/159) [87.5, 95.7]	89 (170/191) [84.3, 93.2]	83.3 (25/30) [48.7, 80]	95.4 (146/153) [91.8, 98.7]

Note: Data in brackets are 95% CIs. Unless otherwise specified, data in parentheses are numbers of images. NPV = negative predictive value, PPV = positive predictive value. † Optimal threshold–indicated cutoff values that maximize the Youden index.

Table S6: Diagnostic performance of MAOSS for grading liver steatosis at different thresholds on external testset (90% sensitivity and 90% specificity)

Hepatic Steatosis Classification	Sensitivity (%)	Specificity (%)	Accuracy (%)	PPV (%)	NPV (%)
S0 vs ≥S1 (n=32 vs. n=445)					
Optimal threshold†	79 (352/445) [75.2, 82.7]	97 (31/32) [89.7, 100]	78 (382/477) [76.5, 83.4]	99.7 (352/353) [99.1, 100]	25 (31/124) [17.7, 32.8]
Threshold for 90% sensitivity	90.1 (401/445) [87.3, 92.7]	56.4 (18/32) [39.6, 75]	87.5 (418/477) [84.3, 90.4]	96.6 (401/415) [94.9, 98.3]	29.2 (18/62) [17.3, 40.6]
Threshold for 90% specificity	79.7 (354/445) [75.9, 100]	90.7 (29/32) [79.4, 100]	78 (382/477) [76.5, 83.4]	99.1 (354/357) [98.0, 100]	24.3 (29/120) [16.5, 32]
≤S1vs ≥S2 (n=223 vs. n=254)					
Optimal threshold†	79.1 (201/254) [73.8, 83.9]	93.3 (208/223) [89.6, 96.3]	85.4 (408/477) [82.2, 88.7]	93.2 (201/216) [89.7, 96.2]	79.6 (208/261) [74.9, 84.2]
Threshold for 90% sensitivity	91.7 (233/254) [88.2, 94.9]	77.9 (174/223) [72.8, 82.9]	85.1 (406/477) [81.9, 88.1]	82.5 (233/282) [77.7, 86.6]	89.3 (174/195) [85, 93.7]
Threshold for 90% specificity	79.6 (202/254) [74.6, 84.6]	92.9 (207/223) [89.4, 95.9]	85.5 (408/477) [82.2, 88.5]	92.7 (202/218) [89.1, 96.2]	79.9 (207/259) [74.5, 84.6]
≤S2vs S3 (n=330 vs. n=147)					
Optimal threshold†	92.4 (136/147) [88, 96.5]	92.7 (306/330) [89.8, 95.3]	92.4 (441/477) [89.9, 94.5]	96.6 (136/160) [79.3, 90.4]	96.5 (306/317) [94.5, 98.4]
Threshold for 90% sensitivity	92.6 (136/147) [88.3, 96.5]	92.7 (306/330) [89.8, 95.3]	92.4 (441/477) [89.9, 94.5]	96.6 (136/160) [79.5, 90.7]	96.5 (306/317) [94.4, 98.4]
Threshold for 90% specificity	92.4 (136/147) [87.8, 96.5]	92.7 (306/330) [89.8, 95.5]	92.4 (441/477) [89.9, 94.5]	96.5 (136/160) [79.4, 90.4]	96.5 (306/317) [94.5, 98.4]

Note: Data in brackets are 95% CIs. Unless otherwise specified, data in parentheses are numbers of images. NPV = negative predictive value, PPV = positive predictive value. † Optimal threshold–indicated cutoff values that maximize the Youden index.

Table S7: Diagnostic performance of MAOSS for grading liver steatosis at different thresholds on MRI-PDFF testset (90% sensitivity and 90% specificity)

Hepatic Steatosis Classification	Sensitivity (%)	Specificity (%)	Accuracy (%)	PPV (%)	NPV (%)
S0 vs ≥S1 (n=75 vs. n=300)					
Optimal threshold†	75.9 (228/300) [70.7, 80.8]	90.8 (68/75) [83.7, 96.9]	78.8 (296/375) [74.4, 83.2]	97 (228/235) [94.7, 99.1]	48.5 (68/140) [40.3, 56.4]
Threshold for 90% sensitivity	90 (270/300) [86.4, 93.2]	70.9 (53/75) [60.7, 80.5]	86.1 (323/375) [82.7, 89.6]	92.4 (270/292) [89.4, 95.3]	64.1 (53/83) [53.2, 74.1]
Threshold for 90% specificity	76.2 (228/300) [70.9, 81]	90.5 (68/75) [83.5, 96.3]	78.9 (338/375) [74.7, 82.9]	97 (228/235) [94.9, 98.8]	48.6 (68/140) [40.5, 56.9]
≤S1vs ≥S2 (n=289 vs. n=86)					
Optimal threshold†	82.5 (71/86) [73.8, 89.9]	94.5 (273/289) [92, 96.9]	91.8 (344/375) [88.8, 94.4]	81.7 (71/87) [73.3, 88.9]	94.8 (273/288) [92.1, 97.2]
Threshold for 90% sensitivity	89.4 (77/86) [82.3, 95.5]	72.4 (209/289) [67.3, 77.3]	76.3 (286/375) [71.7, 80.8]	49.1 (77/157) [41.6, 56.3]	95.9 (209/218) [93.2, 98.6]
Threshold for 90% specificity	83.7 (29/31) [83.3, 100]	94.2 (324/344) [91.7, 96.5]	94.2 (353/375) [91.7, 96.3]	59.5 (29/49) [44.8, 74.3]	99.4 (324/326) [98.5, 100]
≤S2vs S3 (n=344 vs. n=31)					
Optimal threshold†	93.7 (29/31) [83.3, 100]	94.2 (324/344) [91.7, 96.5]	94.2 (353/375) [91.7, 96.3]	59.5 (29/49) [44.8, 74.3]	99.4 (324/326) [98.5, 100]
Threshold for 90% sensitivity	93.5 (29/31) [83.3, 100]	94.2 (324/344) [91.7, 96.5]	94.2 (353/375) [91.7, 96.3]	59.5 (29/49) [44.8, 74.3]	99.4 (324/326) [98.5, 100]
Threshold for 90% specificity	93.5 (29/31) [83.3, 100]	94.2 (324/344) [91.6, 96.8]	94.1 (353/375) [91.7, 96.3]	59.4 (29/49) [45.2, 72.7]	99.4 (324/326) [98.5, 100]

Note:Data in brackets are 95% CIs. Unless otherwise specified, data in parentheses are numbers of images. NPV = negative predictive value, PPV = positive predictive value. † Optimal threshold–indicated cutoff values that maximize the Youden index.

Table S8: Diagnostic performance and accuracy of unimodal-clinical model for grading liver steatosis on internal and external test

Steatosis Grading	AUC	Sensitivity (%)	Specificity (%)	Accuracy (%)	PPV (%)	NPV (%)
S0 vs. S1 or higher						
Independent test	0.866	65.4 (112/171)	100 (20/20)	69.2 (132/191)	100 (112/112)	25.3 (20/79)
(n=20 vs. n=171)	[0.806, 0.922]	[58.1, 72.9]	[100, 100]	[62.8, 75.9]	[100, 100]	[15.8, 35.6]
External test	0.886	68.1 (303/445)	100 (32/32)	70.1 (335/477)	100 (303/303)	18.3 (32/174)
(n=32 vs. n=445)	[0.822, 0.905]	[63.8, 72.6]	[100, 100]	[66, 74]	[100, 100]	[12.9, 23.9]
S0 or S1 vs. S2 or S3						
Independent test	0.886	76.6 (68/89)	88.2 (90/102)	82.7 (158/191)	85.2 (68/80)	81 (90/111)
(n=102 vs. n=89)	[0.835, 0.929]	[67, 85.6]	[81.9, 94]	[77, 88]	[76.6, 92.6]	[73.3, 87.8]
External test	0.887	80.1 (203/254)	86.5 (193/223)	83.1 (396/477)	87.1 (203/233)	79.1 (193/244)
(n=223 vs. n=254)	[0.855, 0.915]	[75.5, 84.7]	[81.9, 90.8]	[79.7, 86.4]	[82.6, 91]	[73.6, 84.2]
S2 or lower vs. S3						
Independent test	0.906	87.7 (28/32)	87.4 (141/159)	87.5 (168/191)	58.2(27/45)	97.3 (141/146)
(n=159 vs. n=32)	[0.833, 0.961]	[75, 97]	[82, 92.3]	[82.7, 91.6]	[44.6, 72.6]	[94.3, 99.3]
External test	0.872	90.4 (133/147)	69.6 (230/330)	76.1 (363/477)	57.2 (133/233)	94.3 (230/244)
(n=330 vs. n=147)	[0.835, 0.906]	[85.6, 94.9]	[64.4, 74.5]	[72.1, 79.9]	[50.7, 63.4]	[91.4, 97]

Note: Data in brackets are 95% CIs. Unless otherwise specified, data in parentheses are numbers of images. Hepatic steatosis grades as represented as none (grade S0), mild (S1), moderate (S2), and severe (S3). AUC = area under the receiver operating characteristic curve, NPV = negative predictive value, PPV = positive predictive value.

Table S9: Diagnostic performance and accuracy of unimodal-image model for grading liver steatosis on internal and external testset.

Steatosis Grading	AUC	Sensitivity (%)	Specificity (%)	Accuracy (%)	PPV (%)	NPV (%)
S0 vs. S1 or higher						
Independent test	0.881	71.5(122/171)	100 (20/20)	79.4 (142/191)	100 (122/122)	28.8 (20/69)
(n=20 vs. n=171)	[0.823, 0.926]	[64.9, 78.2]	[100, 100]	[70.6, 78.8]	[100, 100]	[14.9, 27.9]
External test	0.864	73.2 (325/445)	100 (32/32)	74.7 (357/477)	100 (325/325)	20.9 (32/152)
(n=32 vs. n=445)	[0.827, 0.909]	[69.1, 77.2]	[100, 100]	[70.6, 78.8]	[100, 100]	[14.9, 27.9]
S0 or S1 vs. S2 or S3						
Independent test	0.914	83.2 (74/89)	84.3 (86/102)	83.8 (160/191)	82.1 (74/90)	85.2 (86/101)
(n=102 vs. n=89)	[0.87, 0.952]	[75.3, 90.6]	[76.8, 91.3]	[81.8, 88.1]	[74.2, 89.5]	[76.9, 91.8]
External test	0.916	81.5 (207/254)	88.8 (198/223)	84.9 (405/477)	89.3(207/232)	80.9 (198/245)
(n=223 vs. n=254)	[0.892, 0.939]	[76.4, 86.1]	[84.5, 92.6]	[81.8, 88.1]	[85.3, 93]	[75.5, 85.5]
S2 or lower vs. S3						
Independent test	0.914	87.8(28/32)	86 (137/159)	86.4 (161/191)	55.7 (28/50)	97.1 (137/141)
(n=159 vs. n=32)	[0.848, 0.965]	[75.8, 97.1]	[80.5, 90.9]	[81.2, 91.1]	[41.8, 68.2]	[94.1, 99.3]
External test	0.952	89.1 (131/147)	88.5 (292/330)	88.7 (423/477)	77.3 (131/169)	94.8 (292/308)
(n=330 vs. n=147)	[0.93, 0.971]	[83.9, 93.8]	[84.9, 91.8]	[86, 91.4]	[70.6, 83.5]	[92.3, 97.2]

Note: Data in brackets are 95% CIs. Unless otherwise specified, data in parentheses are numbers of images. Hepatic steatosis grades as represented as none (grade S0), mild (S1), moderate (S2), and severe (S3). AUC = area under the receiver operating characteristic curve, NPV = negative predictive value, PPV = positive predictive value.

Table S10: Diagnostic performance and accuracy of TE-CAP for grading liver steatosis on internal and external testset.

Steatosis Grading	AUC	Sensitivity (%)	Specificity (%)	Accuracy (%)	PPV (%)	NPV (%)
S0 vs. S1 or higher						
Independent test	0.791	62 (106/171)	100 (20/20)	66 (126/191)	100 (106/106)	23.5 (20/85)
(n=20 vs. n=171)	[0.725, 0.852]	[54.7, 69.2]	[100, 100]	[59.2, 72.3]	[100, 100]	[14.7, 32.9]
External test	0.832	59.9 (266/445)	97 (31/32)	62.3 (297/477)	99.6 (266/267)	14.8 (179/210)
(n=32 vs. n=445)	[0.777, 0.882]	[55.3, 64]	[90, 100]	[58.1, 66.7]	[98.8, 100]	[10.1, 19.5]
S0 or S1 vs. S2 or S3						
Independent test	0.842	75.4 (67/89)	87 (87/102)	80.6 (154/191)	81.9 (67/82)	80.1 (87/109)
(n=102 vs. n=89)	[0.78, 0.897]	[66.3, 84.1]	[75.4, 91.6]	[74.9, 85.9]	[73.6, 89.3]	[72.2, 87.3]
External test	0.868	81.1 (206/254)	84.3 (188/223)	82.5 (394/477)	85.4 (206/241)	79.8 (188/236)
(n=223 vs. n=254)	[0.831, 0.901]	[76, 85.9]	[79.5, 89.5]	[79, 85.7]	[80.6, 89.7]	[74.6, 84.8]
S2 or lower vs. S3						
Independent test	0.853	78.5 (25/32)	82.5 (131/159)	81.6 (156/191)	47.3 (25/53)	94.9 (131/138)
(n=159 vs. n=32)	[0.764, 0.925]	[64.1, 92.3]	[75.9, 88.3]	[75.9, 86.9]	[34.4, 60.7]	[91, 98.3]
External test	0.851	93.8 (138/147)	70.9 (234/330)	78 (372/477)	58.8 (138/234)	96.3 (234/243)
(n=330 vs. n=147)	[0.819, 0.885]	[89.8, 97.3]	[66.5, 75.9]	[74, 81.8]	[52.5, 64.7]	[93.7, 98.7]

Note: Data in brackets are 95% CIs. Unless otherwise specified, data in parentheses are numbers of images. Hepatic steatosis grades as represented as none (grade S0), mild (S1), moderate (S2), and severe (S3). AUC = area under the receiver operating characteristic curve, NPV = negative predictive value, PPV = positive predictive value.

Table S11: Diagnostic performance and accuracy of conventional CT imaging biomarkers for detecting liver steatosis on internal, external and MRI-PDFF testset.

	Cutoff	Sensitivity (%)	Specificity (%)	Accuracy (%)	PPV (%)	NPV (%)	BACC (%)	F1-score (%)
Internal test								
Liver attenuation	$\leq$ 40Hu	24.1 (41/171) [17.8, 30.6]	100 (20/20) [100, 100]	31.9 (61/191) [25.1, 38.2]	100 (41/41) [100, 100]	13.2 (20/150) [8.1, 18.5]	61.9 [58.8, 65.1]	36.9 [29.6, 44.7]
Liver-spleen ratio	$\leq$ 0.8	18.1 (31/171) [12.7, 23.9]	100 (20/20)[100, 100]	26.6 (51/191) [20.9, 32.9]	100 (31/31) [100, 100]	12.5 (20/160) [7.8, 17.6]	59.1 [56.3, 62]	29.8 [22.1, 37.7]
Liver-spleen diff	≤ -10Hu	80.6 (138/171) [74.9, 86.1]	74.6 (15/20) [53.6, 93.3]	80.2 (153/191) [74.3, 85.9]	96.5 (138/143) [93.3, 99.3]	31.2 (15/48) [18.7, 44.7]	77.9 [67.7, 87.6]	83.3 [78.6, 87.6]
External test								
Liver attenuation	$\leq$ 40Hu	49.3 (219/445) [44.6, 54.2]	100 (32/32) [100, 100]	52.6 (251/477) [48.2, 57.2]	100 (219/219) [100, 100]	12.4 (32/258) [8.4, 16.7]	74.6 [72.2, 77.1]	62.9 [58.7, 66.9]
Liver-spleen ratio	$\leq$ 0.8	40.6 (181/445) [36.2, 45.4]	100 (32/32) [100, 100]	44.6 (213/477) [40.2, 49.1]	100 (181/181) [100, 100]	10.7 (32/296) [7.4, 14.4]	70.3 [68.1, 72.7]	55.3 [50.9, 59.4]
Liver-spleen diff	≤ -10Hu	85.9 (382/445) [82.4, 89.1]	62.5 (20/32) [46.1, 77.8]	84.3 (402/477) [80.9, 87.4]	96.9 (382/394) [95.1, 98.5]	24.1 (20/83) [14.9, 33.8]	74.2 [65, 82.5]	87.3 [84.7, 89.9]
MRI-PDFF test								
Liver attenuation	$\leq$ 40Hu	15.0 (45/300) [11.3, 19.5]	100 (75/75) [100, 100]	15.0 (120/375) [11.3, 19.5]	100 (45/45) [100, 100]	22.7 (75/330) [18.0, 26.8]	57.6 [55.6, 59.7]	28.4 [22.8, 34.0]
Liver-spleen ratio	$\leq$ 0.8	11.6 (35/300) [8.1, 15.4]	100 (75/75) [100, 100]	29.3 (110/375) [24.8, 34.1]	100 (35/35) [100, 100]	22.0 (75/340) [17.9, 26.5]	55.9 [54.2, 57.8]	23.9 [18.7, 28.9]
Liver-spleen diff	≤ -10Hu	67.2 (202/300) [61.5, 72.5]	81.4 (61/75) [72.1, 89.7]	70.2 (263/375) [65.9, 75.2]	93.5 (202/216) [89.9, 96.6]	38.6 (61/159) [30.8, 46.3]	74.2 [69.1, 79.3]	73.2 [68.9, 77.1]

Note: Data in brackets are 95% CIs. Unless otherwise specified, data in parentheses are numbers of images. NPV = negative predictive value, PPV = positive predictive value, BACC = balanced accuracy, liver-spleen ratio = liver/spleen Hu ratio, liver-spleen diff = liver-spleen Hu.

Table S12: Delong test for comparing the AUCs of MAOSS to unimodal-image, unimodal-clinical and TE-CAP models.

	S0 vs. S1 or higher	**S0 or S1 vs. S2 or S3**	**S2 or lower vs. S3**
Internal Test			
MAOSS vs. TE-CAP	0.002	0.025	0.136
MAOSS vs. Unimodal-clinical	0.033	0.007	0.13
MAOSS vs. Unimodal-image	0.019	0.237	0.203
External Test			
MAOSS vs. TE-CAP	0.004	1.66e-06	1.00e-12
MAOSS vs. Unimodal-clinical	0.009	8.66e-06	3.93e-08
MAOSS vs. Unimodal-image	3.60e-07	0.001	0.019

Note: p-value less than 0.05 was considered statistically significant.

Table S13: Reader Characteristics.

Reader	Speciality	Years of Experience	Training/Expertise
R2	Radiology	1	resident radiologist
R8	Radiology	1	resident radiologist
R11	Radiology	1	resident radiologist
R3	Radiology	2	resident radiologist
R4	Radiology	2	resident radiologist
R6	Radiology	>5	midcareer radiologist
R9	Radiology	>5	midcareer radiologist
R1	Radiology	>5	midcareer radiologist
R10	Radiology	>10	senior radiologist
R7	Radiology	>10	senior radiologist
R5	Radiology	>15	senior radiologist

Table S14: Odds of different clinical pathways for screening patients at risk of steatohepatitis and advanced liver fibrosis.

Clinical Pathways	**Biomarkers**	**Steatohepatitis Detection (NAS≥4+ ≥F2)**	**P Value**	**Advanced Fibrosis or Cirrhosis Detection (≥F3)**	**P Value**
ASSLD Pathway	CAP	1.86 (1.23, 2.79)	0.002	1.56 (0.98, 2.45)	0.04
	CAP+FIB-4	2.95 (1.72, 4.93)	8.2e-06	4.81 (2.85, 8)	6.38e-12
	CAP+FIB-4+LSM	3.38 (1.93, 5.76)	9.10e-07	5.59 (3.25, 9.48)	1.33e-13
MAOSS Pathway	MAOSS	3.37 (2.3, 4.99)	1.40e-11	3.63 (2.39, 5.58)	4.23e-11
	MAOSS+FIB-4	4.89 (3.35, 7.14)	<2.2e-16	6.88 (4.57, 10.46)	<2.2e-16
	MAOSS+FIB-4+FAST	5.08 (3.47, 7.47)	<2.2e-16	6.4 (4.24, 9.69)	<2.2e-16

Note: Data in parentheses are 95% CIs. Odds ratios were calculated for the ablation of biomarkers compared with histology gold standards in ASSLD and MAOSS pathways. P values from χ^2 tests were used to evaluate the statistical significance of differences in odds ratios between the prediction and gold standards. CAP = Controlled attenuation parameter, FIB-4 = Fibrosis-4 Index, LSM = Liver Stiffness Measurements, FAST = FibroScan-AST score.

Table S15: Diagnostic accuracy metrics of different methods for assessing SLD patients at risk of steatohepatitis (NAS≥4 and fibrosis stage ≥F2) of AASLD pathway and MAOSS pathway in primary care screening.

Variables	AASLD Pathway			MAOSS Pathway		
	CAP	CAP+FIB-4	CAP+FIB-4+LSM	MAOSS	MAOSS+FIB-4	MAOSS+FIB-4+FAST
SEN	29.1 (42/145) [21.9, 36.4]	17.2 (25/145) [11.1, 23.3]	16.5 (24/145) [10.6, 22.8]	66.9 (97/145) [59.2, 74.4]	51.6 (75/145) [43.3, 59.9]	48.3 (70/145) [40.3, 56.8]
SPEC	82 (859/1047) [79.7, 84.4]	93.4 (978/1047) [91.9, 94.9]	94.5 (989/1047) [93, 95.8]	62.5 (655/1047) [59.5, 65.5]	82 (859/1047) [79.6, 84.5]	84.5 (885/1047) [82.2, 86.7]
PPV	18.3 (42/230) [13.1, 23.8]	26.6 (25/94) [17.3, 35.8]	20.4 (24/117) [16.8, 40]	19.8 (97/489) [16.4, 23.3]	28.3 (75/263) [22.3, 39.3]	30.2 (70/232) [24.7, 35.9]
NPV	89.3 (859/962) [87.2, 91.1]	89.1 (978/1098) [87.2, 90.9]	89.1 (989/1110) [87.4, 90.9]	93.2 (655/703) [91.3, 95]	92.5 (859/929) [90.6, 94.1]	92.2 (885/960) [90.5, 93.9]
ACC	75.7 (901/1192) [73.1, 77.9]	84.1 (1003/1192) [81.9, 86.1]	84.9 (1013/1192) [82.7, 87.1]	62.9 (752/1192) [60.4, 65.4]	78.4 (934/1192) [76.1, 80.8]	80.2 (955/1192) [77.9, 82.2]

Note: Data in brackets are 95% CIs. Unless otherwise specified, data in parentheses are numbers of patients.
CAP = Controlled attenuation parameter, FIB-4 = Fibrosis-4 Index, LSM = Liver Stiffness Measurements, FAST = FibroScan-AST score, SEN = sensitivity, SPEC = specificity, PPV = positive predictive value, NPV = negative predictive value, ACC = accuracy.

Table S16: Diagnostic accuracy metrics of different methods for assessing steatosis patients at risk of advanced liver fibrosis ≥F3 of AASLD pathway and MAOSS pathway in primary care screening.

Variables	AASLD Pathway			MAOSS Pathway		
	CAP	CAP+FIB-4	CAP+FIB-4+LSM	MAOSS	MAOSS+FIB-4	MAOSS+FIB-4+FAST
SEN	26.1 (32/122) [18.9, 33.9]	23.9 (29/122) [16.7, 31.7]	23 (28/122) [15.6, 30.5]	68.9 (84/122) [60.5, 77.2]	59.8 (73/122) [50.4, 68.8]	54.1 (66/122) [44.6, 62.6]
SPEC	81.6 (872/1070) [79.2, 83.7]	93.9 (1005/1070) [92.4, 95.4]	94.9 (1016/1070) [93.5, 96.2]	62.1 (655/1070) [59.4, 65.1]	82.2 (880/1070) [79.7, 84.7]	84.5 (904/1070) [82.3, 86.6]
PPV	13.8 (32/230) [9.3, 18.6]	30.8 (29/94) [21.9, 40]	34.2 (28/82) [23.9, 45.7]	17.2 (84/489) [13.9, 20.5]	27.8 (73/263) [22.1, 33.3]	28.5 (66/232) [22.6, 34.5]
NPV	90.7 (872/962) [88.8, 92.4]	91.5 (1005/1098) [89.8, 93.2]	91.5 (1016/1110) [89.9, 93]	94.6 (665/703) [92.8, 96.3]	93.7 (880/929) [93.2, 96.2]	94.2 (904/960) [92.8, 95.6]
ACC	75.9 (904/1192) [73.5, 78.4]	86.8 (1034/1192) [84.9, 88.6]	87.6 (1044/1192) [85.7, 89.3]	62.8 (749/1192) [60.1, 65.4]	79.9 (953/1192) [77.6, 82.1]	81.5 (970/1192) [79.2, 83.7]

Note: Data in brackets are 95% CIs. Unless otherwise specified, data in parentheses are numbers of patients. CAP = Controlled attenuation parameter, FIB-4 = Fibrosis-4 Index, LSM = Liver Stiffness Measurements, FAST = FibroScan-AST score, SEN = sensitivity, SPEC = specificity, PPV = positive predictive value, NPV = negative predictive value, ACC = accuracy.

Table S17: MAOSS pathway cox regression and competing risk analysis.

Variable	**n=**	**Univariable Analysis**		**Multivariable Analysis**		**Competing Risk Analysis (Multivariabl)**	
		Harzard Ratio	**P Value**	**Harzard Ratio**	**P Value**	**Harzard Ratio**	**P Value**
MAOSS score							
<1.6	160 (67.8%)	Reference	-		-		
>= 1.6	76 (32.2%)	2.45 (1.29-4.64)	0.006	2.07 (0.77-5.59)	0.15	2.07 (0.82-5.22)	0.12
FIB-4							
<1.3	154 (65.3%)	Reference	-		-		
>= 1.3	82 (34.7%)	3.42 (1.77-6.63)	<0.001	0.52 (0.15-1.77)	0.292	0.52 (0.16-1.7)	0.3
FAST							
<0.35	164 (69.5%)	Reference	-		-		
>= 0.35	72 (30.5%)	3.17 (1.66-6.04)	<0.001	1.51 (0.64-3.59)	0.346	1.51 (0.69-3.33)	0.3
MAOSS Pathway							
Low Risk	203 (86%)	Reference	-		-		
Intermediate-high Risk	33 (14%)	5.54 (2.69-10.32)	<0.001	5.45 (2.28-13.03)	<0.001	5.54 (2.29-13)	<0.001

Note: Data in parentheses are 95% CIs. There were 236 patients (n=236) in total and 39 patients (n=39) developed cirrhosis. FIB-4 = Fibrosis-4 Index, FAST = FibroScan-AST score, MAOSS pathway = MAOSS score + FIB-4 + FAST.

Table 18: ASSLD pathway cox regression and competing risk analysis.

Variable	n=	Univariable Analysis		Multivariable Analysis		Competing Risk Analysis (Multivariable)	
		Harzard Ratio	P Value	Harzard Ratio	P Value	Harzard Ratio	P Value
CAP							
<275	204 (86.4%)	Reference	-		-		
>= 275	32 (13.6%)	1.89 (0.83-4.31)	0.128	2.07 (0.58-7.34)	0.26	2.07 (0.63-6.80)	0.23
FIB-4							
<1.3	160 (67.8%)	Reference	-		-		
>= 1.3	76 (32.2%)	3.58 (1.27-10.1)	0.016	0.66 (0.13-3.36)	0.615	0.66 (0.13-3.43)	0.62
LSM							
<8	120 (50.8%)	Reference	-		-		
>= 8	116 (49.2%)	3.43 (1.62-7.26)	0.001	1.97 (0.9-4.33)	0.091	1.97 (0.91-4.26)	0.084
ASSLD pathway							
Low Risk	230 (97.5%)	Reference	-		-		
Intermediate-high Risk	6 (2.5%)	5.45 (2.32-11.42)	<0.001	4.80 (2.12-10.70)	<0.001	4.80 (2.19-10.48)	<0.001

Note: Data in parentheses are 95% CIs. There were 236 patients (n=236) in total and 39 patients (n=39) developed cirrhosis. CAP = Controlled attenuation parameter, FIB-4 = Fibrosis-4 Index, LSM = Liver Stiffness Measurements, ASSLD pathway = CAP+FIB-4+LSM.

References

1. Sun, M., Jiang, Y. & Guo, H. Semi-supervised detection, identification and segmentation for abdominal organs. In *MICCAI Challenge on Fast and Low-Resource Semi-supervised Abdominal Organ Segmentation*, 35–46 (Springer, 2022).
2. Ma, J. *et al.* Unleashing the strengths of unlabelled data in deep learning-assisted pan-cancer abdominal organ quantification: the flare22 challenge. *The Lancet Digital Health* **6**, e815–e826 (2024). URL https://www.sciencedirect.com/science/article/pii/S2589750024001547.
3. Huo, Y. *et al.* Fully automatic liver attenuation estimation combing cnn segmentation and morphological operations. *Medical physics* **46**, 3508–3519 (2019).
4. Van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *Journal of machine learning research* **9** (2008).
5. Van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in r. *Journal of statistical software* **45**, 1–67 (2011).
6. Troyanskaya, O. *et al.* Missing value estimation methods for dna microarrays. *Bioinformatics* **17**, 520–525 (2001).
7. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
8. Touvron, H. *et al.* Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357 (PMLR, 2021).
9. Zhang, H., Xue, J. & Dana, K. Deep ten: Texture encoding network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 708–717 (2017).
10. Lee, Y.-L., Tsai, Y.-H., Chiu, W.-C. & Lee, C.-Y. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14943–14952 (2023).