Dear editors:

We thank all reviewers for their constructive comments and a chance to improve our manuscript. Below we include all reviewer comments in italics and our response follows in roman.

Reviewer 1

1. The study uses stored images of US to train the deep learning. These images used in the study are liable to have operator and US settings variability. No indices like Hemaguchi score, US-FLI score and hepato renal steatosis score were used to validate these semi quantitative US images.

Thank you for the comments. In fact, based on our hospital's practice, the clinical diagnosis of our big data was generally based on the principles of Hemaguchi score, US-FLI score and hepatorenal steatosis score. Our discussion now includes the following sentence in paragraph 6:

"The clinical diagnosis of our big data was generally based on the principle of Hemaguchi score, US-FLI score and hepatorenal steatosis score[13]."

However, we are unable to explicitly use such scores for the big data because it was a retrospective cohort, and such scores were not recorded in the clinical records. Instead, the US-based clinical diagnosis of our big data cohorts is reflective of what is stored in large-scale clinical archives, e.g., in our case US-based diagnoses that were made by 63 clinicians with different brands of machines. This is what allows us to obtain such a large cohort of training data in the first place and is what also makes it possible for others to readily replicate our workflow. Finally, we note that we only use our big data for algorithmic training and not testing. Our test sets are all labelled with gold-standard histopathology. We demonstrate high performance and high reliability on these test sets, demonstrating that our big-data training cohorts can produce effective and well-validated DL algorithms for steatosis assessment.

2. There is no doubt that 2D US is the modality of choice for diagnosis of Steatosis. US though has many advantages, has its own shortcomings; poor inter observer

variability; poor sensitivity and specificity in discerning mild steatosis; poor utility in obese where steatosis is common etc. These limitations necessitated the use of other US based methodologies like CAP, AC, ASQ, BSC, SWE, QUS Spectroscopy, transient elastography, etc., to quantify steatosis with comparable efficiency. CAP needs different system and even in this there is considerable overlap with grades of steatosis. Various US machines that use advanced software to quantify the steatosis by attenuation coefficient, ATI, backscatter coefficient etc., are not indicated in this study. It may be too much to expect 2D US based DL to nullify these shortcomings.

Thanks for the comments. We agree that there are many shortcomings of 2D US. The main aim of this study was to establish an objective diagnosis algorithm using machine learning which will mitigate some of these shortcomings. We noticed that the DL algorithm is more repeatable at milder and severe steatosis (Figure 3 and supplementary figure 1). We compared our algorithm with CAP of Fibroscan (Table 3). The AUROCs were the same with or better than CAP when pathology was used as a gold standard. Further study is certainly warranted, but we are encouraged by our preliminary findings, and we will compare more extensively with other modalities in the future. Our advantages are that we do not need an additional expensive equipment nor are we limited to specific brands/scanners.

3. Patients with alternative causes of hepatic steatosis other than NAFLD are also included in this study and their images may have liver structural alteration that is caused by factors other than fat, and this may influence the interpretations.

Thank you for this important point. The structural alteration may be induced by other factors, including fibrosis. We are developing an analogous algorithm for fibrosis and testing it under the presence of steatosis. We believe that steatosis and fibrosis may have complex interactions with each other in both directions, which certainly warrants further study. In the limitations section of our Discussion, we now include the following sentence:

“Third, patients in HP-U and HP-T may also have co-occurring liver fibrosis, which likely has complex interactions with hepatic steatosis. The effect of this interaction on interpretations needs further study.”

4. The non-image based noninvasive methodologies like FIB-4 Index, APRT, ELF, NAFLD Score recommended to grade the steatosis in various guidelines of NAFLD, are

not used in this study to compare and validate the US based DL Algorithm's potential.

In this study, we use histology as a gold standard, which follows standard practices. While we do agree with the value in validating against other measures, we view this as outside the scope of this work, as it would significantly increase the complexity of analysis. However, our Discussion now mentions more explicitly the need to validate against both imaging and non-imaging-based tests, which we agree remains an important aspect of future work.

5. The aim to develop quantitative score to assess liver steatosis with 2D US images using DL algorithm is marred by including too many view groups, viewpoints, etc., in this study, leaving one confused.

We prefer to use multiple viewpoints because one hepatic lobe may be absent, or its space may be occupied by a lesion. Thus, performance across multiple viewpoints is an important feature for clinical use. It also allows the algorithm to fill in important gaps left by popular quantitative US measures, which are typically restricted to one lobe. While this does introduce unavoidable complexity in analysis, we view this as justified in light of the considerations above.

Reviewer 2

This study develop a scalable deep learning (DL) algorithm for quantitative scoring of liver steatosis from 2D ultrasound images. It incorporated different 2D US scanner models and brands, different liver viewpoints, and prospectively and retrospectively collected images. Good reliability and performance across different liver viewpoints and scanners were demonstrated. The performance of the deep learning algorithm's quantitative score was comparable to or better than the control attenuation parameter produced by FibroScan. The highlight of this manuscript is the use of big data ultrasound images, which shows that the researchers have been committed to studying liver steatosis for long time and accumulated rich research data. This study covers multiple sub-studies, and it is recommended to make an experimental flow chart for the whole experimental process to facilitate readers' overall understanding of the research process and experimental design.

Thanks for your comments and assessment. We have now added a flowchart as

Figure 1.

Reviewer 3

Well written manuscript, wish you success with your academic carrier.

Thank you so much for your support.

Reviewer 4

Thank you very much for letting me contribute to this scientific achievement. Deep Learning is a technology that has been remarkably developing these years and has been frequently used to analyze images. Since ultrasound is relatively a subjective diagnostic modality, establishing objective tools that provide universal recommendations is of great help to make our healthcare better and more accessible. In that sense, this study provided a high impact on screening choices of hepatologic conditions, such as NAFLD. Overall, this study revealed that an appropriately trained neural network algorithm potentially outperforms humans' ability in diagnosis. However, the potential problem of this study is the selection of the neural network. While the authors adopted ResNet-18 to train the algorithm, the rationale for using that network is not fully explained. In addition, since most of the readers of this journal are not experts of deep learning, the whole concept or benefit of ResNet-18 should be clarified. Moreover, as is always concerned when we discuss machine learning, it would enhance the usefulness of this article if the authors mentioned clinical implications. Machine learning models may detect NAFLD more easily and effectively, but how can we use that technique in daily practice? Is this trained model going to be published as a universal tool? Or do we have to develop a system like this on our own? This is an expected question because all machine learning models work on the premise of interpolations. That is to say, the prediction ability of each model plummets when it comes across a value outside the training set range. Many clinicians, including me, want to hear the authors' opinions and suggestions on this matter.

Thank you for these comments and the positive assessment. We will address each specific point in turn.

Choice of ResNet-18

Our supplementary material now includes an expanded discussion of the rationale behind our choice.

“Using the images from BD-L, we trained a DL classifier using the 2D US diagnoses extracted from the CGMH records. We opted for the ResNet family of DL classifiers [3] given their ubiquity and performance in both natural imaging and medical imaging tasks. The ResNet family of DL classifiers are 2D convolutional neural networks[4] that use the concept of residual connections to reduce the problem of vanishing gradients and improves learning speed. Based on performance on the BD-V validation dataset, we determined that the ResNet-18 variant performed best. The ResNet-18 has the added virtue of being lightweight, reducing overfitting tendencies compared to alternative variants.”

Clinical implications

Thank for you these comments. We agree that more discussion is warranted here. Please see our expanded discussion where we elaborate on these points.

“Our algorithm could be used on different brands of scanners in different hospitals. However, each scanner or hospital may determine their cutoff values in different grade of steatosis based on their preferred gold standard (liver histology or MRI-PDFF). Its potential applications include improved steatosis screening and longitudinal tracking in clinical settings. Our trained algorithm may be shared with researchers for non-commercial use upon reasonable requests to the clinical principal investigator (D.I T.) and subject to IRB approval. We expect it to have good generalizability given the diversity of the BD-L dataset. Because our algorithm is trained only on retrospective US-labeled data, which are readily available in most PACS systems, we also believe the methodology we presented here to be highly generalizable and accessible to institutions wishing to train their own algorithms.”

Reviewer 5

The authors present a method based on deep learning for the quantification of

hepatic steatosis in liver US images. Specifically, a resnet18 is used to classify the images into 4 classes: None, Mild, Moderate and Severe Steatosis. The paper is clear and easy to follow.

Thank you for this assessment.

1) Additional background knowledge should be added within the introduction. I would add that the gold standard method for quantification of steatosis is liver biopsy. In the same section I would mention the several recent methods that have been published in this topic (doi: 10.1016/j.compbiomed.2020.103836, 10.1002/cyto.b.21790). However, liver biopsy is an invasive method and therefore other methods such as MRI and US are used for screening. This makes it easier for the reader to understand the problem and the framing of this work.

Thank for you for these references, which we now include. We did mention that liver biopsy is the gold standard for liver steatosis evaluation along with its downsides, please see this passage from the introduction (with the added references).

“Liver needle biopsy is the gold standard diagnosis at present. However, even with machine-learning based interpretation[11,12], biopsy’s invasiveness severely limits its clinical applicability as a screening and assessment tool, and, at the same time, it is also prone to sampling error. Thus, with the growing prevalence of NAFLD and NASH, accurate, reliable, and accessible non-invasive screening tools are increasingly important to quantify liver steatosis and provide follow-up monitoring[4].

Such tools include magnetic resonance imaging with derived proton density fat fraction (MRI-PDFF), quantitative ultrasound (US), and 2D US diagnoses[13]…”

2) More details should be added on the network training to make the work more reproducible. How many epochs was it trained on? with what optimization algorithm? was it pre-trained? why was the ResNet18 architecture chosen?

Thank for pointing out these suggestions. Please see our response to Reviewer 4’s comments regarding our network choice. Our supplementary material includes a table (Supplementary Table 5) of hyper-parameters we used to train ResNet-18 and

description in the text, which we have updated based on these suggestions. We have also added the following details regarding the epoch number and pretraining in our Supplementary text.

“We use an ImageNet pretrained network[8], as that performed better than random initialization. The stopping criterion was the model checkpoint that performed best on BD-V, based on a rolling average of five epochs.”