Estimating the degree of Non-Alcoholic Fatty Liver Disease (NAFLD) from ultrasound images: preliminary results

S. Kresevic[1], M. Ajcevic[1], M. Giuffrè[2], P. Pupa[3], C. Moretto[3], S. Pennini[3], L.S. Crocè [2] and A. Accardo[1]

[1] *Department of Engineering and Architecture, University of Trieste, Italy*

[2] *Department of Medicine, Surgery and Health Sciences, University of Trieste, Trieste, Italy*

[3] *Prodigys Technology s.r.l., Trieste, Italy*

***Abstract*—The early diagnosis of Non-Alcoholic Fatty Liver Disease (NAFLD) is crucial to prevent fibrosis progression or onset of advanced chronic liver disease. Among the non-invasive methods, ultrasound (US) B-mode imaging is recommended for population screening and follow-up. Hamaguchi's score was proposed to improve the objectivity in the evaluation of the fatty liver US imaging evaluation. In this study, we aimed to preliminary assess the possibility to estimate Hamaguchi's score by means of an advanced ultrasound image semi-automatic analysis.**

The study encompassed a dataset of 220 bariatric patients with NAFLD diagnosed by liver biopsy who underwent ultrasound assessment at the Liver Clinic at Trieste University Hospital. The classification models for the estimation of the three Hamaguchi's sub-scores were produced by semiautomatic US image analysis based on clustering and CNN with transfer learning techniques.

The results showed that the produced models were able to estimate the three sub-scores with high classification accuracy. The predictive models produced for the estimation of liver brightness hepatorenal echo contrast, the diaphragm deep attenuation, and the vessel blurring sub-scores presented a classification accuracy on the validation set of 90.5%, 83.3%, and 84.0%, respectively.

In conclusion, in this preliminary study, the results assessed the possibility to produce the NAFLD computer-aided diagnostic models based on analysis of US images.

***Keywords*—NAFLD, Hamaguchi's score, ultrasound images, Artificial Intelligence.**

I. Introduction

NAFLD represents a spectrum of diseases related to excessive fat deposition in the liver, ranging from simple steatosis (i.e., non-alcoholic fatty liver, NAFL) to non-alcoholic steatohepatitis (NASH), characterized by lobular inflammation and hepatocyte ballooning, which over time can increase liver fibrosis, thus promoting liver cirrhosis or hepatocellular carcinoma [1]. NAFLD is the most common chronic liver disease in the world (approximately 25% of the global population has been affected by NAFLD [2] and is projected to be 33.5 % by 2030 [3]). Current clinical guidelines recommend dietary treatment as the most efficient preventive intervention in addition to early evaluation and monitoring of liver functions to prevent advanced fibrosis or hepatocellular carcinoma [4], [5]. The gold standard for staging is liver biopsy, which is expensive and invasive, making it unsuitable for broad screening at the population level [6].

Among the noninvasive methods, ultrasound B-mode imaging has been recommended as the preferred first-line diagnostic procedure for imaging of NAFLD in adults by the clinical practice guidelines of the European Association for the Study of the Liver released together with the European Association for the Study of Diabetes and the European Association for the Study of Obesity [7].

But this method has the limitation that the evaluation of a US image is operator dependent: several studies conducted have reported significant intra- and inter-observer variability in the assessment of ultrasonographic findings of hepatic steatosis [8]-[10]. In particular, an agreement between pairs of experienced observers of only 47-59% on the first reading and 59-64% on the second reading was reported [9].

In daily practice, physicians evaluate hepatic steatosis on the US image by analyzing some of its features: the echogenicity of the liver parenchyma compared with the echogenicity of the kidney, the visibility of intrahepatic vessels and the diaphragm blurring [11]. These features allow physicians to assess the presence of hepatic steatosis because an excess fat component in the liver would make it shine brighter to a greater extent than in the kidney (normally isoecogenic) and would attenuate the ultrasound probe's beam making visualization of intrahepatic vascular structures and the diaphragm poor.

To improve objectivity, Hamaguchi et al. proposed a semi-quantitative US-based scoring method [12]. In particular, the proposed scoring system is based on the comparison of liver and kidney echogenicity, assessment of liver brightness (scored liver brightness from 0 to 3), of deep attenuation of diaphragmatic contours by the liver (scored from 0 to 2), and of liver vessel blurring (scored from 0 to 1). The sum of the aforementioned sub-scores yields Hamaguchi's score ranging from 0 (corresponding to a healthy liver) to 6 (corresponding to a fatty liver). Although Hamaguchi's score improved the standardization of the degree of hepatic lipid accumulation in the liver, it still suffers from the fact that the assessment is a subjective measure strongly influenced by the expertise of the physician evaluating the US image.

Over the past decade methods based on Artificial Intelligence (AI) are used for identifying and predicting patterns or connections within large datasets in various fields of medicine, demonstrating utility in the diagnostic process. A recent meta-analysis on the AI-based methods for liver diagnosis reported high-accuracy results of the application of these methods on diagnostic imaging for the diagnosis and staging of NAFLD [13]. However, despite the numerous applications of AI in the diagnostic field of NAFLD, methods based on AI have never been applied to the analysis for Hamaguchi's score estimation. The aim of this study is to develop Machine Learning (ML) and Deep Learning (DL) algorithms for the advanced analysis of US images that can estimate the sub-scores of the various categories of Hamaguchi's score and thus propose a new approach for automatic analysis of the same.

II. Materials and Methods

A. *Data acquisition and dataset definition*

The study includes 220 bariatric patients with NAFLD diagnosed by liver biopsy who underwent US assessment at the Liver Clinic at Trieste University Hospital. The clinical and radiological data of included patients were analyzed to create the study dataset. The inclusion criteria were liver biopsy-based NAFLD diagnosis, and US assessment characterized by the visibility of liver and renal parenchyma, visibility of intrahepatic vessels, and visibility of the diaphragm.

To assign the sub-scores constituting the Hamaguchi score, three parameters must be evaluated on the US image, but not in all cases all three parameters are assessable on a single US scan as it depends on the bariatric patient's physical conformations: there are patients who have all three parameters of interest in one US image, patients who have liver and kidney in a single US image while diaphragm and intrahepatic vessels in another, as well as patients who have all three parameters in three different US images. So, with the support of physicians in the selection, three datasets were created: the Hepatorenal dataset, the Diaphragm dataset, and the Vessel dataset.

All three datasets contain 220 US images (one for each patient). For each of the three datasets, the images were appropriately evaluated by four physicians who independently assigned the respective sub-scores constituting the Hamaguchi score, and where ratings were not equal, the scores were reviewed and decided by consensus.

B. *The proposed framework to evaluate automatically Hamaguchi's score*

The proposed approach for automatic estimation of Hamaguchi's score is based on separate estimation of sub-scores by semiautomatic analysis of US images. Considering that the sub-score assessments are based on the identification of different features, three sub-score related algorithms were developed. For the evaluation of the liver brightness hepatorenal echo contrast score, a semi-automatic evaluation based on clustering was adopted, while for the evaluation of the diaphragm deep attenuation score and of the vessel blurring score a semi-automatic methodology based on the development of two CNNs with transfer learning techniques was implemented.

Considering the variability of the US images and moderate sample size, to produce performant classification algorithms, additional pre-processing was performed by manually delineating the ROIs to define Diaphragm dataset and Vessel dataset.

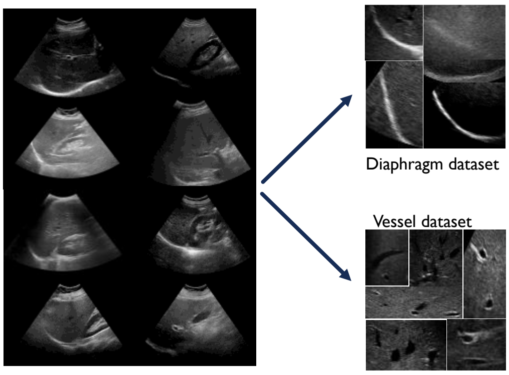
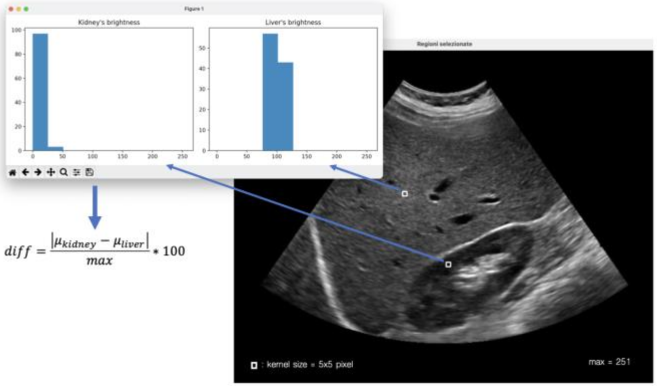


Fig. 1: The ROIs definition from raw US images for the Diaphragm and Vessel datasets.

C. *Estimate of liver brightness hepatorenal echo contrast score*

For the calculation of liver brightness hepatorenal echo contrast, two ROIs were manually identified by the medical team: one in the lighter region of the liver parenchyma, and the other in the darker area of the cortical area of the kidney. A script written in Python allows to select two ROIs (of size 5x5 pixel to ensure that the cortical region of the kidney is selected without including its interface with the liver) and process them to extract the feature of interest. In this way, features related to the echogenicity of the liver were obtained for each US image sample. All previously saved features are later analyzed with K-means by setting the number of four classes (liver brightness hepatorenal echo contrast score could be 0, 1, 2, and 3). In this way, four classes were distinguished.

Based on this clustering, a classification model was built to evaluate the feature extracted from an ultrasound image of the liver by assigning it a score ranging from 0 to 3.



$$diff = \frac{|\mu_{kidney} - \mu_{liver}|}{max} * 100$$

Fig. 2: The feature extraction for the liver brightness hepatorenal echo contrast sub-score. The two ROIs are manually selected by the physician from the raw image. One is in the lighter region of the liver parenchyma, and the other is in the darker region of the cortical zone of the kidney. When these two ROIs are selected, a feature is extracted by calculating the distance between the average intensity level of the area corresponding to the liver (μ_{liver}) and the average intensity level of the area corresponding to the kidney (μ_{kidney}). The difference is divided by the maximum intensity present in the US image.

D. *Estimate of diaphragm deep attenuation score and liver vessel blurring score*

The diaphragm deep attenuation score is evaluated according to how well the diaphragm is visible: the greater the presence of fat in the liver, the greater the absorption of ultrasound signal and thus the lower the visibility of the diaphragm. Similarly for liver vessel blurring score. So, we chose to evaluate these scores using CNNs and transfer learning for each kind of score.

Before implementing and training the CNNs, the image datasets of both the diaphragm's dataset and the vessel's dataset are divided into three sub-sets of images each: training set, validation set, and test set with a ratio of 75%, 15%, and 10% respectively. So, the sub-set consists of 165, 33, and 22 US images on the training set, validation set, and test set respectively. The test set is used only once the model has been fully trained. To give correct assessments, the test set must be well maintained and contain a variety of images covering the various classes of each sub-score.

Although the transfer learning technique is used to produce CNNs with high classification accuracy, many images are still required in the training set, and the sample case histories need to be balanced. So, the training set has been also augmented and balanced via data augmentation and the techniques which consisted of 10° right/left rotation, random in/out zooming ranging from 0.8 to 1.2, and horizontal flip operations in a random manner.

Ten CNNs were implemented with transfer learning techniques (TABLE I) to choose the most suitable and performing network for the diaphragm deep attenuation score and the vessel blurring score. The networks' development was implemented with Python scripts by importing the Tensorflow and Keras libraries. In particular, the pre-trained networks VGG-16 and VGG-19 from Keras are imported as the base model.

TABLE I: ARCHITECTURES OF THE TEN IMPLEMENTED CNNs

#	Base model	Additional convolutional block	Classification block
1	VGG-16	-	Flatten layer Dense layer (512 nodes) Dense layer (256 nodes) Output layer
2	VGG-16	Conv2D layer (512 nodes) Conv2D layer (512 nodes) MaxPooling2D layer	Flatten layer Dense layer (512 nodes) Dense layer (256 nodes) Output layer
3	VGG-16	Conv2D layer (512 nodes) Conv2D layer (256 nodes) MaxPooling2D layer	Flatten layer Dense layer (4096 nodes) Dense layer (512 nodes) Output layer
4	VGG-16	Conv2D layer (512 nodes) Conv2D layer (256 nodes) MaxPooling2D layer	Flatten layer Dense layer (4096 nodes) Dense layer (4096 nodes) Output layer
5	VGG-16	Conv2D layer (512 nodes) Conv2D layer (256 nodes) MaxPooling2D layer	Flatten layer Dense layer (512 nodes) Dense layer (256 nodes) Dense layer (128 nodes) Output layer
6	VGG-16	Conv2D layer (512 nodes) Conv2D layer (256 nodes) Conv2D layer (128 nodes) MaxPooling2D layer	Flatten layer Dense layer (1024 nodes) Dense layer (512 nodes) Dense layer (256 nodes) Output layer
7	VGG-16	Conv2D layer (512 nodes) Conv2D layer (256 nodes) Conv2D layer (128 nodes) MaxPooling2D layer	Flatten layer Dense layer (1024 nodes) Dense layer (512 nodes) Output layer
8	VGG-19	-	Flatten layer Dense layer (512 nodes) Dense layer (256 nodes) Output layer
9	VGG-19	Conv2D layer (512 nodes) Conv2D layer (256 nodes) MaxPooling2D layer	Flatten layer Dense layer (512 nodes) Dense layer (256 nodes) Output layer
10	VGG-19	Conv2D layer (512 nodes) Conv2D layer (256 nodes) Conv2D layer (128 nodes) MaxPooling2D layer	Flatten layer Dense layer (1024 nodes) Dense layer (512 nodes) Dense layer (256 nodes) Output layer

In output to VGG-Nets, a convolutional layers' block is added with trainable parameters to extrapolate additional features. After this block, there is a flatten layer that brings all the outputs of the previous layer onto a one-dimensional array to be given as input to the top layer. The top layer of the different implemented architectures shown in TABLE I change depending on the number of convolutional layers and the number of nodes per layer. All top layers share the fact that they have the same activation function (the Rectified Linear Unit - ReLU). Between the various dense layers of all models, there are dropout layers: these layers keep the overfitting of the model under control.

The output layer of the models differs between the two scores: the diaphragm deep attenuation score could be 0, 1, or 2, so should be able to predict three classes. Instead, the vessel blurring score should only involve two classes (0 or 1).

For the diaphragm deep attenuation score, which is a multi-class problem, then the network outputs an array containing the probability associated with each score value. The array index that will have obtained the highest probability will constitute the final output. The activation function is a softmax and the loss function is the categorical cross entropy. Instead the vessel blurring score is a binary problem and then the output layer has only one neuron. This neuron will report the value of the predicted parameter and the probability of the realized prediction. This is a binary classification problem, so the activation function is a sigmoid, and the loss function calculation is the binary cross entropy. The performances of models were evaluated by classification accuracy, loss, Area under the ROC curve (AUC), and precision.

III. Results

The predictive model produced for the classification of liver brightness hepatorenal echo contrast score presented a classification accuracy of 90.5% considering the sub-scores labeled by physicians. In the misclassified cases the maximum error for this sub-score was one point. Regarding the model for estimation of the diaphragm deep attenuation and the vessel blurring sub-scores, architecture #10 and architecture #2 (TABLE I) showed the best classification performance on the validation dataset, respectively. The classification accuracy, loss, AUC, and precision obtained for these models are reported in TABLE II.

TABLE II: THE PERFORMANCE PARAMETERS OF THE TWO SELECTED MODELS OF EACH SUBSCORES

	Accuracy	Loss	AUC	Precision
Diaphragm deep attenuation	83.25 %	0.48	0.93	83.0 %
Vessel blurring	84.05 %	0.44	0.89	89.0 %

The identified models were subsequently tested on the test dataset. The confusion matrixes for diaphragm deep attenuation and vessel blurring sub-score models are reported in TABLE III-A and III-B, respectively. The algorithm for the estimation of diaphragm deep attenuation showed an accuracy of 81.8%. The misclassified scores were under or over-estimated by one point. Reading the vessel blurring sub-score model the identified model presented a classification accuracy of 86.4% on the test dataset. In both cases, the models maintained the performance obtained during the training and validation process.

TABLE III: COMPARISON OF THE PREDICTED SUBSCORE AND THE SUBSCORE ASSIGNED BY THE PHYSICIANS ON THE TEST SET REPRESENTED BY A NORMALIZED CONFUSION MATRIX

A) DIAPHRAGM DEEP ATTENUATION SUBSCORE

True label	Score 0	0.875	0.125	0
	Score 1	0.125	0.75	0.125
	Score 2	0	0.125	0.875
		Score 0	Score 1	Score 2
		Predicted label		

B) VESSEL BLURRING SUBSCORE

True label	Score 0	0.73	0.27
	Score 1	0	1.00
		Score 0	Score 1
		Predicted label	

IV. Discussion

The early diagnosis of NAFLD is important to prevent fibrosis progression, liver cirrhosis, or hepatocellular carcinoma [4], [5]. Among the noninvasive methods, US B-mode imaging is recommended for population screening and follow-up [7]. But this method has the limitation that the evaluation of a US image is operator dependent [8]-[10]. Hamaguchi's score improved the standardization of the degree of hepatic lipid accumulation in the liver [12], but this score suffers from the fact that the assessment is a subjective measure strongly influenced by the expertise of the physician evaluating the US image. In this study, we aimed to preliminary assess the possibility to estimate Hamaguchi’s score by means of advanced image analysis.

The results of this study showed that methods based on AI can estimate the three sub-scores which determine Hamaguchi’s score. Indeed, the produced models presented a high classification accuracy for all three sub-scores. The results obtained for all three sub-scores are clinically relevant and suggest that such decision support systems in the future may support the diagnosis of liver disease in a way that will reduce intra- and inter-operator assessment error.

The study presents the following limits: the possible reliance on manual segmentations, the use of images acquired from only one type of US scanner, the moderate sample size, and the retrospective nature of the study. The latter limited the possibility to build a dataset with balanced case histories.

These preliminary results should be confirmed and potentially improved on a larger sample size and more clinically balanced dataset. In addition, the presented approach could be additionally improved by developing an automatic tool for ROI selection.

In conclusion, in this preliminary study, the results assessed the possibility to produce the NAFLD computer-aided diagnostic models based on analysis of US images.

References

[1] Brunt EM, Janney CG, Di Bisceglie AM, Neuschwander-Tetri BA, Bacon BR. Nonalcoholic steatohepatitis: A proposal for grading and staging the histological lesions. Am J Gastroenterol 1999;94:2467–74.

[2] Huang DQ, El-Serag HB, Loomba R. Global epidemiology of NAFLD-related HCC: trends, predictions, risk factors and prevention. Nat Rev Gastroenterol Hepatol. 2021;18(4):223-238. doi:10.1038/s41575-020-00381-6

[3] Estes C, Razavi H, Loomba R, Younossi Z, Sanyal AJ. Modeling the epidemic of nonalcoholic fatty liver disease demonstrates an exponential increase in burden of disease. Hepatology. 2018;67(1):123-133. doi:10.1002/hep.29466

[4] Styne DM, Arslanian SA, Connor EL, et al. Pediatric Obesity-Assessment, Treatment, and Prevention: An Endocrine Society Clinical Practice Guideline. J Clin Endocrinol Metab. 2017;102(3):709-757. doi:10.1210/jc.2016-2573

[5] Valerio G, Maffeis C, Saggese G, et al. Diagnosis, treatment and prevention of pediatric obesity: consensus position statement of the Italian Society for Pediatric Endocrinology and Diabetology and the Italian Society of Pediatrics. Ital J Pediatr. 2018;44(1):88. Published 2018 Jul 31. doi:10.1186/s13052-018-0525-6

[6] Ting Soon GS, Wee A. Liver biopsy in the quantitative assessment of liver fibrosis in nonalcoholic fatty liver disease. Indian J Pathol Microbiol. 2021;64(Supplement):S104-S111. doi:10.4103/IJPM.IJPM_947_20

[7] Ferraioli G, Soares Monteiro LB. Ultrasound-based techniques for the diagnosis of liver steatosis. World J Gastroenterol. 2019;25(40):6053-6062. doi:10.3748/wjg.v25.i40.6053

[8] Lee SS, Park SH. Radiologic evaluation of nonalcoholic fatty liver disease. World J Gastroenterol. 2014;20(23):7392-7402. doi:10.3748/wjg.v20.i23.7392

[9] 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. doi:10.2214/AJR.07.2123

[10] Cengiz M, Sentürk S, Cetin B, Bayrak AH, Bilek SU. Sonographic assessment of fatty liver: intraobserver and interobserver variability. Int J Clin Exp Med. 2014;7(12):5453-5460. Published 2014 Dec 15.

[11] Targher G. What's Past Is Prologue: History of Nonalcoholic Fatty Liver Disease. Metabolites. 2020;10(10):397. Published 2020 Oct 8. doi:10.3390/metabo10100397

[12] Hamaguchi M, Kojima T, Itoh Y, et al. The severity of ultrasonographic findings in nonalcoholic fatty liver disease reflects the metabolic syndrome and visceral fat accumulation. Am J Gastroenterol. 2007;102(12):2708-2715. doi:10.1111/j.1572-0241.2007.01526.x

[13] Decharatanachart P, Chaiteerakij R, Tiyarattanachai T, Treeprasertsuk S. Application of artificial intelligence in chronic liver diseases: a systematic review and meta-analysis. BMC Gastroenterol. 2021;21(1):10. Published 2021 Jan 6. doi:10.1186/s12876-020-01585-5