A Deep Learning Approach for Hepatic Steatosis Estimation from Ultrasound Imaging

Sara Colantonio[1][0000−0003−2022−0804], Antonio Salvati[2][0000−0002−9779−2731], Claudia Caudai[1][0000−0002−1590−7890], Ferruccio Bonino[2,3,4,5][0000−0001−9942−0328], Laura De Rosa[2,6][0000−0002−6030−8483], Maria Antonietta Pascali[1][0000−0001−7742−8126], Danila Germanese[1][0000−0002−7814−5280], Maurizia Rossana Brunetto[2,3][0000−0001−8364−9152], and Francesco Faita[6][0000−0002−6201−1843]

[1] Institute of Information Science and Technologies, National Research Council, Pisa, Italy
[2] Hepatology Unit, Pisa University Hospital, Italy
[3] Department of Clinical and Experimental Medicine, University of Pisa, Pisa, Italy
[4] Fondazione Italiana Fegato, AREA Science Park, Campus Basovizza, Trieste, Italy
[5] IRCSS SDN, Naples, Italy,
[6] Institute of Clinical Physiology, National Research Council, Pisa, Italy

Abstract. This paper proposes a simple convolutional neural model as a novel method to predict the level of hepatic steatosis from ultrasound data. Hepatic steatosis is the major histologic feature of non-alcoholic fatty liver disease (NAFLD), which has become a major global health challenge. Recently a new definition for FLD, that take into account the risk factors and clinical characteristics of subjects, has been suggested; the proposed criteria for Metabolic Disfunction-Associated Fatty Liver Disease (MAFLD) are based on histological (biopsy), imaging or blood biomarker evidence of fat accumulation in the liver (hepatic steatosis), in subjects with overweight/obesity or presence of type 2 diabetes mellitus. In lean or normal weight, non-diabetic individuals with steatosis, MAFLD is diagnosed when at least two metabolic abnormalities are present. Ultrasound examinations are the most used technique to non-invasively identify liver steatosis in a screening settings. However, the diagnosis is operator dependent, as accurate image processing techniques have not entered yet in the diagnostic routine. In this paper, we discuss the adoption of simple convolutional neural models to estimate the degree of steatosis from echographic images in accordance with the state-of-the-art magnetic resonance spectroscopy measurements (expressed as percentage of the estimated liver fat). More than 22,000 ultrasound images were used to train three networks, and results show promising performances in our study (150 subjects).

Keywords: ultrasound (US) · medical imaging · convolutional neural network · hepatic steatosis.

1 Introduction

Hepatic steatosis is the major histologic feature of MAFLD, associated with other liver diseases, as well as to type 2 diabetes, cardiovascular disease, chronic kidney disease, and some types of extrahepatic malignancies; MAFLD is diagnosed when at least two metabolic abnormalities are present [7, 6, 18]. Hepatic steatosis is due to the accumulation of fat within the liver and its association with inflammation, steatohepatitis causes the progression of fibrosis, to cirrhosis and hepatocellular carcinoma [8, 11]. Therefore, the early detection and accurate quantification of steatosis is an essential task for prevention and monitoring of disease progression.

Liver biopsy is nowadays the standard reference diagnostic method to assess steatosis [2], though it represents an invasive procedure and it may be prone to errors, due to sampling issues in case of dishomogeneous intrahepatic distribution. To date, among the noninvasive modalities for the quantitative assessment of steatosis, the most reproducible and effective one is based on Magnetic Resonance Spectroscopy (MRS), which provides a sensitive, accurate and quantitative evaluation of liver fat content, the so-called *H-MRS index*, by using non-ionizing radiation [9]. An extensive study has demonstrated a high correlation between H-MRS index and biopsy results (Spearman's nonparametric correlation coefficient rs = 0.9) [5]. Hence, this modality is currently considered as the non-invasive gold-standard. Nevertheless, MRS is an expensive diagnostic procedure, and MRS devices are relatively scarcely available, thus preventing the adoption of MRS in daily clinical settings. On the other hand, UltraSound (US) imaging represents a valid approach for the assessment of liver steatosis, as demonstrated in previous studies [13]. Further to this, US is non-invasive, non-ionizing, inexpensive and widely available modality, which may fit also screening purposes.
Recently, some studies have proposed quantitative assessment of hepatic steatosis based on ultrasound imaging [17, 19, 12]. In these works, the US-based methodologies are generally compared to the H-MRS index. Among these, one of the most interesting result has been published in 2018 by Di Lascio et al. [10]: the authors propose and discuss the Steato-Score, a fat liver index representative of intra-hepatic fat content, which is defined by combining five different ultrasound parameters, and showed a good correlation with the H-MRS index (adjusted coefficient of determination $R^2 = 0.72$).

Artificial intelligence techniques have recently emerged as the leading tool in various research fields, and especially in general imaging analysis and computer vision, for several tasks, such as object detection, segmentation, and classification. Machine and deep learning shows huge potential for enabling the automation of the NAFLD diagnosis and staging, hence providing a viable alternative solution to traditional biomedical image processing [14]. Concerning hepatic steatosis, generally deep learning methods have been used not to estimate the liver fat fraction but to perform a classification, e.g. to discriminate normal liver vs. non-alcoholic fatty liver diseases (NAFLD, fat fraction $> 5\%$) [3, 16, 1, 4, 8]: most of them show quite desirable performances (about 95% of accuracy). Even if a coarse classification (e.g. binary classification) could support screening on

large population, it is not enough accurate to have a clinical impact, e.g. for patient monitoring. Also, some of the architecture proposed are quite complex (training up to 22 layers [1], or using a transfer learning approach [3]) and are trained using small datasets (made of 63 subjects in [1], and of 55 obese patients in [3]); deeper is the learning architecture, larger should be the dataset, and using very few data could lead to unstable models (due to over-parametrization). The work of Han et al. [8], to best of our knowledge, is the most promising in this research line, developing a quantitative analysis of raw radiofrequency (RF) ultrasound signal, based on two one-dimensional CNN algorithms: a binary classifier and a fat fraction estimator. The classifier yielded a classification accuracy (96%); while the fat fraction estimator predicted fat fraction values that positively correlated with proton MRI ($r = 0.85$; $p < .001$).

The aim of the present work is to provide a CNN model able to accurately estimate from US images the liver fat fraction; the accuracy of the fat estimation is assessed with respect to the H-MRS index.

The following Section is devoted to describe the materials and methods of the present investigation, by including a description of the population study, the acquisition protocols used to collect the US data, and the preprocessing steps performed to set the input for the proposed CNN-based method. Section 2.2 provides details of the CNN architectures used in the experimentation, which is presented in Section 2.3. Results are reported and discussed in Section 3.

2 Materials and Methods

The study population included 150 consecutive patients enrolled the Hepatology Unit of the University Hospital of Pisa for evidence of hepatic steatosis at the standard ultrasound examination without with-out increased liver enzymes who gave their informed consent (their characteristics are reported in 1). Proton MRS imaging was performed with a MRI scanners (3.0 T) equipped with 32-channel receiver coils employed on patients in the supine position (Pisa: Philips Ingenia, Philips Healthcare, Best, Netherlands). The percentage of fat (H-MRS index) was assessed for each subject, normalizing the fitted signal amplitude of the fat to the sum of water and fat amplitudes [17]. The population showed different levels of steatosis: the fat percentage, assessed through the H-MRS index, ranges from 0.27% to 50.97%.

A mapping between the H-MRS and the classes routinely used in histology to estrablish the steatosis level has been established in [9] and comprises the following stratification classes:

class S0 corresponds to cases with H-MRS index <= 3.12%
class S1 to cases with H-MRS index > 3.12% and <= 8.77%
class S2 to cases with H-MRS index > 8.77% and <= 13,69%
class S3 to cases with H-MRS index > 13.69%.

In this respect, the distribution of this study population is the following:

S0: 111 subjects, corresponding to the 74% of the population study;

Table 1. Characteristics of the study population. Data were expressed as counts/percentages for categorical variable and mean ± standard deviation (sd) for continuous variables.

	n. of subjects	**% of population**
SEX (M:F)	73 : 77	48.7 : 51.3
	mean ± sd	**min-max**
AGE (years)	53.54 ± 12.66	20.0 - 75.3
BMI (kg/m^2)	24.86 ± 3.69	15.28 - 33.9
Fat (%)	4.50 ± 8.01	0.27 - 50.97

S1: 21 subjects, corresponding to the 14% of the population study;
S2: 6 subjects, corresponding to the 4% of the population study;
S3: 12 subjects, corresponding to the 8% of the population study.

2.1 US data acquisition and preprocessing

The ultrasound data were acquired in the center of Pisa. In particular, several ultrasound images were acquired, using different projections, according to an acquisition protocol that included the acquisition of an intercostal or subcostal longitudinal scan view with subject in supine/left lateral position (named HR) and an oblique subcostal scan view (named AR). The two different scan views were chosen in such a way that in the first scan the level of ecogenicity of the ultrasound beam within both the renal and hepatic parenchyma is appreciable (i.e. the HR view in which the hepatic parenchyma and the right kidney are clearly visualized) and the second scan is the oblique subcostal scan view showing the complete liver parenchyma and diaphragm, i.e. the AR view which provides a view of the attenuation of the ultrasound beam within the liver parenchyma.

Ultrasound clips were acquired at 30 fps and their duration ranges from few frames to 224 frames, with an average of 114 frames. The clips were processed by extracting all the frames as gray images, which were centered and cropped to the size of 360 x 360 pixels. Each frame inherits a label given by the ground truth value (H-MRS index). Some remarks about the dataset:

(i) Due to the different length of the US clips, and to the uneven distribution of the H-MRS index values, the resulting datasets (HR and AR) do not have the same number of images;
(ii) Several images, namely those extracted from the same clip, are very similar because of the centrality of the sonographic cone and the acquisition with fixed probe (Fig. 1);

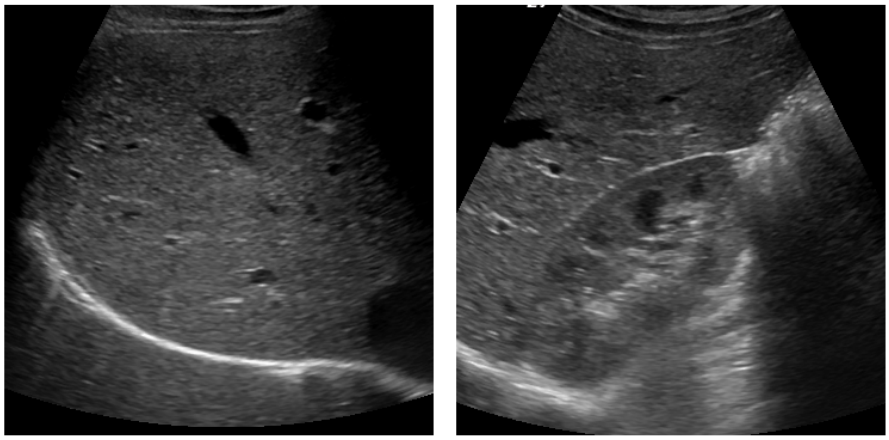

Fig. 1. Example of AR view, on the left and HR view, on the right, of the same patient.

(iii) The ground-truth values are not uniformly distributed in the population study.

As it usually happens, the dataset is quite unbalanced, which makes identifying the less present classes a challenging task. Nevertheless, the regression approach may help coping with this unbalance, as the CNN model learns to approximate the H-MRS index instead of the class.

2.2 The CNN architecture

Each view was treated independently using both a single-view architecture, and coupled into a two-branch architecture, in which both AR and HR views are used to train the predictive model. This was done to understand whether the peculiarity of each single view contributes with informative content to predict the H-MRS index.

We designed two CNN architectures: a 5-layer CNN and a two-branch CNN. The first architecture is used to produce two predictive models: one trained on AR clips, and the other trained on HR clips; the latter architecture is a two-branch CNN which has been trained on both AR and HR clips:

AR CNN: 5-layer CNN trained on AR images;
HR CNN: 5-layer CNN trained on HR images;
AR & HR CNN: two-branch CNN trained on both AR and HR images.

All the CNNs have been developed in Python (version 3.7.9, Jupyter environment), using TensorFlow 2.0.0 and the learning library Scikit-learn 0.23.2.

The first architecture, in Fig. 2, is made of five convolutional layers respectively with 8, 16, 32, 64 and 128 filters. The input images, which are one-channel

gray images of 360 x 360 pixels, are preliminarily regularized through a batch normalization. The padding is used to control the size of the images. The activation function is the REctified Linear Unit (ReLU) for all the convolutional layers and for the first fully connected layer, while the last fully connected layer uses a linear activation function to perform the regression. The maxpooling layers have size (2,2) and stride 2.

The two-branch CNN, (in Fig. 3), is made of three pieces: the two branches, i.e. the AR CNN and the HR CNN, and a final tail, in which the weights associated with the best regression performances of the two branches are imported and concatenated, just before the fully connected layers: the last fully connected layer returns the regression on the subject's steatosis level.

In order to finalize the best architecture, many attempts have been made using: different hyperparameters, such as the learning rate and kernel size; the activation functions relu, elu, prelu, and swish; normalization L1, L2 and dropout; and different kinds of weight initialization. It resulted that the model was not sensitive to such changes. On the other hand, the size and number of convolutional layers and the presence of a fully connected layer before the last regression layer proved to be very important.

2.3 Model Training and Test

To each subject we associate the ultrasound clips (for each view) and the corresponding H-MRS index. This allows us to test the trained model on images extracted from different subjects, hence to assess the model performances on never-before-seen cases.

The test set is made selecting 10 patients out of 150 (resulting in 802 AR and 998 HR images). The selection was performed manually in order to have the same distribution of disease severity as the whole dataset. In more details the test set is made of: 6 cases of S0 class, 2 of S1 class, 1 of S2 and 1 of S3.

The remaining 140 subjects (13,406 AR and 16,496 HR images) were used to train independently the AR CNN and the HR CNN.

Concerning the training, the Adam optimizer, using its default parameters and the mean squared error as loss function, was used to optimize the gradient computation; all the three models have been initialized with random weights, batch size was set to 50; the number of epochs was set to 30, using the early-stopping criterion (patience = 5) to prevent overfitting.

We adopted a stratified 4-fold cross-validation, recommended for regression on unbalanced datasets [15]. The values of the validation loss for the four sets of the 4-fold cross-validation are reported in Table 2. Weights associated with best performances (set 1 for AR and set 4 for HR) have been imported into the two-branch architecture and used for a paired regression on the two views (Fig. 3). Finally, we evaluated the performance of the trained models using the test set of never-seen-before 10 subjects.

Summarizing, we trained three models to predict the hepatic steatosis level from US images: the CNN in Fig. 2 trained on AR data, the CNN in Fig. 2

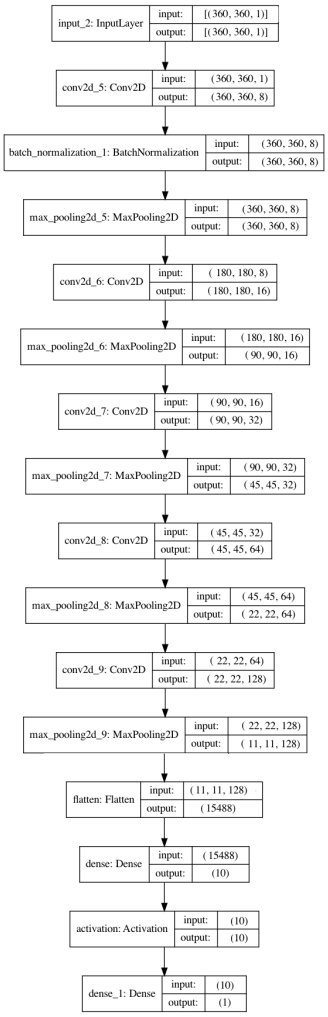

Fig. 2. Sketch of the CNN architecture used for the estimation of hepatic steatosis level from the AR views and from the HR view.

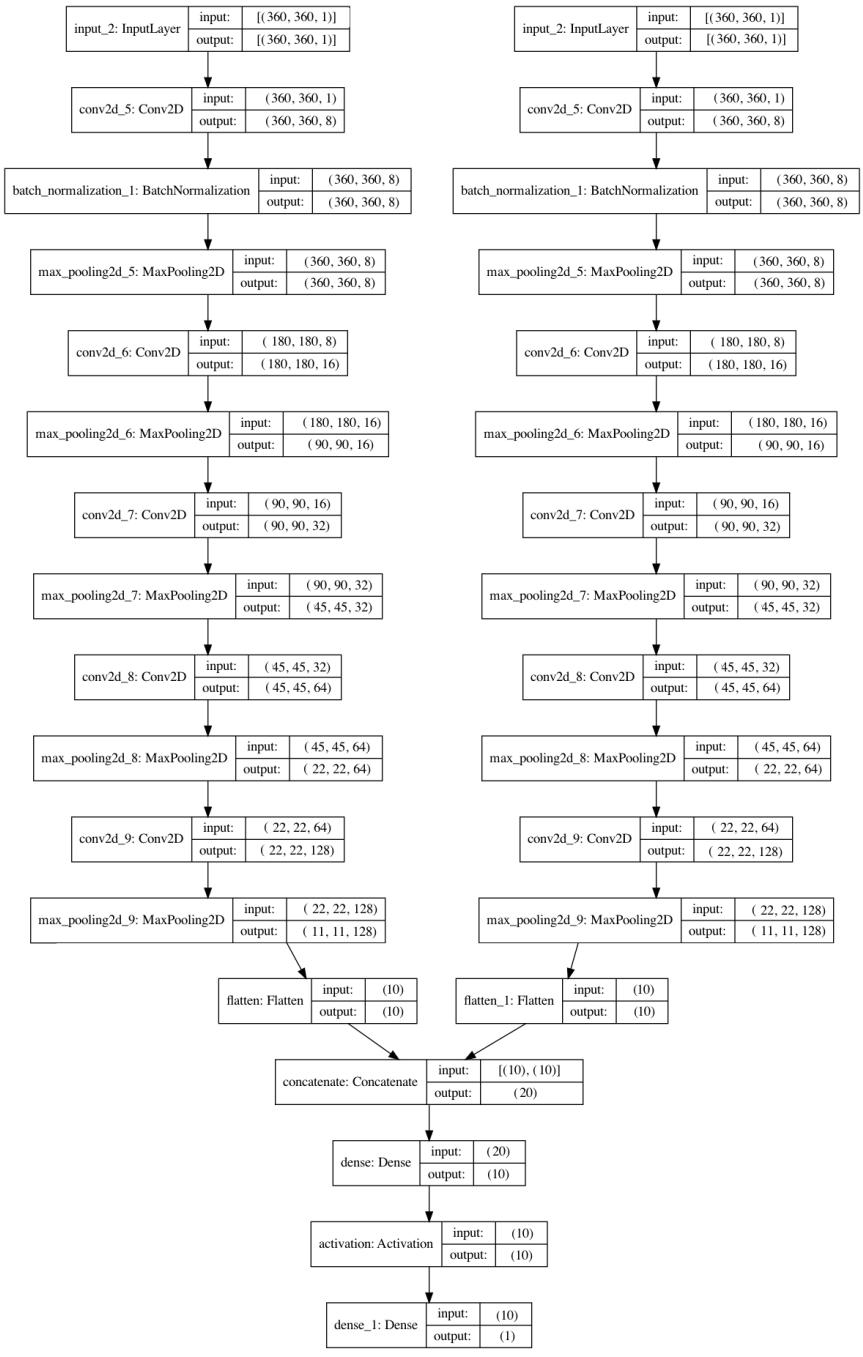

Fig. 3. Sketch of the two-branch architecture used for the estimation of hepatic steatosis level from paired AR and HR views.

trained on HR data, and the two-branch architecture (Fig. 3) trained on paired AR and HR data.

3 Results and Discussion

In this work, we investigate if convolutional neural network architectures could learn to predict the H-MRS index from US sequences. For each subject, two US clips were acquired from different projections, in order to understand which one of the two is the most informative, and if the two projections may convey complementary information about the liver tissue composition.

Performances of the three approaches on the test set have been computed and reported in Table 3. The AR CNN model, trained on images representing full parenchyma, achieves the best regression performances compared to the HR CNN model, with a minimun RMSE of 1.11 and an error standard deviation of 0.77. Also the concatenated architecture, AR & HR CNN, allows to achieve very good results in regression (RMSE is 1.32 with a standard deviation of 1.11), but it does not outperform the results obtained with the AR CNN model.

This leads us to think that the HR clips do not provide additional or complementary information with respect to the information already encoded in the AR images.

A paired T-test has been used to assess the Pearson correlation between the H-MRS index and the fat fraction predicted by the AR CNN model at each frame of the test set, achieving a correlation of 0.983, with a p-value < 0.001.

Table 2. Values of the validation loss for the four sets of the 4-fold cross-validation for the CNN architectures trained on both views.

	Set 1	**Set 2**	**Set 3**	**Set 4**	**Mean**
AR CNN	**1.3274**	3.0840	2.2256	2.6277	2.3162
HR CNN	3.5108	2.9160	3.1472	**2.1721**	2.9365

In order to compare the proposed method with similar approaches, already mentioned in Section 1, we map the fat-liver score computed on the test set through the AR CNN, the HR CNN, and the AR & HR CNN in the classes S0-S3, introduced in [9] and described in Section 2.3. The accuracy of the class-mapping is 90% for all the trained CNNs.We would like to point out that choosing any of the CNN does not affect the class-mapping at all, i.e. all the CNNs predict the same class for each never-seen-before subject, misclassifying only one subject (which belongs to the less represented group, i.e. S2). Also, the classification is performed patient-level; the accuracy of 90% would increase performing a frame

Table 3. The performances of the three CNN models evaluated frame by frame in the test set (RMSE is the root mean square error).

	AR CNN	HR CNN	AR & HR CNN
RMSE	**1.1127**	1.4476	1.3197
Error Std Dev	**0.7701**	1.1757	1.1068

by frame classification, coherently with the metrics showed in Table 3 (RMSE of 1.11, and error standard deviation 0.77).

We can state that our trained CNNs achieve performances comparable to the best results in literature. In addition, we strongly believe that further improvements could be achieved by increasing the dimension of the dataset.

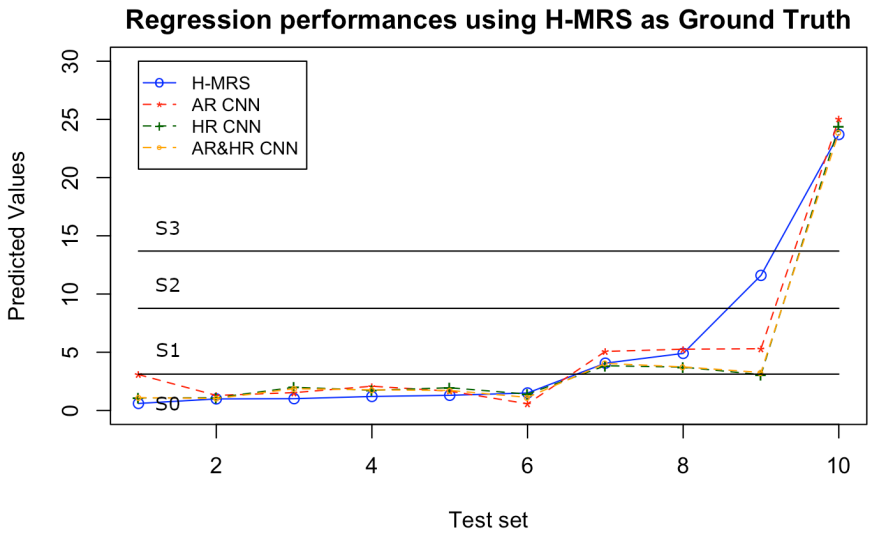

Fig. 4. Comparison of the regression performances on the test set (computed per subject).

4 Conclusions

We propose to use the deep learning to evaluate the non-alcoholic hepatic steatosis starting from US images and using H-MRS as ground truth. We started with

simple CNN models to investigate the efficacy of this approach. Experimental results are encouraging and evidence that even simple models enable quite an accurate steatosis evaluation, even with US images only, without having magnetic resonance available.

In the near future, additional subjects will be included in the study population, allowing for the design of more complex and deep models, as well as for further testing and better validation of the high reliability of our approach.

References

1. Biswas, M., Kuppili, V., Edla, D., Suri, H.S., Saba, L., Marinho, R., Sanches, J., Suri, J.: Symtosis: A liver ultrasound tissue characterization and risk stratification in optimized deep learning paradigm. Computer methods and programs in biomedicine **155**, 165–177 (2018)
2. Bravo, A.A., Sheth, S., Chopra, S.: Liver biopsy. The New England journal of medicine **344 7**, 495–500 (2001)
3. Byra, M., Styczynski, G., Szmigielski, C., Kalinowski, P., Michalowski, L., Paluszkiewicz, R., Piotrzkowska-Wróblewska, H., Zieniewicz, K., Sobieraj, P., Nowicki, A.: Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images. International Journal of Computer Assisted Radiology and Surgery **13**, 1895 – 1903 (2018)
4. 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. Journal of Ultrasound in Medicine **39** (2019)
5. Cowin, G., Jonsson, J.R., Bauer, J., Ash, S., Ali, A., Osland, E., Purdie, D., Clouston, A., Powell, E., Galloway, G.: Magnetic resonance imaging and spectroscopy for monitoring liver steatosis. Journal of Magnetic Resonance Imaging **28** (2008)
6. Eslam, M., Newsome, P.N., Sarin, S.K., Anstee, Q.M., Targher, G., Romero-Gomez, M., Zelber-Sagi, S., Wai-Sun Wong, V., Dufour, J.F., Schattenberg, J.M., Kawaguchi, T., Arrese, M., Valenti, L., Shiha, G., Tiribelli, C., Yki-Järvinen, H., Fan, J.G., Grønbæk, H., Yilmaz, Y., Cortez-Pinto, H., Oliveira, C.P., Bedossa, P., Adams, L.A., Zheng, M.H., Fouad, Y., Chan, W.K., Mendez-Sanchez, N., Ahn, S.H., Castera, L., Bugianesi, E., Ratziu, V., George, J.: A new definition for metabolic dysfunction-associated fatty liver disease: An international expert consensus statement. Journal of Hepatology **73**(1), 202–209 (2020). https://doi.org/https://doi.org/10.1016/j.jhep.2020.03.039
7. Eslam, M., Sanyal, A.J., George, J., Sanyal, A., Neuschwander-Tetri, B., Tiribelli, C., Kleiner, D.E., Brunt, E., Bugianesi, E., Yki-Järvinen, H., Grønbæk, H., Cortez-Pinto, H., George, J., Fan, J., Valenti, L., Abdelmalek, M., Romero-Gomez, M., Rinella, M., Arrese, M., Eslam, M., Bedossa, P., Newsome, P.N., Anstee, Q.M., Jalan, R., Bataller, R., Loomba, R., Sookoian, S., Sarin, S.K., Harrison, S., Kawaguchi, T., Wong, V.W.S., Ratziu, V., Yilmaz, Y., Younossi, Z.: Mafld: A consensus-driven proposed nomenclature for metabolic associated fatty liver disease. Gastroenterology **158**(7), 1999–2014.e1 (2020). https://doi.org/https://doi.org/10.1053/j.gastro.2019.11.312, nonalcoholic Fatty Liver Disease in 2020
8. Han, A., Byra, M., Heba, E., Andre, M., Erdman, J., Loomba, R., Sirlin, C., O'Brien, W.: Noninvasive diagnosis of nonalcoholic fatty liver disease and quan-

tification of liver fat with radiofrequency ultrasound data using one-dimensional convolutional neural networks. Radiology p. 191160 (2020)
9. Karlas, T., Petroff, D., Garnov, N., Böhm, S., Tenckhoff, H., Wittekind, C., Wiese, M., Schiefke, I., Linder, N., Schaudinn, A., Busse, H., Kahn, T., Mössner, J., Berg, T., Troeltzsch, M., Keim, V., Wiegand, J.: Non-invasive assessment of hepatic steatosis in patients with nafld using controlled attenuation parameter and 1h-mr spectroscopy. PLoS ONE **9** (2014)
10. Lascio, N.D., Avigo, C., Salvati, A., Martini, N., Ragucci, M., Monti, S., Prinster, A., Chiappino, D., Mancini, M., D'Elia, D., Ghiadoni, L., Bonino, F., Brunetto, M., Faita, F.: Steato-score: Non-invasive quantitative assessment of liver fat by ultrasound imaging. Ultrasound in medicine and biology **44 8**, 1585–1596 (2018)
11. Loomba, R., Sanyal, A.: The global nafld epidemic. Nature Reviews Gastroenterology and Hepatology **10**, 686–690 (2013)
12. Machado, M., Cortez-Pinto, H.: Non-invasive diagnosis of non-alcoholic fatty liver disease. a critical appraisal. Journal of hepatology **58 5**, 1007–19 (2013)
13. Mancini, M., Prinster, A., Annuzzi, G., Liuzzi, R., Giacco, R., Medagli, C., Cremone, M., Clemente, G., Maurea, S., Riccardi, G., Rivellese, A., Salvatore, M.: Sonographic hepatic-renal ratio as indicator of hepatic steatosis: comparison with (1)h magnetic resonance spectroscopy. Metabolism: clinical and experimental **58 12**, 1724–30 (2009)
14. Popa, S.L., Ismaiel, A., Cristina, P., Cristina, M., Chiarioni, G., David, L., Dumitrascu, D.L.: Non-alcoholic fatty liver disease: Implementing complete automated diagnosis and staging. a systematic review. Diagnostics **11**(6) (2021)
15. Purushotham, S., Tripathy, B.: Evaluation of classifier models using stratified tenfold cross validation techniques (2011)
16. Reddy, D.S., Bharath, R., Rajalakshmi, P.: A novel computer-aided diagnosis framework using deep learning for classification of fatty liver disease in ultrasound imaging. 2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom) pp. 1–5 (2018)
17. Reeder, S., Cruite, I., Hamilton, G., Sirlin, C.: Quantitative assessment of liver fat with magnetic resonance imaging and spectroscopy. Journal of Magnetic Resonance Imaging **34** (2011)
18. Targher, G., Tilg, H., Byrne, C.D.: Non-alcoholic fatty liver disease: a multisystem disease requiring a multidisciplinary and holistic approach. The Lancet Gastroenterology & Hepatology (2021)
19. Xia, M., mei Yan, H., He, W., Li, X., Li, C., zhong Yao, X., kun Li, R., Zeng, M., Gao, X.: Standardized ultrasound hepatic/renal ratio and hepatic attenuation rate to quantify liver fat content: An improvement method. Obesity (Silver Spring, Md.) **20**, 444 – 452 (2012)