

● Original Contribution

COMPARATIVE STUDY OF RAW ULTRASOUND DATA REPRESENTATIONS IN DEEP LEARNING TO CLASSIFY HEPATIC STEATOSIS

SERGIO J. SANABRIA,^{*,†} AMIR M. PIRMOAZEN,^{*} JEREMY DAHL,^{*} AYA KAMAYA,^{*} and AHMED EL KAFFAS^{*}

^{*} Department of Radiology, Stanford University, Stanford, California, USA; and [†] Deusto Institute of Technology, University of Deusto/Ikerbasque, Basque Foundation for Science, Bilbao, Spain

(Received 14 December 2021; revised 23 May 2022; in final form 24 May 2022)

Abstract—Adiposity accumulation in the liver is an early-stage indicator of non-alcoholic fatty liver disease. Analysis of ultrasound (US) backscatter echoes from liver parenchyma with deep learning (DL) may offer an affordable alternative for hepatic steatosis staging. The aim of this work was to compare DL classification scores for liver steatosis using different data representations constructed from raw US data. Steatosis in $N = 31$ patients with confirmed or suspected non-alcoholic fatty liver disease was stratified based on fat-fraction cutoff values using magnetic resonance imaging as a reference standard. US radiofrequency (RF) frames (raw data) and clinical B-mode images were acquired. Intermediate image formation stages were modeled from RF data. Power spectrum representations and phase representations were also calculated. Co-registered patches were used to independently train 1-, 2- and 3-D convolutional neural networks (CNNs), and classifications scores were compared with cross-validation. There were 67,800 patches available for 2-D/3-D classification and 1,830,600 patches for 1-D classification. The results were also compared with radiologist B-mode annotations and quantitative ultrasound (QUS) metrics. Patch classification scores (area under the receiver operating characteristic curve [AUROC]) revealed significant reductions along successive stages of the image formation process ($p < 0.001$). Patient AUROCs were 0.994 for RF data and 0.938 for clinical B-mode images. For all image formation stages, 2-D CNNs revealed higher patch and patient AUROCs than 1-D CNNs. CNNs trained with power spectrum representations converged faster than those trained with RF data. Phase information, which is usually discarded in the image formation process, provided a patient AUROC of 0.988. DL models trained with RF and power spectrum data (AUROC = 0.998) provided higher scores than conventional QUS metrics and multiparametric combinations thereof (AUROC = 0.986). Radiologist annotations indicated lower hepatic steatosis classification accuracies (Acc = 0.914) with respect to magnetic resonance imaging proton density fat fraction that DL models (Acc = 0.989). Access to raw ultrasound data combined with artificial intelligence techniques may offer superior opportunities for quantitative tissue diagnostics than conventional sonographic images. (E-mail: sergio.sanabria@deusto.es) © 2022 World Federation for Ultrasound in Medicine & Biology. All rights reserved.

Key Words: Liver steatosis, Quantitative ultrasound, Convolutional neural network, Raw radiofrequency data, Power spectrum, Phase information.

INTRODUCTION

Non-alcoholic fatty liver disease (NAFLD) is the most common liver disorder, affecting one in four adults (Marjot et al. 2020). NAFLD is characterized by the accumulation of adipose cells in the liver of causes other than alcohol. Although strongly associated with obesity, NAFLD also affects around 11% of the lean population and is generally asymptomatic. Fat accumulation (steatosis) can progress to liver inflammation (steatohepatitis), which can lead to severe complications such as cirrhosis,

liver carcinoma, liver failure and cardiovascular disease. The estimated direct health care burden of NAFLD is estimated at around 100 billion annually in the United States. Low-cost high-yield methods for lifelong NAFLD screening are therefore a clinical priority (Younossi et al. 2016; Marjot et al. 2020; Young et al. 2020).

The current gold standard for NAFLD diagnosis and classification is liver biopsy, which is limited by sampling error and potential patient complications. With the development of highly accurate quantitative sequences, magnetic resonance imaging (MRI) has been postulated as a non-invasive alternative to biopsy for macroscopic assessment of liver adipositas (Tang et al.

Address correspondence to: Sergio J. Sanabria, Deusto Institute of Technology, University of Deusto, Avenida de las Universidades 24, 48007 Bilbao, Spain. E-mail: sergio.sanabria@deusto.es

2013; Wong *et al.* 2018). Yet MRI is impractical for general screening because of its high costs and limited availability. Ultrasound (US) is widely used for liver examination because of its wide accessibility, non-invasive nature and low cost. Medical probes emit sound waves at megahertz frequencies—three orders of magnitude over the human audible range—and record echoes reflected from tissue as a series of time-domain signals (also known as radiofrequency [RF] data). A series of signal processing steps are used to create a human-interpretable image (B-mode or echography) (Hoskins 2003). Fatty liver is recognized by experienced sonographers in B-mode images with distinctive qualitative features, for example, an increase in echogenicity (“bright liver”) of liver parenchyma and decreased visualization of vasculature and diaphragm (Hernaiz *et al.* 2011). The broader use of ultrasound is currently limited by a steep learning curve and lack of quantification (Strauss *et al.* 2007; Ferraioli and Monteiro 2019).

Significant efforts are ongoing to develop and validate quantitative ultrasound (QUS) metrics, which are displayed as adjunct modalities to B-mode images for objective liver tissue characterization (Imbault *et al.* 2017; Ferraioli *et al.* 2019; Li *et al.* 2019; Chen *et al.* 2020; Basavarajappa *et al.* 2021; Gao *et al.* 2022; Pirmoazen *et al.* 2022; Wen *et al.* 2022). Elastography measures tissue deformation in response to mechanical excitation as a quantitative surrogate of tissue stiffness, and is well established for fibrosis staging (Barr *et al.* 2015). Analysis of shear wave attenuation shows promise for steatosis quantification (Shi *et al.* 2014; Ferraioli *et al.* 2019; Ormachea and Parker 2020). Viscosity, for instance, estimated from tissue shear wave dispersion (Nightingale *et al.* 2015; Bhatt *et al.* 2019), may provide an indication of liver inflammation, with reported more moderate correlates than stiffness to liver fibrosis (Sugimoto *et al.* 2020; Popa *et al.* 2021) and inconsistent correlates with liver fat fraction (Pirmoazen *et al.* 2022) and steatosis (Deffieux *et al.* 2015; Pearson *et al.* 2022).

Alternative QUS parameters have recently been proposed for liver steatosis quantification based on the parametrization of tissue echogenicity from raw RF data. These include normalized brightness metrics such as the hepatorenal index; attenuation parameters, which model the damping of ultrasound signal with depth; statistical models of tissue reflectivity texture such as the Nakagami parameter or entropy metrics; and quantitative spectroscopy metrics, among others (Lee *et al.* 2020; Pirmoazen *et al.* 2020, 2022; Jeon *et al.* 2021). Multiparametric combinations of these metrics have indicated improved diagnostic outcomes for steatosis and NAFLD (Han *et al.* 2019; Sanabria *et al.* 2020a; Sugimoto *et al.* 2020; Gao *et al.* 2022; Wen *et al.* 2022).

Deep learning (DL) methods based on multilayered neural networks efficiently learn features directly from annotated data, typically images or raw signals, and can classify data without feature extraction. By use of radiologist-labeled annotations, DL has achieved human-comparable performance in classification of numerous pathologies in radiological images, with promise for companion diagnostics (Saba *et al.* 2019; Zhou *et al.* 2019). Self-supervised models based on annotations obtained from quantitative reference standards, such as MRI and biopsy, exhibit the potential to augment quantitative capacities of ultrasound systems beyond what is today possible with human annotation. To date, most of the available DL literature for US is based on B-mode images, which are widely available in the clinical setting and easily annotated by radiologists (Liu *et al.* 2019; Zhou *et al.* 2019; Yang *et al.* 2020; Cha *et al.* 2021; Colantonio *et al.* 2021; Byra *et al.* 2022). A key aspect in developing DL techniques in ultrasound tissue characterization is to find adequate ultrasound data or adequate representations to train the neural network. In analogy to model-based QUS metrics, it is reasonable to believe that DL based on RF time-series data can provide additional learning information than the final post-processed B-mode images. Recent DL work has indicated the potential of RF data for hepatic steatosis staging (Chen *et al.* 2020; Han *et al.* 2020; Sanabria *et al.* 2020b).

In this study, we compare different US data representations with respect to their ability to classify steatotic livers. We capture the ultrasound data at different stages of the B-mode image formation process (RF, in-phase and quadrature [IQ], envelope, log compression, denoising) and extract co-registered patches for each representation, which we use to train convolutional neural network classifiers of different complexities. We compare the classification outcomes with patches extracted from clinical-grade B-mode images extracted from the same US system. We also explore classification based on spectral representations of the data, which are commonly employed in the estimation of QUS parameters. Finally, we investigate classification based on phase representations of ultrasound data, which are less conventional and usually discarded during the B-mode image formation process.

METHODS

Clinical data acquisition and annotation

In this prospective, HIPAA-compliant study, ultrasound RF data and B-mode images were acquired for 31 patients with confirmed or suspected NAFLD. The study was approved by the Stanford Institutional Review Board (IRB), and written informed consent was obtained

prior to the patient's participation in the study. The patient population was screened from a larger 211-patient population, which underwent clinical liver MRI at our institution for 11 mo. The same patient population had been examined in a previous study with serology and QUS techniques (Pirmoazen et al. 2022), where patient demographics were detailed. Inclusion criteria were confirmed or suspected NAFLD, including at least one independent risk factor. Exclusion criteria were moderate to heavy alcohol consumption, history of other chronic liver diseases and suspected drug-related hepatotoxicity.

Liver MRI was carried out with a 3.0-T scanner (Discovery MR750, GE Medical Systems, Waukesha, WI, USA) within a maximum of 2 wk from the US examination. To provide an extensive evaluation of liver parenchyma, a body radiologist (10 y of experience) defined regions of interest in each of the liver segments 5 to 8, excluding potentially artifactual contributions, major blood vessels, bile ducts, cysts or any other incidental findings. Four MRI proton density fat fraction (PDFF) measurements for each segment of the right hepatic lobe were averaged and used as reference for hepatic fat quantification. Liver steatosis staging (S0 = none, S1 = mild, S2 = moderate and S3 = severe) was performed with MRI-PDFF cutoff values at 6.4%, 17.4% and 22.1%, respectively, for stages S1, S2 and S3 (Tang et al. 2015) (Table 1). These cutoff values were used as a reference to stratify patients.

Clinical B-mode US data acquisition was carried out with a high-end Philips EPIQ 7 US system (Philips Healthcare, Bothell, WA, USA) and a C5-1 curvilinear transducer (bandwidth: 1-5 MHz, aperture: 55.5 mm, field of view: 111°, 160 crystal elements). Simultaneously recorded, beamformed, ultrasound RF data were made available by the manufacturer for each acquired US view. A total of 9–12 US views were acquired per patient including both intercostal views with liver and kidney and subcostal sagittal views. RF data were applied to a simplified image formation process (Fig. 1a) to generate intermediate “images” from the various US

data representations. The liver region was manually segmented in clinical B-mode views by two readers with at least 10 y of experience in reading ultrasound images (Fig. 1b). For average registration to the MRI-PDFF reference, the liver views in ultrasound also covered segments 5 to 8 of the right liver lobe. The readers also performed qualitative liver steatosis grading of the B-mode images on the S0–S3 scale. The same annotations were used for the corresponding spatial locations across all US data representations.

Ultrasound data representations through image formation process

The investigated ultrasound data representations are listed in Table 2. The ultrasound RF data were acquired immediately after the beamformer stage (Fig. 1a.1) of the ultrasound system. The RF data were digitized with a high sampling frequency (20.7 MHz, equivalent to a 37- μ m spatial step). A total of 250 RF lines were acquired for each ultrasound frame (Fig. 1c.1).

After beamforming, a series of signal processing steps was carried out to emulate the conversion of RF data into a visually representable grayscale image. The calculations are detailed in Appendix S1 (online only). In the first step (Fig. 1a.2), RF data were demodulated to shift the information bandwidth from the transducer peak frequency (here 1.8 MHz) to baseband. This process removes fast oscillations from the RF data and allows decimation of the depth axis by a factor of 8 (300- μ m spatial step) without information loss. The resulting demodulated data representation (IQ data) is complex valued (Fig. 1c.2). Next, phase information is eliminated, and an amplitude image (also known as envelope or ENV) is calculated as the absolute value of the IQ data (Fig. 1c.3). Because of acoustic damping, the amplitude of the ultrasound signal is reduced exponentially with depth (Hoskins 2003). This is partially accounted for by time-gain compensation performed by the ultrasound system before RF data recording, thereby using nominal attenuation values in soft tissue (~ 0.5 dB/cm/MHz). Moreover, the range of echo amplitudes detected from different targets is wide (Hoskins 2003). A logarithmic compression of ENV data (LOG) is further applied to normalize the dynamic range of the data (see Appendix S1, online only) so that lower amplitude signals (the majority of ultrasound reflections in the liver) are visualized alongside the high-amplitude signal, and a uniform grayscale representation of tissue reflectivity is achieved for all depths (Fig. 1c.4).

In a final step, a combination of denoising algorithms with edge preservation are applied to improve visualization of tissue structures (DEN in Fig. 1c.5). The DEN image is a pseudo-B-mode representation, which is

Table 1. Categorization of study patient population in steatosis stages according to MRI-PDFF

	Number of patients	MRI-PDFF (%)		
		Cutoff	Mean	Range
S0 (none)	13	<6.4	3.4	1.3–4.9
S1 (mild)	7	[6.4, 17.4]	12.9	8.6–14.5
S2 (moderate)	3	[17.4, 22.1]	18.9	17.6–19.9
S3 (severe)	8	≥ 22.1	29.5	23.1–42.9
All	31		13.8	1.3–42.9

MRI-PDFF = magnetic resonance imaging proton density fat fraction.

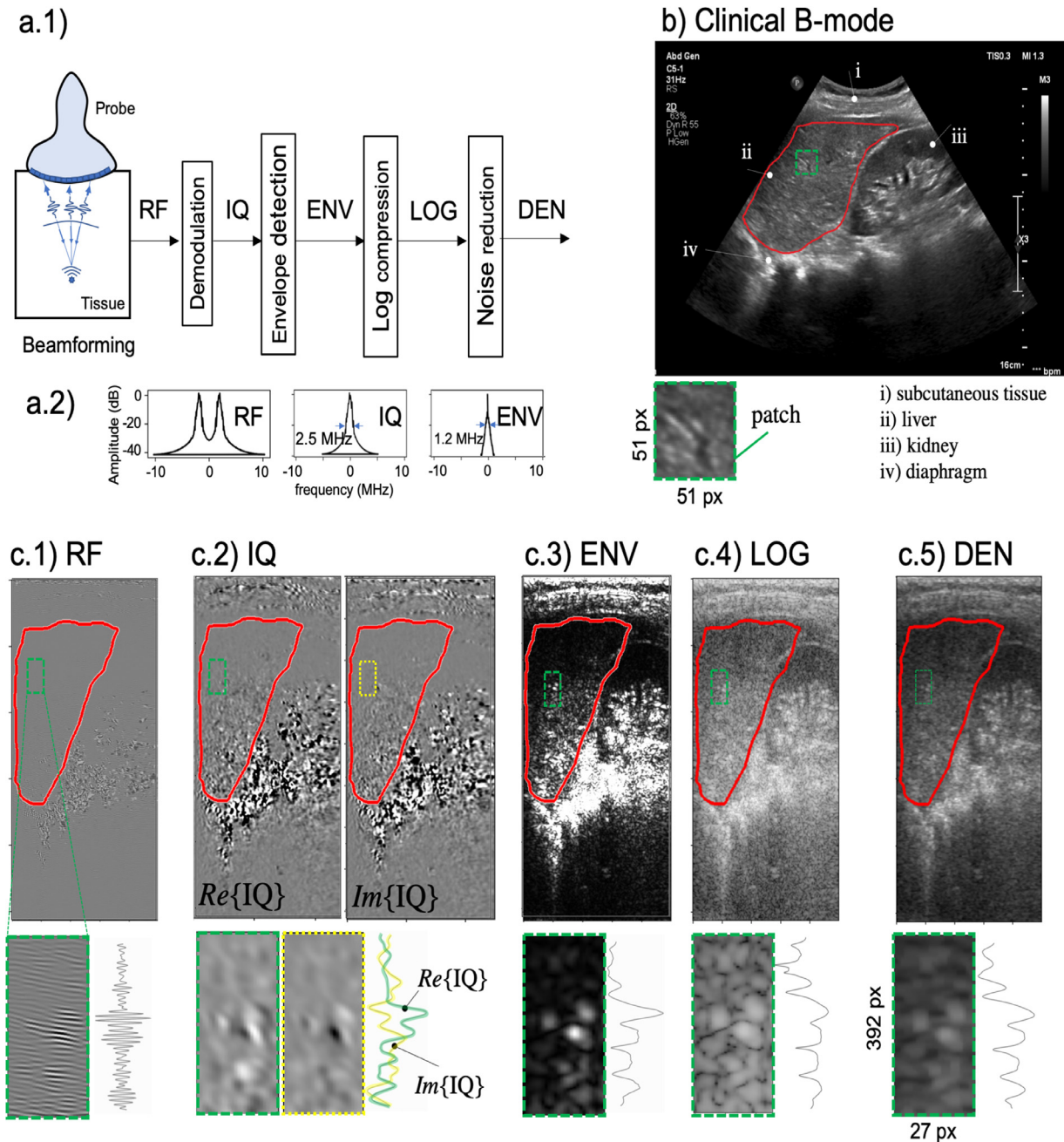


Fig. 1. Representations of ultrasound data during stages of the image formation pipeline. (a.1) Image formation steps from beamformed RF data to denoised brightness image (DEN). (a.2) Spectrum of RF, baseband IQ and ENV signals. (b) Clinical B-mode image. (c) Output images at intermediate stage together with exemplary co-registered image patches. Data are shown for imaging pipeline for healthy 66-y-old female patient with MRI-PDFF = 2.45% (S0 stage). ENV = envelope; IQ = in phase and quadrature; MRI-PDFF = magnetic resonance imaging proton density fat fraction; RF = radiofrequency.

visually comparable to the clinical B-mode image of Figure 1b. In comparison, the clinical B-mode image incorporates more sophisticated noise reduction techniques to improve the delineation of tissue structures. This can include, for instance, spatial compounding techniques combining envelope images obtained from different

steering angles or signal decompositions to separately reduce noise at different frequencies (Mohd Sagheer and George 2020). Generally, noise reduction techniques are the proprietary information of manufacturers and are not disclosed. The clinical B-mode images also incorporated scan conversion into a Cartesian coordinate system with

Table 2. Ultrasound data representations evaluated in the study

Data representation	Description
RF	Beamformed radiofrequency US data
IQ	Complex-valued base-band representation
ENV	Envelope signal, absolute value of IQ
LOG	Logarithmic-compressed ENV
DEN	Noise-filtered LOG
Clinical B-mode	Clinical-grade US images
PS	Power spectrogram of RF
PS_LOG	Logarithmic-compressed PS
PS_LOG_NORM	PS_LOG normalized separately for each frequency
PS_LOG_NORM_DEN	Noise-filtered PS_LOG_NORM
CARRIER	Carrier signal obtained by dividing RF by ENV
ANGLE	Angle of IQ complex phasor
PHASE	Unwrapped phase of RF spectrogram
NN model	Description
_1D	One-dimensional Convolutional NN (Fig. 4a)
_2D	Two-dimensional Convolutional NN (Fig. 4b)
_3D	Three-dimensional Convolutional NN (Fig. 4c)

respect to the probe, here a curvilinear transducer, which was not applied to DEN images.

Power spectrum representation of ultrasound data for QUS

Alternative representations of RF data have been proposed for tissue quantification. QUS metrics are calculated from spectral representations of the RF data, which allow analysis of tissue properties at different frequencies (Mamou and Oelze 2013; Pirmoazen et al. 2020). Here, the power spectrum (PS) was calculated in a non-parametric manner with a digital Fourier transform. Using a moving temporal window along the RF line (short-time Fourier transform [STFT]) allowed us to capture local spectral properties as a function of tissue depth. STFT is a popular PS representation in deep learning models applied to audio signals (Adavanne et al. 2019; Gogate et al. 2020; Arias-Vergara et al. 2021). To minimize spectral leakage and preserve resolution, a Hanning window with 50% overlap was used in the STFT. The length of the temporal window of $2.5 \mu\text{s}$ (52 px) provided an effective depth resolution of 2 mm, and is associated with the effective frequency resolution $1/2.5 \mu\text{s} = 400 \text{ kHz}$ (Rao and Swamy 2018). Power spectral density frequency components were calculated for each pixel between 0 and 8.7 MHz with a frequency step of 320 kHz (Fig. 2b). This led to a 3-D PS representation (axial, RF line index, frequency).

A selection of post-processed variations of PS were evaluated for DL classification (Fig. 2a). This included logarithmic compression of the full spectrogram

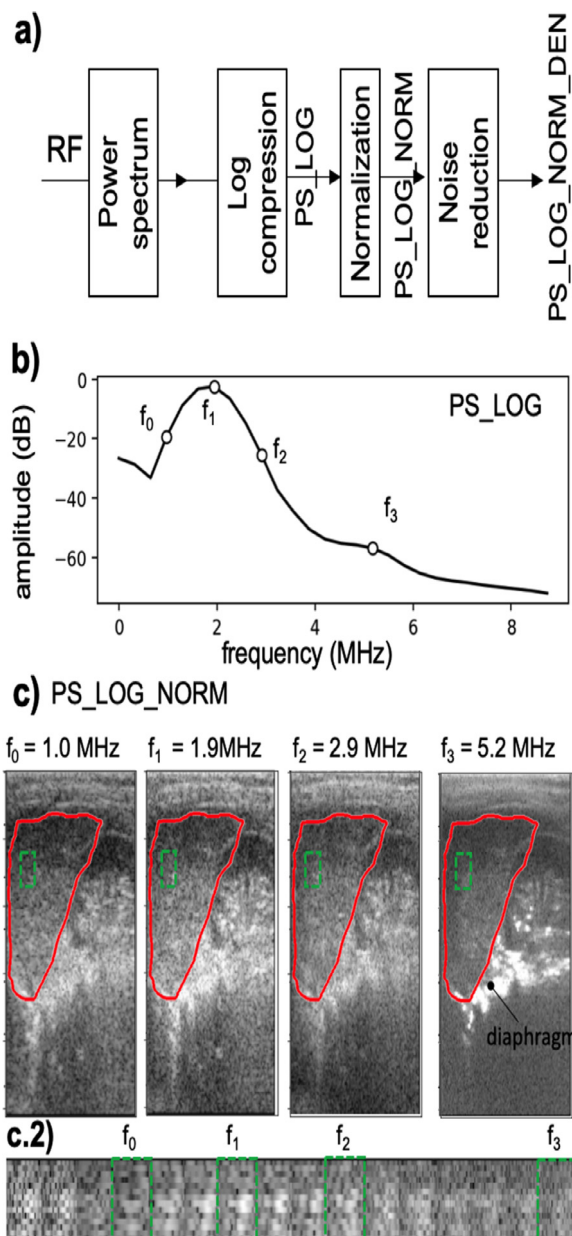


Fig. 2. Power spectrum representation of ultrasound data for same participant as in Figure 1. (a) Flow diagram for evaluated PS representations. (b) Sample log-compressed power spectrogram LOG_PS at liver parenchyma region. (c) Image representation normalized in same log amplitude scale for different frequencies LOG_PS_NORM. (c.2) Exemplary stacked patches at same coordinates for successive frequencies. PS = power spectrum; RF = radiofrequency.

(PS_LOG), individual normalization and image compression of the US images for each frequency component (PS_LOG_NORM, Fig. 2c) and additional noise reduction for each frequency component (PS_LOG_NORM_DEN, not shown) with the same approach as in the earlier subsection Ultrasound Data Representations

through Image Formation Process (see details in [Appendix S1](#), online only).

Phase representations of ultrasound data

As described under Ultrasound Data Representations through Image Formation Process and illustrated in [Figure 1c](#), phase information in the RF data is generally discarded in ultrasound systems during the I/Q-to-envelope ENV signal processing step. In this work, we tentatively explored the potential of RF phase information for DL hepatic steatosis classification. Three different phase representations were tested based on the workflow illustrated in [Figure 3a](#). The CARRIER representation removes the amplitude information and *retrieves the oscillatory component* of the RF data by normalizing the RF signal with the envelope ENV. This results in a sinusoidal phase oscillatory pattern with amplitude in a range $[-1, 1]$ ([Fig. 3a](#)). A second ANGLE representation was obtained by calculating the phase of the complex IQ signal in the range $[-\pi, \pi]$ rad ([Fig. 3b](#)). Finally, a PHASE representation is calculated as the unwrapped complex phase for each frequency component of the spectrogram discussed in the subsection Power Spectrum Representation of Ultrasound Data for QUS ([Fig. 3c](#)).

Convolutional neural network architectures and training process

The US data representations calculated in the previous three subsections were independently used as input data to train convolutional neural network (CNN) architectures according to [Figure 4](#). Because of the limited patient sample available in this study, lightweight neural networks inspired by the AlexNet architecture were used to avoid data overfitting ([Krizhevsky et al. 2017](#)). Our model consisted of up to three convolutional layers, with pooling stages in-between to reduce data dimensionality and encode image features, followed by two fully connected layers and a SoftMax layer for multiclass hepatic steatosis classification in the S0, S1, S2 or S3 stage. Both 1-D ([Fig. 4a](#)) and 2-D CNNs were tested ([Fig. 4b](#), [Table 2](#)). One-dimensional CNNs trained based on separate RF lines, with an architecture similar to that of [Han et al. \(2020\)](#). Two-dimensional CNNs included multiple RF lines as input data, similarly to [Sanabria et al. \(2020b\)](#). For 3-D spectrogram representations, additionally, a lightweight 3-D CNN architecture, which had previously been successfully applied for ultrasound classification of prostate cancer ([El Kaffas et al. 2021](#)), was tested (time, RF line index, frequency) ([Fig. 4c](#)).

To achieve a sufficiently large data set for training, a total of 300 patches were extracted from each

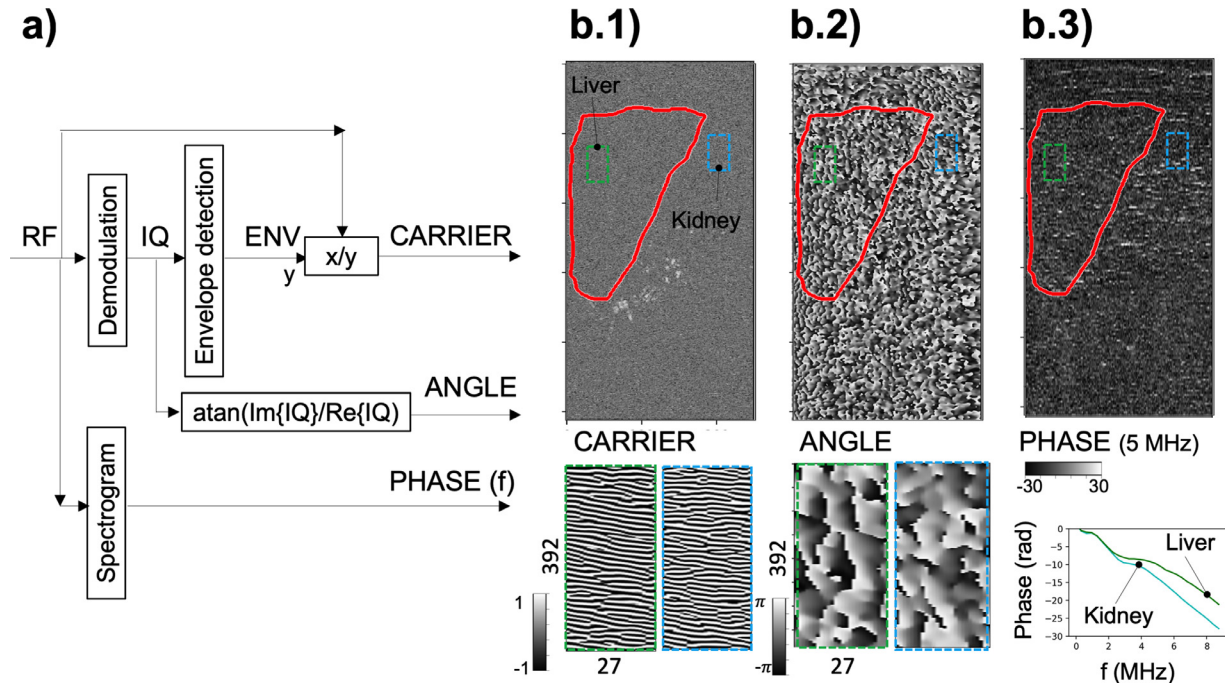


Fig. 3. Proposed phase representations of ultrasound data for same participant as in [Figure 1](#). (a) Flow diagram. (b.1) Carrier frequency representation obtained by normalizing RF data with envelope ENV signal. (b.2) Phasor angle of base-band IQ representation. (c) Unwrapped phase of spectrogram plotted for frequency of 5 MHz. Exemplary patches are shown for each data representation in liver parenchyma, together with the patch in the kidney region for comparison. ENV = envelope; IQ = in-phase and quadrature; RF = radiofrequency.

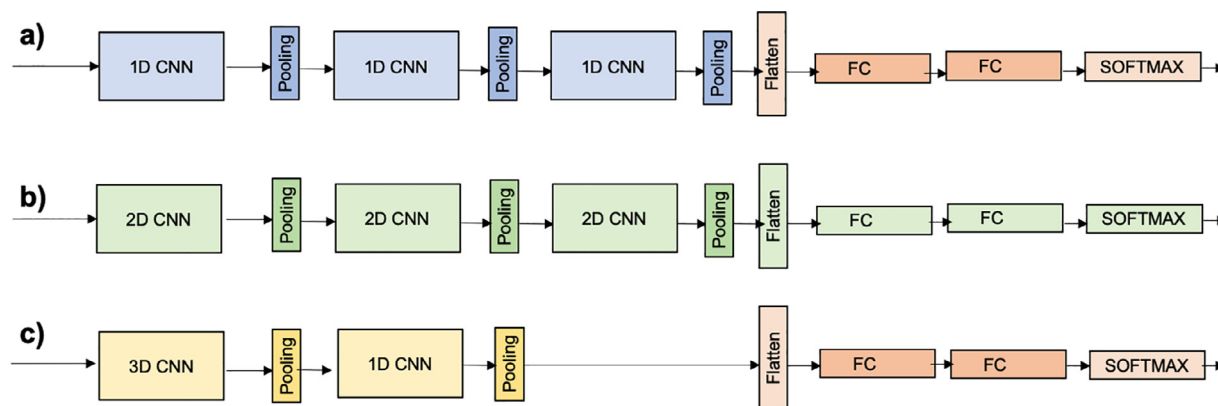


Fig. 4. Convolutional neural network architectures used in this work. (a) One-dimensional architecture (_1D). (b) Two-dimensional architecture (_2D). (c) Three-dimensional architecture (_3D). CNN = convolutional neural network, FC = fully connected layer.

ultrasound view at uniformly distributed random positions within the segmentation mask and fed independently into the CNN for classification. Each patch was labeled with the reference steatosis stage (S0, S1, S2 or S3) according to the average MRI-PDFF value of the patient. Each patch represented a region of interest (ROI) of $15 \times 15 \text{ mm}^2$, corresponding to a lateral dimension of 20 wavelengths. The ROI was chosen as a compromise between a sufficiently large size to provide a representative texture of liver parenchyma and a sufficiently small size to fit within the edges of the segmented liver masks. The selected patch size of $15 \times 15 \text{ mm}^2$ is in the range of typical regions of interest in US elastography (Barr et al. 2015) and allowed for a coverage of 88.1% of the liver parenchyma region. It should be noted that patch extraction was performed in the pre-scan converted images. The scan-converted patch for the curvilinear probe was fan-shaped, and its width increased with depth. The patch dimensions were adjusted to achieve an average area of $15 \times 15 \text{ mm}^2$ in a typical liver parenchyma depth range (30–80 mm).

The patches were randomly shuffled prior to training. Eighty percent of the available data were used for training and 20% for validation. Patches from different patients were used in the training and validation sets. A total of 67,800 patches were available for 2-D/3-D classification and 1,830,600 patches for 1-D classification. Patient probability scores were calculated by averaging the probability scores of the corresponding patches.

For direct comparability between US data representations, patch center positions were co-registered along different US representations, and the same CNN architectures were used to compare different US data representation stages. The same loss function (categorical cross-entropy), optimization algorithm (Adadelta with a decay rate of 0.95), batch size (30) and learning rate

(0.01) were used in all cases. All NNs were trained for 100 epochs, with early stopping occurring after five consecutive iterations without improvement of the validation loss. To minimize overfitting on the validation data set, the last NN architecture before early stopping was stored. The same patch input size (392, 27 px [axial, lateral]), corresponding to the patch size before scan conversion, was used for all calculated data representations (in spectral representations it was reshaped as [14, 27, 28 px]). The following hyperparameter tuning measures were allowed to maximize the CNN classification performance for each data representation: (i) optimizing non-linearity activation functions (favoring elu for bipolar data representations and relu for unipolar data representations); (ii) reducing the complexity of the network for demodulated data representations by applying average pooling at the beginning of the network; (iii) introducing dropout and batch normalization for regularization. For each case, the best performing network was selected. Because of the scan conversion process, the input patch size (51, 51 px) was slightly larger for clinical B-mode patches for the same equivalent geometrical size as DEN patches, while an equivalent network architecture was preserved. The detailed NN architectures are provided in Figure S1 (online only). More than 150 iterations of networks were carried out to arrive at the final architectures and hyperparameter settings for each model. Neural networks were implemented in Python 3.8.5 using deep learning libraries (TensorFlow 2.0) with the high-level application programming interface Keras.

Comparison with quantitative ultrasound metrics

The CNN hepatic steatosis staging performance was retrospectively compared with classification based on a comprehensive list of QUS metrics available for the same population from a previous study (Pirmoazen et al.

2022). Three backscattering metrics were included: hepato-renal index (HRI), attenuation coefficient (AC) and Nakagami parameter (NP). Additionally, three shear wave elastography parameters were measured: shear wave elasticity (SWE), viscosity (SWV) and dispersion (SWV). Finally, three spectroscopy metrics were considered: mid-band fit (MBF), spectral intercept (SI) and spectral slope (SS). Multiparametric descriptors based on the combination of the nine QUS metrics described above were also evaluated. Four machine learning regression models—random forests, Gaussian process, multilayered perceptron and multilinear model—were implemented according to the approach previously described by Sanabria *et al.* (2020a).

Statistics

To accommodate the small sample size in the study, a stratified fivefold cross-validation approach was used. The patient data were divided into five folds with balanced representation of hepatic steatosis categories. The fold division was performed on the patient level, and data from different patients were included in different folds. The same fold stratification was used for all US data representations, as well as for the QUS multiparametric models described in the previous subsection. For five separate training iterations, four folds were used for training the model and the remaining fold for validation. To avoid network overfitting and provide generalizable results, the validation fold contained unseen patients in the training fold. Then, for each patient belonging to the training folds, all corresponding image patches were used as independent training instances. Similarly, for each patient belonging to the validation fold, all corresponding image patches were used as independent validation instances. Multivariate analysis of variance (MANOVA) was performed to identify differences among US data representations on a patch classification level across each CNN inference class (S0, S1, S2 and S3). A paired *t*-test was then used to compare AUROC outcomes for different US data representations across validation folds and output classes on a patch classification level. Conformity of AUROCs to a Gaussian distribution was tested with both the Shapiro–Wilk and Kolmogorov–Smirnov tests. Adjustment of significance for multiple paired *t*-tests was performed with the Benjamini and Hochberg procedure (Benjamini and Hochberg 1995; Groppe 2022) for controlling the false discovery rate (FDR). FDR is the expected proportion of “discoveries” (rejected null hypothesis) that are false (incorrect rejections).

Patient analysis was performed to stratify patients with steatosis based on different MRI-PDFF cutoff values according to Table 1. For this purpose, three binary classifiers of patients were built based on the multiclass

patch classification outputs. S0 versus S1, S2 and S3 differentiate, with an MRI-PDFF cutoff of 6.4%, patients with no steatosis (S0) from patients with steatosis (S1 = mild, S2 = moderate, S3 = severe). S0, S1 versus S2 differentiates, with an MRI-PDFF cutoff of 17.4%, patients with either steatosis stage S0 or S1 (none, mild) from patients with more advanced steatosis (S2 = moderate or S3 = severe). Similarly, S0–S2 versus S3 differentiates, with an MRI-PDFF cutoff of 22.1%, patients with either steatosis stage S0, S1 or S2 (none, mild or moderate) from patients with severe steatosis (S3). Patient scores were calculated by averaging the probability scores of the corresponding patches. To build binary classifiers, SoftMax probability scores corresponding to the steatosis stages of each binary patient class were added up. Then the probability scores obtained in each validation fold were grouped, and a single ROC analysis was performed to obtain an AUROC for the entire patient population. The cutoff was calculated as that providing the maximum Youden index. Sensitivity, specificity and accuracy were calculated at the optimum cutoff for each US data representation. As a comparison reference of the merit of the classifiers, accuracy values were calculated for the majority classifier (all instances assigned to the same classifier). Accuracy values were also calculated from radiologists’ qualitative hepatic steatosis staging from B-mode images. For comparisons between data representations, average AUROCs of all MRI-PDFF cutoff values were reported. Statistics were calculated with SPSS Statistics version 28.0.1.1 (IBM Corp., Armonk, NY, USA) and MATLAB version 2018b (The MathWorks, Inc., Natick, MA, USA).

RESULTS

Visualization of US data representations of liver exams and textural differences between healthy and steatotic livers

Figures 1 to 3 provide exemplary images and co-registered liver patches for each of the ultrasound representations considered in this study, all corresponding to an intercostal view with the liver and kidney of a healthy patient. In Figure 1, successive image formation steps from RF data to B-mode from (c.1) to (c.5) improve the visualization and interpretability of the images. Conversely, they reduce the complexity and thus potentially the information content of the patch representations. Early stages of image formation, including RF (Fig. 1c.1) and IQ data (Fig. 1c.2), exhibit oscillatory patterns, where the identification of tissue boundaries is challenging. On the other hand, both RF and IQ patches exhibit rich interference patterns and contain phase information of the ultrasound signal. RF is a passband

signal with high-frequency oscillations, while IQ is a complex-valued baseband signal, where phase-shifted low-frequency oscillations are observed for both real ($\text{Re}[IQ]$) and imaginary ($\text{Im}[IQ]$) components). The transformation of IQ data into the envelope signal ENV leads to a reduction of the textural information in the patch 1. c.3, as phase information is removed. Moreover, the available signal bandwidth at the -20 -dB cutoff is reduced from 2.5 to 1.2 MHz, as illustrated in Figure 1a.2. The envelope image (Fig 1c.3) delineates subcutaneous layers, but the strongly reflecting diaphragm below the liver leads to a large signal range and a saturated grayscale, which constrain the visualization of the liver. LOG compression (Fig. 1c.4) balances the grayscale and allows resolution of the liver boundaries, while denoising DEN (Fig. 1c.5) reduces speckle noise, thus highlighting the texture of liver parenchyma. On the other hand, the gray level contrast of the LOG and DEN is reduced with respect to ENV because of the limitation of the dynamic range.

Figure 2 is a PS representation of the RF data. For each frequency of the PS (Fig. 2b), separate images are obtained (Fig. 2c), which illustrate varying tissue information content. A stacked representation of a liver patch as a function of frequency (Fig. 2c.2) indicates variations in the texture of liver parenchyma, with grainier texture at higher frequencies. High-frequency harmonics (f_3) of the transducer carrier frequency (f_1) are observed, which encode non-linear information (Fig. 2b), for instance, highlighting reflections in the right hemidiaphragm. Figure 3 contains additional representations that attempt to recover and isolate the phase information. Overall, these representations are challenging to interpret visually. Neither CARRIER, ANGLE nor PHASE allow visual delineation of tissue structures. Patches in Figure 3 exhibit complex structured interference patterns. Side-by-side comparison of patches in liver and kidney tissues reveal perceptible differences in frequency-dependent PHASE trends.

Figure 5 compares patches at selected US data representations for a healthy participant (female, 51 y, MRI_PDFP = 3.3%) (Fig. 5a) and a patient with severe liver steatosis (female, 55 y, MRI_PDFP = 42.9%) (Fig. 5b), both in the subcostal sagittal view. Observation of the complete liver structure allows separation of both cases based on distinctive features, such as decreased visibility of hepatic vasculature and right hemidiaphragm for the steatotic liver case. The textural differences between individual extracted patches (Fig. 5a.2, 5b.2) are subtler. A more homogeneous texture is discerned in the grayscale maps of ENV, DEN of patients with severe liver steatosis, while the healthy participant exhibits regional variations of echogenicity within the same patch. The same pattern can be visually recognized in the RF patches, where isolated bright spots are more

frequent with respect to the more iso-echogenic interference pattern observed for the patient with liver steatosis. The phase patterns (CARRIER) are difficult to visually interpret, yet it seems that for the healthy case, more discontinuous patterns with less connectivity of equal phase surfaces are generally observed. Overlapped comparison of normalized power spectra (Fig. 5c) reveals for the cases with severe liver steatosis a shift of the spectrum peak toward lower frequencies and a reduction of amplitude at higher frequencies, both effects increasing for higher patch depths. These effects are indicative of an increase in acoustic attenuation in steatotic livers, which comparatively has a larger effect on higher-frequency components (Amin 1989). Therefore, differences in attenuation can be observed in the power spectra despite time-gain compensation of attenuation and the normalization of dynamic range in the log-compression step and provide potentially useful information for patient classification.

Comparison of classifications of hepatic steatosis grades with different US data representations and NN architectures

In Figure 6 are exemplary loss-accuracy plots for one fold of 2-D CNN classification trained with different US data representations to separate healthy (S0) and steatotic livers (S1, S2, S3). For all US data representations, the loss function decreases smoothly for both the training and validation sets until it eventually flattens. Similarly, for both training and validation sets, accuracy increases to a point where they flatten out. At this point, early stopping is applied after five consecutive iterations without improvement of the validation loss to avoid overfitting in the validation set. Comparison between US data representations showed lower values of the minimum of the loss function for early US imaging stages (RF, IQ) than for latter stages (ENV, LOG, DEN). Spectral representations (PS_LOG_NORM_DEN) achieved lower loss and converged in fewer epochs than RF data.

Table 3 compares validation scores at patch-level classification of all hepatic steatosis categories (S0, S1, S2, S3) for different CNN models and US data representations. Model scores were globally compared in terms of the AUROC with respect to the probability score output by the NN. Average AUROCs and standard deviations for five cross-validation folds are plotted. A paired test was used to compare differences in AUROCs for each model with respect to both RF_2D and clinical B-mode classifiers. For all data representations, normality could not be rejected with both the Kolmogorov–Smirnov and Shapiro–Wilk tests at a significance level of $p > 0.05$ and with either test at $p < 0.01$. Therefore, Gaussianity of the ROC AUC data distribution was an acceptable approximation for the paired tests.

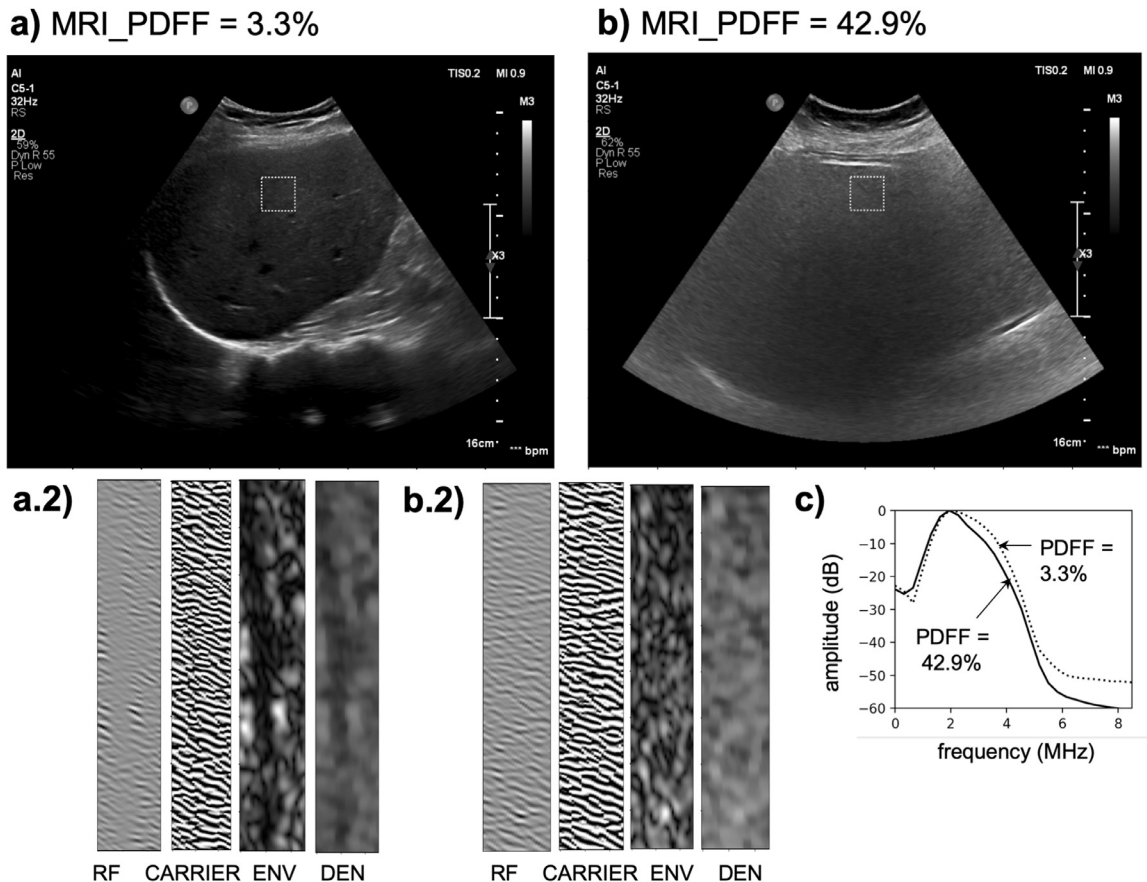


Fig. 5. Ultrasound views and exemplary patches for healthy participant (a) and patient with severe steatosis (b). (a.2, b.2) Patches were extracted for both patients at a depth of 40 mm for different ultrasound representations. (c) Log-compressed average power spectrum profiles for both patients. IQ = in-phase and quadrature; MRI-PDF = magnetic resonance imaging proton density fat fraction; RF = radiofrequency.

The MANOVA revealed significant differences among AUROCs obtained with different US data representations; the hypothesis of the means being the same was rejected at $p < 0.001$. Similarly, considering 2-D and 1-D representations separately, MANOVA had $p < 0.001$ and $p = 0.02$, respectively. With respect to 2-D patches, the RF_2D data representation exhibited significantly higher AUROCs in Table 3 than all other successive imaging stages (IQ_2D, ENV_2D, LOG_2D, and DEN_2D, $p < 0.001$) and the clinical B-mode image representation ($p = 0.001$). Also, two-dimensional patches RF_2D showed significantly higher AUROCs than one-dimensional patches RF_1D. A decreasing AUROC trend was observed along successive stages of the image formation process in Figure 1a.1, which flattened out in the later stages LOG and DEN. This trend was consistently observed for all hepatic steatosis categories and both 1-D and 2-D CNN models. For the same image classification stage, all 2-D CNN architectures achieved significantly larger AUROCs than the corresponding 1-D CNN architecture ($p < 0.001$ for RF, IQ

and ENV, $p = 0.006$ for LOG and $p = 0.01$ for DEN). With respect to 2-D patches, clinical B-mode revealed lower AUROCs than the RF_2D and IQ_2D representations, non-significant differences with ENV_2D and LOG_2D data and higher AUROCs than DEN_2D representations. Two-dimensional power spectrum data representations of RF data (PS_2D) and variants thereof (PS_LOG_2D, PS_LOG_NORM_2D, PS_LOG_NORM_DEN_2D) achieved higher scores than the originating RF_2D data. Log-compressed PS representations (PS_LOG, PS_LOG_NORM, PS_LOG_NORM_DEN) had superior classification outcomes than PS representation in linear scale (PS_) ($p < 0.03$). The 2-D CNN architecture outperformed the 3-D CNN architecture for PS patch classification ($p < 0.03$). Classifiers trained on phase information (CARRIER_, ANGLE_, PHASE_) were able to learn hepatic steatosis classification, achieving AUROCs comparable to their amplitude-based ENV_2D counterpart. PHASE_3D achieved superior AUROCs to the clinical B-mode US data representation and non-significant differences with RF data.

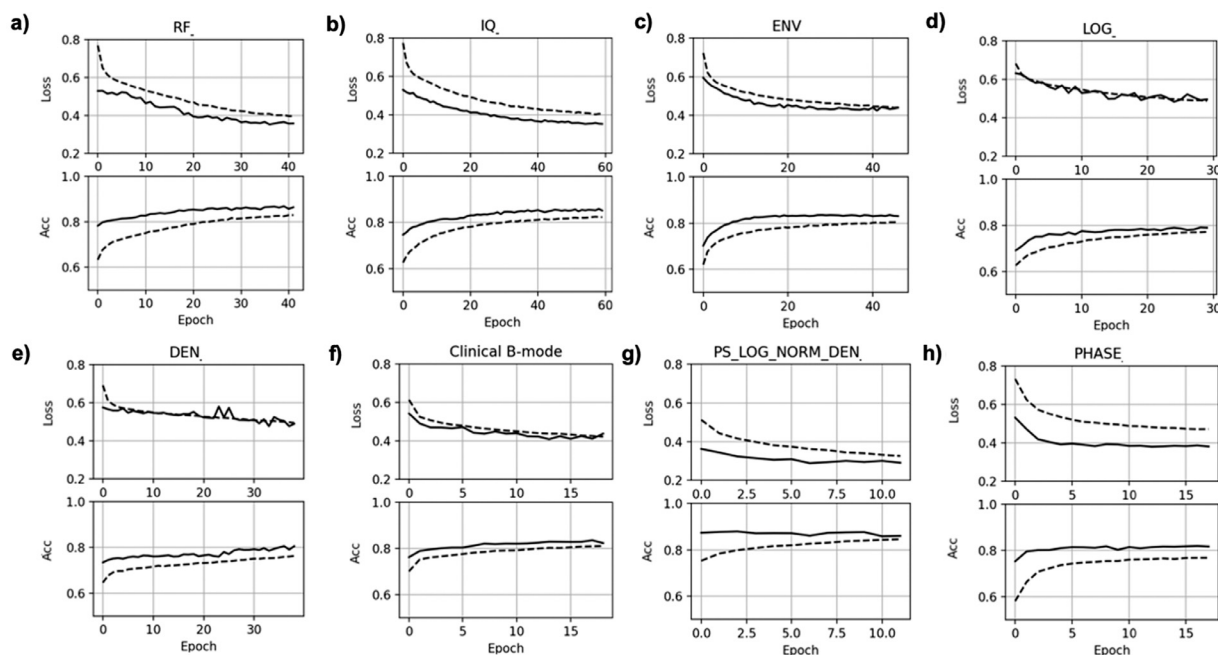


Fig. 6. Exemplary loss-validation plots for training one fold of binary steatosis classification model S0 versus S1S2S3 with 2-D CNN architecture. Training accuracy and loss are plotted with *black dashed lines*, while validation accuracy and loss are plotted with *black solid lines*. ENV = envelope; IQ = in-phase and quadrature; RF = radiofrequency.

Differences in AUROCs between the training and validation sets were small (root mean square error = 0.02), which revealed a good generalization for models and US data representations.

Figure 7 illustrates selected ROC plots for classification of patients with liver steatosis with different MRI-PDFF cutoff values. In trend, early steatosis stages with lower MRI-PDFF cutoff values had ROC curves closer to the ideal classifier. An improvement in ROC curves can be visually assessed for RF, IQ and PS_LOG_NORM_DEN representations with respect to ENV, LOG, DEN, clinical B-mode and PHASE. Figure 8 summarizes the corresponding AUROCs. Leveraging multiple patch classifications in one patient's score increases the overall observed AUROCs. The comparative classification performance trends observed at patch-level classification are also reproduced at patient-level classification, with AUROC being reduced at successive steps of the image formation process. Performance differences between 1-D and 2-D CNN models were reduced for RF US data representation, but were noticeable for ENV, LOG and DEN stages. In comparison, differences between 2-D and 3-D models were less noticeable. Performance differences between RF (AUROC = 0.994) and PS data representations become less marked, the best performing model being PS_LOG_NORM_DEN_2D (AUROC = 0.998). All phase representations (CARRIER_, ANGLE, PHASE_) had a mean AUROC > 0.983 across liver steatosis

categories, outperforming both clinical B-mode (AUROC = 0.938) and demodulated amplitude representations (ENV, LOG, DEN).

Table 4 outlines the sensitivity, specificity and accuracy metrics at the optimum cutoff of the ROC curve. The cutoff scores values were determined by the MRI PDFF threshold and did not exhibit large variations across different US data representations and NN models. Overall, the sensitivity at cutoff improved for PS and phase-based representations. The comparative trends of accuracy across US data representations and NN models at optimum cutoff were consistent with the AUROC trends.

Comparison of NN classifications with multiparametric models based on quantitative ultrasound metrics

Figure 9 illustrates the AUROCs for selected quantitative ultrasound metrics extracted from the same patient population, together with multiparametric combinations. Table 5 summarizes the QUS metrics considered. For individual QUS metrics, monotonically increasing trends with increasing hepatic fat accumulation have previously been observed that linearly correlate with MRI-PDFF. Therefore, cutoff thresholds were used to classify liver steatosis stages (Pirmoazen et al. 2022). The best-performing metrics in terms of average AUROCs were quantitative ultrasound spectroscopy metrics (SI: AUROC = 0.981, SS: AUROC = 0.963), statistical speckle distributions (NP: AUROC = 0.964), attenuation coefficient (AUROC = 0.965) and HRI (liver

Table 3. Patch-based results: Comparison of AUROC curves on validation set for multiclass classification of image patches into steatosis categories*

	AUROC				Paired test <i>p</i> value [†]	
	S0	S1	S2	S3	Ref: RF_2D	Ref: Clinical B-mode
RF_1D	0.876 (0.03)	0.679 (0.05)	0.726 (0.02)	0.800 (0.05)	<0.001 (+)	0.467 (?)
IQ_1D	0.842 (0.03)	0.641 (0.04)	0.722 (0.05)	0.769 (0.06)	<0.001 (+)	0.346 (?)
ENV_1D	0.743 (0.07)	0.576 (0.05)	0.541 (0.02)	0.628 (0.03)	<0.001 (+)	<0.001 (+)
LOG_1D	0.746 (0.07)	0.608 (0.07)	0.573 (0.02)	0.636 (0.03)	<0.001 (+)	<0.001 (+)
DEN_1D	0.761 (0.08)	0.603 (0.09)	0.580 (0.04)	0.705 (0.08)	<0.001 (+)	0.006 (+)
RF_2D	0.910 (0.04)	0.724 (0.04)	0.783 (0.05)	0.857 (0.03)	—	0.001 (—)
IQ_2D	0.888 (0.03)	0.719 (0.04)	0.759 (0.03)	0.841 (0.04)	0.001 (+)	0.015 (—)
ENV_2D	0.862 (0.03)	0.669 (0.06)	0.741 (0.04)	0.767 (0.06)	<0.001 (+)	0.971 (?)
LOG_2D	0.831 (0.03)	0.586 (0.10)	0.712 (0.06)	0.748 (0.07)	<0.001 (+)	0.052 (?)
DEN_2D	0.808 (0.05)	0.612 (0.05)	0.609 (0.07)	0.748 (0.07)	<0.001 (+)	0.012 (+)
Clinical B-mode_2D	0.880 (0.06)	0.660 (0.08)	0.688 (0.08)	0.792 (0.05)	0.001 (+)	—
PS_2D	0.916 (0.06)	0.767 (0.06)	0.802 (0.03)	0.870 (0.04)	0.012 (—)	<0.001 (—)
PS_LOG_2D	0.928 (0.07)	0.812 (0.07)	0.832 (0.02)	0.885 (0.06)	0.002 (—)	<0.001 (—)
PS_LOG_NORM_2D	0.949 (0.10)	0.811 (0.10)	0.874 (0.07)	0.887 (0.09)	0.008 (—)	<0.001 (—)
PS_LOG_NORM_DEN_2D	0.922 (0.08)	0.777 (0.08)	0.836 (0.03)	0.920 (0.03)	0.001 (—)	<0.001 (—)
PS_3D	0.839 (0.06)	0.664 (0.06)	0.726 (0.07)	0.775 (0.02)	<0.001 (+)	0.579 (?)
PS_LOG_3D	0.905 (0.07)	0.807 (0.07)	0.781 (0.04)	0.876 (0.05)	0.062 (?)	<0.001 (—)
PS_LOG_NORM_3D	0.907 (0.06)	0.718 (0.06)	0.774 (0.04)	0.842 (0.08)	0.579 (?)	0.003 (—)
PS_LOG_NORM_DEN_3D	0.903 (0.07)	0.728 (0.07)	0.808 (0.03)	0.901 (0.04)	0.178 (?)	0.001 (—)
CARRIER_2D	0.830 (0.05)	0.666 (0.05)	0.682 (0.02)	0.809 (0.05)	<0.001 (+)	0.676 (?)
ANGLE_2D	0.770 (0.04)	0.609 (0.06)	0.687 (0.01)	0.756 (0.05)	<0.001 (+)	0.015 (+)
PHASE_2D	0.827 (0.05)	0.673 (0.02)	0.748 (0.02)	0.738 (0.05)	<0.001 (+)	0.579 (?)
PHASE_3D	0.886 (0.04)	0.753 (0.06)	0.767 (0.02)	0.803 (0.05)	0.346 (?)	0.018 (—)

AUROC = area under the receiver operating characteristic curve; Ref = reference.

* Averages of five cross-validation folds are provided, with the standard deviations in parentheses. A paired *t*-test was used to compare the AUROC for each data representation with both RF and clinical B-mode data representations (+ = reference had superior performance, — = reference had inferior performance, ? = no difference).

† Computed for AUROCs across all classification classes and five validation folds. *p* Values are adjusted using the [Benjamini and Hochberg \(1995\)](#) procedure to control the false discovery rate (FDR) in the multiple paired *t*-tests. A *p* value <0.05 is set as the cutoff for significant differences.

parenchyma echogenicity normalized with respect to kidney, AUROC = 0.903). Elastography metrics (elastic modulus, Voigt viscosity and shear wave dispersion) had lower AUROCs (<0.680).

Four QUS metrics in [Table 5](#) had higher average AUROCs (SI: 0.981, NP: 0.964, AC: 0.965 and SS: 0.963) than B-mode 2-D CNN classifications in [Table 4](#) (AUROC = 0.938). All QUS metrics had lower AUROCs than 2-D CNN classifications with RF US data (AUROC = 0.994 in [Table 4](#)). Multiparametric combinations of QUS metrics exhibited AUROC performance gains with respect to individual metrics, consistent with the results for MRI-PDFF regression previously discussed in [Sanabria et al. \(2020a\)](#), where a comprehensive list of machine learning combinations were tested for liver steatosis classification. The best multiparametric results were obtained with a random forest classifier. In this case, average AUROC (0.986) was lower than those for CNN classifiers with RF_2D (AUROC = 0.994) and PS_2D (AUROC = 0.995) representations.

DISCUSSION

In this work, we have exhaustively compared US data representations along the image formation process as inputs to machine learning models, and specifically NN algorithms in terms of classification performance for the same steatotic liver populations. Although CNN architectures for hepatic steatosis classification based on B-mode images have been explored ([Han et al. 2020](#); [Cha et al. 2021](#)), and the potential use of RF data for liver steatosis classification has recently been illustrated ([Chen et al. 2020](#); [Han et al. 2020](#); [Sanabria et al. 2020b](#); [Pirmoazen et al. 2022](#)), to the best of our knowledge this is the first work to comprehensively explore the effect of US data representation in DL classification performance.

The study yields several learnings. First, the proposed patch extraction approach, which was already described in [Sanabria et al. \(2020b\)](#), augments ultrasound training instances that can be extracted from small patient populations and is generalizable to new patient

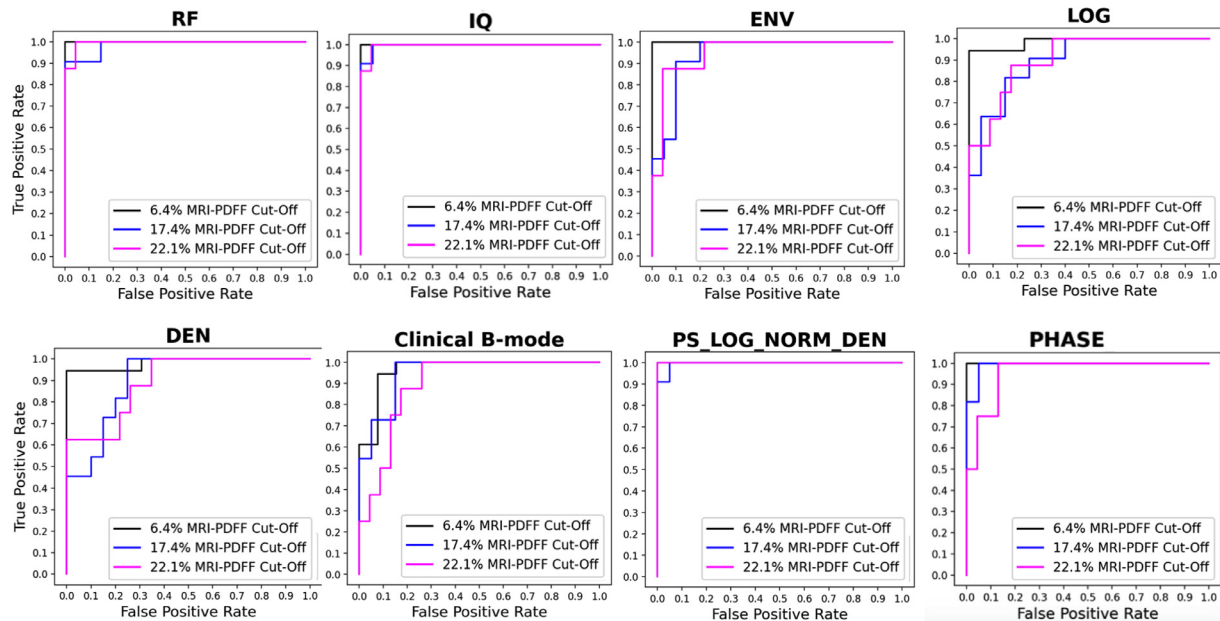


Fig. 7. Receiver operating characteristic curve plots for stratification of patients with mild steatosis (S0 vs. S1–S3, 6.4% MRI cutoff), moderate steatosis (S0, S1 vs. S2, S3, 17.4% MRI-PDFF cutoff), and severe steatosis (S0–S2 vs. S3, 22.1% MRI-PDFF cutoff) with different ultrasound data representations and _2D CNN architecture. CNN = convolutional neural network; IQ = in-phase and quadrature; MRI-PDFF = magnetic resonance imaging proton density fat fraction; RF = radiofrequency.

populations. To avoid network overfitting and provide generalizable results, the validation data set contains patches from unseen patients in the training data set. Here, of 31 patient examinations, a total of 67,800 2-D patches were extracted. As each patch is fed into the CNN as an independent instance, enough training

samples are available to achieve satisfactory model convergence. This is probably related to the strong degree of variability of US data within a single examination and with multiple views from the same patient, which provides natural data augmentation. To bring instances together on a patient level, the patch probability scores

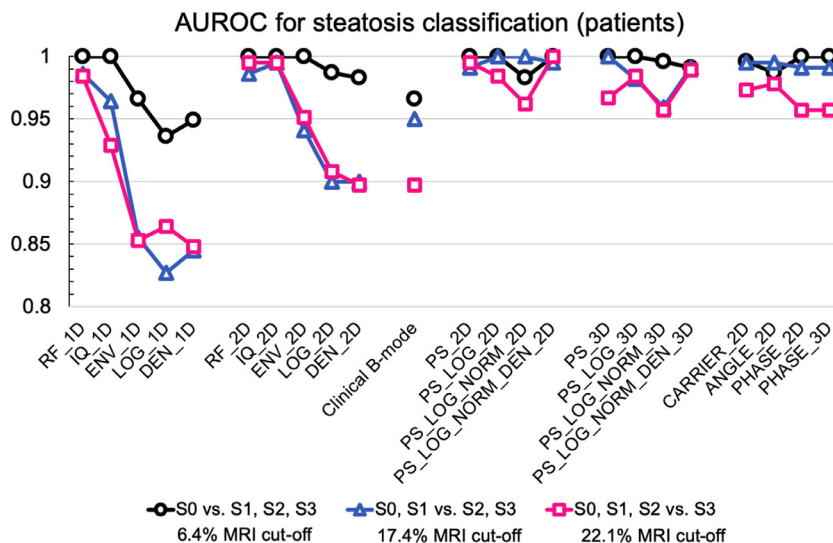


Fig. 8. Comparison of areas under the receiver operating characteristic curves for patient stratification in mild, moderate and severe steatosis with different ultrasound data representations and convolutional neural network architectures. AUROC = area under the receiver operating characteristic curve; ENV = envelope; IQ = in-phase and quadrature; MRI = magnetic resonance imaging; PS = power spectrum; RF = radiofrequency.

Table 4. Patient-based results: Comparison of scores for patient stratification for different data representations and MRI-PDFF cutoffs*

	AUROC				Cutoff according to Youden index			Sensitivity at cutoff				Specificity at cutoff				Accuracy at cutoff†			
	S0vs. S1S2S3	S0S1 vs.S2S3	S0S1S2 vs.S3	Average	S0vs. S1S2S3	S0S1 vs. S2S3	S0S1S2 vs.S3	S0vs. S1S2S3	S0S1 vs. S2S3	S0S1S2 vs.S3	Average	S0vs. S1S2S3	S0S1vs. S2S3	S0S1S2 vs.S3	Average	S0vs. S1S2S3	S0S1vs. S2S3	S0S1S2 vs.S3	Average
Human annotation	—	—	—	—	—	—	—	1.0	1.0	0.875	0.958	0.923	0.800	0.913	0.879	0.968	0.871	0.903	0.914
RF_1D	1	0.986	0.984	0.990	0.703	0.601	0.323	1	0.909	1	0.970	1	1	0.957	0.986	1	0.968	0.968	0.979
IQ_1D	1	0.964	0.929	0.964	0.692	0.543	0.264	1	0.909	1	0.970	1	0.9	0.739	0.880	1	0.903	0.806	0.903
ENV_1D	0.966	0.855	0.853	0.891	0.751	0.501	0.264	0.889	0.909	0.75	0.849	1	0.75	0.913	0.888	0.935	0.806	0.871	0.871
LOG_1D	0.936	0.827	0.864	0.876	0.751	0.53	0.273	0.8333	0.727	0.75	0.770	0.923	0.85	0.913	0.895	0.871	0.806	0.871	0.849
DEN_1D	0.949	0.845	0.848	0.881	0.754	0.507	0.278	0.8333	0.818	0.625	0.759	1	0.75	0.957	0.902	0.903	0.774	0.871	0.849
RF_2D	1	0.986	0.995	0.994	0.686	0.621	0.332	1	0.909	1	0.970	1	1	0.957	0.986	1	0.968	0.968	0.979
IQ_2D	1	0.995	0.995	0.997	0.67	0.516	0.297	1	1	1	1.000	1	0.95	0.957	0.969	1	0.968	0.968	0.979
ENV_2D	1	0.941	0.951	0.964	0.756	0.6	0.314	1	0.909	0.875	0.928	1	0.9	0.957	0.952	1	0.903	0.935	0.946
LOG_2D	0.987	0.9	0.908	0.932	0.774	0.611	0.309	0.944	0.818	0.875	0.879	1	0.85	0.826	0.892	0.968	0.839	0.839	0.882
DEN_2D	0.983	0.9	0.897	0.927	0.761	0.515	0.268	0.944	1	1	0.981	1	0.75	0.652	0.801	0.968	0.839	0.742	0.850
Clinical B-mode	0.966	0.95	0.897	0.938	0.783	0.554	0.314	0.944	1	1	0.981	0.923	0.85	0.739	0.837	0.935	0.903	0.806	0.881
PS_2D	1	0.991	0.995	0.995	0.727	0.566	0.363	1	1	1	1.000	1	0.95	0.957	0.969	1	0.968	0.968	0.979
PS_LOG_2D	1	1	0.984	0.994	0.663	0.601	0.352	1	1	1	1.000	1	1	0.57	0.857	1	1	0.968	0.989
PS_LOG_NORM_2D	0.983	1	0.962	0.982	0.672	0.651	0.391	0.944	1	0.875	0.940	0.923	1	0.957	0.960	0.935	1	0.935	0.957
PS_LOG_NORM_DEN_2D	1	0.995	1	0.998	0.669	0.619	0.4	1	1	1	1.000	1	0.95	1	0.983	1	0.968	1	0.989
PS_3D	1	1	0.967	0.989	0.685	0.533	0.284	1	1	1	1.000	1	1	0.913	0.971	1	1	0.935	0.978
PS_LOG_3D	1	0.982	0.984	0.989	0.697	0.552	0.349	1	1	1	1.000	1	0.9	0.957	0.952	1	0.935	0.968	0.968
PS_LOG_NORM_3D	0.996	0.959	0.957	0.971	0.663	0.529	0.221	0.944	1	1	0.981	1	0.9	0.783	0.894	0.968	0.935	0.839	0.914
PS_LOG_NORM_DEN_3D	0.991	0.991	0.989	0.990	0.722	0.522	0.347	0.944	1	1	0.981	1	0.95	0.957	0.969	0.968	0.968	0.968	0.968
CARRIER_2D	0.996	0.995	0.973	0.988	0.726	0.535	0.291	1	1	1	1.000	1	0.85	0.957	0.936	1	0.903	0.968	0.957
ANGLE_2D	0.987	0.995	0.978	0.987	0.728	0.557	0.267	0.944	1	1	0.981	1	0.95	0.913	0.954	0.968	0.968	0.935	0.957
PHASE_2D	1	0.991	0.957	0.983	0.757	0.557	0.262	1	1	1	1.000	1	0.95	0.87	0.940	1	0.968	0.903	0.957
PHASE_3D	1	0.991	0.957	0.983	0.769	0.545	0.298	1	1	1	1.000	1	0.95	0.87	0.940	1	0.968	0.903	0.957

MRI-PDFF = magnetic resonance imaging proton density fat fraction.

* S0 vs. (S1, S2, S3); (S0, S1) vs. (S2, S3); and (S0, S1, S2) vs. S3, respectively, correspond to 6.4%, 17.4% and 22.1% MRI-PDFF cutoff values.

† Assigning all instances to the majority class (no merit classifier) results in accuracies of 0.581 (S0 vs. S1, S2, S3), 0.645 (S0, S1 vs. S2, S3) and 0.742 (S0, S1, S2 vs. S3). Human annotation scores based on B-mode images are provided as a reference.

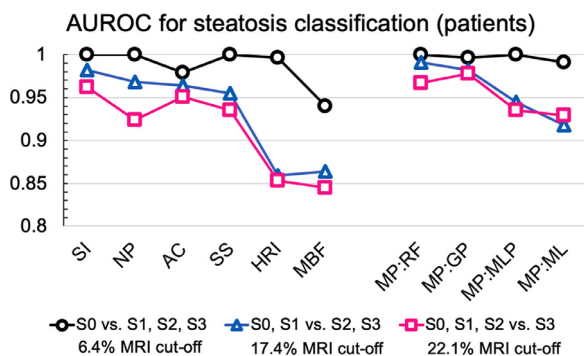


Fig. 9. AUROC comparison for patient stratification based on quantitative ultrasound metrics. SI, NP, AC, SS, HRI, MBF, and MP combinations are provided. AC = attenuation coefficient; AUROC = area under the receiver operating characteristic curve; GP = Gaussian process; HRI = hepatorenal index; MBF = midband fit; ML = multilinear; MLP = multilayer perceptron; MP = multiparametric model; MRI-PDFF = magnetic resonance imaging proton density fat fraction; NP = Nakagami parameter; RF = random forest; SI = spectral intercept; SS = spectral slope.

of each patient were averaged before performing AUROC analysis.

Second, CNN classification performance decreases as we progress into the image formation process, with a significant performance drop between RF data representation and B-mode images. This may appear counterintuitive to the fact that B-mode images are visually more informative in terms of hepatic steatosis classification performance, while RF data are difficult to graphically represent and even more challenging to comprehend. On the other hand, it can be established that there is information loss through the image formation process through demodulation, envelope detection and dynamic range limitation, with B-mode images being synthesizable from RF data but not conversely. Another indicator of information reduction is the reduced frequency bandwidth of envelope signals from 2.5 to 1.2 MHz with respect to corresponding demodulated RF data (IQ representation). The CNN experiments suggest that the additional information contained in RF data may contribute to improve liver steatosis classification. These results support recent observations that RF data may contain more useful data than B-mode for liver tissue characterization (Li et al. 2019; Chen et al. 2020; Han et al. 2020).

Third, we observe that CNN-based hepatic steatosis classification is feasible for phase representations of US data, which are usually discarded during the image formation process. This includes the carrier signal of RF data, the phase of the IQ signal or the phase component of the RF data spectrum. Our results indicate that classification based on phase representations could potentially outperform B-mode data classification. This further supports the hypothesis that RF data are more useful than B-

mode data for liver steatosis classification. The value of phase information has been acknowledged in recent works related to signal coherence in beamforming and speed of sound estimation. The speed of sound (SoS) has shown quantitative potential as a biomarker to differentiate liver steatosis stages (Imbault et al. 2017; Popa et al. 2021). Several methods of SoS estimation rely on the measurement of decoherence in US speckle, which is encoded in the phase of the RF data (Imbault et al. 2017; Jakovljevic et al. 2018; Zubajlo et al. 2018). In a more general framework, we hypothesize that changes in the relative position between tissue microstructures affect the interference pattern of tissue backscatter, which consequently affect the phase signal.

We next observe that transforming US data in certain favorable ways for CNN learning may facilitate neural network convergence and classification outcomes. PS representations have been widely used in deep learning literature for classification of wave-based signals, with proposed applications for audio classification, seismography and non-destructive testing (Ebrahimkhanlou and Salamone 2018; Huang et al. 2018; Sinha et al. 2020; Arias-Vergara et al. 2021). The PS is also a fundamental tool in most QUS tissue characterization methods for extracting attenuation measurements and estimating scatterer properties (Mamou and Oelze 2013; Oelze and Mamou 2016). In this study, we observed that PS representations reduced the number of epochs needed for NN convergence with respect to training directly on RF data, translating also into AUROC improvements. PS representations have also more explainable interpretations with respect to the linear system models typically applied in quantitative ultrasound spectroscopy. For instance, in Figure 5c it is possible to interpret PS variations for healthy and steatotic livers in terms of increases in acoustic attenuation. Acoustic attenuation is a known quantitative biomarker for liver steatosis (Ferraioli and Monteiro 2019). Additional time–frequency representation models to STFT need to be explored in future work for PS calculation, for instance, by introducing wavelet decompositions or multitapers to allow for more control of time and frequency smoothing (Arias-Vergara et al. 2021), including parametric or semiparametric representations of the spectrum to further reduce noise in PS estimates (Rao and Swamy 2018), or by introducing higher-dimensional frequency representations.

Finally, we compare the performance of CNN classifiers with quantitative ultrasound metrics. In this study, potential performance improvements were observed for CNN-based RF with respect to individual QUS metrics and multiparametric combinations thereof. Considering the small patient size, it is a remarkable result, which should be verified for larger patient populations. QUS metrics are beneficial because they are explainable and

Table 5. Patient-based results: Scores for patient stratification with QUS metrics and multiparametric (MP) combinations*

	AUROC				Cutoff value according to Youden index			Sensitivity at cutoff				Specificity at cutoff				Accuracy at cutoff†			
	S0 vs. S1S2S3	S0S1 vs. S2S3	S0S1S2 vs. S3	Average	S0 vs. S1S2S3	S0S1 vs. S2S3	S0S1S2 vs. S3	S0 vs. S1S2S3	S0S1 vs. S2S3	S0S1S2 vs. S3	Average	S0 vs. S1S2S3	S0S1 vs. S2S3	S0S1S2 vs. S3	Average	S0 vs. S1S2S3	S0S1 vs. S2S3	S0S1S2 vs. S3	Average
Human annotation	—	—	—	—	—	—	—	1.0	1.0	0.875	0.958	0.923	0.800	0.913	0.879	0.968	0.871	0.903	0.914
Spectral intercept	1.0	0.982	0.962	0.981	25.70	33.98	33.98	1.0	0.909	1.0	0.970	1.0	0.95	0.869	0.940	1.0	0.935	0.903	0.946
Nakagami parameter	1.0	0.968	0.924	0.964	0.874	0.923	0.923	1.0	1.0	1.0	1.0	1.0	0.9	0.783	0.894	1.0	0.935	0.839	0.925
Attenuation coefficient	0.979	0.964	0.951	0.965	0.644	0.711	0.867	0.944	0.909	0.875	0.909	1.0	0.950	1.0	0.983	0.968	0.935	0.968	0.957
Spectral slope	1.0	0.955	0.935	0.963	-5.11E-6	-6.09E-6	-6.73E-6	1.0	0.909	0.875	0.928	1.0	0.9	0.957	0.952	1.0	0.903	0.935	0.946
Hepatorenal index	0.996	0.859	0.853	0.903	2.47	2.638	2.652	0.944	1.0	1.0	0.981	1.0	0.75	0.696	0.815	0.968	0.839	0.774	0.860
Midband fit	0.940	0.864	0.845	0.884	7.96	8.60	9.35	0.833	0.818	0.875	0.842	0.923	0.8	0.826	0.850	0.871	0.806	0.839	0.839
Elastic modulus§	0.745	0.670	0.625	0.680	4.684	4.780	4.780	0.889	0.909	1.0	0.933	0.615	0.5	0.478	0.531	0.774	0.645	0.613	0.677
Voigt viscosity§	0.722	0.665	0.648	0.678	1.45	1.5	1.5	0.833	0.909	0.875	0.872	0.692	0.6	0.522	0.605	0.774	0.710	0.613	0.699
Dispersion§	0.588	0.512	0.443	0.514	14.3	14.3	14.3	0.389	0.364	0.375	0.376	0.923	0.8	0.783	0.835	0.613	0.645	0.677	0.645
MP: Trees/Random Forest	1.0	0.991	0.967	0.986	—	—	—	1.0	0.909	1.0	0.970	1.0	1.0	0.957	0.986	1.0	0.968	0.968	0.979
MP: Gaussian process	0.996	0.982	0.978	0.985	—	—	—	0.944	0.909	1.0	0.951	1.0	0.913	1.0	0.971	0.968	0.968	0.935	0.957
MP: Multilayer perceptron	1.0	0.945	0.935	0.960	—	—	—	1.0	0.909	1.0	0.970	1.0	0.95	0.870	0.940	1.0	0.935	0.903	0.946
MP: Multilinear	0.991	0.918	0.929	0.946	—	—	—	0.944	1.0	1.0	0.981	1.0	0.75	0.739	0.830	0.968	0.839	0.806	0.871

AUROC = area under the receiver operating characteristic curve; MRI-PDFF = magnetic resonance imaging proton density fat fraction.

* S0 vs. (S1, S2, S3); (S0, S1) vs. (S2, S3); and (S0, S1, S2) vs. S3, respectively, correspond to 6.4%, 17.4% and 22.1% magnetic resonance imaging proton density fat fraction cutoffs.

† Assigning all instances to the majority class (no merit classifier) results in accuracies of 0.581 (S0 vs. S1, S2, S3), 0.645 (S0, S1 vs. S2, S3) and 0.742 (S0, S1, S2 vs. S3). Human annotation scores based on B-mode images are provided as a reference.

§ Attenuation coefficient is measured in dB/cm/MHz, elastic modulus in kPa, Voigt viscosity in Pa·s and dispersion in m/s per 100–150 Hz.

generalizable across ultrasound systems and configurations. On the other hand, they require dedicated efforts in data acquisition, often requiring phantoms to establish a baseline for each system and data acquisition protocol, and are sensitive to ROI selection (Pirmoazen et al. 2020). The use of CNN architectures based on RF data representations may alleviate these issues, although for generalizability to different ultrasound systems, variations of RF data with respect to different equipment still need to be considered, for instance, because of the use of different beamforming approaches or different noise statistics in the data acquisition (Han et al. 2020; El Kaffas et al. 2021).

The data set analyzed in this work represents a highly curated case, with high-end ultrasound system hardware and image acquisitions and system configurations and with QUS metrics carefully reviewed by experienced radiologists. In real clinical scenarios, the quality of the data may exhibit greater variability. In this sense, the use of raw data representations, such as US RF data, may provide additional benefits for quantitative US applications based on CNNs, as they are less dependent on the proprietary filters applied by each manufacturer for B-mode imaging and less affected by the gain adjustment settings applied to visualize the final B-mode image.

Limitations

There are several limitations to this study. First, the small patient population and class imbalance make it difficult to establish comparative statements about hepatic steatosis classification performance for different MRI-PDFF cutoff thresholds or to compare the performance of different US data representations for MRI-PDFF regression. Second, individual US patches were labeled to the average MRI-PDFF value of the patient, and patient scores were calculated by adding up individual patch classification probabilities. This approximation was supported by considering the same cumulative liver region in both US and MRI-PDFF measurements and because the MRI-PDFF variability among liver segments in NAFLD is not large (Bonekamp et al. 2014). In particular, the standard deviation of MRI-PDFF among liver segments 5 to 8 was on average 1.2%, which corresponds to <3% of the MRI-PDFF range of the patient population. To further improve MRI-US agreement, a registration between MRI and ultrasound slices can be considered in future work. Third, it is difficult to establish that the investigated CNN networks will provide the best possible performances for each given data set stage. We have alleviated this issue by independently testing CNN networks of different dimensionality (1-D, 2-D and 3-D), by using equivalent patch sizes and NN architectures for compared US data representations, by

controlling overfitting between training and validation sets and by performing a consistent hyperparameter to optimize each US data representation stage. Finally, although our observations are highly suggestive of performance differences between US data representations for hepatic steatosis diagnostics, it is also necessary to study how they generalize to other typical US CNN-based classification use cases, such as tumor diagnostics, or other liver pathology such as fibrosis.

CONCLUSIONS

Our study indicates that US data representations may have an impact on CNN hepatic steatosis classification performance. Raw data representations close to the beamformer, such as RF data and the usually discarded phase information, had superior scores compared with the final clinical B-mode images and conventional quantitative ultrasound metrics. Power spectrum representations improved neural network convergence and further improved classification outcomes. Our study suggests that we potentially need to rethink US data representations for tissue quantification in the deep learning era, favoring access to raw data in ultrasound systems for clinical research.

Acknowledgments—S.J.S. was funded by the Gottfried and Julia Bangerter-Rhyner Foundation (Grant No. 0141/2019). A.E.K. is the recipient of a Grand Challenges Canada Rising Star award. We thank Philips for its continued support with data collection on its research systems. The authors also thank Dongwoon Hyun, Leandra Brickson and Raul Blázquez García for the useful deep learning discussions and their support in setting up the computing infrastructure.

SUPPLEMENTARY MATERIALS

Supplementary material associated with this article can be found in the online version at [doi:10.1016/j.ultrasmedbio.2022.05.031](https://doi.org/10.1016/j.ultrasmedbio.2022.05.031).

REFERENCES

- Adavanne S, Politis A, Nikunen J, Virtanen T. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE J Select Top Signal Process* 2019;13:34–48.
- Amin V. Ultrasonic attenuation estimation for tissue characterization. PhD thesis. : Iowa State University; 1989.
- Arias-Vergara T, Klumpp P, Vasquez-Correa JC, Nöth E, Orozco-Arroyave JR, Schuster M. Multi-channel spectrograms for speech processing applications using deep learning methods. *Pattern Anal Appl* 2021;24:423–431.
- Barr RG, Ferraioli G, Palmeri ML, Goodman ZD, Garcia-Tsao G, Rubin J, Garra B, Myers RP, Wilson SR, Rubens D, Levine D. Elastography assessment of liver fibrosis: Society of Radiologists in Ultrasound consensus conference statement. *Radiology* 2015;276:845–861.
- Basavarajappa L, Baek J, Reddy S, Song J, Tai H, Rijal G, Parker KJ, Hoyt K. Multiparametric ultrasound imaging for the assessment of normal versus steatotic livers. *Sci Rep* 2021;11:2655.
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Ser B* 1995;57:289–300.

- Bhatt M, Moussu MAC, Chayer B, Destremes F, Gesnik M, Allard L, Tang A, Cloutier G. Reconstruction of viscosity maps in ultrasound shear wave elastography. *IEEE Trans Ultrason Ferroelectr Freq Control* 2019;66:1068–1078.
- Bonekamp S, Tang A, Mashhood A, Wolfson T, Changchien C, Middleton MS, Clark L, Gamst A, Loomba R, Sirlin CB. Spatial distribution of MRI-determined hepatic proton density fat fraction in adults with nonalcoholic fatty liver disease. *J Magn Reson Imaging* 2014;39:1525–1532.
- Byra M, Han A, Boehringer AS, Zhang YN, O'Brien WD, Erdman JW, Loomba R, Sirlin CB, Andre M. Liver fat assessment in multiview sonography using transfer learning with convolutional neural networks. *J Ultrasound Med* 2022;41:175–184.
- Cha DI, Kang TW, Min JH, Joo I, Sinn DH, Ha SY, Kim K, Lee G, Yi J. Deep learning-based automated quantification of the hepatorenal index for evaluation of fatty liver by ultrasound. *Ultrasonography* 2021;40:565–574.
- Chen JR, Chao YP, Tsai YW, Chan HJ, Wan YL, Tai DI, Tsui PH. Clinical value of information entropy compared with deep learning for ultrasound grading of hepatic steatosis. *Entropy* 2020;22:1006.
- Colantonio S, Salvati A, Caudai C, Bonino F, De Rosa L, Pascali MA, Germanese D, Brunetto MR, Fatta F. A deep learning approach for hepatic steatosis estimation from ultrasound imaging. *Comput Inf Sci* 2021;1463:703–714.
- Deffieux T, Gennisson JL, Bousquet L, Corouge M, Coscinea S, Amroun D, Tripon S, Terris B, Mallet V, Sogni P, Tanter M, Pol S. Investigating liver stiffness and viscosity for fibrosis, steatosis and activity staging using shear wave elastography. *J Hepatol* 2015;62:317–324.
- Ebrahimkhanlou A, Salamone S. Single-sensor acoustic emission source localization in plate-like structures using deep learning. *Aerospace* 2018;5:50.
- El Kaffas A, Wodlinger B, Rusu M, Kamaya A, Fan R, Sonn G. Raw micro-ultrasound tissue characterization using convolutional neural networks to differentiate benign tissue from clinically significant prostate cancer. *J Urol* 2021;206(Suppl 3):e1008.
- Ferraioli G, Soares Monteiro LB. Ultrasound-based techniques for the diagnosis of liver steatosis. *World J Gastroenterol* 2019;25:6053–6062.
- Ferraioli G, Maiocchi L, Raciti MV, Tinelli C, De Silvestri A, Nichetti M, De Cata P, Rondanelli M, Chiovato L, Calliada F, Filice C. Detection of liver steatosis with a novel ultrasound-based technique: a pilot study using MRI-derived proton density fat fraction as the gold standard. *Clin Transl Gastroenterol* 2019;10:e00081.
- Gao F, He Q, Li G, Huang OY, Tang LJ, Wang XD, Targher G, Byrne CD, Luo JW, Zheng MH. A novel quantitative ultrasound technique for identifying non-alcoholic steatohepatitis. *Liver Int* 2022;42:80–91.
- Gogate M, Dashtipour K, Adeel A, Hussain A. CochleaNet: A robust language-independent audio-visual model for real-time speech enhancement. *Inf Fusion* 2020;63:273–285.
- Groppe D, Benjamini & Hochberg/Yekutieli false discovery rate control procedure for a set of statistical tests. MATLAB Central File Exchange. Available at: https://www.mathworks.com/matlabcentral/fileexchange/27418-fdr_bh. Accessed July, 2022.
- Han A, Boehringer AS, Zhang YN, Montes V, Andre MP, Erdman JW, Loomba R, Sirlin CB, O'Brien WD. Improved assessment of hepatic steatosis in humans using multi-parametric quantitative ultrasound. *Proc IEEE Int Ultrason Symp* 2019;1819–1822.
- Han A, Byra M, Heba E, Andre MP, Erdman JW, Loomba R, Sirlin CB, O'Brien WD. Noninvasive diagnosis of nonalcoholic fatty liver disease and quantification of liver fat with radiofrequency ultrasound data using one-dimensional convolutional neural networks. *Radiology* 2020;295:342–350.
- Hernaez R, Lazo M, Bonekamp S, Kamel I, Brancati FL, Guallar E, Clark JM. Diagnostic accuracy and reliability of ultrasonography for the detection of fatty liver: A meta-analysis. *Hepatology* 2011;54:1082–1090.
- Hoskins PR. Diagnostic ultrasound: Physics and equipment. London: Greenwich Medical Media Limited; 2003.
- Huang L, Li J, Hao H, Li X. Micro-seismic event detection and location in underground mines by using convolutional neural networks (CNN) and deep learning. *Tunnelling Underground Space Technol* 2018;81:265–276.
- Imbault M, Faccineto A, Osmanski BF, Tissier A, Deffieux T, Gennisson JL, Vilgrain V, Tanter M. Robust sound speed estimation for ultrasound-based hepatic steatosis assessment. *Phys Med Biol* 2017;62:3582–3598.
- Jakovljevic M, Hsieh S, Ali R, Chau Loo Kung G, Hyun D, Dahl JJ. Local speed of sound estimation in tissue using pulse-echo ultrasound: Model-based approach. *J Acoust Soc Am* 2018;144:254–266.
- Jeon SK, Joo I, Kim SY, Jang JK, Park J, Park HS, Lee ES, Lee JM. Quantitative ultrasound radiofrequency data analysis for the assessment of hepatic steatosis using the controlled attenuation parameter as a reference standard. *Ultrasonography* 2021;40:136–146.
- Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM* 2017;60:84–90.
- Lee DH, Cho EJ, Bae JS, Lee JY, Yu SJ, Kim H, Lee KB, Han JK, Choi BI. Accuracy of two-dimensional shear wave elastography and attenuation imaging for evaluation of patients with nonalcoholic steatohepatitis. *Clin Gastroenterol Hepatol* 2020;19:797–805.
- Li W, Huang Y, Zhuang BW, Liu GJ, Hu HT, Li X, Liang JY, Wang Z, Huang XW, Zhang CQ, Ruan SM, Xie XY, Kuang M, De Lu M, Da Chen L, Wang W. Multiparametric ultrasonics of significant liver fibrosis: A machine learning-based analysis. *Eur Radiol* 2019;29:1496–1506.
- Liu S, Wang Y, Yang X, Lei B, Liu L, Li SX, Ni D, Wang T. Deep learning in medical ultrasound analysis: A review. *Engineering* 2019;5:261–275.
- Mamou J, Oelze ML. Quantitative ultrasound in soft tissues. Dordrecht: Springer; 2013.
- Marjot T, Moolla A, Cobbolt JF, Hodson L, Tomlinson JW. Nonalcoholic fatty liver disease in adults: Current concepts in etiology, outcomes, and management. *Endocr Rev* 2020;41:66–117.
- Mohd Sagheer SV, George SN. A review on medical image denoising algorithms. *Biomed Signal Process Control* 2020;61:102036.
- Nightingale K, Rouze N, Rosenzweig S, Wang M, Abdelmalek M, Guy C, Palmeri M. Derivation and analysis of viscoelastic properties in human liver: impact of frequency on fibrosis and steatosis staging. *IEEE Trans Ultrason Ferroelectr Freq Control* 2015;62:165–175.
- Oelze M, Mamou J. Review of quantitative ultrasound: envelope statistics and backscatter coefficient imaging and contributions to diagnostic ultrasound. *IEEE Trans Ultrason Ferroelectr Freq Control* 2016;63:336–351.
- Ormachea J, Parker KJ. Comprehensive viscoelastic characterization of tissues and the inter-relationship of shear wave (group and phase) velocity, attenuation and dispersion. *Ultrasound Med Biol* 2020;46:3448–3459.
- Pearson A, Dujardin PA, d'Alterroche L, Patat F, Scotto B, Dujardin F, Bastard C, Miette V, Sandrin L, Remenieras JP. Vibration-controlled transient elastography for noninvasive evaluation of liver steatosis. *Med Phys Med Phys* 2022;49:1507–1521.
- Pirmoazen AM, Khurana A, Kaffas A El, Kamaya A. Quantitative ultrasound approaches for diagnosis and monitoring hepatic steatosis in nonalcoholic fatty liver disease. *Theranostics* 2020;10:4277–4289.
- Pirmoazen AM, Khurana A, Loening AM, Liang T, Shamdasani V, Xie H, ElKaffas A, Kamaya A. Diagnostic performance of 9 quantitative ultrasound parameters for detection and classification of hepatic steatosis in nonalcoholic fatty liver disease. *Invest Radiol* 2022;57:23–32.
- Popa A, Bende F, Şirli R, Popescu A, Bâldea V, Lupuşoru R, Cotrău R, Fofiu R, Foncea C, Sporea I. Quantification of liver fibrosis, steatosis, and viscosity using multiparametric ultrasound in patients with non-alcoholic liver disease: A “real-life” cohort study. *Diagnostics* 2021;11:783.
- Rao KD, Swamy MNS. Spectral analysis of signals. Digit signal processing. Singapore: Springer; 2018 Springer.

- Saba L, Biswas M, Kuppli V, Cuadrado Godia E, Suri HS, Edla DR, Omerzu T, Laird JR, Khanna NN, Mavrogeni S, Protogerou A, Sfrikakis PP, Viswanathan V, Kitas GD, Nicolaides A, Gupta A, Suri JS. The present and future of deep learning in radiology. *Eur J Radiol* 2019;114:14–24.
- Sanabria SJ, Dahl J, Pirmoazen A, Kamaya A, ElKaffas A. Multiparametric ultrasound tissue characterization (MUTC) as a surrogate to magnetic resonance imaging (MRI) for non-alcoholic fatty liver disease (NAFLD) characterization. *Proc IEEE Int Ultrason Symp* 2020; 9251805.
- Sanabria SJ, Dahl J, Pirmoazen A, Kamaya A, ElKaffas A. Learning steatosis staging with two-dimensional convolutional neural networks: comparison of accuracy of clinical B-mode with a co-registered spectrogram representation of RF data. *Proc IEEE Int Ultrason Symp* 2020; 9251329.
- Shi KQ, Tang JZ, Zhu XL, Ying L, Li DW, Gao J, Fang YX, Li GL, Song YJ, Deng ZJ, Wu JM, Tang KF. Controlled attenuation parameter for the detection of steatosis severity in chronic liver disease: A meta-analysis of diagnostic accuracy. *J Gastroenterol Hepatol* 2014;29:1149–1158.
- Sinha H, Awasthi V, Ajmera PK. Audio classification using braided convolutional neural networks. *IET Signal Process* 2020;14:448–454.
- Strauss S, Gavish E, Gottlieb P, Katsnelson L. Interobserver and intraobserver variability in the sonographic assessment of fatty liver. *AJR Am J Roentgenol* 2007;189:1449.
- Sugimoto K, Moriyasu F, Oshiro H, Takeuchi H, Abe M, Yoshimasu Y, Kasai Y, Sakamaki K, Hara T, Itoi T. The role of multiparametric US of the liver for the evaluation of nonalcoholic steatohepatitis. *Radiology* 2020;296:532–540.
- Tang A, Tan J, Sun M, Hamilton G, Bydder M, Wolfson T, Gamst AC, Middleton M, Brunt EM, Loomba R, Lavine JE, Schwimmer JB, Sirlin CB. Nonalcoholic fatty liver disease: MR imaging of liver proton density fat fraction to assess hepatic steatosis. *Radiology* 2013;267:422–431.
- Tang A, Desai A, Hamilton G, Wolfson T, Gamst A, Lam J, Clark L, Hooker J, Chavez T, Ang BD, Middleton MS, Peterson M, Loomba R, Sirlin CB. Accuracy of MR imaging-estimated proton density fat fraction for classification of dichotomized histologic steatosis grades in nonalcoholic fatty liver disease. *Radiology* 2015;274:416–425.
- Wen H, Zheng W, Li M, Li Q, Liu Q, Zhou J, Liu Z, Chen X. Multiparametric quantitative us examination of liver fibrosis: A feature-engineering and machine-learning based analysis. *IEEE J Biomed Heal Inform* 2022;26:715–726.
- Wong VWS, Adams LA, de Lédinghen V, Wong GLH, Sookoian S. Noninvasive biomarkers in NAFLD and NASH—Current progress and future promise. *Nat Rev Gastroenterol Hepatol* 2018;15:461–478.
- Yang Q, Wei J, Hao X, Kong D, Yu X, Jiang T, Xi J, Cai W, Luo Y, Jing X, Yang Y, Cheng Z, Wu J, Zhang H, Liao J, Zhou P, Song Y, Zhang Y, Han Z, Cheng W, Tang L, Liu F, Dou J, Zheng R, Yu J, Tian J, Liang P. Improving B-mode ultrasound diagnostic performance for focal liver lesions using deep learning: A multicentre study. *EBioMedicine* 2020;56 102777.
- Young S, Tariq R, Provenza J, Satapathy SK, Faisal K, Choudhry A, Friedman SL, Singal AK. Prevalence and profile of nonalcoholic fatty liver disease in lean adults: Systematic review and meta-analysis. *Hepatol Commun* 2020;4:953–972.
- Younossi ZM, Blissett D, Blissett R, Henry L, Stepanova M, Younossi Y, Racila A, Hunt S, Beckerman R. The economic and clinical burden of nonalcoholic fatty liver disease in the United States and Europe. *Hepatology* 2016;64:1577–1586.
- Zhou LQ, Wang JY, Yu SY, Wu GG, Wei Q, Bin Deng Y, Wu XL, Cui XW, Dietrich CF. Artificial intelligence in medical imaging of the liver. *World J Gastroenterol* 2019;25:672–682.
- Zubajlo RE, Benjamin A, Grajo JR, Kaliannan K, Kang JX, Bhan AK, Thomenius KE, Anthony BW, Dhyani M, Samir AE. Experimental validation of longitudinal speed of sound estimates in the diagnosis of hepatic steatosis (Part II). *Ultrasound Med Biol* 2018;44:2749–2758.