Radiology

ORIGINAL RESEARCH • GASTROINTESTINAL IMAGING

Comparison of Radiologists and Deep Learning for US Grading of Hepatic Steatosis

Pedro Vianna, MSc • Sara-Ivana Calce • Pamela Boustros, MD • Cassandra Larocque-Rigney, LLB • Laurent Patry-Beaudoin • Yi Hui Luo • Emre Aslan, MD • John Marinos, MD • Talal M. Alamri, MD • Kim-Nhien Vu, MD • Jessica Murphy-Lavallée, MD • Jean-Sébastien Billiard, MD, MSc • Emmanuel Montagnon, PhD • Hongliang Li, PhD • Samuel Kadoury, PhD • Bich N. Nguyen, MD • Shanel Gauthier, MSc • Benjamin Therien, BSc • Irina Rish, PhD • Eugene Belilovsky, PhD • Guy Wolf, PhD • Michaël Chassé, MD, PhD • Guy Cloutier, PhD • An Tang, MD, MSc

From the Department of Imaging and Engineering (P.V., S.I.C., C.L.R., L.P.B., E.M., H.L., S.K., M.C., G.C., A.T.), Laboratory of Biorheology and Medical Ultrasonics (P.V., G.C.), and Clinical Laboratory of Image Processing (E.M., A.T.), Centre de Recherche du Centre Hospitalier de l'Université de Montréal (CRCHUM), Montréal, Canada; Institute of Biomedical Engineering (P.V., G.C.) and Department of Computer Science and Operations Research (S.G., I.R., G.W.), Université de Montréal, Montréal, Canada; Departments of Radiology (S.I.C., P.B., C.L.R., L.P.B., Y.H.L., E.A., J.M., T.M.A., K.N.V., J.M.L., J.S.B., A.T.) and Pathology (B.N.N.), Centre Hospitalier de l'Université de Montréal (CHUM), 1058 rue Saint-Denis, Montréal, QC, Canada H2X 3J4; Department of Computer Engineering, École Polytechnique de Montréal, Montréal, Canada (S.K.); Mila–Quebec Artificial Intelligence Institute, Montréal, Canada (S.G., B.T., I.R., E.B., G.W.); and Department of Computer Science and Software Engineering, Concordia University, Montréal, Canada (B.T., E.B.). Received March 15, 2023; revision requested May 15; final revision received August 13; accepted August 22. **Address correspondence to** A.T. (email: *an.tang@umontreal.ca*).

Fonds de Recherche du Québec–Santé (FRQ-S) and Fondation de l'Association des Radiologistes du Québec (FARQ) Clinical Research Scholarship–Senior Salary Award (FRQS-ARQ no. 298509) to A.T. Institute of Data Valorization (IVADO) operating grant awarded to A.T.

Conflicts of interest are listed at the end of this article.

See also the editorial by Tuthill in this issue.

Radiology 2023; 309(1):e230659 • https://doi.org/10.1148/radiol.230659 • Content codes: GI US AI

Background: Screening for nonalcoholic fatty liver disease (NAFLD) is suboptimal due to the subjective interpretation of US images.

Purpose: To evaluate the agreement and diagnostic performance of radiologists and a deep learning model in grading hepatic steatosis in NAFLD at US, with biopsy as the reference standard.

Materials and Methods: This retrospective study included patients with NAFLD and control patients without hepatic steatosis who underwent abdominal US and contemporaneous liver biopsy from September 2010 to October 2019. Six readers visually graded steatosis on US images twice, 2 weeks apart. Reader agreement was assessed with use of κ statistics. Three deep learning techniques applied to B-mode US images were used to classify dichotomized steatosis grades. Classification performance of human radiologists and the deep learning model for dichotomized steatosis grades (S0, S1, S2, and S3) was assessed with area under the receiver operating characteristic curve (AUC) on a separate test set.

Results: The study included 199 patients (mean age, 53 years ± 13 [SD]; 101 men). On the test set (n = 52), radiologists had fair interreader agreement (0.34 [95% CI: 0.31, 0.37]) for classifying steatosis grades S0 versus S1 or higher, while AUCs were between 0.49 and 0.84 for radiologists and 0.85 (95% CI: 0.83, 0.87) for the deep learning model. For S0 or S1 versus S2 or S3, radiologists had fair interreader agreement (0.30 [95% CI: 0.27, 0.33]), while AUCs were between 0.57 and 0.76 for radiologists and 0.73 (95% CI: 0.71, 0.75) for the deep learning model. For S2 or lower versus S3, radiologists had fair interreader agreement (0.37 [95% CI: 0.33, 0.40]), while AUCs were between 0.52 and 0.81 for radiologists and 0.67 (95% CI: 0.64, 0.69) for the deep learning model.

Conclusion: Deep learning approaches applied to B-mode US images provided comparable performance with human readers for detection and grading of hepatic steatosis.

Published under a CC BY 4.0 license.

Supplemental material is available for this article.

Nonalcoholic fatty liver disease (NAFLD) is the most common cause of chronic liver disease, with an estimated global prevalence of 38% (1). NAFLD can progress to nonalcoholic steatohepatitis (NASH) in 16% of patients (1) and to cirrhosis in 3% of patients (2). NAFLD is associated with obesity and type 2 diabetes, which are globally increasing (3). The anticipated increase in NAFLD will have implications on health care systems, as the higher demand for imaging studies will require additional equipment, trained personnel, infrastructure, and likely assistance using artificial intelligence.

Hepatic steatosis is defined by the presence of vacuoles of fat within hepatocytes, which constitutes a hallmark histopathologic feature. Increased grades of steatosis are associated with worse outcomes in liver diseases (4). Currently, liver biopsy and MRI are respectively considered the historical and noninvasive reference standard techniques for assessment of hepatic steatosis (5–7). Both techniques have limitations associated with cost-effectiveness and availability, hampering their use for screening and monitoring of patients (8).

B-mode US is often used for screening and monitoring of hepatic steatosis due to its wide availability, low cost, and lack of radiation (9,10). However, B-mode US has some limitations. In the setting of chronic liver disease, it may be difficult to attribute the echogenicity of the liver to steatosis,

This copy is for personal use only. To order copies, contact *reprints@rsna.org*

Abbreviations

AUC = area under the receiver operating characteristic curve, NAFLD = nonalcoholic fatty liver disease, NASH = nonalcoholic steatohepatitis

Summary

Deep learning methods applied to B-mode US images showed comparable performance with six human readers in grading hepatic steatosis and may be used as a valuable tool for screening patients.

Key Results

- In a retrospective study of 199 patients with a spectrum of nonalcoholic fatty liver disease, six radiologists visually graded hepatic steatosis on B-mode US images; interreader agreement was fair in detecting patients with any degree of steatosis (S0 vs S1 or higher, 0.34 [95% CI: 0.31, 0.37]), for differentiating no or mild steatosis from those with moderate or more severe steatosis (S0 or S1 vs S2 or S3, 0.30 [95% CI: 0.27, 0.33]), and for grading patients with severe steatosis against all other grades (S2 or lower vs S3, 0.37 [95% CI: 0.33, 0.40]).
- Individual readers graded hepatic steatosis, with areas under the receiver operating characteristic curve (AUC) ranging from 0.49 (95% CI: 0.46, 0.51) to 0.84 (95% CI: 0.70, 0.98) for classifying S0 versus S1 or higher, from 0.57 (95% CI: 0.43, 0.71) to 0.76 (95% CI: 0.64, 0.87) for S0 or S1 versus S2 or S3, and from 0.52 (95% CI: 0.43, 0.60) to 0.81 (95% CI: 0.67, 0.95) for S2 or lower versus S3.
- A deep learning model graded hepatic steatosis with AUCs of 0.85 (95% CI: 0.83, 0.87) for classifying S0 versus S1 or higher, 0.73 (95% CI: 0.71, 0.75) for S0 or S1 versus S2 or S3, and 0.67 (95% CI: 0.64, 0.69) for S2 or lower versus S3.

fibrosis, or both, given that both conditions affect the brightness on gray-scale images (11). Additionally, image acquisition may be affected by technical sources of variability, such as scanner, probe, and settings (9). Moreover, grading of hepatic steatosis at US is reader-dependent (9), with high intraobserver and interobserver variability (10,12). Improving the ability to objectively grade hepatic steatosis on B-mode US images would provide a scalable option for the diagnosis and follow-up of hepatic steatosis.

The continuous development of artificial intelligence technologies, specifically advances in deep learning in the past decade, has shown potential for diverse applications (13–15). Deep learning techniques have been increasingly popular in medical imaging (16), aiming to automate computer vision tasks such as classification, detection, and segmentation. Deep learning has been recently investigated for steatosis detection and classification (17,18). However, to our knowledge, prior studies have not compared its performance with human readers, including fellowship-trained abdominal radiologists, on the same test data set.

The purpose of this study was to evaluate the classification agreement and diagnostic performance of radiologists and deep learning models applied to B-mode US images for grading hepatic steatosis in NAFLD, using biopsy as the reference standard.

Materials and Methods

Study Design

The Centre Hospitalier de l'Université de Montréal institutional review board approved this retrospective, cross-sectional, case-control, diagnostic, single-site model creation study. Requirement for informed consent was waived, and consent was obtained for data access.

Patient Selection

Patients were identified in a convenience series among those who underwent B-mode abdominal US and liver biopsy within 1 year of each other for suspected or confirmed chronic liver disease. The study period was across 9 years (September 2010 to October 2019). Patients were included if they had a histopathologic diagnosis of NAFLD, NASH, or NASH-related cirrhosis and excluded if they had any other causes of chronic liver disease. Control patients were included when there was no other lesion present or if they had a minor nonspecific lesion (such as cholestasis, dysplasia, nonspecific inflammation, and microgranulomas). Therefore, they had less than 5% hepatocytes containing macrovesicular steatosis and no inflammation or fibrosis. To protect health information, patient identifiers were encrypted using salt and pepper cryptographic hashing, nominal information and headers were cropped from images, and pathology reports were anonymized using randomization algorithms.

Index Test

The index test was B-mode abdominal US from seven different scanners, including iU22 (Philips), Aplio 500 and i800 (Canon Medical Systems), Acuson S2000 and S3000, (Siemens Healthineers) Sequoia (Siemens Healthineers), and LOGIQ E9 (GE HealthCare). Images were acquired according to the institutional clinical US protocol. Image analysts (S.I.C., Y.H.L., C.L.R., and L.P.B., medical students in training with 1–3 years of experience) reviewed images associated with selected pathology reports and excluded images in Doppler mode or dual-display view or images with severe visualization limitations according to the Liver Imaging Reporting and Data System (19).

Reference Standard

Liver biopsies were interpreted using NASH score (20), including steatosis grade, ballooning grade, lobular inflammation grade, and fibrosis stage. Biopsies were performed via a transcutaneous approach with US guidance using 16- or 18-gauge core needles (21). Medical students (S.I.C. and Y.H.L.) extracted steatosis grades from pathology reports (22,23), which determined steatosis as the percentage of hepatocytes containing fat macrovesicles according to the following ordinal scale: none (<5%, grade S0), mild (5%–33%, grade S1), moderate (34%–66%, grade S2), and severe (>66%, grade S3). Fat fraction is a common data element for radiology, and histologic scores were dichotomized as follows: S0 versus S1 or higher; S0 or S1 versus S2 or S3; and S2 or lower versus S3.

Reader Subset Selection

Considering all eligible patients, a subset was created for reader assessment by selecting those with at least one good quality image showing *(a)* the liver and kidney, *(b)* portal vein, and *(c)* hepatic vein. Image quality was considered good when there were no or minimal limitations (US visualization score A) or moderate limitations (US visualization score B) according to the Liver Imaging Reporting and Data System US (19). Not all patients

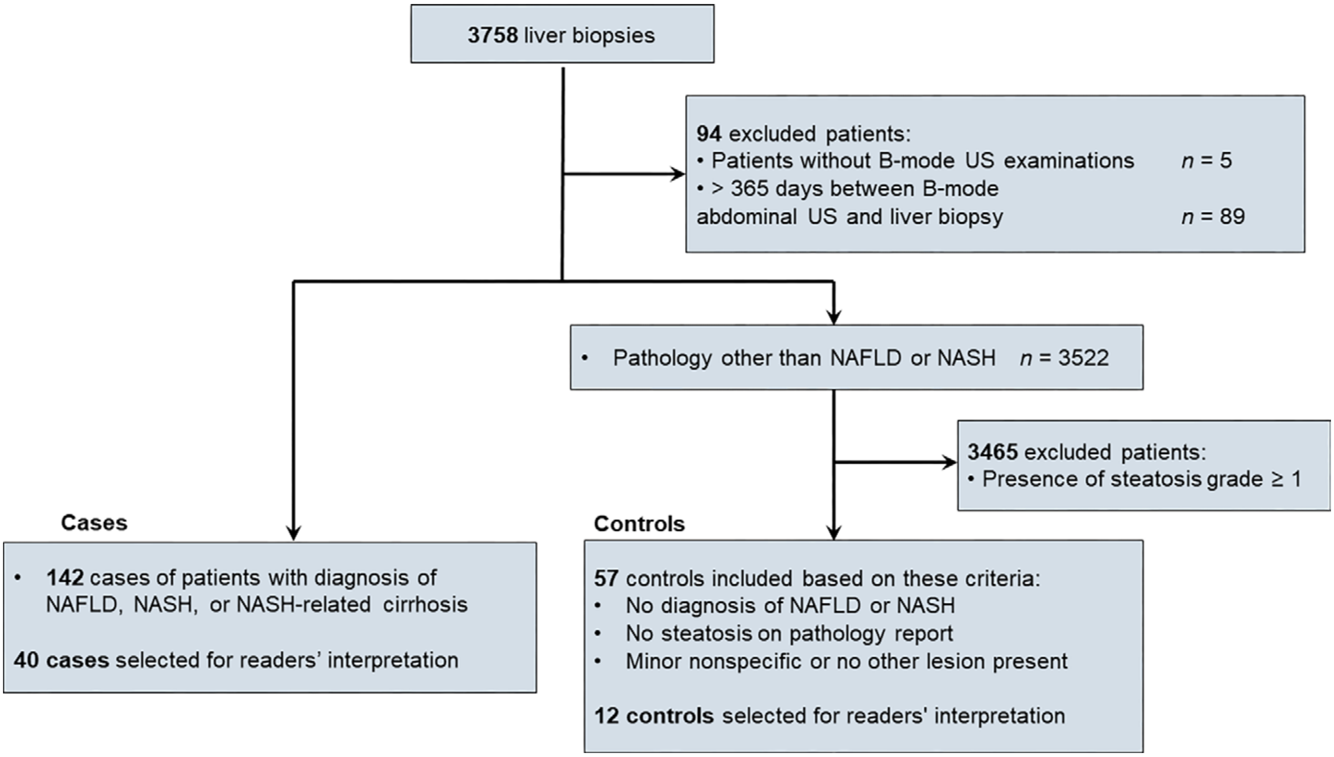
3758 liver biopsies
94 excluded patients:
• Patients without B-mode US examinations n = 5
• > 365 days between B-mode abdominal US and liver biopsy n = 89
• Pathology other than NAFLD or NASH n = 3522
3465 excluded patients:
• Presence of steatosis grade ≥ 1
Cases
• 142 cases of patients with diagnosis of NAFLD, NASH, or NASH-related cirrhosis
40 cases selected for readers' interpretation
Controls
57 controls included based on these criteria:
• No diagnosis of NAFLD or NASH
• No steatosis on pathology report
• Minor nonspecific or no other lesion present
12 controls selected for readers' interpretation

Figure 1: Flowchart of patient selection. NAFLD = nonalcoholic fatty liver disease, NASH = nonalcoholic steatohepatitis.

in the database had appropriate images meeting the three requirements. The sampling maintained a similar proportion of steatosis grades across the selected patients. The final number of patients included in the study was based on feasibility and convenience for the readers.

Image Interpretation

To assess intra- and interobserver agreement and diagnostic performance of steatosis grading by humans, six readers independently graded steatosis. The six readers had increasing levels of expertise in abdominal sonography: junior resident, senior resident, medical fellow, junior, midcareer, and senior fellowship-trained abdominal radiologists (J.M., E.A., T.M.A., K.N.V., J.M.L., and J.S.B., respectively a 2nd-year resident, 5th-year resident, 1st-year abdominal fellow, and radiologists with 4, 13, and 29 years of experience). Before image interpretation, a radiologist (A.T., with 18 years of experience) provided instructions on how to grade liver steatosis according to features such as degree of ultrasound attenuation, echogenicity of liver parenchyma relative to the adjacent kidney, visibility and clarity of vessel border definition, and visualization of the diaphragm. The readers graded steatosis on an ordinal scale from S0 to S3. Steatosis grading was based on visual assessment of figure collages presenting three views per patient that included the liver, right kidney, right portal vein, and hepatic veins. To minimize recall bias, readers graded steatosis severity again 2 weeks later, but with figure collages of the same patients presented in a randomized order. During both reading sessions, the readers were blinded to the corresponding pathology reports as well as to the steatosis grading assigned in other readings.

Data Set Selection for Deep Learning

The entire data set was split into training, validation, and test sets. The independent test set contained 15% of the data, while the remainder was used for fivefold cross-validation. For each fold, 15% of the data were used for validation, with the remaining 85% were used for training. The data were partitioned at the patient level to prevent data leakage, and the proportion of steatosis grades was similar for each set. Preprocessing was applied by cropping the display to remove all screen information and by resizing images to a determined input size with use of bicubic interpolation.

To allow a direct comparison between the deep learning model and readers, a second split of the data set was created. Specifically, the test set was the same subset evaluated by the readers, while the remainder was used for training, without cross-validation. Data partitioning methods are presented in Figure S1.

Deep Learning Models

Training sets containing images and biopsy scores were used to train binary classifiers for predicting hepatic steatosis grades. To standardize the scale and distribution of the input data, the mean and SD were computed across all images in the training set. Standardization of all sets was performed by subtracting the mean and dividing by the SD of the training set. No data augmentation was used in this study.

The VGG16 (24) architecture was used for the binary classification task, with transfer learning and dropout layers (Table S1). VGG16 is a widely adopted architecture for classification, featuring a moderate depth and availability of pretrained weights on the ImageNet data set (24). VGG16 performance was compared with that of ResNet-50 and Inception-v3.

During fivefold cross-validation, after each training epoch, the validation set provided an evaluation of the models. Hyperparameter tuning and configuration changes were evaluated in the cross-validation setting by varying input size, batch size, optimizer, learning rate, loss function, and dropout to determine the optimal settings (Table S2). To monitor training performance, the lowest validation loss was used, and the epoch with the lowest average loss across the five folds was selected. Subsequently, only the best-performing configuration was used for the test sets.

All layers of the network were allowed to update during training. Input size was 128 × 128, and batch size was 32. Stochastic gradient descent with a learning rate of 0.0001 was used as the optimizer. Binary cross-entropy was selected as the loss function, and the dropout rate was set at 0.5 on the fully connected layers before the softmax.

Class activation mapping (25) was implemented to investigate the influence of areas of an image on the output of a trained classifier. A weighted activation map was generated for any given image, allowing for localizable deep features that can be interpreted to better understand the class identified by the trained model.

Software was developed in Python (version 3.7; Python Software Foundation) using NumPy (version 1.19.2), Pillow (version 6.2.0), Pandas (version 0.25.1), scikit-learn (version 1.0.2), Keras (version 2.2.4), and TensorFlow (version 1.14.0) libraries. The code is publicly available at *https://github.com/LCTI-AnTang/binary_steatosis_classifier*.

Statistical Analysis

The kappa coefficient, κ, was used to measure intra- and interreader agreement. Multireader Fleiss κ was used for the interreader agreement between multiple readers, and Cohen κ was used when performing paired comparisons. A κ value less than 0.21 was considered poor agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81–1.00, excellent agreement. To investigate the relationship between the years of experience of the readers and their accuracy and intrareader agreement, linear regression was used. Steatosis grade predictions of readers and deep learning models were compared with reference standard values, with measurements of sensitivity, specificity, positive predictive value, negative predictive value, and area under the receiver operating characteristic curve (AUC). AUC was the preferred metric for comparisons, as it is less sensitive to class imbalance and allows for comparisons in different thresholds. Metrics calculated for deep learning architectures considered the cutoff point with the highest Youden index on the receiver operating characteristic curve. For the independent test set and the reader set, results were obtained for all images in single tests. All statistical analyses were performed and implemented in Python by using the SciPy package (version 1.3.1), and 95% CIs for AUC were calculated with the DeLong method. To compare the results of the deep learning model and the six readers, DeLong tests were conducted, pairing each reading session with the model. The power of the test was evaluated by using the fixed sample size and the effect size for the comparisons. To investigate the potential confounding effects of fibrosis, inflammation, and ballooning on the steatosis classification, Spearman rank correlation was used. The level of significance was set at $P < .05$ for all tests. Bonferroni correction was applied for the comparison between the deep learning model and reading sessions.

Table 1: Characteristics of the 199 Patients Included in the Study

Characteristic	Value
Sex	
M	101 (50.8)
F	98 (49.2)
Age (y)	
Mean ± SD	53 ± 13
Median and range	55 (20–81)
IQR	45–62
Body mass index*	30.5 ± 7.8
Time between examination and biopsy (d)	
Mean ± SD	76 ± 110
Median and range	15 (0–364)
IQR	1–135
Steatosis grade	
S0 (<5% of hepatocytes involved)	57 (28.6)
S1 (5%–33% of hepatocytes involved)	87 (43.7)
S2 (34%–66% of hepatocytes involved)	25 (12.6)
S3 (>66% of hepatocytes involved)	30 (15.1)
Lobular inflammation	
0 (no foci)	3 (1.5)
1 (<2 foci per 200 × field)	98 (49.2)
2 (2–4 foci per 200 × field)	28 (14.1)
3 (>4 foci per 200 × field)	0 (0)
Not reported	70 (35.2)
Hepatocellular ballooning	
0 (none)	13 (6.5)
1 (few balloon cells)	84 (42.2)
2 (many cells or prominent)	32 (16.1)
Not reported	70 (35.2)
Fibrosis	
F0 (none)	21 (10.6)
F1 (perisinusoidal or periportal)	27 (13.6)
F2 (perisinusoidal and periportal)	18 (9.0)
F3 (bridging fibrosis)	28 (14.1)
F4 (cirrhosis)	39 (19.6)
Not reported	66 (33.2)

Note.—Unless otherwise specified, data are numbers of patients, with percentages in parentheses.

* Body mass index was calculated as patient weight in kilograms divided by patient height in meters squared.

Results

Patient Characteristics

Figure 1 shows the study flowchart. The initial database contained 3758 biopsy reports. Ninety-four patients without US examinations linked to the reports or with a delay longer than 1 year between the biopsy and US examination were excluded, and 3465 reports were excluded for reporting a pathologic

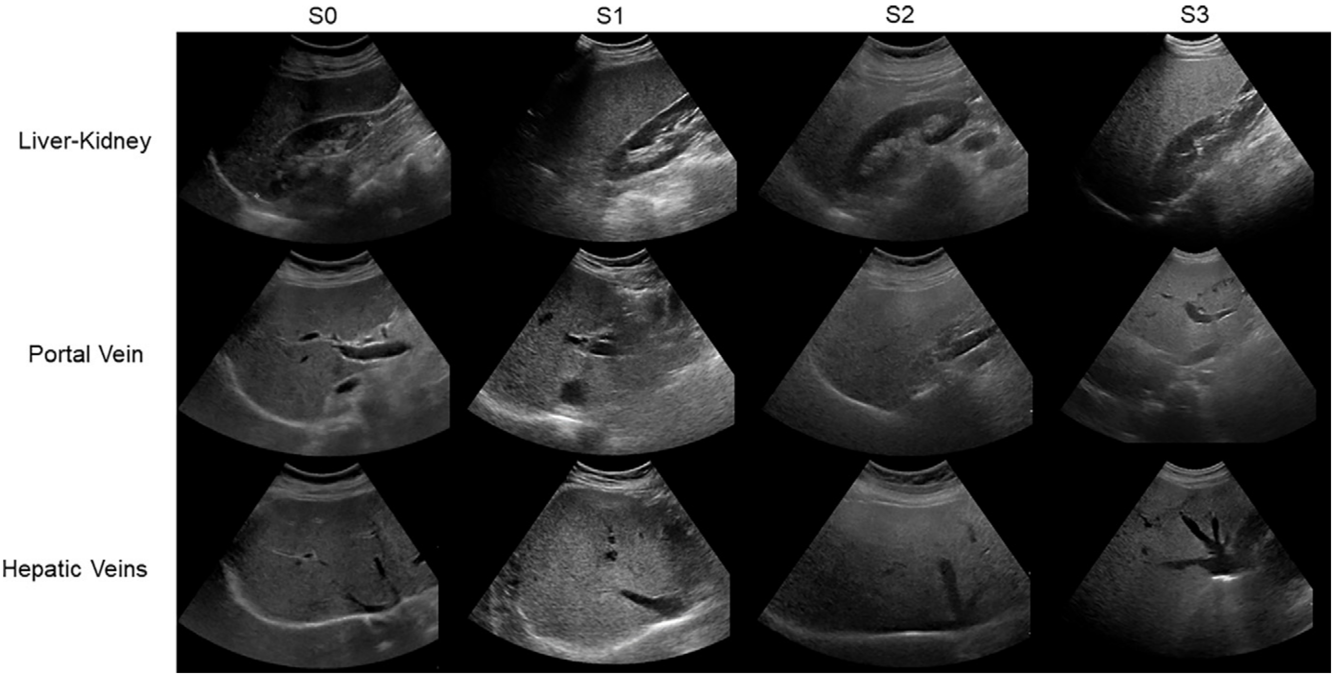

Figure 2: Examples of representative B-mode US images. Hepatic steatosis grades are classified as none (grade S0), mild (S1), moderate (S2), and severe (S3). Biopsy-proven steatosis grades, from S0 to S3, indicate the amount of fat according to pathology reports, which were used as the reference standard. S0 images are in a 50-year-old female patient, S1 images are in a 34-year-old female patient, S2 images are in a 46-year-old female patient, and S3 images are in a 35-year-old female patient.

abnormality other than NAFLD or NASH. Table 1 summarizes patient characteristics. Table S3 presents the steatosis distributions across the sets, and Table S4 describes the patient distributions across the US scanners. The data set for this study consisted of 7529 B-mode US images from 142 patients (mean age, 53 years ± 13 [SD]; 70 men) with hepatic steatosis and 57 control patients (54 years ± 13; 31 men). The data set included 57 patients and 966 images for grade S0, 87 patients and 3312 images for grade S1, 25 patients and 1683 images for grade S2, and 30 patients and 1568 images for grade S3. The majority of the images (5765 of 7529 [77%]) were collected using transducers C5–1, PVI-475BX, or PVT-375BT, which have frequency ranges from 1.0 to 6.2 MHz. Most images in the data set were acquired using frequencies of 2.5 MHz, 4.0 MHz, or 4.5 MHz. The average body mass index (calculated as patient weight in kilograms divided by patient height in meters squared) was 30.5 ± 7.8, and 44% of the patients (87 of 199) had mild steatosis. Figure 2 shows representative B-mode images of different steatosis grades. The subset selected for reader assessment contained 52 patients, including 12 patients with grade S0, 17 patients with grade S1, 11 patients with grade S2, and 12 patients with grade S3.

Intra- and Interreader Agreement

Table 2 shows intra- and interreader agreements for classification of visual grading of steatosis. Based on κ values, intrareader and interreader agreements were moderate (0.45 [95% CI: 0.32, 0.59]) and fair (0.34 [95% CI: 0.31, 0.37]), respectively, for classifying steatosis as S0 versus S1 or higher; moderate (0.56 [95% CI: 0.45, 0.66]) and fair (0.30 [95% CI: 0.27, 0.33]) for S0 or S1 versus S2 or S3; and moderate (0.44 [95% CI: 0.28, 0.60]) and fair (0.37 [95% CI: 0.33, 0.40]) for S2 or lower versus S3. Table S5 shows the κ values for all reading pairs. The impact of years of experience on intrareader agreement was weak ($R^2 = 0.09$) (Appendix S1). Across all steatosis grades, the mean interreader κ was 0.25 (95% CI: 0.19, 0.32) for the two residents and 0.17 (95% CI: 0.10, 0.23) for the three radiologists. The mean interreader κ was 0.24 (95% CI: 0.21, 0.27) between the medical fellow and radiologists and 0.19 (95% CI: 0.15, 0.23) between the medical fellow and residents. There was no relationship between the level of training and interreader agreements.

Diagnostic Performance of Readers

Table 3 shows the diagnostic performance of the readers. For classifying S0 versus S1 or higher, readers had AUCs ranging from 0.49 (95% CI: 0.46, 0.51) to 0.84 (95% CI: 0.70, 0.98). For classifying S0 or S1 versus S2 or S3, readers had AUCs ranging from 0.57 (95% CI: 0.43, 0.71) to 0.76 (95% CI: 0.64, 0.87). For classifying S2 or lower versus S3, readers had AUCs ranging from 0.52 (95% CI: 0.43, 0.60) to 0.81 (95% CI: 0.67, 0.95). Table S6 and Figure S2 show detailed results for each reader. There was no strong relationship between the level of training and the accuracy of readers ($R^2 < 0.15$).

Diagnostic Performance of Deep Learning

In this study, VGG16 outperformed ResNet-50 and Inception-v3 in terms of AUC. The deep learning model performance was assessed on the independent test set and on the reader set. Detailed results for both test splits can be found in Table 4. Figure 3 shows representative class activation maps generated using trained models on images with different outcomes on the three dichotomized classification tasks (Appendix S1), and receiver operating characteristic curves are given in Figure 4. Confusion matrixes are presented in Figure S3.

Comparing the results of the reading sessions and deep learning (Tables 3, 4), in S0 versus S1 or higher, the model significantly outperformed 11 of 12 readings (*P* < .001 for all), except for the highest-performing reading (*P* = .84). In S0 or S1 versus S2 or S3, the model outperformed only the lowest-performing reading (*P* = .03) when not adjusted for multiple comparisons. When Bonferroni correction was applied for 12 comparisons, no statistical difference was observed. In S2 or lower versus S3, the model outperformed only one reading (*P* = .002; *P* ≥ .04 for all remaining comparisons). For all comparisons with significant difference, the power of the test was higher than 80%. Secondary analysis was performed to observe the influence of time delays between US examinations and biopsies and is presented in Table S7.

Confounding Variables

Univariable correlation coefficients demonstrated that steatosis classification was correlated with fibrosis stage in S2 or lower versus S3 (*P* < .001), but not for S0 versus S1 or higher (*P* = .40) or for S0 or S1 versus S2 or S3 (*P* = .61). Unweighted sum of inflammation and ballooning was correlated with steatosis classification for S2 or lower versus S3 (*P* = .01) and for S0 versus S1 or higher (*P* = .02) while not showing any correlation in S0 or S1 versus S2 or S3 (*P* = .48). The detailed results for confounding variables are presented in Table S8.

Table 2: Intra- and Interreader Agreement for Visual Grading of Steatosis on B-Mode US Images

Steatosis Grade Classification and Comparison	κ Statistic	Agreement
Four grades separately		
Intrareader	0.38 (0.26, 0.49)	Fair
Interreader	0.22 (0.20, 0.24)	Fair
S0 vs S1 or higher		
Intrareader	0.45 (0.32, 0.59)	Moderate
Interreader	0.34 (0.31, 0.37)	Fair
S0 or S1 vs S2 or S3		
Intrareader	0.56 (0.45, 0.66)	Moderate
Interreader	0.30 (0.27, 0.33)	Fair
S2 or lower vs S3		
Intrareader	0.44 (0.28, 0.60)	Moderate
Interreader	0.37 (0.33, 0.40)	Fair

Note.—Data are average κ statistics, with 95% CIs in parentheses. Hepatic steatosis grades are classified as none (grade S0), mild (S1), moderate (S2), and severe (S3). Multireader Fleiss κ was used for the interreader agreement across all readers, and Cohen κ was used for intrareader agreement in paired readings. A κ value of less than 0.21 was considered poor agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81–1.00, excellent agreement.

Discussion

This study evaluated intra- and interreader agreements and diagnostic performance between six readers with different levels of training and performed a head-to-head comparison with a deep learning model in grading hepatic steatosis on B-mode US images in patients with a spectrum of nonalcoholic fatty liver disease and control patients without steatosis, using histopathologic findings as the reference standard for both groups. For S0 versus S1 or higher, the readers achieved fair interreader agreement (κ = 0.34) and performance ranging between 0.49 and 0.84 in area under the receiver operating characteristic curve (AUC), while the deep learning model achieved an AUC of 0.85, which was better than 11 of 12 readings (*P* < .001). In S0 or S1 versus S2 or S3, the readers achieved fair interreader agreement (0.30) and performance ranging between 0.57 and 0.76 in AUC, while the deep learning model achieved an AUC of 0.73, with no statistically significant difference. In S2 or lower versus S3, the readers achieved fair interreader agreement (0.37) and performance ranging between 0.52 to 0.81 in AUC, while the deep learning model achieved an AUC of 0.67, which was better than one reading (*P* = .002). For human visual classification, the sensitivity was the highest when distinguishing patients with steatosis from those without steatosis. This could be helpful in the clinical setting to screen patients with any level of fat.

Automatic deep learning classification of steatosis grades at B-mode US achieved a higher AUC for detection of steatosis than for grading severity of steatosis (ie, S0 or S1 vs S2 or S3 and S2 or lower vs S3). This is consistent with the prior literature reporting high performance for detection of hepatic steatosis (17,18). Using US images obtained with a single system and using biopsy reports as the reference standard, Byra et al (17) reported an AUC of 0.98 in 55 patients with obesity, and using visual assessment as reference, Cao et al (18) reported an AUC of 0.93 in 240 patients. Direct comparison with other studies is limited owing to the differences in the data sets, although the results for steatosis detection by the deep learning model in the two separate test sets in our study (AUC, 0.85 for the reader set and 0.98 for the independent test set) are similar to those reported in the literature.

While there was substantial variation between readers, there was no evidence of a difference across levels of experience for all readers. Prior studies reporting κ for readers assessing fatty liver at B-mode US addressed either detection or grading of steatosis. For detection of steatosis, prior studies reported intrareader κ values ranging from 0.54 to 0.89 (12,26–28) and interreader κ from 0.40 to 1.00 (12,27,29–31). For grading of steatosis, prior studies reported intrareader κ from 0.58 to 0.68 (12,26) and interreader κ from 0.49 to 0.93 (12,32–34). However, most studies used readers with similar levels of expertise, except one (28), which was the only study with more than three readers.

Our study presents several advantages over the current literature, including the use of images from multiple equipment types and biopsy reports as the reference standard. Additionally, model results are compared with various human readers. A recent study used radiofrequency-based techniques for steatosis detection and grading (35), but our proposed method uses existing B-mode images from various manufacturers and produces real-time predictions without the need for dedicated software or calibration phantoms.

Our study had limitations. First, the potential confounding effect of fibrosis, inflammation, and ballooning on the

Table 3: Diagnostic Performance of Readers Grading Hepatic Steatosis on B-Mode US Images

Steatosis Grade Classification*	A1	A2	B1	B2	C1	C2	D1	D2	E1	E2	F1	F2
S0 (*n* = 12) vs S1 or higher (*n* = 40)	0.57 (0.44, 0.69)	0.54 (0.41, 0.67)	0.58 (0.46, 0.70)	0.52 (0.43, 0.62)	0.55 (0.43, 0.68)	0.84 (0.70, 0.98)	0.52 (0.43, 0.62)	0.58 (0.46, 0.70)	0.53 (0.44, 0.63)	0.53 (0.44, 0.63)	0.49 (0.46, 0.51)	0.53 (0.44, 0.63)
S0 or S1 (n = 29) vs S2 or S3 (*n* = 23)	0.64 (0.54, 0.74)	0.67 (0.56, 0.78)	0.65 (0.53, 0.77)	0.60 (0.46, 0.73)	0.72 (0.60, 0.84)	0.76 (0.64, 0.87)	0.73 (0.62, 0.83)	0.71 (0.61, 0.81)	0.60 (0.46, 0.73)	0.57 (0.43, 0.71)	0.63 (0.50, 0.75)	0.66 (0.52, 0.79)
S2 or lower (*n* = 40) vs S3 (*n* = 12)	0.71 (0.55, 0.86)	0.81 (0.67, 0.95)	0.64 (0.48, 0.80)	0.71 (0.56, 0.87)	0.69 (0.53, 0.84)	0.72 (0.57, 0.88)	0.63 (0.46, 0.79)	0.69 (0.54, 0.85)	0.54 (0.39, 0.68)	0.52 (0.43, 0.60)	0.60 (0.43, 0.76)	0.64 (0.49, 0.80)

Note.—Each letter from A to F denotes a different reader in order of experience from least experienced, and 1 and 2 denote the reading session. Hepatic steatosis grades are classified as none (grade S0), mild (S1), moderate (S2), and severe (S3). Values reported are the areas under the receiver operating characteristic curve (AUCs) for each individual reading, with 95% CIs in parentheses.

* Parentheses indicate the numbers of patients in each dichotomized steatosis class.

Table 4: Diagnostic Performance and Accuracy of Deep Learning for Grading Hepatic Steatosis on B-Mode US Images

Steatosis Grade Classification and Set (*n* = 199)*	AUC†	Sensitivity (%)	Specificity (%)	Accuracy (%)	PPV (%)	NPV (%)
S0 vs S1 or higher						
Independent test set (*n* = 8 vs *n* = 22)	0.98 (0.98, 0.99)	88 (962/1099)	98 (172/175)	89 (1134/1274)	>99 (962/965)	56 (172/309)
Reader set (*n* = 12 vs *n* = 40)	0.85 (0.83, 0.87)	79 (1920/2441)	78 (262/336)	79 (2182/2777)	96 (1920/1994)	33 (262/783)
S0 or S1 vs S2 or S3						
Independent test set (*n* = 21 vs *n* = 9)	0.67 (0.64, 0.70)	67 (362/541)	58 (423/733)	62 (785/1274)	54 (362/672)	70 (423/602)
Reader set (*n* = 29 vs *n* = 23)	0.73 (0.71, 0.75)	76 (917/1214)	58 (912/1563)	66 (1829/2777)	58 (917/1568)	75 (912/1209)
S2 or lower vs S3						
Independent test set (*n* = 24 vs *n* = 6)	0.66 (0.63, 0.69)	74 (215/291)	54 (531/983)	59 (746/1274)	32 (215/667)	87 (531/607)
Reader set (*n* = 40 vs *n* = 12)	0.67 (0.64, 0.69)	82 (488/592)	47 (1034/2185)	55 (1522/2777)	30 (488/1639)	91 (1034/1138)

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

* Parentheses indicate the numbers of patients in each dichotomized steatosis class.

† Data in parentheses are 95% CIs.

performance of steatosis classification could not be fully explored given the available data. In our results, the relationship between steatosis classification and the remaining features is inconsistent across different grade comparisons. Therefore, there is no evidence to suggest that these effects had a significant influence on the overall results. Second, we did not address the NASH diagnosis, as we were only focusing on steatosis. Third, we assessed the performance of our model on a split test set from a single-center study, which limits generalizability, as the data set is reflective of the patient population at our institution and limited by the eligibility criteria. Finally, readers did not have access to image settings such as transmit frequency, harmonics, compounding, and time gain compensation, which could hide the influence of attenuation.

In conclusion, this cross-sectional, case-control, single-center study demonstrates that deep learning approaches may provide similar classification accuracies compared with human readers for steatosis detection and grading. The performance of our model suggests that deep learning may be used for opportunistic screening of steatosis with use of B-mode US across scanners from different manufacturers or even for epidemiologic studies at a populational level if deployed on large regional imaging

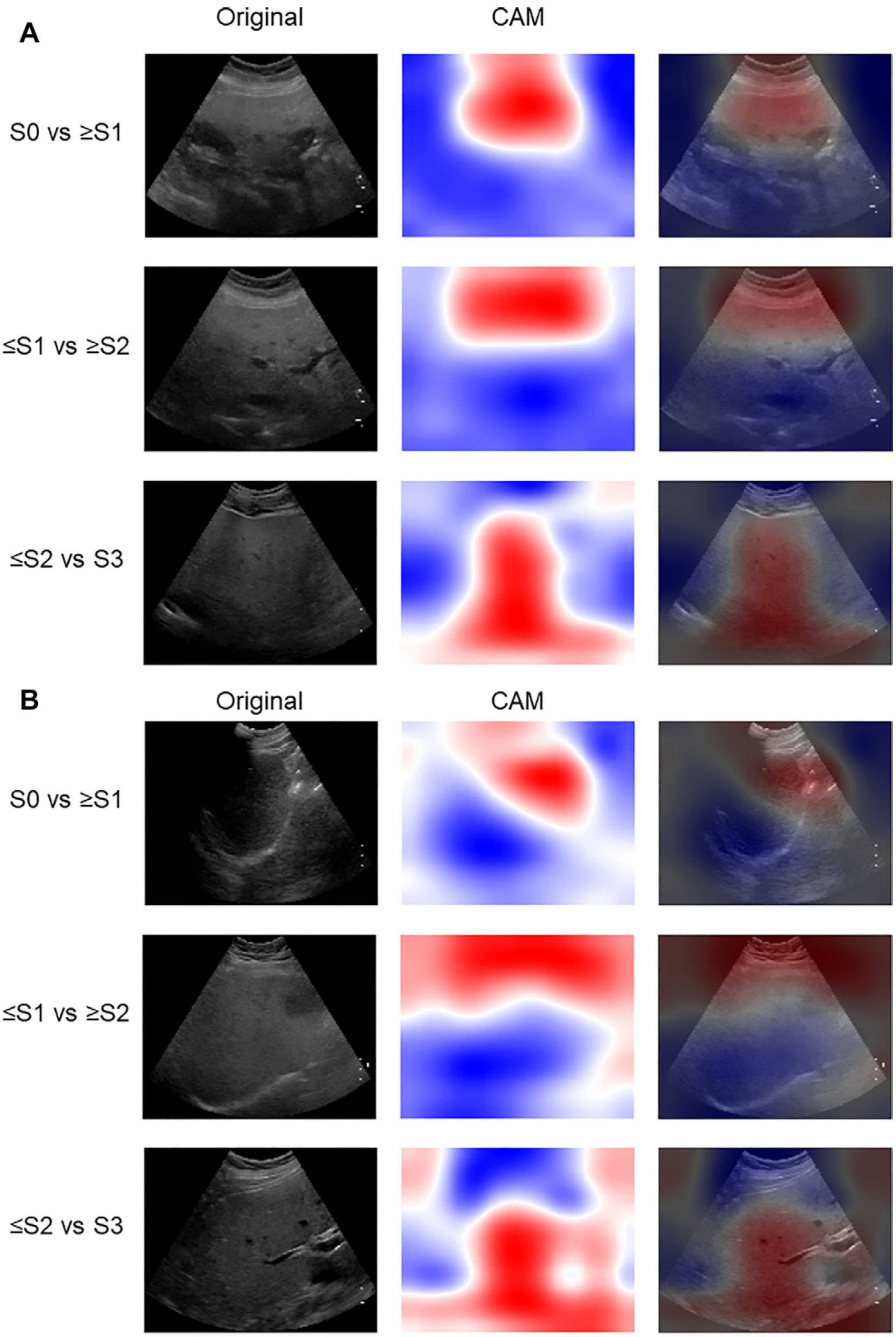

Figure 3: Class activation maps (CAM) for **(A)** correctly classified and **(B)** incorrectly classified images during the cross-validation procedure. Red areas are the most relevant regions of the image for the model's prediction. Hepatic steatosis grades are classified as none (grade S0), mild (S1), moderate (S2), and severe (S3). Representative images for the three dichotomizations of hepatic steatosis grades (S0 vs S1 or higher, S0 or S1 vs S2 or S3, and S2 or lower vs S3) are displayed from left to right: B-mode US image, class activation map, and class activation map overlaid on B-mode image. True-positive images are in, from top to bottom, a 46-year-old female patient with S2, 46-year-old female patient with S2, and 60-year-old male patient with S3. False-positive images are in, from top to bottom, a 46-year-old female patient with S0, 67-year-old female patient with S1, and 55-year-old female patient with S2.

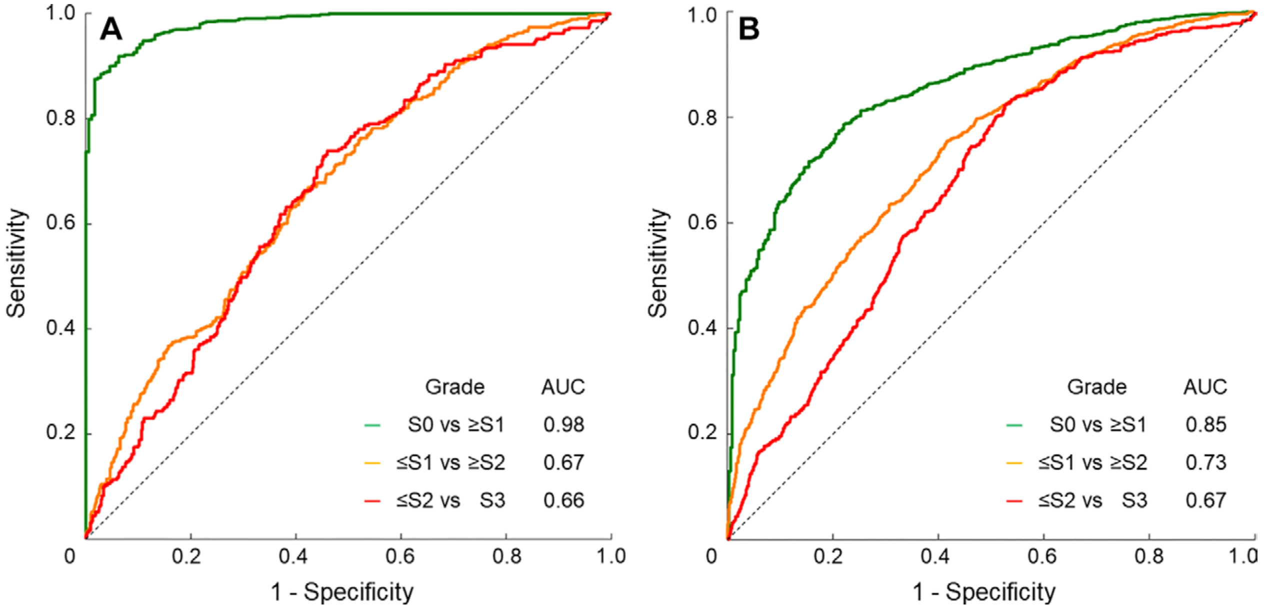

Figure 4: Receiver operating characteristic analysis of deep learning for classification of dichotomized histologically determined hepatic steatosis grades for **(A)** the independent test set and **(B)** reader set. Hepatic steatosis grades are classified as none (grade S0), mild (S1), moderate (S2), and severe (S3). AUC = area under the receiver operating characteristic curve.

repositories. These results support the need to conduct further multicenter studies to validate deep learning models for hepatic steatosis screening with use of B-mode US.

Author contributions: Guarantors of integrity of entire study, **P.V., E.A., J.S.B., B.T., A.T.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **P.V., S.I.C., C.L.R., L.P.B., Y.H.L., E.A., J.S.B., G.W., G.C., A.T.**; clinical studies, **P.V., S.I.C., C.L.R., Y.H.L., E.A., J.M., K.N.V., J.M.L., J.S.B., B.N.N., M.C., A.T.**; experimental studies, **P.V., L.P.B., E.A., T.M.A., H.L., S.G., B.T.**; statistical analysis, **P.V., L.P.B., E.A., J.M., J.S.B., B.T., E.B., G.W., M.C., A.T.**; and manuscript editing, **P.V., S.I.C., P.B., C.L.R., E.A., K.N.V., J.S.B., E.M., S.K., B.N.N., I.R., G.W., M.C., G.C., A.T.**

Disclosures of conflicts of interest: P.V. No relevant relationships. **S.I.C.** No relevant relationships. **P.B.** No relevant relationships. **C.L.R.** No relevant relationships. **L.P.B.** No relevant relationships. **Y.H.L.** No relevant relationships. **E.A.** No relevant relationships. **J.M.** No relevant relationships. **T.M.A.** No relevant relationships. **K.N.V.** No relevant relationships. **J.M.L.** No relevant relationships. **J.S.B.** No relevant relationships. **E.M.** No relevant relationships. **H.L.** No relevant relationships. **S.K.** No relevant relationships. **B.N.N.** No relevant relationships. **S.G.** No relevant relationships. **B.T.** Recipient of NSERC USRA and FRQNT supplement from the Canadian government. **I.R.** No relevant relationships. **E.B.** Grants from the Institute of Data Valorization (IVADO). **G.W.** No relevant relationships. **M.C.** No relevant relationships. **G.C.** No relevant relationships. **A.T.** Equipment loan from Siemens Healthcare.

References

1. Younossi ZM, Golabi P, Paik JM, Henry A, Van Dongen C, Henry L. The global epidemiology of nonalcoholic fatty liver disease (NAFLD) and nonalcoholic steatohepatitis (NASH): a systematic review. Hepatology 2023;77(4):1335–1347.
2. Loomba R, Friedman SL, Shulman GI. Mechanisms and disease consequences of nonalcoholic fatty liver disease. Cell 2021;184(10):2537–2564.
3. Estes C, Anstee QM, Arias-Loste MT, et al. Modeling NAFLD disease burden in China, France, Germany, Italy, Japan, Spain, United Kingdom, and United States for the period 2016-2030. J Hepatol 2018;69(4):896–904.
4. Qayyum A, Nystrom M, Noworolski SM, Chu P, Mohanty A, Merriman R. MRI steatosis grading: development and initial validation of a color mapping system. AJR Am J Roentgenol 2012;198(3):582–588.
5. Qu Y, Li M, Hamilton G, Zhang YN, Song B. Diagnostic accuracy of hepatic proton density fat fraction measured by magnetic resonance imaging for the evaluation of liver steatosis with histology as reference standard: a meta-analysis. Eur Radiol 2019;29(10):5180–5189.
6. Yokoo T, Serai SD, Pirasteh A, et al. Linearity, bias, and precision of hepatic proton density fat fraction measurements by using MR imaging: a meta-analysis. Radiology 2018;286(2):486–498.
7. Tang A, Desai A, Hamilton G, et al. Accuracy of MR imaging-estimated proton density fat fraction for classification of dichotomized histologic steatosis grades in nonalcoholic fatty liver disease. Radiology 2015;274(2):416–425.
8. Zhang E, Wartelle-Bladou C, Lepanto L, Lachaine J, Cloutier G, Tang A. Cost-utility analysis of nonalcoholic steatohepatitis screening. Eur Radiol 2015;25(11):3282–3294.
9. Zwiebel WJ. Sonographic diagnosis of diffuse liver disease. Semin Ultrasound CT MR 1995;16(1):8–15.
10. Hernaez R, Lazo M, Bonekamp S, et al. Diagnostic accuracy and reliability of ultrasonography for the detection of fatty liver: a meta-analysis. Hepatology 2011;54(3):1082–1090.
11. Fetzer DT, Rosado-Mendez IM, Wang M, et al. Pulse-echo quantitative US biomarkers for liver steatosis: toward technical standardization. Radiology 2022;305(2):265–276.
12. 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(6):W320–W323.
13. Chartrand G, Cheng PM, Vorontsov E, et al. Deep learning: a primer for radiologists. RadioGraphics 2017;37(7):2113–2131.
14. Cheng PM, Montagnon E, Yamashita R, et al. Deep learning: an update for radiologists. RadioGraphics 2021;41(5):1427–1445.
15. Acosta JN, Falcone GJ, Rajpurkar P. The need for medical artificial intelligence that incorporates prior images. Radiology 2022;304(2):283–288.
16. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.
17. Byra M, Styczynski G, Szmigielski C, et al. Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images. Int J CARS 2018;13(12):1895–1903.
18. Cao W, An X, Cong L, Lyu C, Zhou Q, Guo R. Application of deep learning in quantitative analysis of 2-dimensional ultrasound imaging of nonalcoholic fatty liver disease. J Ultrasound Med 2020;39(1):51–59.
19. Morgan TA, Maturen KE, Dahiya N, Sun MRM, Kamaya A; American College of Radiology Ultrasound Liver Imaging and Reporting Data System (US LI-RADS) Working Group. US LI-RADS: Ultrasound Liver Imaging Reporting and Data System for screening and surveillance of hepatocellular carcinoma. Abdom Radiol (NY) 2018;43(1):41–55.

20. Boyd A, Cain O, Chauhan A, Webb GJ. Medical liver biopsy: background, indications, procedure and histopathology. Frontline Gastroenterol 2020;11(1):40–47.
21. Neuberger J, Patel J, Caldwell H, et al. Guidelines on the use of liver biopsy in clinical practice from the British Society of Gastroenterology, the Royal College of Radiologists and the Royal College of Pathology. Gut 2020;69(8):1382–1403.
22. Bedossa P, Poitou C, Veyrie N, et al. Histopathological algorithm and scoring system for evaluation of liver lesions in morbidly obese patients. Hepatology 2012;56(5):1751–1759.
23. Kleiner DE, Brunt EM, Van Natta M, et al. Design and validation of a histological scoring system for nonalcoholic fatty liver disease. Hepatology 2005; 41(6):1313–1321.
24. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 1409.1556 [preprint] https://arxiv.org/abs/1409.1556. Posted September 4, 2014. Accessed January 2023.
25. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning Deep Features for Discriminative Localization. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016; 2921–2929.
26. Brouwers MC, Bilderbeek-Beckers MA, Georgieva AM, van der Kallen CJ, van Greevenbroek MM, de Bruin TW. Fatty liver is an integral feature of familial combined hyperlipidaemia: relationship with fat distribution and plasma lipids. Clin Sci (Lond) 2007;112(2):123–130.
27. Liang RJ, Wang HH, Lee WJ, Liew PL, Lin JT, Wu MS. Diagnostic value of ultrasonographic examination for nonalcoholic steatohepatitis in morbidly obese patients undergoing laparoscopic bariatric surgery. Obes Surg 2007;17(1):45–56.
28. Riley TR 3rd, Mendoza A, Bruno MA. Bedside ultrasound can predict nonalcoholic fatty liver disease in the hands of clinicians using a prototype image. Dig Dis Sci 2006;51(5):982–985.
29. Chang Y, Ryu S, Sung E, Jang Y. Higher concentrations of alanine aminotransferase within the reference interval predict nonalcoholic fatty liver disease. Clin Chem 2007;53(4):686–692.
30. Jun DW, Han JH, Kim SH, et al. Association between low thigh fat and nonalcoholic fatty liver disease. J Gastroenterol Hepatol 2008;23(6):888–893.
31. Liu KH, Chan YL, Chan JCN, Chan WB, Kong WL. Mesenteric fat thickness as an independent determinant of fatty liver. Int J Obes 2006;30(5):787–793.
32. Chan DFY, Li AM, Chu WCW, et al. Hepatic steatosis in obese Chinese children. Int J Obes 2004;28(10):1257–1263.
33. Fishbein M, Castro F, Cheruku S, et al. Hepatic MRI for fat quantitation: its relationship to fat morphology, diagnosis, and ultrasound. J Clin Gastroenterol 2005;39(7):619–625.
34. Holt HB, Wild SH, Wood PJ, et al. Non-esterified fatty acid concentrations are independently associated with hepatic steatosis in obese subjects. Diabetologia 2006;49(1):141–148.
35. Cloutier G, Destrempes F, Yu F, Tang A. Quantitative ultrasound imaging of soft biological tissues: a primer for radiologists and medical physicists. Insights Imaging 2021;12(1):127.