This is a repository copy of *Machine learning models predict liver steatosis but not liver fibrosis in a prospective cohort study*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/201949/

Version: Published Version

Article:
Mamandipoor, B. orcid.org/0000-0001-9441-3815, Wernly, S., Semmler, G. et al. (6 more authors) (2023) Machine learning models predict liver steatosis but not liver fibrosis in a prospective cohort study. Clinics and Research in Hepatology and Gastroenterology, 47 (7). 102181. ISSN 2210-7401

https://doi.org/10.1016/j.clinre.2023.102181

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Clinics and Research in Hepatology and Gastroenterology 47 (2023) 102181

Contents lists available at ScienceDirect

Clinics and Research in Hepatology and Gastroenterology

journal homepage: www.elsevier.com/locate/clinre

Original article

Machine learning models predict liver steatosis but not liver fibrosis in a prospective cohort study

Behrooz Mamandipoor [a], Sarah Wernly [b], Georg Semmler [c], Maria Flamm [d], Christian Jung [e], Elmar Aigner [f], Christian Datz [b], Bernhard Wernly [b,d,#], Venet Osmani [g,#,*]

[a] Fondazione Bruno Kessler Research Institute, Trento, Italy
[b] Department of Internal Medicine, General Hospital Oberndorf, Teaching Hospital of the Paracelsus Medical University, Salzburg, Austria
[c] Division of Gastroenterology and Hepatology, Department of Medicine III, Medical University of Vienna, Vienna, Austria
[d] Institute of general practice, family medicine and preventive medicine, Paracelsus Medical University, Salzburg, Austria
[e] Department of Cardiology, Pulmonology and Vascular Medicine, Medical Faculty, Heinrich-Heine-University Düsseldorf, Germany
[f] Clinic I for Internal Medicine, University Hospital Salzburg, Paracelsus Medical University, Salzburg, Austria
[g] Information School, University of Sheffield, United Kingdom

ARTICLE INFO

Keywords:
Steatosis
Liver fibrosis
Machine learning
Predictive modelling
Gender differences
Patient self-reported outcomes

ABSTRACT

Introduction: Screening for liver fibrosis continues to rely on laboratory panels and non-invasive tests such as FIB-4-score and transient elastography. In this study, we evaluated the potential of machine learning (ML) methods to predict liver steatosis on abdominal ultrasound and liver fibrosis, namely the intermediate-high risk of advanced fibrosis, in individuals participating in a screening program for colorectal cancer.

Methods: We performed ultrasound on 5834 patients admitted between 2006 and 2020, and transient elastography on a subset of 1240 patients. Steatosis on ultrasound was diagnosed if liver areas showed a significantly increased echogenicity compared to the renal parenchyma. Liver fibrosis was defined as a liver stiffness measurement $\geq$8 kPa in transient elastography. We evaluated the performance of three algorithms, namely Extreme Gradient Boosting, Feed-Forward neural network and Logistic Regression, deriving the models using data from patients admitted from January 2007 up to January 2016 and prospectively evaluating on the data of patients admitted from January 2016 up to March 2020. We also performed a performance comparison with the standard clinical test based on Fibrosis-4 Index (FIB-4).

Results: The mean age was 58$\pm$9 years with 3036 males (52%). Modelling laboratory parameters, clinical parameters, and data on eight food types/dietary patterns, we achieved high performance in predicting liver steatosis on ultrasound with AUC of 0.87 (95% CI [0.87–0.87]), and moderate performance in predicting liver fibrosis with AUC of 0.75 (95% CI [0.74–0.75]) using XGBoost machine learning algorithm. Patient-reported variables did not significantly improve predictive performance. Gender-specific analyses showed significantly higher performance in males with AUC of 0.74 (95% CI [0.73–0.74]) in comparison to female patients with AUC of 0.66 (95% CI [0.65–0.66]) in prediction of liver fibrosis. This difference was significantly smaller in prediction of steatosis with AUC of 0.85 (95% CI [0.83–0.87]) in female patients, in comparison to male patients with AUC of 0.82 (95% CI [0.80–0.84]).

Conclusion: ML based on point-prevalence laboratory and clinical information predicts liver steatosis with high accuracy and liver fibrosis with moderate accuracy. The observed gender differences suggest the need to develop gender-specific models.

Introduction

Non-alcoholic fatty liver disease (NAFLD) is a highly prevalent disease, affecting up to 25% of the global population worldwide and more than 40% in some countries [1]. The development of NAFLD is closely related to obesity and metabolic syndrome, creating a vicious

* Corresponding author.
E-mail address: v.osmani@sheffield.ac.uk (V. Osmani).
# Contributed equally.

https://doi.org/10.1016/j.clinre.2023.102181

Available online 17 July 2023
2210-7401/© 2023 The Authors. Published by Elsevier Masson SAS. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

circle with NAFLD [2]. For this reason, it is recommended in the European guidelines that patients with risk factors for NAFLD (such as diabetes and obesity) should be examined for the presence of hepatic steatosis, e.g., using simple abdominal ultrasound [3] since patients with NAFLD not only suffer significant hepatologic and liver-specific morbidity and mortality, but also have increased cardiovascular risk [4].

To predict liver-specific outcome, close evaluation with special attention to development of liver fibrosis is necessary [4,5]. Up to 20% of patients with NAFLD experience disease progression to NASH [22]. Also, patients with NASH are at significantly increased risk for cirrhosis and hepatocellular carcinoma and should be managed by hepatologists [6].

For evaluation of liver fibrosis, biomarker-based scores for broad screening, transient elastography and liver biopsy as the gold standard are available in the armamentarium [4]. If biomarkers do not exclude liver fibrosis, further evaluation with transient elastography is recommended. However, transient elastography can also be performed as a screening tool, depending on local availability [23], also when considering low specificity of blood-based biomarkers [24]. The European Association for the Study of the Liver (EASL) and German Society for Gastroenterology, Digestive and Metabolic Diseases (DGVS) recommend screening of high-risk patients, particularly for the presence of steatosis using ultrasound or FLI if ultrasound is not available. Further evaluation is recommended with Fib-4 and, if necessary, additional hepatological work-up in case of a positive finding.

Machine learning uses algorithms to detect patterns in large, heterogeneous data sets. These patterns would be difficult or impossible to identify even for experts with many years of training. Especially in view of the large amount of data generated in modern medicine, the application of machine learning in numerous medical fields seems promising. On the other hand, machine learning should be held to the same standards as other medical innovations and its benefits should first be established by means of exploratory studies and ultimately proven by means of randomised trials [7].

Machine learning has already been applied to predict NAFLD and showed good sensitivity [8–10]. Conversely, much simpler and thus more user-friendly scores also showed similar sensitivity [11]. Especially the prediction of relevant liver fibrosis [12] would be of high interest for clinical practice.

Therefore, the objective of this work was to evaluate performance of machine learning models in predicting liver steatosis and liver fibrosis based on cross-sectional data in a Central European cohort of symptomatic patients.

Main contributions

The main contributions of our work are as follows: i) our study is the first to prospectively evaluate predictive performance of steatosis and liver fibrosis from only routinely collected clinical data; ii) we show that steatosis can be predicted with a high performance, paving the way towards a wide-scale screening instrument; iii) we also show that routinely collected clinical data is not sufficient for a high-performance prediction of liver fibrosis; iv) we establish clinically meaningful cut-offs of clinical parameters where the estimated risk of steatosis significantly increases; v) we show that patient-reported outcomes related to nutrition and lifestyle do not significantly contribute to risk estimation and, vi) we show significant differences in risk estimation between genders, particularly in prediction of liver fibrosis, suggesting the necessity to develop gender-specific models.

Methods

Ethics statement

The study and data acquisition was performed according to the Helsinki Declaration and was approved by the local ethics committee (Ethics Commission for the Province of Salzburg, committee approval no. 415-E/1262/2–2010). Written informed consent was obtained from every participant and all assessments were performed according to national or international guidelines.

Setting and study population

We included 5834 participants from the Salzburg Colon Cancer Prevention Initiative (Sakkopi) in this analysis. The participants were asymptomatic subjects who underwent opportunistic colon cancer screening in Austria. As part of this study, anthropometric, laboratory, and sonographic additional findings were collected. The study cohort consists of asymptomatic subjects screened for colorectal cancer between January 2007 and March 2020. The study was performed in Austria at a single centre hospital. We obtained anthropomorphic, clinical as well as laboratory parameters in all participating subjects [7, 13]. The study participants completed a questionnaire about their medical history. We defined and calculated body mass index (BMI), arterial hypertension, smoking status, dyslipidemia, as well as metabolic syndrome in accordance with published guidelines [14,15]. We performed abdominal ultrasonography in all patients. The liver was considered normal if the echogenicity was similar to the renal parenchyma. The liver was considered steatotic if areas showed a significantly increased echogenicity compared to the renal parenchyma.

Measurement protocol

A liver stiffness measurement by FibroScan is a non-invasive test that assesses the degree of fibrosis (scarring) in the liver. The protocol typically involves the following steps: The patient is asked to lie on their back with their right arm raised behind their head; A gel is applied to the skin over the liver area, which helps to conduct sound waves; The FibroScan probe, which is a small hand-held device, is placed on the skin over the liver area; The probe sends low-frequency sound waves through the liver tissue, which bounce back at different speeds depending on the stiffness of the liver; The FibroScan machine measures the speed of the sound waves and calculates a liver stiffness measurement (in kilopascals, kPa). We obtained 10 measurements and only accepted measurements with a variation range of $<15\%$. All patients had an overnight fast; all operators were board-certified gastroenterologists or supervised by a board-certified gastroenterologist and board certification was obtained according to the Austrian law.

Missing values

To handle the missing values, we used mean and mode imputation for continuous and categorical variables, respectively. The percentage of missing values for the main variables were as follows: Amylase 4.5%, INR 4.7%, Transferrin saturation 7%, Lipase 9%, HOMA-IR 17%, Albumin 38%, Baso 72%, while the rest of variables had less than 3% missing values. We normalised the variables by scaling them between zero and one. We excluded patients with known liver disease and alcohol abuse (>1 standard drink for women, >2 for men). For a subgroup ($n = 1240$), liver stiffness measured by transient elastography was available.

Outcomes definition

The primary outcome was the diagnosis of liver steatosis as defined above, namely areas with a significantly increased echogenicity compared to the renal parenchyma. For this purpose, the echogenicity of the liver is dichotomized by the examiner. Further, we defined liver fibrosis as a binary outcome using LSM $\geq$8 kPa [25]. Throughout this work, fibrosis is meant to signify intermediate-high risk of advanced fibrosis.

Table 1
Variables used in the development and validation of the predictive models for steatosis and liver fibrosis.

Type	Variables
Demographic	Age, gender, weight, height, BMI
Laboratory	ALT, AST, AP, Bilirubin, Cholinestherase, GGT, Blood_glucose, Cholesterol, C-reactive protein, Uric acid, HDL cholesterol, hemoglobin, LDH, LDL cholesterol, Platetelets, Triglycerides, TSH, INR, Lipase, Amylase, Baso, Blood sedation rate, Iron, Ferritin, Transferrin saturation, Transferrin, Albumin, HOMA-IR, Oral glucose tolerance test
Patient-reported (interval is per week, if not specified otherwise)	Alcohol per day, coffee cups, vegetable portions, sugar sweetened beverages, red meat consumption, fast food meals, fruits porition, current or past smoker
Other measurements	Systolic arterial pressure, diastolic arterial pressure, ACE Inhibitors, Statins, Family history, aspirin use

Study design and sensitivity analysis

We prospectively evaluated the performance of machine learning methods to predict the outcomes of interest. Namely, the dataset was divided into a cohort of patients admitted in the first 8 years (from January 2007 up to January 2016) which comprised the model derivation cohort. The validation cohort comprised the patients admitted in the last 4 years (from January 2016 up to March 2020), in which the performance of the model for predicting steatosis and liver fibrosis was evaluated on.

Furthermore, we performed sensitivity analysis to investigate predictive performance of objective laboratory variables only and compare these results with the performance of our method when subjective, patient-reported variables are included. The patient-reported variables measured the amount of alcohol consumption, coffee, sweetened drinks, highly processed food, smoking, red meat, vegetables and fruit.

Finally, we also perform analyses to investigate the performance of our models in predicting the outcomes of interest in male and female patients separately.

Variables of interest

We collected several demographic variables, including age, gender, weight, height and Body Mass Index (BMI) for each patient visit. Additionally, we also collected several laboratory variables, as well as patent reported variables, as shown in Table 1.

Statistical analysis

We analysed baseline characteristics of patients using medians (IQRs) for continuous variables and frequencies (percentages) for categorical variables. We used the Kruskal–Wallis test for continuous variables and the chi-square test for categorical variables to compare subgroups of alive and deceased patients.

Machine learning model development and validation

We compare the predictive performance of three algorithms, namely, Extreme Gradient Boosting (XGBoost) [16], Feed-Forward (FF) neural network and Logistic Regression (LR) [17,18] (details shown in Appendix 5) .

All the three models were tuned for the best hyperparameters on the internal evaluation cohorts in each study design and outcome definitions. The models' hyperparameters were optimised through exhaustive grid-search for maximising the F-1 score metric and set for the final prospective validation.

Table 2
Baseline characteristics of the steatosis patient cohort, including interquartile range in square brackets. Measurement units are provided in round brackets.

Patients	5834	Cholesterol, median [Q1,Q3] (mg/dL)	219.0 [191.0,248.0]
Age, median [Q1, Q3] (years)	57.8 [51.8,65.7]	C-reactive protein, median [Q1,Q3] (mg/dL)	0.2 [0.1,0.4]
Gender, n (%) male	3036 (52.0)	Uric acid, median [Q1,Q3] (mg/dL)	5.6 [4.6,6.7]
Gender, n (%) female	2798 (48.0)	HDL Cholesterol, median [Q1,Q3] (mg/dL)	56.0 [47.0,67.0]
Weight, median [Q1,Q3] (kg)	78.0 [68.0,89.0]	Haemoglobin, median [Q1,Q3] (g/dL)	14.6 [13.8,15.4]
Height, median [Q1,Q3] (cm)	170.0 [164.0,178.0]	Lactate dehydrogenase (LDH), median [Q1,Q3] (IU/L)	170.0 [153.0,191.0]
BMI, median [Q1, Q3]	26.6 [24.0,29.7]	LDL Cholesterol, median [Q1,Q3] (mg/dL)	140.0 [116.0,167.0]
Systolic Blood Pressure, median [Q1,Q3] (mmHg)	130.0 [120.0,140.0]	Platelets, median [Q1,Q3] (x 109/L)	230.0 [197.0,268.0]
Diastolic Blood Pressure, median [Q1,Q3] (mmHg)	80.0 [75.0,90.0]	Triglycerides, median [Q1,Q3] (mg/dL)	107.0 [79.0,149.0]
ALT, median [Q1, Q3] (IU/L)	21.0 [15.0,30.0]	TSH, median [Q1, Q3] (mIU/L)	1.4 [1.0,2.0]
AST, median [Q1, Q3] (IU/L)	21.0 [17.0,26.0]	INR, median [Q1, Q3]	1.0 [1.0,1.0]
Alkaline Phosphatase (AP), median [Q1,Q3] (U/L)	63.0 [53.0,76.2]	Lipase, median [Q1,Q3] (U/L)	28.0 [20.0,38.0]
Bilirubin, median [Q1,Q3] (umol/L)	0.7 [0.5,0.9]	Amylase, median [Q1,Q3] (U/L)	25.0 [20.0,32.0]
Cholinesterase, median [Q1,Q3] (U/mL)	9900.0 [8414.5,11,369.0]	Basophils, median [Q1,Q3] (x 109/L)	0.6 [0.4,0.8]
GGT, median [Q1, Q3] (IU/L)	26.0 [17.0,43.0]	Blood sed. rate, median [Q1,Q3] (mm/hr)	6.0 [3.0,10.0]
Blood Glucose, median [Q1,Q3] (mg/dL)	97.0 [90.0,106.0]	Iron, median [Q1, Q3] (mcg/dL)	104.0 [81.0,130.0]

Performance metrics

We evaluated predictive performance of the models using area under the receiver operator characteristic curve (AUC) and area under the precision-recall curve (AUPRC). Furthermore, since machine learning models can be discriminative but with low calibration quality, we also plotted the calibration curve for all the analyses. The calibration curve shows the actual class probabilities (x-axis) against the models' probability predictions (y-axis) and is evaluated using Brier scores (a lower Brier score indicates higher calibration quality). To assess the predictive performance, we calculated additional metrics including Positive Predictive Value (PPV), Negative Predictive Value (NPV), F-1 score, and Matthews correlation coefficient (MCC), shown in Appendix 1 to Appendix 4. We note that in addition to MCC that considers the class imbalance [27], other methods could also be applicable such as subgroup analysis [28,29].

Predictive model interpretability

To increase the transparency of our predictive models, we also investigated model interpretability based on Shapley Additive exPlanations (SHAP). SHAP method deconstructs each prediction into a sum

Table 3
Baseline characteristics of the steatosis patient cohort divided by outcome and derivation and validation cohort.

	Model derivation cohort			Prospective validation cohort		
	Negative	Positive	p-value	Negative	Positive	p-value
Patients	2168	1930	–	948	788	–
Age, median [Q1,Q3]	56.6 [50.7,65.2]	59.9 [53.1,67.1]	<0.001	56.1 [51.7,62.7]	58.8 [53.3,65.7]	<0.001
Gender Male, (%)	946 (43.6)	1201 (62.2)	<0.001	388 (40.9)	501 (63.6)	<0.001
Weight, median [Q1,Q3]	72.0 [63.0,81.0]	86.0 [77.0,96.0]	<0.001	70.0 [62.0,80.0]	87.0 [77.0,96.0]	<0.001
Height, median [Q1,Q3]	169.0 [163.0,176.0]	172.0 [165.0,178.0]	<0.001	169.5 [164.0,176.0]	173.0 [166.0,179.0]	<0.001
BMI, median [Q1,Q3]	25.0 [22.8,27.3]	29.0 [26.4,32.4]	<0.001	24.4 [22.3,27.0]	28.7 [26.3,31.7]	<0.001
Systolic Blood Pressure, median [Q1,Q3]	125.0 [120.0,140.0]	130.0 [120.0,150.0]	<0.001	130.0 [120.0,140.0]	140.0 [130.0,150.0]	<0.001
Diastolic Blood Pressure, median [Q1,Q3]	80.0 [70.0,80.0]	80.0 [80.0,90.0]	<0.001	80.0 [72.8,85.0]	80.0 [80.0,90.0]	<0.001
ALT, median [Q1,Q3]	17.0 [14.0,23.0]	26.0 [19.0,37.0]	<0.001	18.0 [14.0,23.0]	27.0 [20.0,39.0]	<0.001
AST, median [Q1,Q3]	19.0 [17.0,24.0]	22.0 [19.0,29.0]	<0.001	19.0 [17.0,23.0]	22.0 [18.0,28.0]	<0.001
AP, median [Q1,Q3]	62.0 [51.0,74.0]	66.0 [55.0,79.0]	<0.001	62.0 [51.0,75.0]	64.0 [54.0,78.0]	0.001
Bilirubin, median [Q1,Q3]	0.7 [0.5,0.9]	0.7 [0.5,0.9]	0.065	0.7 [0.6,1.0]	0.7 [0.6,1.0]	0.515
Cholinesterase, median [Q1,Q3]	9370.5 [7885,10,833]	10,028.0 [8528,11,622]	<0.001	9860.0 [8601,11,074]	10,888.0 [9689,12,176]	<0.001
GGT, median [Q1,Q3]	21.0 [14.8,33.0]	33.0 [22.0,55.0]	<0.001	21.0 [16.0,30.0]	35.5 [24.0,55.0]	<0.001
Blood Glucose, median [Q1,Q3]	94.0 [88.0,102.0]	102.0 [94.0,115.0]	<0.001	92.0 [87.0,98.0]	101.0 [94.0,112.0]	<0.001
Cholesterol, median [Q1,Q3]	222.0 [195.0,249.0]	215.0 [186.0,245.0]	<0.001	223.0 [197.0,253.0]	216.0 [188.5,246.5]	<0.001
C-reactive protein, median [Q1,Q3]	0.1 [0.1,0.3]	0.2 [0.1,0.5]	<0.001	0.1 [0.1,0.2]	0.2 [0.1,0.4]	<0.001
Uric acid, median [Q1,Q3]	5.1 [4.2,6.1]	6.2 [5.2,7.2]	<0.001	5.2 [4.4,6.1]	6.3 [5.4,7.4]	<0.001

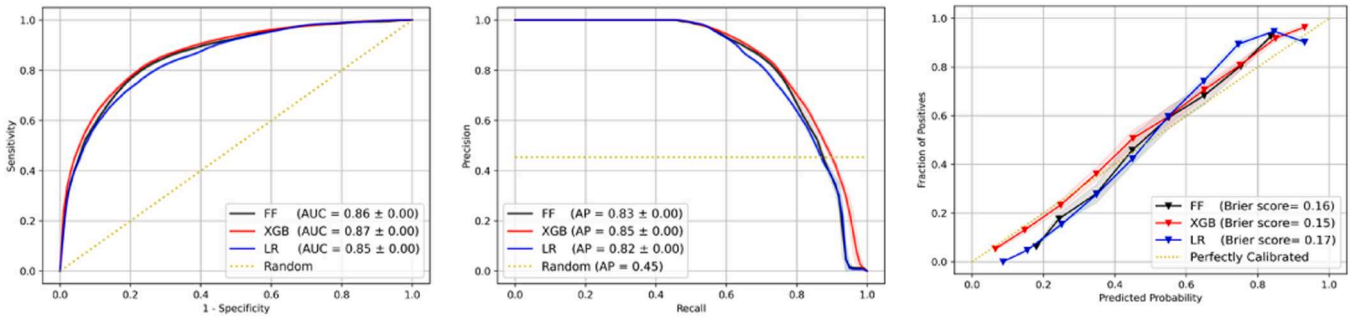

Fig. 1. From left to right, AUC, AUPRC and calibration curve performance of the steatosis model derived from the patient cohort admitted from January 2007 up to January 2016 and prospectively evaluated on the patient cohort admitted from January 2016 up to March 2020. The model was derived and evaluated on all the available variables, including both objective and self-reported.

of individual contributions from each variable known as SHAP values. SHAP values reveal how the input variables influence the model's predictions, both at the instance level and throughout the entire population. We calculated SHAP values for each run of the 5-fold cross validation to precisely capture the influence of each variable on the outcome prediction, plotting them into a Bee swarm plot. Furthermore, we also used SHAP dependence to show how the actual value of a variable influences the predicted outcome.

Results

The overall patient cohort included 5834 patients with a steatosis prevalence of 47%, median age of 58 years (IQR [52–65.5]) and 48% female. The overall transient elastography cohort included 1240 patients with a fibrosis prevalence of 7%, median age of 57 years (IQR [52–63]) and 48% female as shown in Table 2 and Table 3.

Prediction of steatosis including objective and self-reported variables

Initially we evaluated the predictive performance of our machine

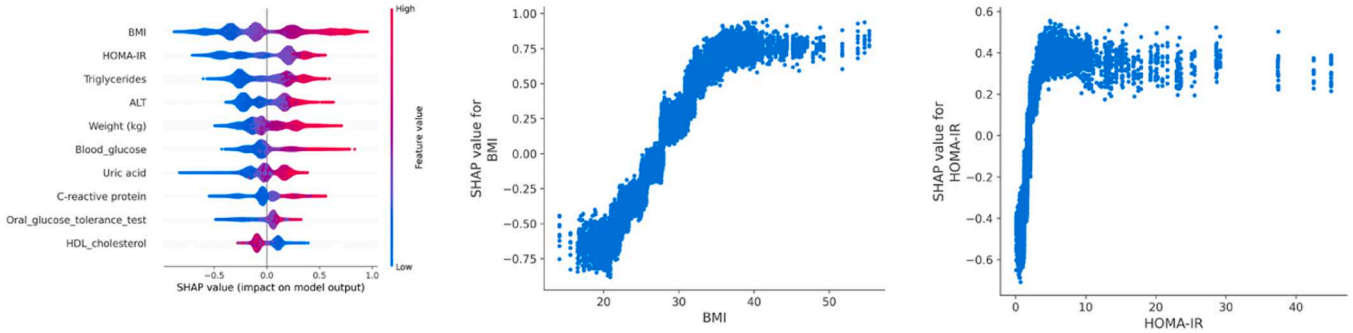

Fig. 2. From left to right: i) SHAP variable ranking graph, with blue colour representing a low value for a variable, while red the opposite indicated by the colour ramp, ii) SHAP dependency plot showing relationship between BMI values and the estimated risk of steatosis, and iii) SHAP dependency plot showing relationship between values of HOMA-IR and the estimated risk of steatosis. Each dot represents measurements from a single patient. Red colour represents a high value for a particular variable, while blue.

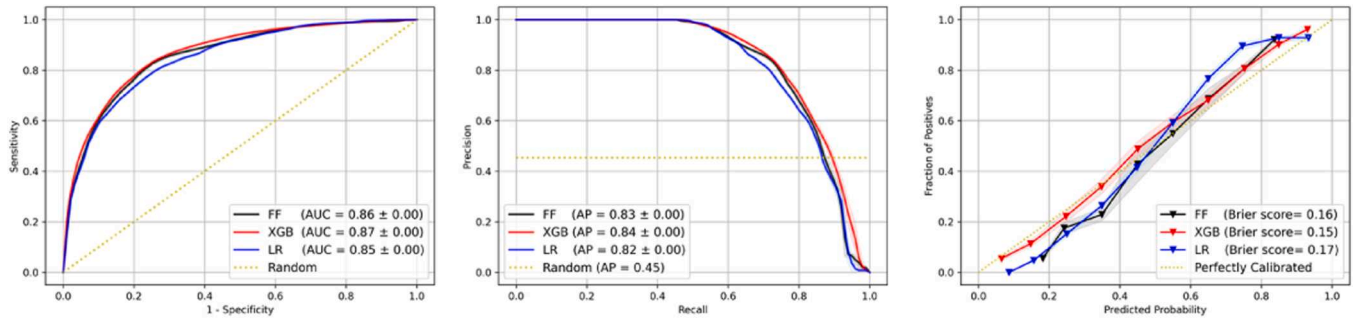

Fig. 3. From left to right, AUC, AUPRC and calibration curve performance of the steatosis model derived from the patient cohort admitted from January 2007 up to January 2016 and prospectively evaluated on the patient cohort admitted from January 2016 up to March 2020. This model was derived and evaluated on objectively collected variables only, while excluding self-reported variables.

learning model using all the available variables, namely both objective and patient self-reported variables. In this respect our model derived using XGBoost algorithm, prospectively predicted steatosis with AUC of 0.87 (95% CI [0.87–0.87]) based on the evaluation of the prospective patient cohort as shown in Fig. 1. Furthermore, area under precision-recall curve, positive predictive value and negative predictive value are also high with 0.85 (95% CI [0.85–0.85]), 0.80 (95% CI [0.79–0.80]) and 0.78 (95% CI [0.78–0.78]) as shown in Appendix 1, while our model showed high quality calibration with a low Brier score of 0.16 (a lower Brier score indicates higher calibration quality). These results are shown in Fig. 1. It should be noted that FF and LR algorithms performed just as well, with no statistically significant differences between them ($p = 0.43$ and $p = 0.12$ respectively) based on the Delong test.

Variable saliency analysis in the prediction of steatosis

In addition to the predictive performance, we also performed saliency analysis using SHAP. The results, outlined in Fig. 2 show the top 10 variables with the highest predictive influence, namely BMI, HOMA-IR, triglycerides, ALT, weight, blood glucose, uric acid, C-reactive protein, glucose tolerance and HDL cholesterol. Furthermore, we also devised dependency plots outlining the relationship between the actual values of the variables (x-axis) and risk of predicted outcome (y-axis) expressed in terms of SHAP values for the top 2 variables, namely BMI and HOMA-IR. We found that for values of BMI of around 25 the predicted risk of steatosis remains low, while it increases gradually for the values above 25 up to BMI of 40. For the BMI values over 40 the predicted risk of steatosis remains consistently high. Similar pattern is seen also with the HOMA-IR variable, where at values above 3 the risk sharply increases and remains consistently high.

Prediction of steatosis excluding self-reported variables

To investigate the influence of self-reported variables on predictive performance of steatosis, we derived and prospectively validated models in the same manner as above, however excluding the self-reported variables (shown in Table 1). We found that there were no differences between predictive performance when excluding and when including self-reported variables, with AUC of 0.87 (95% CI [0.87–0.87]) as shown in Fig. 3. These results, and those shown in Appendix 2, indicate that in predicting steatosis, objectively collected variables had a significant role, while self-reported variables had an ancillary role .

Sensitivity analysis on gender-specific predictive performance of steatosis

We also investigated performance of our models in gender specific subgroups in predicting steatosis. We found that the models performed slightly better in female patients with AUC of 0.85 (95% CI [0.83–0.87]), in comparison to male patients with AUC of 0.82 (95% CI [0.80–0.84]).

Prediction of liver fibrosis using objective and self-reported variables

We then focused on investigating performance of the models in prospectively predicting patients with liver fibrosis (Fig. 4). The overall transient elastography cohort included 1240 patients with a fibrosis prevalence of 7%, median age of 57 years (IQR [52–63]) and 48% female. Cohort baseline characteristics are shown in Table 4 .

We evaluated the predictive performance of our machine learning model using all the available variables, namely both objective and patient self-reported variables. In this respect our model derived using XGBoost algorithm, prospectively predicted fibrosis with AUC of 0.75 (95% CI [0.74–0.75]). In contrast the AUC of FIB-4 was 0.61, however

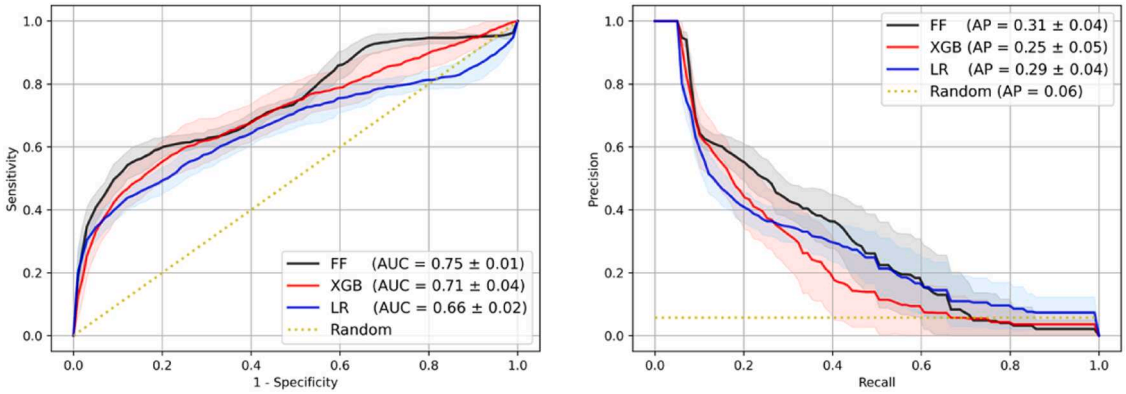

Fig. 4. From left to right, AUC, AUPRC performance of the liver fibrosis model derived from the patient cohort admitted from January 2007 up to January 2016 and prospectively evaluated on the patient cohort admitted from January 2016 up to March 2020. The model was derived and evaluated on all the available variables, including both objective and self-reported.

Table 4
Baseline characteristics of transient elastography cohort.

	Model derivation cohort			Prospective validation cohort		
	Negative	Positive	p-value	Negative	Positive	p-value
Patients	808	68		343	21	
Age, median [Q1,Q3]	56.7 [52.2,63.0]	56.1 [51.5,63.8]	0.935	57.4 [52.5,63.7]	61.8 [56.3,65.5]	0.035
Gender Male (%)	411 (50.9)	53 (77.9)	<0.001	160 (46.6)	15 (71.4)	0.048
Weight, median [Q1,Q3]	77.0 [67.0,87.0]	87.0 [74.0,100.0]	<0.001	76.0 [66.0,86.0]	95.0 [76.0,108.0]	<0.001
Height, median [Q1,Q3]	171.0 [165.0,178.0]	174.0 [169.0,179.2]	0.016	170.0 [164.0,178.0]	173.0 [168.0,180.0]	0.203
BMI, median [Q1,Q3]	26.0 [23.3,29.0]	28.4 [24.7,32.5]	<0.001	25.6 [23.9,28.7]	30.4 [28.1,34.4]	<0.001
Systolic Blood Pressure, median [Q1,Q3]	135.0 [125.0,150.0]	140.0 [123.0,150.0]	0.768	140.0 [125.0,150.0]	145.0 [140.0,160.0]	0.004
Diastolic Blood Pressure, median [Q1,Q3]	80.0 [76.0,90.0]	80.0 [77.5,89.2]	0.725	80.0 [79.0,90.0]	90.0 [80.0,92.0]	0.001
ALT, median [Q1,Q3]	21.0 [16.0,29.0]	29.0 [19.8,43.2]	<0.001	21.0 [16.0,27.0]	31.0 [25.0,37.0]	<0.001
AST, median [Q1,Q3]	20.0 [17.0,24.0]	26.0 [18.8,36.0]	<0.001	19.0 [17.0,24.0]	25.0 [20.0,33.0]	0.004
AP, median [Q1,Q3]	63.0 [53.0,75.0]	60.0 [49.8,81.2]	0.85	63.0 [52.0,76.0]	63.0 [50.0,87.0]	0.542
Bilirubin, median [Q1,Q3]	0.7 [0.6,1.0]	0.8 [0.6,1.0]	0.528	0.7 [0.5,0.9]	0.7 [0.6,0.8]	0.747
Cholinesterase, median [Q1,Q3]	10,143.5 [8913,11,449]	10,013.5 [8463,11,369]	0.259	10,328.0 [9043,11,655]	10,248.0 [8982,11,839]	0.868
GGT, median [Q1,Q3]	25.0 [17.8,40.0]	42.5 [24.0,74.8]	<0.001	22.0 [17.0,34.0]	44.0 [33.0,68.0]	<0.001
Blood Glucose, median [Q1,Q3]	96.0 [89.0,104.0]	98.5 [90.0,122.2]	0.034	93.0 [87.0,99.0]	106.0 [89.0,127.0]	0.008
Cholesterol, median [Q1,Q3]	220.0 [192.0,248.0]	192.0 [161.5,219.5]	<0.001	223.0 [199.5,254.0]	207.0 [191.0,249.0]	0.294
C-reactive protein, median [Q1,Q3]	0.1 [0.1,0.3]	0.2 [0.1,0.3]	0.604	0.1 [0.1,0.3]	0.4 [0.3,0.8]	<0.001
Uric acid, median [Q1,Q3]	5.6 [4.8,6.7]	6.3 [5.6,7.3]	<0.001	5.6 [4.7,6.6]	6.7 [5.8,7.8]	0.001

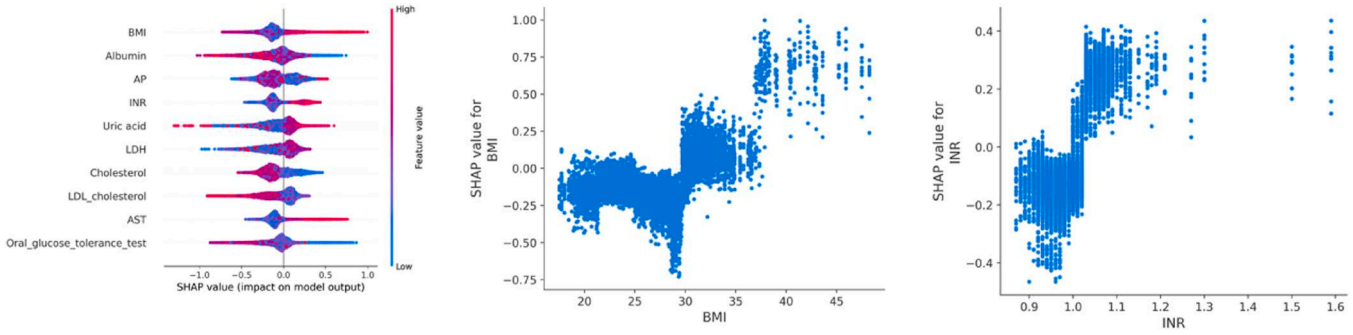

Fig. 5. From left to right: i) SHAP variable ranking graph, with blue colour representing a low value for a variable, while red the opposite indicated by the colour ramp, ii) SHAP dependency plot showing relationship between BMI values and the estimated risk of liver fibrosis, and iii) SHAP dependency plot showing relationship between values of INR and the estimated risk of liver fibrosis. Each dot represents measurements from a single patient.

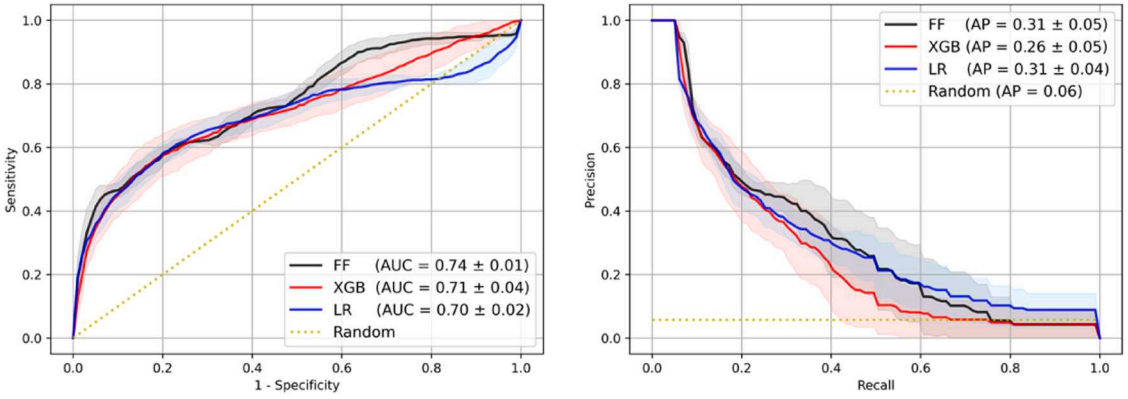

Fig. 6. From left to right, AUC and AUPRC performance of the liver fibrosis model derived from the patient cohort admitted from January 2007 up to January 2016 and prospectively evaluated on the patient cohort admitted from January 2016 up to March 2020. This model was derived and evaluated on objectively collected variables only, while excluding self-reported variables.

these findings could be explained by several factors, including dilution of performance.

Variable saliency analysis in the prediction of liver fibrosis

In addition to the predictive performance of our models, we also analyse salient variables that contribute to the prediction of liver fibrosis. Our analysis is based on SHAP and shown in Fig. 5. BMI is the highest-ranking predictor variable, followed by levels of albumin and AP.

Prediction of liver fibrosis excluding self-reported variables

To investigate the influence of self-reported variables on predictive performance of liver fibrosis, we derived and prospectively validated models in the same manner as above, however excluding the self-

reported variables, shown in Table 1. We find that there are no significant differences when excluding and including self-reported variables, with AUC of 0.74 (95% CI [0.74–0.75]) versus AUC of 0.75 (95% CI [0.74–0.75]) when including all the variables. These results, shown in Fig. 6 and Appendix 4, indicate that in predicting liver fibrosis, self-reported variables do not contribute to improving predictive performance.

Sensitivity analysis on gender-specific predictive performance of liver fibrosis

We investigated performance of our models in gender specific subgroups in predicting liver fibrosis. Our results show that there is a significant difference between predicting liver fibrosis in male patients in comparison to female patients. Namely, predictive performance in prospective validation of male patients is AUC of 0.74 (95% CI [0.73–0.74]) in comparison to female patients with AUC of 0.66 (95% CI [0.65–0.66]), $p = 0.03$.

Discussion

We evaluated the capabilities of machine learning to predict hepatic steatosis and hepatic fibrosis in a cohort of asymptomatic subjects. For this purpose, we examined anthropometric, and laboratory cross-sectional data using different algorithms and found that although machine learning was able to detect hepatic steatosis with relatively high performance in our cohort, it was unsuitable for predicting liver fibrosis.

We therefore consider our finding to be primarily negative. In our opinion, machine learning is unsuitable in the setting we evaluated (as a screening tool for clinically relevant NAFLD in asymptomatic subjects). We also consider our result to be explicitly in line with the relatively abundant prior literature that has already investigated the diagnostic power of machine learning for NAFLD. For example, Ma et al. [10] were able to show that machine learning can predict NAFLD with high accuracy in nearly 10.000 Chinese patients. For this study, NAFLD was diagnosed by ultrasound and several machine learning algorithms were evaluated. A similar conclusion was also reached by Liu et al. in more than 15,000 Chinese patients [19]. Atsawarungruangkit et al. [11] investigated the predictive capabilities of machine learning in the NHANES database. The authors concluded that a simple algorithm consisting of two covariates (fasting C-peptide and waist circumference) had high predictive power. However, in this study, NAFLD was defined based on biomarkers, so a biochemical surrogate parameter and not a clinical finding was defined as the endpoint, which significantly limits its clinical applicability. On the other hand, although formally this criticism that only a surrogate parameter, and not the diagnostic gold standard, was used as endpoint also applies to our study: we diagnosed (clinically relevant) liver fibrosis by transient elastography, and not by liver biopsy. However, for our setting, where population-based screening was evaluated, we think that liver biopsy would not be ethically justifiable, and we therefore think that - in contrast to the previous studies - we evaluated the best endpoint in this situation. Even beyond that, our negative results could provide valuable input for future studies.

First, it could be that machine learning has much higher predictive power when data of similar granularity but from multiple time points is included in the algorithms. For example, the results of several routine blood draws documented over several years, as well as the history of anthropometric data such as weight, could significantly increase the diagnostic capabilities of machine learning, as machine learning particularly benefits from complex data sets.

Second, we were able to define variables associated with hepatic steatosis and fibrosis using SHAP analysis. While many expected variables such as BMI, but also age and gender were detected, it was of interest to us that uric acid emerged as a relevant covariate. While literature on the association of uric acid and NAFLD [20] is already available, our data may underline the potential relevance of this association.

Third, we found a significant gender-specific difference in the diagnostic power of machine learning in detecting NAFLD. This, of course, underscores the well-established and documented gender difference in the distribution of risk factors of NAFLD and NAFLD itself [21] in the literature. On the other hand, this result could also be interpreted to suggest that in the future machine learning algorithms should be developed in an explicitly gender-specific manner.

In conclusion, in our analysis, machine learning was able to predict hepatic steatosis with high accuracy, as in the previous literature. However, the diagnosis of hepatic steatosis is easily made by abdominal ultrasonography, and the clinical utility of an algorithm predicting hepatic steatosis is low in our opinion. This is underscored by the questionable prognostic relevance of simple hepatic steatosis. Conversely, hepatic fibrosis is clearly associated with an unfavourable prognosis, and the diagnosis of higher-grade hepatic fibrosis is also associated with specific medical management. However, in predicting a well-established surrogate parameter of liver fibrosis (measurement by transient elastography), we find a low predictive power of machine learning. Therefore, we think that machine learning is not a substitute for clinical work-up of patients with NAFLD and questionable liver fibrosis.

Limitations

Our work includes some study-specific limitations. Some of the variables were self-reported, as such may be subject to recall bias. Given the specific patient population, generalisability of these results to other patient populations remains to be investigated. NAFLD was diagnosed by abdominal ultrasound in our study. This is a limitation since hepatic steatosis can only be graded qualitatively, with particular weakness for lower degrees of liver fat. In addition, the absence or presence of NAFLD was determined on clinical grounds. This was because more elaborate and sensitive non-invasive tests such as liver stiffness measurement, including continuation attenuation parameter determination and magnetic resonance imaging (MRI), were either not available at the time our study began or were not part of the study protocol. However, our study was still in line with international guidelines, which state that normal transaminase and ultrasound findings in the absence of MetS components are sufficient to rule out clinically relevant NAFLD. Possible explanations for the high prevalence of steatosis might be underreporting of alcohol abuse [26], predominantly sedentary lifestyle, and dietary factors, however an exact cause would be difficult to determine given the observational nature of our study.

Author contributions

BW and VO conceptualised the work. BM and GS analysed the data. BM, GS, BW, VO wrote the draft of the manuscript. All authors contributed to the conception and design of the study, analysis and interpretation of data, and revising the article critically for important intellectual content. All authors approved the final version of the manuscript.

Funding

C.D. is a member of the scientific advisory board of SPAR Österreichische Warenhandels AG and funding by SPAR AG to C.D. is greatly appreciated. The funders had no role in this work.

Declaration of Competing Interest

Authors declare no competing interests.

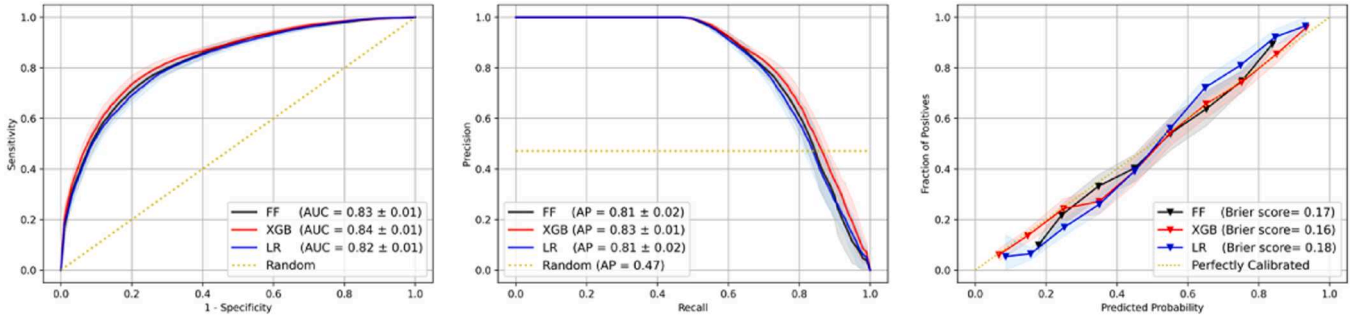

Fig. 7. From left to right, AUC, AUPRC and calibration curve performance of the model derived from the patient cohort admitted from January 2007 up to January 2016 based on internal 5-fold cross validation. The model is derived and evaluated on all the available variables, including both objective and self-reported.

Table 5
Internal evaluation based on objective and self-reported variables in prediction of steatosis.

	AUC	AP	PPV	NPV	F-1	MCC	Brier
LR	0.82 [0.82–0.83]	0.81 [0.80–0.81]	0.74 [0.73–0.74]	0.76 [0.76–0.77]	0.74 [0.73–0.74]	0.50 [0.49–0.51]	0.18 [0.17–0.18]
XGB	**0.84** [0.84–0.84]	**0.83** [0.82–0.83]	**0.75** [0.75–0.76]	**0.78** [0.78–0.79]	**0.76** [0.75–0.76]	**0.54** [0.53–0.55]	**0.16** [0.16–0.16]
FF	0.83 [0.83–0.83]	0.81 [0.81–0.82]	0.74 [0.74–0.75]	0.77 [0.77–0.77]	0.74 [0.74–0.75]	0.51 [0.51–0.52]	0.17 [0.17–0.17]

Table 6
Prospective evaluation based on objective and self-reported variables in prediction of steatosis.

	AUC	AP	PPV	NPV	F-1	MCC	Brier
LR	0.85 [0.85–0.85]	0.82 [0.82–0.82]	0.77 [0.76–0.77]	0.77 [0.77–0.77]	0.73 [0.73–0.74]	0.53 [0.53–0.53]	0.17 [0.17–0.17]
XGB	**0.87** [0.87–0.87]	**0.85** [0.85–0.85]	**0.80** [0.79–0.80]	0.78 [0.78–0.78]	**0.75** [0.75–0.75]	**0.57** [0.57–0.57]	**0.15** [0.15–0.15]
FF	0.86 [0.86–0.86]	0.83 [0.83–0.83]	0.77 [0.77–0.78]	**0.79** [0.78–0.79]	**0.75** [0.75–0.76]	0.56 [0.55–0.56]	0.16 [0.16–0.16]

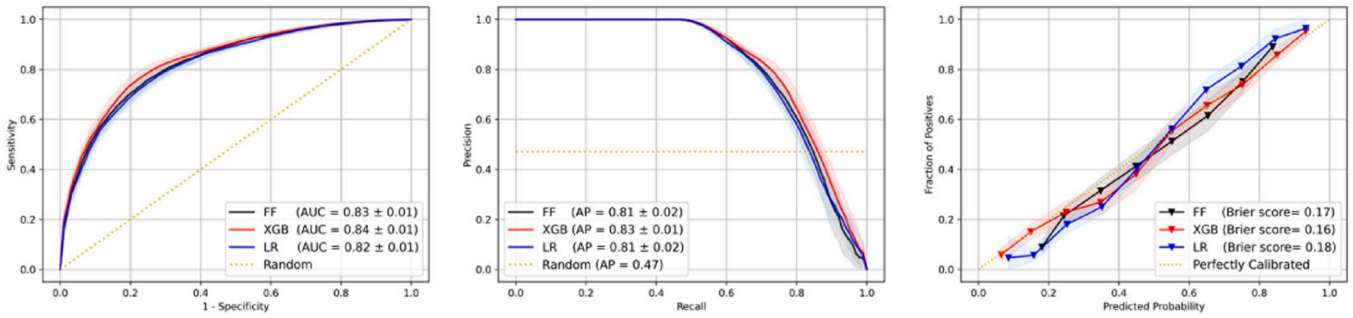

Fig. 8. From left to right, AUC, AUPRC and calibration curve performance of the model derived from the patient cohort admitted from January 2007 up to January 2016 based on internal 5-fold cross validation in prediction of steatosis. This model is derived and evaluated on objectively collected variables only, while excluding self-reported variables.

Appendix 1. Evaluation metrics for the internal and prospective validation cohort using both objective and self-reported variables in prediction of steatosis

Fig. 7 and Tables 5, 6

Appendix 2. Evaluation metrics for the internal and prospective validation cohorts when excluding self-reported variables in prediction of steatosis

Fig. 8

Appendix 3. Evaluation metrics for the internal and prospective validation cohorts using both objective and self-reported variables in prediction of liver fibrosis

Fig. 9 and Tables 7, 8

Appendix 4. Evaluation metrics for the internal and prospective validation cohorts when excluding self-reported variables in prediction of liver fibrosis

Fig. 10, Tables 9, 10

Appendix 5. Details of machine learning algorithms used in this study

XGBoost is an ensemble of decision trees that provides robust predictive performance using an iterative learning process that sequentially builds many models that correct the deficiencies of the preceding model. Even though deep neural networks provide better predictive performance in unstructured datasets, XGBoost has shown great predictive performance for structured, tabular data [18].

To compare the performance of XGBoost, we also implemented Feed-Forward as a deep neural network and Logistic Regression as a statistical

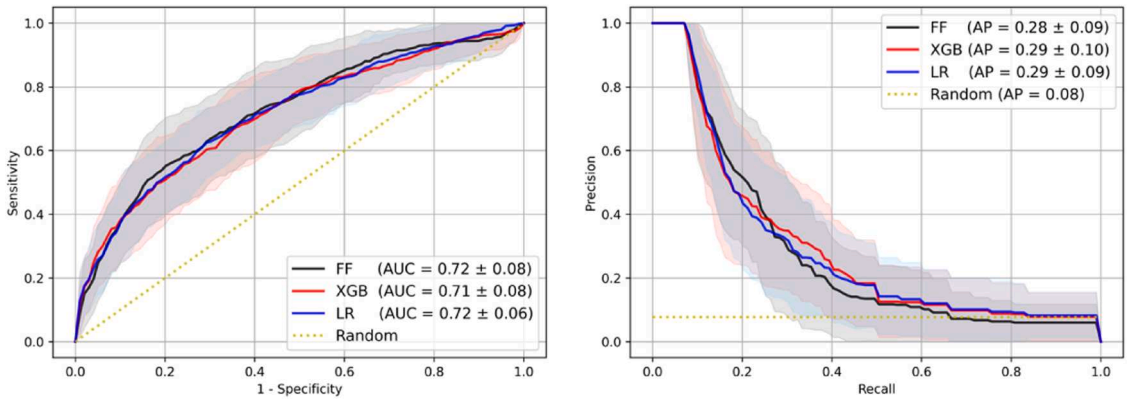

Fig. 9. From left to right, AUC and AUPRC curve performance of the model derived from the patient cohort admitted from January 2007 up to January 2016 based on internal 5-fold cross validation to predict liver fibrosis. The model is derived and evaluated on all the available variables, including both objective and self-reported.

Table 7
Internal evaluation based on objective and self-reported variables in prediction of liver fibrosis. The model is derived and evaluated on all the available variables, including both objective and self-reported.

	AUC	AP	PPV	NPV	F-1	MCC	Brier
LR	**0.72** [0.70–0.73]	**0.29** [0.27–0.32]	0.29 [0.26–0.33]	**0.94** [0.94–0.94]	0.27 [0.25–0.30]	0.22 [0.19–0.25]	**0.09** [0.09–0.09]
XGB	0.71 [0.69–0.74]	**0.29** [0.26–0.32]	0.29 [0.26–0.32]	**0.94** [0.94–0.94]	**0.30** [0.27–0.32]	**0.23** [0.20–0.26]	0.10 [0.09–0.10]
FF	**0.72** [0.70–0.75]	0.28 [0.25–0.30]	**0.33** [0.04–0.62]	0.92 [0.92–0.93]	0.21 [0.16–0.26]	0.02 [0.01–0.04]	0.11 [0.10–0.13]

Table 8
Prospective evaluation based on objective and self-reported variables in prediction of liver fibrosis. The model is derived and evaluated on all the available variables, including both objective and self-reported.

	AUC	AP	PPV	NPV	F-1	MCC	Brier
LR	0.66 [0.66–0.67]	0.29 [0.28–0.30]	**0.36** [0.35–0.38]	**0.96** [0.96–0.96]	**0.33** [0.32–0.34]	**0.30** [0.29–0.31]	**0.07** [0.07–0.07]
XGB	0.71 [0.70–0.72]	0.25 [0.23–0.26]	0.31 [0.30–0.33]	**0.96** [0.96–0.96]	0.30 [0.28–0.32]	0.26 [0.24–0.28]	**0.07** [0.07–0.08]
FF	**0.75** [0.74–0.75]	**0.30** [0.29–0.31]	0.26 [0.07–0.46]	0.94 [0.94–0.95]	0.26 [0.16–0.37]	0.03 [0.01–0.05]	0.10 [0.09–0.12]

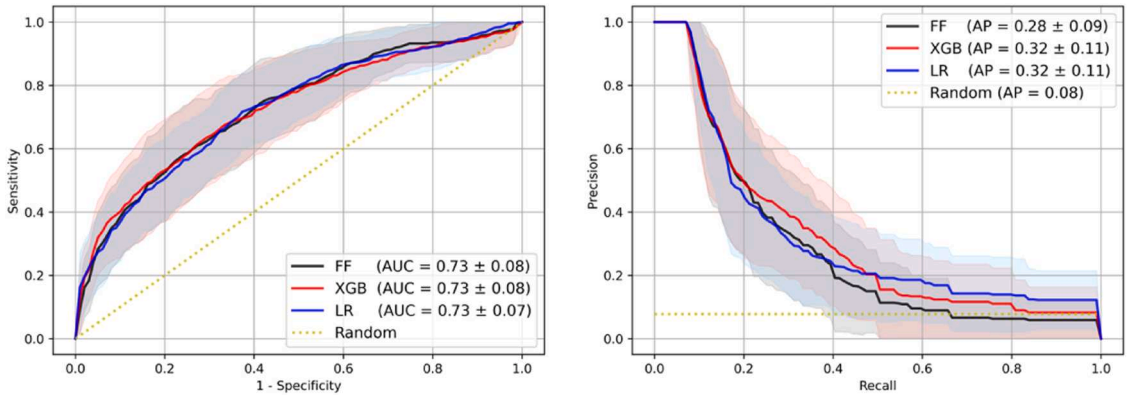

Fig. 10. From left to right, AUC and AUPRC curve performance of the model derived from the patient cohort admitted from January 2007 up to January 2016 based on internal 5-fold cross validation to predict liver fibrosis. This model is derived and evaluated on objectively collected variables only, while excluding self-reported variables.

Table 9
Internal evaluation based on objective and self-reported variables in prediction of liver fibrosis. The model is derived and evaluated on objectively collected variables only, while excluding self-reported variables.

	AUC	AP	PPV	NPV	F-1	MCC	Brier
LR	**0.73** [0.71–0.75]	**0.32** [0.29–0.35]	**0.33** [0.28–0.37]	**0.94** [0.94–0.94]	0.28 [0.25–0.32]	0.24 [0.20–0.27]	**0.09** [0.09–0.09]
XGB	**0.73** [0.70–0.75]	0.31 [0.28–0.34]	0.31 [0.28–0.34]	**0.94** [0.94–0.95]	**0.32** [0.29–0.35]	**0.26** [0.23–0.29]	0.10 [0.09–0.10]
FF	**0.73** [0.70–0.75]	0.28 [0.26–0.31]	0.17 [0.06–0.28]	0.93 [0.92–0.93]	0.19 [0.14–0.24]	0.02 [0.01–0.04]	0.13 [0.11–0.15]

Table 10
Prospective evaluation based on objective and self-reported variables in prediction of liver fibrosis. The model is derived and evaluated on objectively collected variables only, while excluding self-reported variables.

	AUC	AP	PPV	NPV	F-1	MCC	Brier
LR	0.70 [0.69–0.70]	**0.31** [0.30–0.32]	**0.39** [0.37–0.41]	**0.96** [0.96–0.96]	**0.34** [0.33–0.35]	**0.31** [0.30–0.32]	**0.07** [0.07–0.07]
XGB	0.71 [0.70–0.72]	0.26 [0.25–0.27]	0.33 [0.32–0.35]	**0.96** [0.96–0.96]	0.31 [0.30–0.33]	0.27 [0.26–0.29]	**0.07** [0.07–0.07]
FF	**0.74** [0.74–0.75]	0.30 [0.29–0.32]	0.21 [0.05–0.37]	0.94 [0.94–0.95]	0.19 [0.10–0.28]	0.02 [0.01–0.04]	0.12 [0.10–0.14]

baseline comparator. The Feed-ForwardFeed-Forward model used a two-layer neural network with 64 and 16 neurons in the first and second layer respectively, using the sigmoid activation function. Model parameters were randomly initialised based on Xavier normal method, trained for 100 epochs with batch size 32, and optimised using the Adam optimizer algorithm. Logistic Regression is a statistical method that investigates relationships between the outcome variable and the input variables and is typically considered as a baseline algorithm in clinical classification tasks.

References

[1] Younossi Z, Anstee QM, Marietti M, et al. Global burden of NAFLD and NASH: trends, predictions, risk factors and prevention. Nat Rev Gastroenterol Hepatol 2018;15:11–20.

[2] Stefan N, Cusi K. A global view of the interplay between non-alcoholic fatty liver disease and diabetes. Lancet Diabetes Endocrinol 2022;10:284–96.

[3] European Association for the Study of the Liver (EASL), European Association for the Study of Diabetes (EASD), European Association for the Study of Obesity (EASO). EASL-EASD-EASO Clinical Practice Guidelines for the management of non-alcoholic fatty liver disease. J Hepatol 2016;64:1388–402.

[4] Kanwal F, Shubrook JH, Adams LA, et al. Clinical care pathway for the risk stratification and management of patients with nonalcoholic fatty liver disease. Gastroenterology 2021;161:1657–69.

[5] Loomba R, Lim JK, Patton H, et al. AGA clinical practice update on screening and surveillance for hepatocellular carcinoma in patients with nonalcoholic fatty liver disease: expert review. Gastroenterology 2020;158:1822–30.

[6] Bugianesi E, Petta S. NAFLD/NASH. J Hepatol 2022;77:549–50.

[7] Semmler G, Wernly S, Wernly B, et al. Machine learning models cannot replace screening colonoscopy for the prediction of advanced colorectal adenoma. J Pers Med 2021;11. https://doi.org/10.3390/jpm11100981.

[8] Aggarwal P, Alkhouri N. Artificial intelligence in nonalcoholic fatty liver disease: a new frontier in diagnosis and treatment. Clin Liver Dis 2021;17:392–7.

[9] Yip TC-F, Ma AJ, Wong VW-S, et al. Laboratory parameter-based machine learning model for excluding non-alcoholic fatty liver disease (NAFLD) in the general population. Aliment Pharmacol Ther 2017;46:447–56.

[10] Ma H, Xu C-F, Shen Z, et al. Application of machine learning techniques for clinical predictive modeling: a cross-sectional study on nonalcoholic fatty liver disease in China. Biomed Res Int 2018;2018:4304376.

[11] Atsawarungruangkit A, Laoveeravat P, Promrat K. Machine learning models for predicting non-alcoholic fatty liver disease in the general United States population: NHANES database. World J Hepatol 2021;13:1417–27.

[12] Taylor-Weiner A, Pokkalla H, Han L, et al. A Machine learning approach enables quantitative measurement of liver histology and disease monitoring in NASH. Hepatology 2021;74:133–47.

[13] Wernly S, Semmler G, Völkerer A, et al. Cardiovascular risk assessment by SCORE2 predicts risk for colorectal neoplasia and tumor-related mortality. J Pers Med 2022;12. https://doi.org/10.3390/jpm12050848.

[14] Grundy SM, Brewer Jr HB, Cleeman JI, et al. Definition of metabolic syndrome: report of the National Heart, Lung, and Blood Institute/American Heart Association conference on scientific issues related to definition. Circulation 2004;109:433–8.

[15] Alberti KGMM, Zimmet P, Shaw J, et al. The metabolic syndrome–a new worldwide definition. Lancet 2005;366:1059–62.

[16] Chen T, Guestrin C. XGBoost: a Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery; 2016. p. 785–94.

[17] Neuhaus J, McCulloch C. Generalized linear models. WIREs Computational Statistics 2011;3:407–13.

[18] Shwartz-Ziv R, Armon A. Tabular data: deep learning is not all you need. Inf Fusion 2022;81:84–90.

[19] Liu Y-X, Liu X, Cen C, et al. Comparison and development of advanced machine learning tools to predict nonalcoholic fatty liver disease: an extended study. Hepatobiliary Pancreat Dis Int 2021;20:409–15.

[20] Jaruvongvanich V, Ahuja W, Wirunsawanya K, et al. Hyperuricemia is associated with nonalcoholic fatty liver disease activity score in patients with nonalcoholic fatty liver disease: a systematic review and meta-analysis. Eur J Gastroenterol Hepatol 2017;29:1031–5.

[21] Lonardo A, Nascimbeni F, Ballestri S, et al. Sex differences in nonalcoholic fatty liver disease: state of the art and identification of research gaps. Hepatology 2019;70:1457–69.

[22] Loomba R, Adams LA. The 20% rule of NASH progression: the natural history of advanced fibrosis and cirrhosis caused by NASH. Hepatology 2019;70(6):1885–8.

[23] Berzigotti A, Tsochatzis E, Boursier J, Castera L, Cazzagon N, Friedrich-Rust M, Petta S, Thiele M. EASL Clinical Practice Guidelines on non-invasive tests for evaluation of liver disease severity and prognosis –2021 update. In: Journal of Hepatology. 75. Elsevier BV; 2021. p. 659–89. https://doi.org/10.1016/j.jhep.2021.05.025.

[24] Graupera I, Thiele M, Serra-Burriel M, Caballeria L, Roulot D, Wong GL-H, Fabrellas N, Guha IN, Arslanow A, Expósito C, Hernández R, Aithal GP, Galle PR, Pera G, Wong VW-S, Lammert F, Ginès P, Castera L, Krag A. Low accuracy of FIB-4 and NAFLD fibrosis scores for screening for liver fibrosis in the population. In: Clinical Gastroenterology and Hepatology. 20. Elsevier BV; 2022. p. 2567–76. https://doi.org/10.1016/j.cgh.2021.12.034. e6.

[25] Friedrich-Rust M, Ong MF, Martens S, Sarrazin C, Bojunga J, Zeuzem S, Herrmann E. Performance of transient elastography for the staging of liver fibrosis: a meta-analysis. Gastroenterology 2008;134(4):960–74. https://doi.org/10.1053/j.gastro.2008.01.034.

[26] Staufer K, Huber-Schönauer U, Strebinger G, Pimingstorfer P, Suesse S, Scherzer TM, Paulweber B, Ferenci P, Stimpfl T, Yegles M, Datz C, Trauner M. Ethyl glucuronide in hair detects a high rate of harmful alcohol consumption in presumed non-alcoholic fatty liver disease. J Hepatol 2022;77(4):918–30. https://doi.org/10.1016/j.jhep.2022.04.040.

[27] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics Dec. 2020;21(1):6. https://doi.org/10.1186/s12864-019-6413-7. PMID: 31898477.

[28] Carrington AM, Manuel DG, Fieguth PW, Ramsay T, Osmani V, Wernly B, Bennett C, Hawken S, Magwood O, Sheikh Y, McInnes M, Holzinger A. Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation. IEEE Trans Pattern Anal Mach Intell 2023;45(1):329–41. https://doi.org/10.1109/TPAMI.2022.3145392.

[29] Carrington, A.M., Manuel, D.G., Fieguth, P.W., Ramsay, T., Osmani, V., Wernly, B., Bennett, C., Hawken, S., McInnes, M., Magwood, O., Sheikh, Y., & Holzinger, A. (2021). Deep ROC analysis and AUC as balanced average accuracy to improve model selection, understanding and interpretation. arXiv. https://doi.org/10.48550/ARXIV.2103.11357.