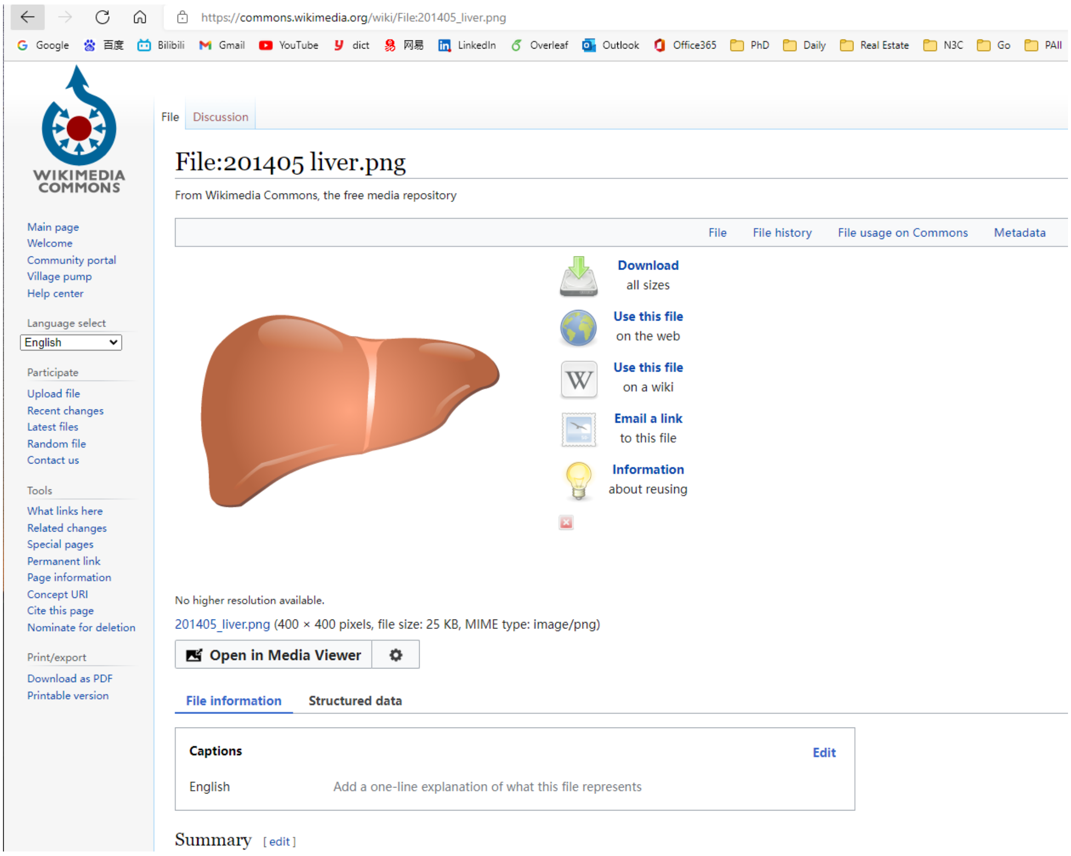
Supplementary material

(1) Copyright permission for Figure 2

Please refer to the Creative Commons Attribution 4.0 International license, which the authors release their figure under. This expressly allows reproductions and modifications without permissions. For further questions, you may consult the CC documentation and/or click the link: https://commons.wikimedia.org/wiki/File:201405_liver.png



https://commons.wikimedia.org/wiki/File:201405_liver.png

File:201405 liver.png

From Wikimedia Commons, the free media repository

No higher resolution available.

201405_liver.png (400 × 400 pixels, file size: 25 KB, MIME type: image/png)

Captions

English — Add a one-line explanation of what this file represents

Summary [edit]

Attribution 4.0 International (Attribution 4.0 International (CC BY 4.0)

This is a human-readable summary of (and not a substitute for) the license. Disclaimer.

You are free to:

- Share — copy and redistribute the material in any medium or format
- Adapt — remix, transform, and build upon the material
- for any purpose, even commercially.
- The licensor cannot revoke these freedoms as long as you follow the license terms.

(2) Supplementary contents

Image preprocessing

As shown in Supplementary Figure 4, all US images were preprocessed to remove any image regions outside of the actual scan area and to also detect and split images depicting dual US beams. The liver ultrasound image preprocessing pipeline includes 3 steps: image deidentification, background removal, and dual image detection. In the first step, ultrasound images were converted from the DICOM files to PNG files and cropped slightly, to remove protected health information in the DICOM headers and on the boundary of the images. Then in the second step, the most frequent pixel intensity value less than 50 was calculated to identify the background for each ultrasound image, which was then removed. To further crop out the background, for each image after filtering, the largest connected component (LCC) was calculated, and the image was cropped by the smallest square which can hold the LCC. By the end of this step, only the area within the ultrasound region was kept for each image. It is common to see that two ultrasound beams are combined in one saved image, so in the third step, we detect whether dual beams exist in one file. The image was first filtered by the Canny edge filter[1], so only edges were kept, and then a Hough filter[2] was applied to detect the top 8 line segments in the edge map in order to find the borders of the US beams. The intersections between the lines were then calculated, and if an intersection was found that lied near the horizontal center of the image, the image was considered a dual-beam image. This process can be somewhat noisy. However, in a US study there are typically many images of the same type. Therefore, we perform dual-beam detection for each image individually, then we aggregate the results across all images in a study using majority voting. If the majority of images were found to have a dual beam, we split all images in the study using the average intersection location. In Figure 5, an example of a dual-image file is presented, and the intersection point (yellow point in Step 3.b) was used to split the image. For all evaluation datasets, *i.e.*, HP-U, TM, and HP-T datasets, all images were manually verified as being preprocessed correctly.

Image selection

We are interested in investigating performance and reliability across viewpoints. Thus, for all our datasets, we only included US images from the viewpoints shown in Figure 1, which can be labelled as a. left lobe longitudinal, b. left lobe transverse, c. right lobe intercostal, d. lower right lobe intercostal (depicting liver/kidney contrast), e. subcostal depicting liver/kidney contrast, and f. subcostal with hepatic veins views. For the prospective TM dataset, we aimed to acquire two US images for each of the six viewpoints of Figure 1, except for the right lobe intercostal viewpoint, where we aimed to acquire four. Occasionally conditions did not allow us to collect certain viewpoints. For HP-U, and HP-T, we only included studies that had >=10 images of any of the studied viewpoints. As shown in Figure 1, we categorized these six viewpoints into four *view groups*: left liver lobe (LLL), right liver lobe (RLL), liver/kidney contrast (LKC), and subcostal (SC).

Categorizing the view for each image is not necessary for the developmental datasets (BD-L and BD-V), as the DL algorithm just trains on each image independently without considering the view. However, even though the specific view for each image need not be categorized, ideally the training set only includes images from the four view groups. Because the BD-L and BD-V big-data datasets were extracted directly from the CGMH PACS, their US studies may contain images unsuited for liver steatosis analysis, *e.g.*, images of organs other than the liver, liver viewpoints other than those of Figure 1, poor quality images, and even non-US images. So that these non-qualifying images did not impact the training of our DL model, we applied an additional filtering step to remove as many of these images as possible. Given the scale of data, it was not feasible to perform this filtering manually. Instead, we performed this semi-automatically by training a binary DL classifier, using the PyTorch library with hyper-parameters listed in Supplementary Table 4. We first randomly selected 44 US studies (696 images) from BD-L, and manually identified the corresponding US images as "qualifying", *i.e.*, belonging to one of the liver viewpoints of Figure 1, or "non-qualifying". We also supplemented the positive training examples using the images within the HP-U and TM datasets. We then measured the sensitivity and specificity of the trained binary classifier using a mini-validation dataset of 175 images from BD-L and chose the operating point corresponding to 95% specificity. Note, this filtering process was only used to clean the big-data cohorts and was not used for any of the evaluation datasets.

Training Steatosis Assessment DL Algorithm

Using the images from BD-L, we trained a DL classifier using the 2D US diagnoses extracted from the CGMH records. We opted for the ResNet family of DL classifiers[3] given their ubiquity and performance in both natural imaging and medical imaging tasks. The ResNet family of DL classifiers are 2D convolutional neural networks[4] that use the concept of residual connections to reduce the problem of vanishing gradients and improves learning speed. Based on performance on the BD-V validation dataset, we determined that the ResNet-18 variant performed best. The ResNet-18 has the added virtue of being lightweight, reducing overfitting tendencies compared to alternative variants. The US diagnoses are ordinal labels ranging from 0 to 3 corresponding to None; Mild; Moderate; and Severe[5] steatosis Consequently, the learning task is an ordinal regression problem. We treat each image independently in training and follow the well-known binary decomposition approach to ordinal classification of Frank and Hall[6]. As shown in Figure 2, instead of directly regressing the images to a numeric scale or training a four-class classifier, we decompose the problem into three binary classification tasks: estimating the probability the image represents >= mild, >= moderate, or = severe steatosis. Practically, this means that a three-output classification head is used on top of the ResNet-18 backbone. Under this scheme, the scalar labels for None, Mild, Moderate and Severe would be, respectively, converted to (0,0,0), (1,0,0), (1,1,0), and(1,1,1) multi-label vectors. Training is then conducted using standard cross-entropy loss. After training, a simple transformation produces a *continuous* score[7] for each image that ranges from 0 to 1, with higher scores corresponding to more severe steatosis. For a single image, if the model confidences in the Frank and Hall labels are denoted $\hat{y}_i$, where$i$ indexes whether the label is for >=mild,

>=moderate, or =severe, then the following formulation produces a severity assessment ∈ [0,1]:

$$\hat{p} = \sum_{i} \hat{y}/3.0,$$

where $\hat{p}$ represents the image-wise confidence. As Figure 2 indicates, during inference, after feeding the model individual images to obtain image-wise scores, we then take the mean of image-wise scores across each view group to produce a single score for each view group. Additionally, we can also produce an “All View Groups” score, which is the mean score across all view groups in the study.

The hyper-parameters were selected to optimize our algorithm’s performance on BD-V. We use an ImageNet pretrained network[8], as that performed better than random initialization. The stopping criterion was the model checkpoint that performed best on BD-V, based on a rolling average of five epochs. Including the convolutional neural network architecture and model optimizer, other hyper- parameters that we tuned include initial learning rate, L2 regularization weight, image size and batch size. The details of these hyper-parameters are specified in Supplementary Table 5. We also applied an aggressive augmentation scheme to increase the variability in the image distribution presented to the network. These include additive Gaussian noise, brightness and contrast jittering, and random rotations. Each augmentation was applied on-the-fly to an image with a 50% probability. We also executed an aggressive cropping augmentation. Finally, all images were resampled to 256x256 pixels before being inputted into the deep neural network.

More Details on the Reliability Study

Repeatability Study (Experiment 1)

We used ***TM*** and ***HP-U*** to assess how many images are needed per view group to achieve repeatability. Note, for the ***TM*** dataset we randomly selected only one US study for each patient to avoid sampling the same patient more than once. Typically, to calculate repeatability one simply acquires repeated measurements and performs an accepted repeatability metric. However, in our case each measurement can itself consist of the mean measurement across several image-wise scores. For example, if we are interested in the repeatability when averaging the score across three images to calculate a view group score, then two view-group measurements would require acquiring six images. This is an onerous data collection requirement. Instead of doing this, we simply first calculate the within-subject standard deviation, s, of the image-wise scores. We do this for each US study, which gives us a set of s values across different mean severity measurements. If we are then taking the mean across k images to obtain a view-group score, the resulting within-subject standard deviation is simply $\sqrt{1/k} \times s$. Finally, the within-subject standard deviation of differences between repeated measurements can be estimated as $s_k = \sqrt{2/k} \times s$. The advantage of such an approach is that the within-subject standard deviation can be calculated for any k without requiring the collection of more images.

As advocated by Bland and Altman[9,10], these within subject standard deviations were then graphed across different view-group steatosis scores. Typically the repeatability coefficient (RC) could then be calculated using a mean s_k value across all US studies[10]: the difference between two repeated measurements should be within the RC value for 95% of the US studies. However, because s_k is not uniform (typically greater variability at moderate steatosis levels), a uniform RC is not appropriate[10]. Instead, we modelled the heteroskedasticity by regressing the within-subject standard deviation on mean severity scores[10,11] using a cubic regression. We chose a cubic regression because there is a skew in the distribution of s_k values (see Supplementary Figure 1). We then used the worst-case RC value (max RC) as a summary statistic, with 95% confidence intervals computed using percentile bootstrap (1000 bootstrap samples)[12]. We conducted this for k = {1,2,3,4} and for every view group.

Cross-Scanner Agreement (Experiment 2)

We evaluated agreement across scanners using the ***TM*** dataset, which consists of multiple studies taken on the same day of the same patient. A Bland-Altman analysis[9,10] was performed for assessing cross-scanner agreement. This was simpler than what was done for repeatability, since for a chosen view group we just computed the mean score across all available images in a study. However, based on the repeatability measurements of Experiment 1, we only included view group scores with >=3 images. Thus, for two studies of the same patient across two different scanners, there are only two observations to compare. We calculated the bias and LOAs, where the latter are the limits by which 95% of the disagreements fall under[10]. To deal with the same heteroskedasticity faced by the repeatability experiment, we regressed non-uniform limits of agreement (LOAs)[10] and used the maximum upper LOAs and minimum lower LOAs as summary statistics. 95% confidence intervals were computed using the same bootstrap approach as in Experiment 1.

References

1 Canny J. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1986; **PAMI-8**: 679–698. [DOI: 10.1109/TPAMI.1986.4767851]

2 Duda RO, Hart PE. Use of the Hough transformation to detect lines and curves in pictures. *Commun ACM* 1972; **15**: 11–15. [DOI: 10.1145/361237.361242]

3 He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image

Recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.2016: 770–778.

4 Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge, MA, USA: MIT Press

5 Scatarige J, Scott W, Donovan P, Siegelman S, Sanders R. Fatty infiltration of the liver: ultrasonographic and computed tomographic correlation. *Journal of Ultrasound in Medicine* 1984; **3**: 9–14.

6 Frank E, Hall M. A Simple Approach to Ordinal Classification. In: De Raedt L, Flach P, editors. *Machine Learning: ECML 2001*. Berlin, Heidelberg: Springer, 2001: 145–156.

7 Fürnkranz J, Hüllermeier E, Vanderlooy S. Binary Decomposition Methods for Multipartite Ranking. In: Buntine W, Grobelnik M, Mladenić D, Shawe-Taylor J, editors. *Machine Learning and Knowledge Discovery in Databases*. Berlin, Heidelberg: Springer, 2009: 359–374.

8 Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*.2009: 248–255.

9 Bland JM, Altman D. STATISTICAL METHODS FOR ASSESSING AGREEMENT BETWEEN TWO METHODS OF CLINICAL MEASUREMENT. *The Lancet* 1986; **327**: 307–310. [DOI: 10.1016/S0140-6736(86)90837-8]

10 Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999; **8**: 135–160. [PMID: 10501650 DOI: 10.1177/096228029900800204]

11 Altman DG. Construction of age-related reference centiles using absolute residuals. *Stat Med* 1993; **12**: 917–924. [PMID: 8337548 DOI: 10.1002/sim.4780121003]

12 Efron B, Tibshirani RJ. An Introduction to the Bootstrap. CRC Press

(3) Supplementary Tables

Supplementary Table 1a Demographic Features of Each Cohort

	BD-L	***BD-V***	***TM***	***HP-U***	***HP-T***
Number of Patients	2899	411	246	147	112
Number of Studies	17149	2364	733	147	112
Number of Images	200654	27421	9215	1647	1996
Mean Age at Scan	56.5	56.9	56.6	49.1	50.0
Male, n (%)	1752 (60.4)	248 (60.3)	157 (63.8)	93 (63.3)	66 (58.9)
Female, n (%)	1147 (39.6)	163 (39.7)	89 (36.2)	54 (36.7)	46 (41.1)
NBNC, n (%)	353 (12.2)	51 (12.4)	56 (22.8)	103 (70.1)	63 (56.2)
HBV, n (%)	1050 (36.2)	145 (35.3)	125 (50.8)	35 (23.8)	46 (41.1)
HCV, n (%)	1322 (45.6)	190 (46.2)	65 (26.4)	9 (6.1)	3 (2.7)
Others/Unknown, n (%)	174 (6.0)	25 (6.1)	0 (0.0)	0 (0.0)	0 (0.0)
Steatosis Grade					
US grade 0, n (%)	2529 (87.2)	352 (85.6)	N/A	N/A	N/A
US grade 1, n (%)	314 (10.8)	50 (12.2)	N/A	N/A	N/A
US grade 2, n (%)	50 (1.7)	8 (1.9)	N/A	N/A	N/A

US grade 3, n (%)	6 (0.3)	1 (0.3)	N/A	N/A	N/A

Supplementary Table 1b. Additional Clinicopathologic Features of *HP-U* and *HP-T*

	HP-U			*HP-T*		
	NBNC	***HBV***	***HCV***	***NBNC***	***HBV***	***HCV***
Number of Patients	103	35	9	63	46	3
Mean Age at Scan	47.5	51.8	56.6	48.8	52.2	43.2
Male, n (%)	71 (68.9)	18 (51.4)	4 (44.4)	28 (44.4)	36 (78.3)	2 (66.7)
Female, n (%)	32 (31.1)	17 (48.6)	5 (55.6)	35 (55.6)	10 (21.7)	1 (33.3)
Mean BMI	27.5	25.4	27.1	25.7	25.8	27.5
Mean AST U/L	64.9	64.1	58.0	115.4	87.0	71.7
Mean ALT U/L	110.2	92.7	76.4	213.4	151.8	128.0
Mean PLT 10^3/mm^3	246.9	201.4	207.3	248.8	186.2	179.7
Steatosis Grade						
grade 0, n (%)	10 (9.7)	11 (31.4)	3 (33.3)	22 (34.9)	21 (45.7)	1 (33.3)
grade 1, n (%)	18 (17.5)	14 (40.0)	4 (44.4)	13 (20.6)	15 (32.6)	1 (33.3)

grade 2, n (%)	31 (30.1)	4 (11.4)	0 (0.0)	6 (9.6)	8 (17.4)	0 (0.0)
grade 3, n (%)	44 (42.7)	6 (17.1)	2 (22.2)	22 (34.9)	2 (4.3)	1 (33.3)

Abbreviation: BD-L (big data learning group); BD-V (Big data validation group); HP-U (Histopathology Unblinded Test Group); TM (trimachine group); HP-T (Histopathology blinded Test Group); ***AST*** (aspartate aminotransferase); ***ALT*** (alanine aminotransferase); ***HBV*** (hepatitis B); ***HCV*** (hepatitis C); ***NBNB*** (non-HBV, non-HCV and excluded other liver diseases, E.g. alcoholic, autoimmune, etc); ***PLT*** (platelet)

Supplementary Table 2 Scanner brands, number of studies, and time ranges (if information is available in de-identified DICOM headers)

	BD-L, BD-V		***HP-U***		***HP-T***	
Scanner Brand	**Studies**	**Time Range**	**Studies**	**Time Range**	**Studies**	**Time Range**
ATL: HDI 5000	2865	1/3/2011 – 4/13/2015	--	--	16	3/18/2011 – 9/2/2014
GE Healthcare: LOGIQ E9	2	11/11/2014– 11/14/2014	--	--	--	--
GE Healthcare: LOGIQ S8	19	8/29/2012 – 9/5/2012	--	--	--	--
Aloka Medical,Ltd.: SSD 5500	4273	1/3/2011 – 10/17/201	--	--	2	5/2/2012 –

		4				3/19/2013
Hitachi Medical Corporation: HI VISION Avius	16	8/27/2012 – 8/31/2012	--	--	--	--
Hitachi Medical Corporation: HI VISION Preirus	20	7/18/2012 – 9/25/2018	--	--	--	--
Philips Medical Systems: EPIQ 7G	2	11/21/2014 – 7/24/2018	--	--	--	--
Philips Medical Systems: HD15	4	11/17/2014 – 11/20/2014	--	--	--	--
Philips Medical Systems: iU22	8827	1/3/2011 – 9/28/2018	7	11/27/2014 – 6/19/2019	12	9/9/2011 – 9/4/2020
Siemens: S2000	193	1/6/2011 – 9/28/2018	117	7/12/2012 – 9/26/2019	78	8/14/2012 – 1/29/2021
SuperSonic Imagine SA: Aixplorer	72	5/14/2012 – 7/24/2012	--	--	--	--
Toshiba MEC US: TUS-A300	3145	11/20/2014 – 9/28/2018	23	7/14/2015 – 7/2/2019	4	6/9/2020 – 12/16/20

						20
Toshiba MEC: Xario	26	8/24/2012 – 8/31/2012	--	--	--	--
Unknown *	49	1/4/2011- 9/28/2018	--	--	--	--

* Unknown: Toshiba SSA-370A or Toshiba SSA-700A, the exact model used was not recorded.

Supplementary Table 3 Performance Statistics for All Experiments Described in This Article. All experiments evaluated the same model, trained on the *BD-L* dataset.

ID	Experiment description	Result statistics
1	Estimate repeatability across view groups and different numbers of images per view group using two ***TM*** and ***HP-U*** cohorts	Max repeatability coefficient (RC), RC graphs
2	Estimate consistency across scanners and view groups using ***TM*** cohort	Bias, upper and lower limits of agreement, Bland-Altman graphs, % Agreement
3a	Estimate diagnostic performance across views using histology proven cohort ***HP-U***	AUCROC (fatty % >=5%; >=33%; and >=66%), ROC Curves, Accuracy
3b	Compare diagnostic performance of DL model to FibroScan using studies with associated FibroScan scores from the ***HP-U*** cohort	AUCROC (fatty % >=5%; >=33%; and >=66%), ROC curves, Accuracy
4a	Estimate diagnostic performance across views using histology proven cohort ***HP-T***	AUCROC (fatty % >=5%; >=33%; and >=66%), ROC curves, Accuracy

Abbreviation: BD-L (big data learning group); BD-V (Big data validation group); HP-U (Histopathology Unblinded Test Group); TM (trimachine group); HP-T (Histopathology blinded Test Group).

Supplementary Table 4 Description and values of all hyperparameters and properties of the image quality binary classifier. This deep learning (DL) model was used to automatically filter out non-qualifying images from the *BD-L* and *BD-V* dataset.

Hyperparameter	**Description**	**Value**
Network architecture	Deep neural network layout	ResNet-18
Image size	Size of image as the network input (in pixel)	256×256
Maximum Epochs	Maximum number each image is shown to the network during training	100
Graphics Processing Unit	Graphics processing unit hardware	NVIDIA Titan V
Initial Learning Rate	Network learning rate during training	0.0001
L2 Regularization	Weight decay (L2 penalty)	0.0005
Batch Size	Number of images processed in parallel	16
Solver	Optimizer to update weights and biases	SGD

Abbreviation: BD-L (big data learning group); BD-V (Big data validation group);

Supplementary Table 5 Description and values of all hyperparameters and properties of the deep learning workflow for steatosis severity assessment

Hyperparameter	Description	Value
Network architecture	Deep neural network layout	ResNet-18
Image size	Size of image as the network input (in pixel)	256×256
Maximum Epochs	Maximum number each image is shown to the network during training	120
Optimization Algorithm	Stochastic gradient descent	
Graphics Processing Unit	Graphics processing unit hardware	NVIDIA Titan V
Initial Learning Rate	Network learning rate during training	0.0005
L2 Regularization	Weight decay (L2 penalty)	0.0001
Batch Size	Number of images processed in parallel	32
Solver	Optimizer to update weights and biases	SGD
Gaussian Noise	Standard deviation upper bound	0.01
Color Jittering	Brightness/Contrast change upper bound	0.2
Rotation	Affine transformation rotation upper bound	10 Degrees
Scaling	Affine transformation ratio bound	[0.9, 1,1]
Augmentation Probability	The possibility to apply each augmentation technique to a single image	50%

Supplementary Table 6 The max repeatability coefficient (RC) is tabulated across different view groups for the *TM* and *HP-U* datasets. Parentheses enclose bootstrapped 95% confidence intervals

View	1 Image	2 Images	3 Images	4 images
LLL	0.46 (0.42, 0.51)	0.33 (0.30, 0.36)	0.27 (0.24, 0.29)	0.23 (0.21, 0.26)
RLL	0.37 (0.34, 0.40)	0.26 (0.24, 0.28)	0.21 (0.20, 0.23)	0.18 (0.17, 0.20)
LKC	0.53 (0.47, 0.58)	0.37 (0.33, 0.41)	0.30 (0.27, 0.34)	0.26 (0.24, 0.29)
SC	0.46 (0.42, 0.50)	0.32 (0.30, 0.36)	0.27 (0.24, 0.29)	0.23 (0.21, 0.26)

Abbreviation: LLL (left liver lobe); RLL (right liver lobe); LKC (liver/kidney contrast); SC (subcostal); HP-U (Histopathology Unblinded Test Group); TM (Trimachine Group)

Supplementary Table 7 Literature review of works applying deep learning techniques for assessing hepatic steatosis using 2D US images. To be included, the works must be using deep learning models and only the deep learning results are highlighted here.

Reference	Byra *et al.*[22]	Chen *et al.*[20]	Cao *et al.*[21]	Han *et al.*[19]	Gummadi *et al.*[17]	Byra *et al.*[14]	Biswas *et al.*[15]	Ours
Reference in Main Body	22	20	21	19	17	14	15	
Publication Year	2021	2020	2020	2020	2020	2018	2018	
Evaluation Studies (case/control)	135‡	41	240 (138/106)	204 (140/64)	Unclear patient or study-wise split	55 (38/17)‡	63 (36/27) ‡	147+112
Training Studies	Leave-one-out cross validation	164	?		Unclear patient or study-wise split	Leave-one-out cross validation	Ten-fold cross validation, unclear if split across patients	19,513
Etiology	NBNC	NBNC	NAFLD	NAFLD	NAFLD/NA	Severely obese	NAFLD	HBV/HCV/

					SH			NBNC
US Scanner	Siemens S3000	Terason M3000	Mindray Resona 7	Siemens S2000	6 models	GE	US Scanner	13 models (training); 5 models (evaluation)
Image type	Grayscale	Grayscale	Grayscale	RF data	Grayscale	Grayscale	Grayscale	Grayscale
Evaluation Images/Case	4	5	?	10	~5	10	?	>= 10
Total Evaluation Images	540	205	?	2040	78	550‡	?	1647+1996
Total Training Images	Leave-one-out cross validation	820	852		725	Leave-one-out cross validation	Ten-fold cross validation, unclear if split across patients	228075
Area of interest	Cropped 224x224 pixel	Manual 3.5*3.5 cm	Manual 224*224 pixels ROI	256 RF signals	Manual Crop	434×636 pixel Image	Auto-cropped 128*128 Image	Auto-cropped 256*256 pixel

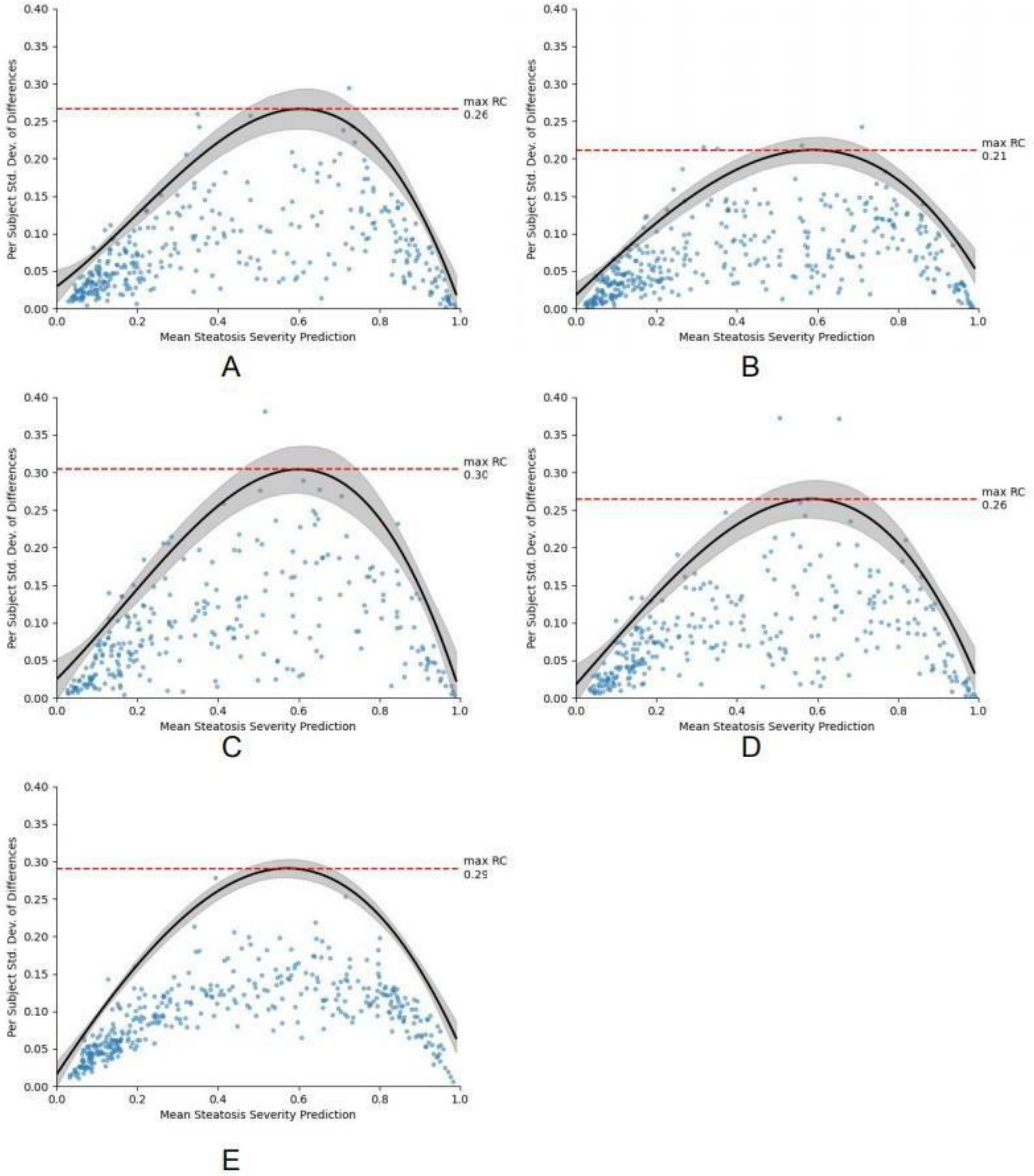
	image	ROI						Image
Gold Standard	MRI PDFF	Histology	2D-US	MRI PDFF	Histology/MRI PDFF/2D-US & Patient History	Histology	Normal control	Histology
View	RLL and LKC	RLL	All Views	RLL	Unspecified	RLL/Kidney	RLL	All Views
Machine Learning Model	ResNet-50	VGG-16	Custom CNN	CNN	Unspecified CNN	Pretrained Inception-ResNetv2 CNN+SVM	SVM/ELM/CNN	ResNet-18
Results AUCROC (mild)		0.71†	0.933*					0.85-0.95
AUCROC (moderate)		0.75†	0.692*					

AUCROC (severe)		0.88†	0.958*					0.91-0.92
AUCROC Binary	0.86-0.91		0.98	0.98	89% Se. & 95% Sp.	0.98	1.0	0.87-0.93

* Based on 2D-US diagnosis; † Separate data splits for each cut-off; ‡ Cross validated

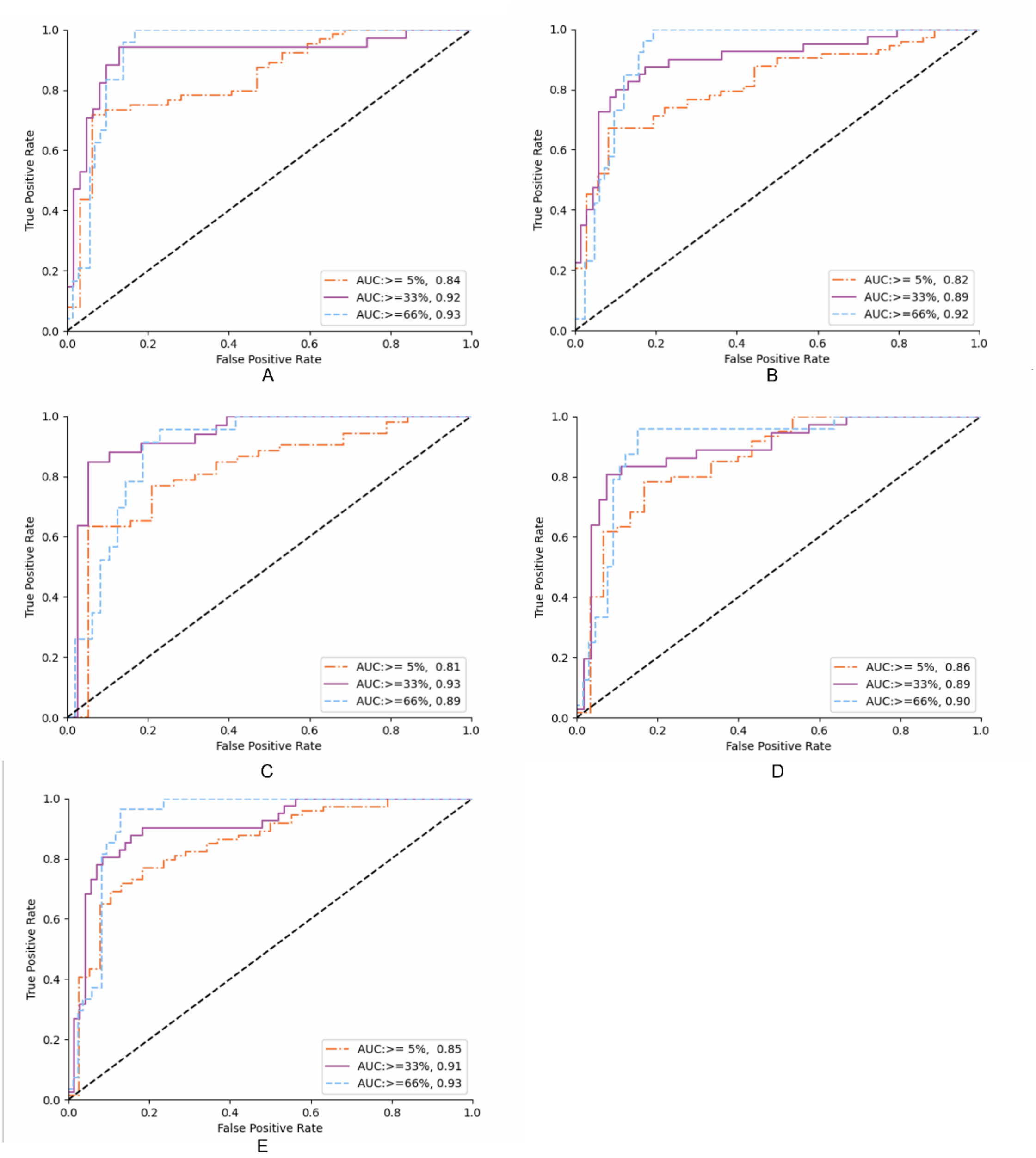
Abbreviations: AUCROC: area under the curve of receiver operating characteristic; CNN: convolutional neural network; HBV: hepatitis virus B; HCV: hepatitis virus C; MRI PDFF magnetic resonance imaging derived proton density fat fraction; NBNC: non-hepatitis B/non-hepatitis C; RF: Radiofrequency; RLL: right liver lobe; LKC: liver kidney contrast; SVM: support vector machine

(4) Supplementary Figures



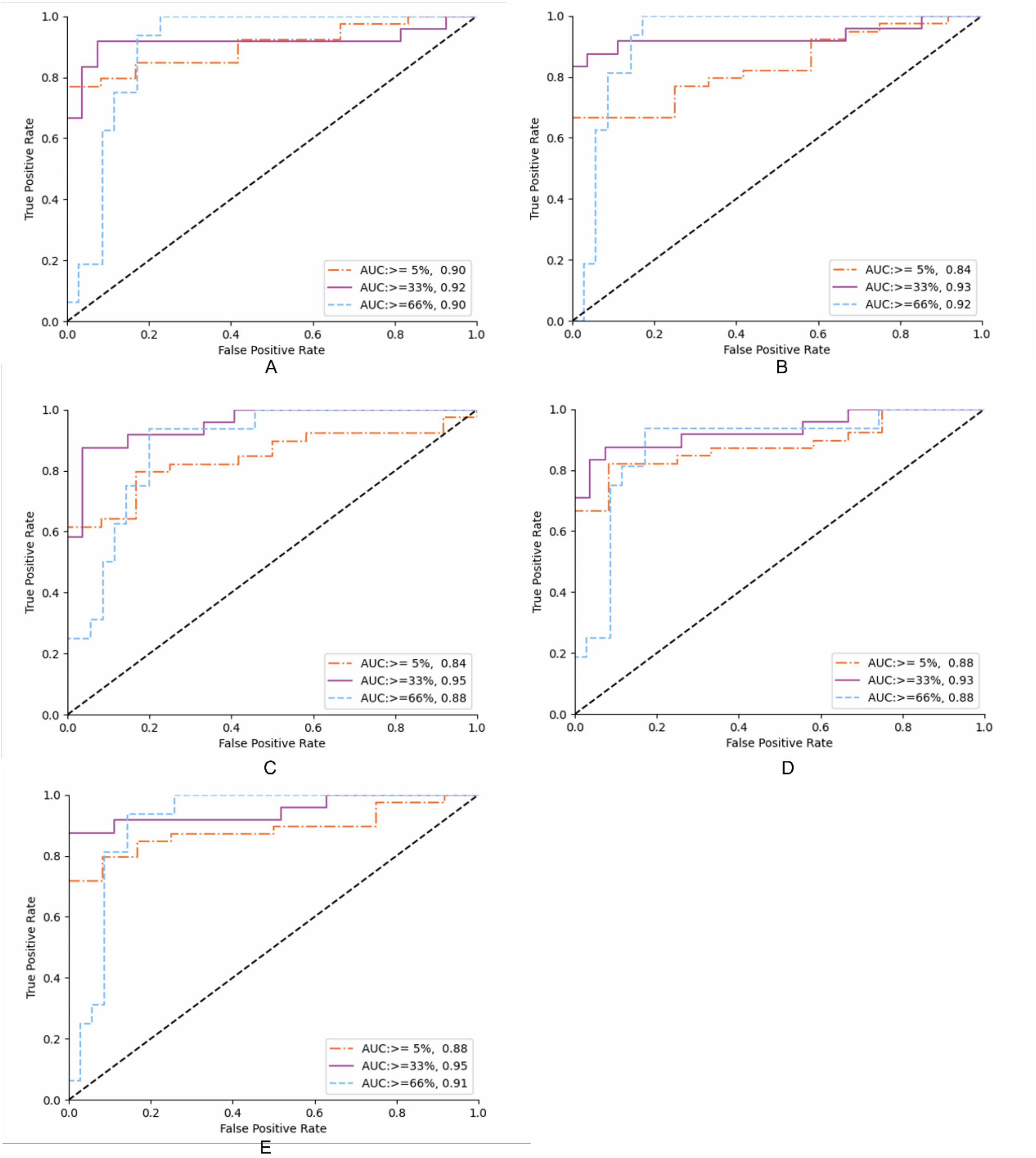
Supplementary Figure 1 Repeatability coefficient (RC) plot across different 2D US viewpoints. (A) to (E) represents LLL, RLL, LKC, SC, and "All View Groups" respectively. Repeatability is measured when taking the mean score across three images per view group. "All View Groups" represents the score after taking the mean each resulting viewpoint score to create one score for each study.

Abbreviation: LLL (left liver lobe); RLL (right liver lobe); LKC (liver/kidney contrast); SC (subcostal);



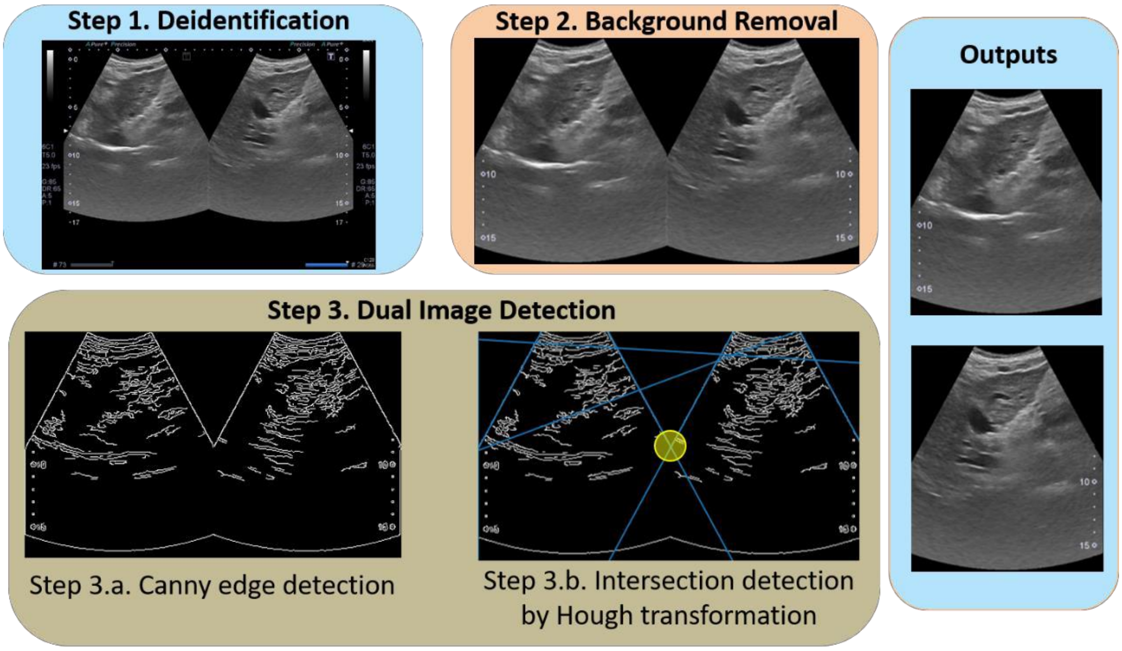
Supplementary Figure 2 ROC analysis of ***HP-T*** (Individual view group setting). (A) to (E) shows ROC curves of the deep learning (DL) model for diagnosing hepatic steatosis grades on HP-T with LLL, RLL, LKC, SC, and "All View Groups", respectively.

Abbreviation: HP-T (Histopathology blinded Test Group); LLL (left liver lobe); RLL (right liver lobe); LKC (liver/kidney contrast); SC (subcostal);



Supplementary Figure 3 ROC analysis of ***HP-T*** (Complete view group setting). (A) to (E) shows ROC curves of the deep learning (DL) model for diagnosing hepatic steatosis grades on HP-T with LLL, RLL, LKC, SC, and "All View Groups", respectively.

Abbreviation: HP-T (Histopathology blinded Test Group); LLL (left liver lobe); RLL (right liver lobe); LKC (liver/kidney contrast); SC (subcostal);


Step 1. Deidentification
Step 2. Background Removal
Outputs
Step 3. Dual Image Detection
Step 3.a. Canny edge detection
Step 3.b. Intersection detection by Hough transformation

Supplementary Figure 4 Liver ultrasound image preprocessing pipeline includes 3 steps: image deidentification, background removal, and dual image detection. In "Step 3.b", the figure is showing the top 8 lines detected by the Hough transform (in blue, two lines are along the boundaries and might not be seen), and the detected intersection point (in yellow).