Original Research Article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 10: 1–19

DOI: 10.1177/20552076241272535
journals.sagepub.com/home/dhj

Sage

Automated machine learning models for nonalcoholic fatty liver disease assessed by controlled attenuation parameter from the NHANES 2017–2020

Lihe Liu[1,*], Jiaxi Lin[1,*], Lu Liu[1], Jingwen Gao[1], Guoting Xu[1], Minyue Yin[2], Xiaolin Liu[1], Airong Wu[1] and Jinzhou Zhu[1]

Abstract

Background: Nonalcoholic fatty liver disease (NAFLD) is recognized as one of the most common chronic liver diseases worldwide. This study aims to assess the efficacy of automated machine learning (AutoML) in the identification of NAFLD using a population-based cross-sectional database.

Methods: All data, including laboratory examinations, anthropometric measurements, and demographic variables, were obtained from the National Health and Nutrition Examination Survey (NHANES). NAFLD was defined by controlled attenuation parameter (CAP) in liver transient ultrasound elastography. The least absolute shrinkage and selection operator (LASSO) regression analysis was employed for feature selection. Six algorithms were utilized on the H2O-automated machine learning platform: Gradient Boosting Machine (GBM), Distributed Random Forest (DRF), Extremely Randomized Trees (XRT), Generalized Linear Model (GLM), eXtreme Gradient Boosting (XGBoost), and Deep Learning (DL). These algorithms were selected for their diverse strengths, including their ability to handle complex, non-linear relationships, provide high predictive accuracy, and ensure interpretability. The models were evaluated by area under receiver operating characteristic curves (AUC) and interpreted by the calibration curve, the decision curve analysis, variable importance plot, SHapley Additive exPlanation plot, partial dependence plots, and local interpretable model agnostic explanation plot.

Results: A total of 4177 participants (non-NAFLD 3167 vs NAFLD 1010) were included to develop and validate the AutoML models. The model developed by XGBoost performed better than other models in AutoML, achieving an AUC of 0.859, an accuracy of 0.795, a sensitivity of 0.773, and a specificity of 0.802 on the validation set.

Conclusions: We developed an XGBoost model to better evaluate the presence of NAFLD. Based on the XGBoost model, we created an R Shiny web-based application named Shiny NAFLD (http://39.101.122.171:3838/App2/). This application demonstrates the potential of AutoML in clinical research and practice, offering a promising tool for the real-world identification of NAFLD.

Keywords

Automated machine learning (AutoML), nonalcoholic fatty liver disease (NAFLD), controlled attenuation parameter, artificial intelligence, shiny application

Submission date: 18 December 2023; Acceptance date: 9 July 2024

[1]Department of Gastroenterology, The First Affiliated Hospital of Soochow University, Suzhou, China
[2]Department of Gastroenterology, Beijing Friendship Hospital, Capital Medical University, Beijing, China
[*]These authors contributed equally to this work.

Corresponding authors:
Airong Wu and Jinzhou Zhu, Department of Gastroenterology, The First Affiliated Hospital of Soochow University, #188 Shizi Street, Suzhou 215006, China.
Emails: arwu@suda.edu.cn; jzzhu@zju.edu.cn

Introduction

Nonalcoholic fatty liver disease (NAFLD) has become a major global health concern, affecting approximately 25% of the population worldwide. It is one of the leading causes of chronic liver diseases, which can progress to severe complications such as cirrhosis and hepatocellular carcinoma within 10 to 20 years post-diagnosis.[1] Despite less than 10% of NAFLD patients developing these severe outcomes, the absolute numbers are still considerable, considering the high prevalence of NAFLD.[2] With the global incidence of NAFLD escalating swiftly, there is an imperative need to intensify efforts towards the development of precise, non-invasive diagnostic techniques and the formulation of efficacious prevention strategies for those at elevated risk of NAFLD and its associated progressive liver diseases. Furthermore, the majority of patients with NAFLD are asymptomatic.[3] The timely and accurate identification of individuals predisposed to NAFLD is crucial, facilitating the implementation of targeted interventions that can halt disease advancement, prevent complications, and ultimately alleviate the strain on healthcare infrastructures.[4]

While liver biopsy remains the gold standard for diagnosing NAFLD, its invasiveness and potential for complications, including pain, infection, and bleeding, limit its practicality for widespread screening.[5,6] The diagnosis of NAFLD via ultrasound examination is often hampered by numerous factors, particularly the subjectivity of the examiner.[5] Previous studies have analyzed NAFLD in the United States population using data from the National Health and Nutrition Examination Survey (NHANES).[7] These studies have relied on diagnostic methods with inherent limitations, such as standard ultrasound and noninvasive biomarkers. Furthermore, liver transient ultrasound elastography (LUTE) (FibroScan), which is based on ultrasonic attenuation of the echo wave measurement (controlled attenuation parameter [CAP]), has emerged as a promising non-invasive diagnostic tool.[8–10] Data from transient elastography are now available in the NHANES database (2017–2020), allowing for real-world analysis of NAFLD within the United States population. Traditional scoring systems such as the Fatty Liver Index (FLI), Lipid Accumulation Product (LAP), Hepatic Steatosis Index (HSI), Fatty Liver Disease Index (FLD Index), NAFLD Index, ZJU Index, and Framingham Steatosis Index (FSI) enable non-invasive detection of NAFLD.[11–17] However, their clinical utility is compromised by limitations inherent in these traditional algorithms, which constrain further performance enhancements. Advancements in the management and diagnostic accuracy of NAFLD hinge upon the development of sophisticated analytical tools.

Machine learning (ML), a burgeoning field of medicine combining computer science and statistics into medical problems, is being widely used based on its efficient computing algorithms and its ability to deal with massive clinical data.[18,19] Developing prediction models based on statistical associations among features from a given input data is one of the most common objectives of ML in medicine. Notably, recent literature[20–30] corroborates the formidable capacity of ML in crafting diagnostic models for fatty liver disease. However, the deployment of ML extends beyond mere algorithmic applications; it necessitates a full spectrum of methodical steps, including data pretreatment, feature engineering, ML algorithm selection, and hyperparameter tuning. These steps demand substantial programming experience and ML knowledge, posing substantial hurdles for clinicians. This gap has led to the rise of Automated Machine Learning (AutoML), a significant breakthrough in artificial intelligence that minimizes human oversight and automatically selects optimal algorithms, tunes hyperparameters, and generates robust models.

Zhang et al. developed a rapid and cost-effective tool to enhance the detection of clinically significant prostate cancer using an AutoML platform, leveraging data from routine clinical examinations.[31] Wang et al. developed and validated models to predict 12-month esophageal variceal bleeding using the H2O-automated machine learning platform (algorithms include DL, XGBoost, GLM, GBM, RF, and stacking).[32] Liu et al. utilized AutoML to predict liver metastasis in patients with gastrointestinal stromal tumors, based on an analysis of SEER data.[33] The ability of AutoML to streamline the development of diagnostic tools presents a promising solution for clinicians, particularly in the context of NAFLD. Early detection and accurate diagnosis of NAFLD are critical yet challenging due to the disease's asymptomatic nature. Despite its potential, to the best of our knowledge, there have been few reports on the application of AutoML in diagnosing NAFLD, highlighting this as a nascent yet promising area of research. This study aims to address this gap by investigating the effectiveness of AutoML in identifying the presence of NAFLD using patient demographic data, laboratory results, and physical examination data in the NHANES database, thereby providing a novel approach to improving clinical outcomes for patients with NAFLD.

Utilizing the H2O AutoML platform, this research is designed to leverage data from the NHANES to develop and validate a series of machine learning models for the identification of NAFLD, as determined by FibroScan CAP measurements. By deploying these models and creating the subsequent R Shiny web-based application, Shiny NAFLD, based on the optimal model, we aim to significantly improve clinicians' diagnostic capabilities. This approach promises a more precise, efficient, and less invasive method for identifying NAFLD compared to existing diagnostic techniques.

Materials and methods

Data source

NHANES, a nationally representative survey of the United States, is administered by the National Center for Health Statistics (NCHS). The NHANES program is notable for its comprehensive approach, integrating both interviews and physical examinations to assess the health and nutritional status of the United States population. Conducted annually since 1999, the survey systematically evaluates a representative national sample of approximately 5000 individuals. The NHANES interview covers a broad spectrum of topics including demographics, socioeconomic status, dietary habits, and health. The examination component of the survey comprises thorough medical assessments, physiological measurements, and laboratory tests, all performed by highly trained medical professionals. Comprehensive details on survey variables can be accessed at https://wwwn.cdc.gov/nchs/nhanes/Default.aspx. The survey was approved by the NCHS research ethics review board, and the consent of all participants was recorded.

Participants

The data was obtained from the NHANES January 2017 to March 2020 database, which contained data on LUTE. A total of 15,560 participants were included. We excluded 5862 participants without CAP data, 677 with ineligible FibroScan data (either < 10 complete FibroScan readings, a fasting time <3 hours, or a liver stiffness interquartile (IQRe) range/median stiffness >30%), and 4515 participants who were not included in the fasting subsample, which chose some participants aged 12 and older to fast for 8 to 24 hours in preparation for the examination the following morning. Additional exclusions were applied to high alcohol consumers, defined in the NHANES alcohol use survey as having an average daily intake of ≥20 g/day and ≥30 g/day for women and men from the NHANES alcohol use survey, or if there were any additional potential factors for liver disease, including viral hepatitis (defined as positive for serum hepatitis B surface antigen or hepatitis C antibody or if hepatitis B or C was reported). The final sample size for our analysis was therefore 4177. The flowchart of the inclusion of this study participants is shown in Figure 1.

Fibroscan CAP

By measuring the ultrasonic attenuation of the echo wave, also known as CAP, LUTE can quantify hepatic steatosis.[9,10] FibroScan model 502 V2 Touch fitted with a medium (M) or extra-large (XL) wand (probe) was utilized in the NHANES database to measure CAP value. In addition, liver steatosis was assessed using the mean CAP value in more than ten complete measurements taken throughout the examination. Individuals were identified as having NAFLD if the CAP values were ≥302 dB/m, which is regarded as the best cutoff for the detection of hepatic steatosis.[8]

Variables

The machine learning models in this study were developed using data from laboratory examinations, anthropometric measurements, and demographic variables from NHANES. The

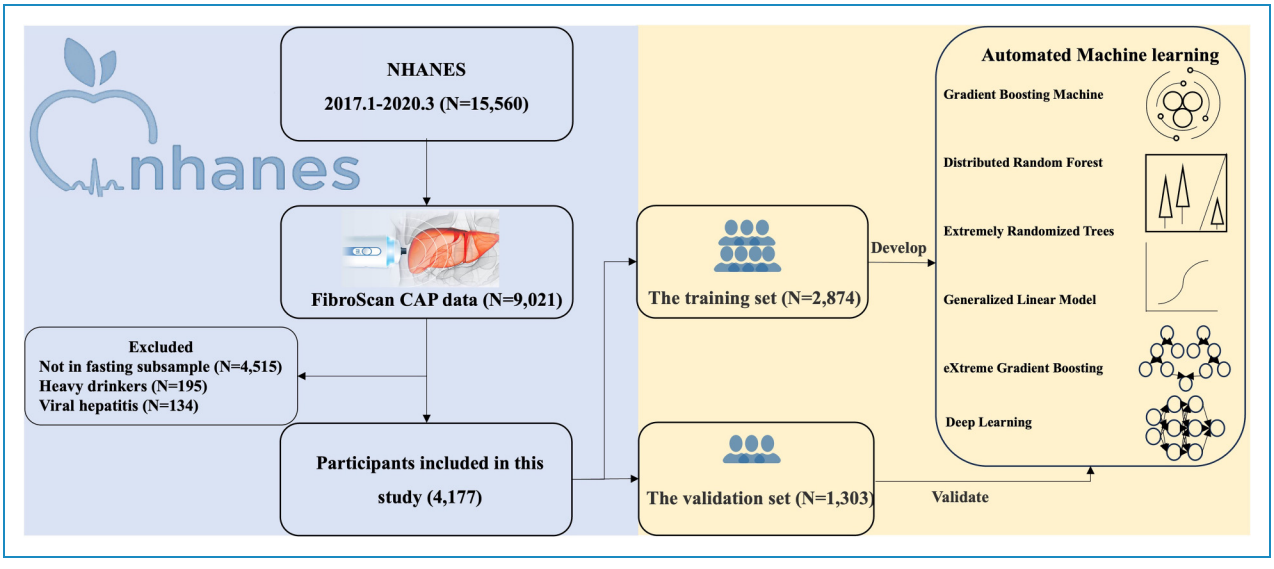

Figure 1. Flowchart of the inclusion of study participants.
NHANES: National Health and Nutrition Examination Survey; CAP: controlled attenuation parameter.

demographic characteristics included gender, age, race, and ratio of family income to poverty (PIR), and anthropometric parameters included weight, body mass index (BMI), arm circumference, waist circumference, hip circumference, systolic pressure (SBP), and diastolic pressure (DBP). The laboratory parameters were composed of alanine aminotransferase (ALT), albumin (ALB), alkaline phosphatase (ALP), aspartate aminotransferase (AST), creatinine (CR), globulin (GLB), gamma-glutamyl transferase (GGT), lactate dehydrogenase (LDH), phosphorus (P), total bilirubin (STB), total calcium (Ca), total protein (TP), uric acid (UA), platelet count (PLT), ferritin, iron, total iron binding capacity (TIBC), transferrin saturation (TSF), glycohemoglobin (HbA1c), fasting plasma glucose (FPG), high-density lipoprotein cholesterol (HDL-C), high-sensitivity c-reactive protein (hs-CRP), insulin (INS), total cholesterol (TC), triglyceride (TG), and low-density lipoprotein-cholesterol (LDL-C) levels. Additionally, the study's main outcome was determined by the presence or absence of hepatic steatosis, indicated by a CAP value of ≥302 dB/m.

Missing data handling

Variables with more than 30% missing values were removed, while those with less than 30% missing values were interpolated using an appropriate technique. The remaining missing data, identified as missing at random, were addressed using the random forest algorithm for multiple imputation and interpolation, as implemented in the R package "mice" (version 3.15.0).[34]

Feature selection

Feature selection was conducted through the least absolute shrinkage and selection operator (LASSO) regression analysis, which strategically penalizes coefficient magnitude to streamline the number of predictive variables. The fine-tuning of the regularization parameter, λ, was meticulously carried out via a 10-fold cross-validation method, ensuring a rigorous optimization process. The coefficients of the variables from the lasso regression models were arranged in ascending order. Variables with nonzero coefficients were selected due to their significant contribution to the model's predictive accuracy. LASSO penalizes less important features, reducing model complexity and enhancing interpretability. By focusing on variables with the strongest predictive power, we ensured that only the most relevant and impactful features were included in the final model.

Automated machine learning

The AutoML analysis was implemented using the H2O package (version 3.40.0.1) installed from the H2O.ai platform (http://www.h2o.ai/). For binary classification problems, this platform's AutoML feature automatically employs six distinct algorithms: Gradient Boosting Machine (GBM), Distributed Random Forest (DRF), Extremely Randomized Trees (XRT), Generalized Linear Model (GLM), eXtreme Gradient Boosting (XGBoost), and Deep Learning (DL). XGBoost is a composite method that integrates several decision-tree classifiers. It reduces the discrepancy between predicted and actual values during training by utilizing objective functions. These functions include differentiable convex loss components and regularization terms to enhance model robustness and prevent overfitting. The descriptions of the remaining algorithms can be found in the supplementary documentation. Participants were randomly allocated into two groups, with a training-to-validation set ratio of 7:3. A 5-fold cross-validation was conducted on the training dataset to assess model performance. AutoML used these evaluations to rank the models based on their area under the ROC curve (AUC), considering various combinations of hyperparameters across six distinct algorithms. Following this ranking, the models were evaluated on the validation set to determine their generalization capabilities. The model with the highest AUC on the validation set was selected as the optimal model. Based on the optimal model, an R Shiny web-based application, Shiny NAFLD, was developed to facilitate the practical identification of NAFLD by clinicians. (http://39.101.122.171:3838/App2/).

Evaluation and interpretation of models

To evaluate the performance of the models on the validation set, a confusion matrix was compiled, which included true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). These metrics were utilized to calculate sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratios (LR+), negative likelihood ratios (LR−), and AUCs. This comprehensive analysis allowed for a detailed assessment of the models' discrimination capabilities. Formulas were as follows: $ACC = (TP + TN)/(TP + FP + FN + TN)$; $PPV = TP/(TP + NP)$; $NPV = TN/(TN + FN)$; $LR+ =$ sensitivity/(1-specificity); $LR- =$ (1-sensitivity)/specificity. The calibration curve was applied to evaluate the model's calibration, while the decision curve analysis (DCA) provided insight into the clinical net benefit. Models interpretability was presented in the form of variable importance, SHapley Additive Explanations (SHAP) partial dependence plot (PDP), and Local Interpretable Model-Agnostic Explanations (LIME). Variable importance is used to assess the statistical significance and impact of each feature within the model. The SHAP analysis is an approach that elucidates the separate contributions of each feature in the development of a prediction model while retaining consistency and local accuracy for a particular prediction.[35] In addition, the marginal influence of features on the predicted outcome can also be displayed in PDP, and the LIME

analysis provides an understanding of the influence of main features on predictions for randomly selected examples from the validation set.

Statistical analysis

In our study, all statistical analyses were performed using R (version 4.2.2). To assess the normality of the data, the Shapiro–Wilk test was employed, which tests the null hypothesis that the data follow a normal distribution. Continuous data were presented as medians with interquartile ranges (IQR), while categorical variables were reported as counts (percentages). The Wilcoxon rank-sum test, a non-parametric test suitable for continuous variables that are not normally distributed, was applied to the continuous data. This test assumes independent groups and similar distribution shapes. For categorical variables, Pearson's Chi-square test was used to determine associations, assuming that the data are in the form of counts or frequencies, with independent samples. These tests were chosen for their robustness and appropriateness given our data characteristics. A two-sided p-value of less than 0.05 was considered statistically significant, indicating a low probability that the observed differences were due to chance.

Results

Demographic and clinical characteristics

The study included 4177 participants, with data randomly divided into a training set and a validation set at a ratio of 7:3 (2874 in the training set and 1303 in the validation set). After excluding participants with hepatitis and excessive alcohol consumption, 1010 participants (24.2%) were diagnosed with NAFLD. Significant differences were observed between the groups in various demographic, biochemical, clinical, and anthropometric variables. The CAP $\geq$ 302 dB/m group had a higher proportion of males (training set $p <$ 0.001 and validation set $p = 0.043$) and a higher median age ($p < 0.001$ for both sets). Racial distribution also differed significantly, with more non-Hispanic whites in the CAP $\geq$ 302 dB/m group ($p < 0.001$ for both sets). No significant differences were found in the ratio of family income to poverty (training set $p = 0.804$ and validation set $p = 0.868$). Biochemical measurements such as ALT, ALB, AST, Cr, GLB, GGT, LDH, phosphorus, calcium, total protein, UA, HbA1c, FPG, HDL-C, hs-CRP, insulin, TC, TG, and LDL-C showed significant differences ($p < 0.05$) between the groups in both sets. Anthropometric measurements, including weight, BMI, arm circumference, waist circumference, and hip circumference, were significantly higher in the CAP $\geq$ 302 dB/m group ($p < 0.05$). Blood pressure measurements (SBP and DBP) were also significantly higher in the CAP $\geq$ 302 dB/m group ($p < 0.001$). Detailed characteristics are presented in Table 1.

Performance of AutoML models and existing scoring systems

The procedures for the selection of variables are shown in Supplemental Figure 1, with seven key variables with nonzero coefficients identified via the "λ_1se" criterion of LASSO regression analysis, including ALT, waist circumference, HbA1c, FPG, HDL-C, insulin, and TG. The detailed coefficient values of the LASSO regression models are listed in Supplemental Table 1. The AutoML in our study developed a total of 29 models by several machine learning algorithms, including GBM, DRF, XRT, GLM, XGBoost, and DL. The hyperparameters of the optimal XGBoost model are listed in Table 2. The hyperparameters of the remaining assessment models are detailed in Supplemental Table 2.

The AUC of each model was tested with the validation set, as shown in Figure 2. According to Table 3, the XGBoost model performed best with the highest AUC value of 0.859 on the validation set. The AUC values obtained by the other models were 0.858 for DL, 0.857 for GBM, 0.853 for GLM, 0.850 for DRF, 0.849 for XRT, 0.838 for FLI, 0.814 for LAP,0.811 for HSI, 0.747 for FLD index, 0.699 for NAFLD index, 0.824 for ZJU index, and 0.844 for FSI.

In terms of sensitivity, the XGBoost model achieved a sensitivity of 0.773, indicating strong performance in identifying true positives. This was followed by DL and GBM with sensitivity rates of 0.747 and 0.753, respectively. XRT achieved a sensitivity of 0.687, while DRF had the lowest sensitivity at 0.568. Among the existing scoring systems, the ZJU index and FLI showed higher sensitivities of 0.860 and 0.803, respectively, while the NAFLD index had the lowest sensitivity at 0.463.

For specificity, GLM recorded the highest value at 0.855, demonstrating strong performance in identifying true negatives. XRT and DRF followed with specificities of 0.846 and 0.840, respectively. XGBoost and DL had slightly lower specificities of 0.802 and 0.808. In comparison, the NAFLD index showed the highest specificity among existing scoring systems at 0.839, while the ZJU index had the lowest specificity at 0.640.

Regarding overall accuracy, GLM achieved the best value of 0.816, indicating balanced performance in both sensitivity and specificity. DRF and XRT followed with accuracy values of 0.808 and 0.809, respectively, while XGBoost and DL had accuracies of 0.795 and 0.794. Among the existing scoring systems, the FSI showed the highest accuracy at 0.758, while the FLD index had the lowest accuracy at 0.703. In summary, XGBoost was chosen as the best model due to its highest AUC, strong sensitivity and NPV, and overall balanced performance across key metrics. Other predictive models for fatty liver as detailed in recent studies are catalogued in Supplemental Tables 3 and 4.

Table 1. Demographic and clinical characteristics of participants in the training and validation set.

	Training set, $N = 2874$			Validation set, $N = 1303$		
Variables	CAP < 302 dB/m, $N = 2164$[#]	CAP ≥ 302 dB/m, $N = 710$[#]	*p*-value*	CAP < 302 dB/m, $N = 1003$[#]	CAP ≥ 302 dB/m, $N = 300$[#]	*p*-value*
Gender			<0.001			0.043
Male	1006 (46%)	393 (55%)		475 (47%)	162 (54%)	
Female	1158 (54%)	317 (45%)		528 (53%)	138 (46%)	
Age	41.00 [22.00,60.00]	53.00 [37.00,64.00]	<0.001	39.00 [22.00,60.00]	54.00 [41.00,64.25]	<0.001
Race			<0.001			<0.001
Mexican American	261 (12%)	124 (17%)		118 (12%)	63 (21%)	
Other Hispanic	219 (10%)	71 (10%)		107 (11%)	31 (10%)	
Non-Hispanic White	698 (32%)	270 (38%)		300 (30%)	98 (33%)	
Non-Hispanic Black	602 (28%)	132 (19%)		268 (27%)	56 (19%)	
Other Race	384 (18%)	113 (16%)		210 (21%)	52 (17%)	
Family income to poverty ratio			0.804			0.868
ratio < 1	432 (20%)	133 (19%)		195 (19%)	58 (19%)	
1 ≤ ratio < 2	559 (26%)	197 (28%)		265 (26%)	82 (27%)	
2 ≤ ratio < 3	368 (17%)	119 (17%)		162 (16%)	50 (17%)	
3 ≤ ratio < 5	432 (20%)	134 (19%)		190 (19%)	61 (20%)	
ratio ≥ 5	373 (17%)	127 (18%)		191 (19%)	49 (16%)	
ALT (U/L)	15.00 [11.00,21.00]	22.00 [16.00,34.00]	<0.001	16.00 [11.00,22.00]	23.00 [16.00,30.25]	<0.001
ALB (mg/L)	41.00 [39.00,43.00]	40.00 [38.00,42.00]	<0.001	41.00 [39.00,43.00]	40.00 [38.00,42.00]	<0.001
ALP (IU/L)	76.00 [61.00,96.00]	77.50 [65.00,94.00]	0.166	76.00 [63.00,98.00]	80.00 [65.75,96.00]	0.344
AST (U/L)	18.00 [15.00,22.00]	20.00 [16.00,26.00]	<0.001	18.00 [16.00,22.00]	20.00 [17.00,25.00]	<0.001
Cr (μmol/L)	70.72 [59.23,84.86]	74.26 [61.00,85.75]	0.007	72.49 [61.00,85.30]	71.60 [59.89,87.52]	0.738
GLB (g/dL)	3.00 [2.80,3.30]	3.10 [2.90,3.40]	<0.001	3.10 [2.80,3.30]	3.10 [2.80,3.40]	0.098
GGT (IU/L)	17.00 [12.00,25.00]	26.00 [18.00,40.00]	<0.001	17.00 [13.00,26.00]	26.00 [18.00,41.00]	<0.001
LDH (IU/L)	152.00 [134.00,173.00]	154.00 [139.00,174.00]	0.018	153.00 [136.00,173.00]	154.00 [137.00,174.25]	0.959
P (mmol/L)	1.16 [1.03,1.29]	1.10 [1.00,1.23]	<0.001	1.16 [1.03,1.29]	1.10 [0.99,1.23]	<0.001
STB (μmol/L)	6.84 [5.13,10.26]	6.84 [5.13,10.26]	0.101	6.84 [5.13,10.26]	6.84 [5.13,8.55]	0.485
Ca (mmol/L),	2.33 [2.27,2.38]	2.30 [2.25,2.35]	<0.001	2.33 [2.25,2.38]	2.30 [2.25,2.38]	0.012
TP (g/L)	72.00 [69.00,74.00]	71.00 [69.00,74.00]	0.021	72.00 [69.00,74.00]	71.00 [68.00,74.00]	0.030
UA (μmol/L)	303.30 [249.80,356.90]	345.00 [291.50,404.50]	<0.001	303.30 [249.80,356.90]	339.00 [279.60,398.50]	<0.001

(continued)

Table 1. Continued.

	Training set, $N = 2874$			Validation set, $N = 1303$		
Variables	CAP < 302 dB/m, $N = 2164$#	CAP ≥ 302 dB/m, $N = 710$#	p-value*	CAP < 302 dB/m, $N = 1003$#	CAP ≥ 302 dB/m, $N = 300$#	p-value*
Weight (kg)	72.10 [60.30,85.80]	96.25 [82.30,111.97]	<0.001	72.40 [60.95,84.70]	92.00 [77.30,107.50]	<0.001
BMI (kg/m^2)	26.05 [22.40,30.32]	33.70 [29.70,38.87]	<0.001	26.10 [22.60,29.90]	32.30 [28.70,37.52]	<0.001
Arm circumference (cm)	31.40 [27.90,34.70]	36.90 [33.50,40.27]	<0.001	31.20 [28.10,34.50]	35.60 [33.00,39.20]	<0.001
Waist circumference (cm)	91.20 [79.70,102.62]	112.50 [102.60,123.27]	<0.001	91.90 [80.50,101.30]	110.35 [99.97,121.23]	<0.001
Hip circumference (cm)	100.25 [93.40,109.00]	113.00 [105.30,124.20]	<0.001	100.30 [94.20,107.30]	110.00 [102.77,121.93]	<0.001
SBP (mmHg)	116.00 [106.67,129.00]	123.00 [112.67,135.92]	<0.001	116.00 [107.33,127.67]	125.00 [114.67,139.08]	<0.001
DBP (mmHg)	70.33 [63.33,77.67]	76.67 [69.33,83.67]	<0.001	70.33 [63.67,77.33]	75.67 [68.00,83.75]	<0.001
PLT (1000 cells/μL)	241.00 [202.00,285.00]	238.00 [203.00,284.00]	0.834	241.00 [205.00,280.00]	237.00 [198.75,285.50]	0.948
Ferritin (μg/L)	82.20 [37.60,156.00]	126.00 [59.05,232.00]	<0.001	83.40 [40.80,161.50]	131.50 [65.25,225.50]	<0.001
Iron (μmol/L)	15.00 [11.10,19.90]	14.90 [11.30,18.80]	0.214	15.00 [10.90,19.70]	14.90 [11.45,17.90]	0.358
TIBC (μg/dL)	57.85 [52.30,64.12]	58.03 [52.52,63.76]	0.849	57.85 [52.30,63.94]	58.12 [53.37,62.51]	0.978
TSF (%)	26.50 [19.00,35.00]	26.00 [20.00,34.00]	0.333	27.00 [20.00,34.00]	26.00 [20.00,32.00]	0.340
HbA1c (%)	5.40 [5.20,5.70]	5.80 [5.50,6.60]	<0.001	5.40 [5.20,5.80]	5.85 [5.50,6.70]	<0.001
FPG (mmol/L)	5.55 [5.22,5.94]	6.16 [5.62,7.33]	<0.001	5.55 [5.22,5.94]	6.16 [5.66,7.50]	<0.001
HDL-C (mg/dL)	1.37 [1.14,1.63]	1.14 [0.98,1.34]	<0.001	1.37 [1.16,1.63]	1.16 [1.01,1.38]	<0.001
hs-CRP (mg/L)	1.27 [0.57,3.21]	3.13 [1.41,6.26]	<0.001	1.33 [0.57,3.13]	2.89 [1.35,6.71]	<0.001
Insulin (μU/mL)	8.84 [5.89,13.34]	17.00 [11.27,25.66]	<0.001	8.26 [5.55,13.18]	17.78 [10.38,25.77]	<0.001
TC (mg/dL)	172.00 [149.00,199.25]	179.00 [156.00,207.00]	<0.001	174.00 [150.00,200.00]	183.00 [155.00,210.25]	0.004
TG (mmol/L)	0.83 [0.58,1.25]	1.38 [0.96,1.95]	<0.001	0.86 [0.60,1.21]	1.29 [0.93,1.87]	<0.001
LDL-C (mmol/L)	2.59 [2.04,3.21]	2.77 [2.20,3.41]	<0.001	2.61 [2.10,3.21]	2.81 [2.22,3.39]	0.009

CAP: controlled attenuation parameter; ALT: alanine aminotransferase; ALB: albumin; ALP: alkaline phosphatase; AST: aspartate aminotransferase; Cr: creatinine; GLB: globulin; GGT: gamma glutamyl transferase; LDH: lactate dehydrogenase; P: phosphorus; STB: total bilirubin; Ca: total calcium; TP: total protein; UA: uric acid; BMI: body mass index; SBP: systolic pressure; DB: diastolic pressure; PLT: platelet count; TIBC: total iron binding capacity; TSF: transferrin saturation; HbA1c: glycohemoglobin; FPG: fasting plasma glucose; HDL-C: high-density lipoprotein cholesterol; hs-CRP: high-sensitivity c-reactive protein; TC: total cholesterol; TG: triglyceride; LDL-C: low-density lipoprotein-cholesterol.
#n (%); Median [25%,75%].
*Pearson's *Chi*-squared test; Wilcoxon rank sum test.

An intuitive comparison of the consistency risk assessment model was conducted between the ideal calibration curve and the calibration curves. The evaluation of consistency was further refined by analyzing the slope, ideally valued at 1, and the Brier score, where the deal value is 0, and values exceeding 0.3 indicate poor calibration. The calibration curves presented in Figure 3 demonstrate a robust calibration of the XGBoost model, affirming its reliability. This is evidenced by the close alignment of predicted probabilities with the actual observed probabilities. The model's precision is further highlighted by Brier scores of 0.112 for the training set and 0.121 for the validation set, along with a calibration slope of 1.110 in the training phase and 0.935 in the validation phase. These metrics underscore the model's consistent performance across both the development and application stages.

To evaluate the clinical utility of the XGBoost model on NAFLD identification, DCA was conducted. The clinical decision curves, illustrated in Figure 4, revealed that the net benefit of using the model for assessing NAFLD surpassed that of the "no assessment" or "all assessment" regimens. This analysis confirms the XGBoost model as an

Table 2. Hyperparameters of AutoML classification algorithms.

	Hyperparameter	Argument
XGBoost	max_depth	3
	min_child_weight	20
	subsample	0.6
	colsample_bytree	0.8
	reg_alpha	0.5
	reg_lambda	100
	stopping_metric	logloss
	stopping_tolerance	0.01865334
	tree_method	exact
	calibration_method	PlattScaling
	categorical_encoding	OneHotInternal
	seed	670

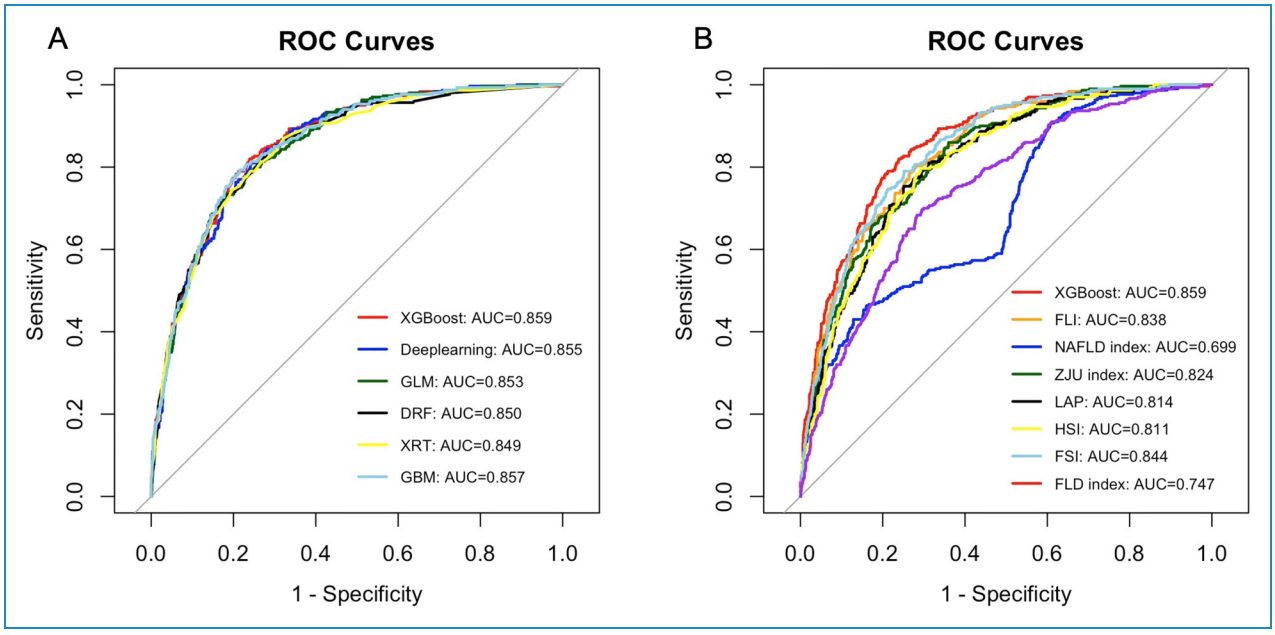

Figure 2. ROC curves of different models and traditional scoring systems on the validation dataset. (A) ROC curves comparing the performance of various machine learning models. (B) ROC curves comparing the performance of the XGBoost model with traditional scoring systems.
XGBoost: eXtreme Gradient Boosting; GBM: Gradient Boosting Machine; GLM: Generalized Linear Model; DRF: Distributed Random Forest; XRT: Extremely Randomized Trees; FLI: fatty liver index; LAP: lipid accumulation product; HSI: hepatic steatosis index; FSI: Framingham steatosis index; FLD index: fatty liver disease index.

Table 3. Performance of models on the validation set.

	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV	LR+	LR−
AutoML								
XGBoost	**0.859**	**0.795**	**0.773**	**0.802**	**0.538**	**0.922**	**3.898**	**0.283**
DL	0.858	0.794	0.747	0.808	0.537	0.914	3.880	0.314
GBM	0.857	0.803	0.753	0.818	0.553	0.917	4.129	0.302
GLM	0.853	0.816	0.683	0.855	0.586	0.900	4.727	0.370
DRF	0.850	0.808	0.568	0.840	0.568	0.903	4.389	0.357
XRT	0.849	0.809	0.687	0.846	0.572	0.900	4.472	0.370
Existed scoring systems								
FLI[11]	0.838	0.738	0.803	0.719	0.461	0.924	2.857	0.274
LAP[12]	0.814	0.750	0.753	0.749	0.473	0.910	2.998	0.329
HSI[13]	0.811	0.728	0.797	0.708	0.449	0.921	2.727	0.287
FLD index[14]	0.747	0.703	0.697	0.705	0.414	0.886	2.361	0.430
NAFLD index[15]	0.699	0.753	0.463	0.839	0.463	0.839	2.886	0.639
ZJU index[16]	0.824	0.691	0.860	0.640	0.417	0.939	2.389	0.219
FSI[17]	0.844	0.758	0.790	0.749	0.485	0.923	3.144	0.280

AUC: area under receiver operating characteristic curves; PPV: positive predictive value; NPV: negative predictive value; LR+: positive likelihood ratio; LR−: negative likelihood ratio; XGBoost: eXtreme Gradient Boosting; DL: Deep Learning; GBM: Gradient Boosting Machine; GLM: Generalized Linear Model; DRF: Distributed Random Forest; XRT: Extremely Randomized Trees; FLI: fatty liver index; LAP: lipid accumulation product; HSI: hepatic steatosis index; FSI: Framingham steatosis index; FLD index: fatty liver disease index.

effective diagnostic tool for NAFLD, demonstrating significant clinical net benefits in identifying the condition.

Interpretation of AutoML models

Furthermore, we identified the most significant clinical features contributing to the identification of NAFLD. The variable importance plot was constructed using XGBoost algorithms. The variable importance plot, shown in Figure 5, ranks these features in descending order of relevance. As depicted in Figure 5, waist circumference emerges as the most crucial risk factor for NAFLD. It is followed by insulin, TG, ALT, FPG, HbA1c, and HDL-C levels. This ranking highlights the relative importance of each variable in the predictive model, with waist circumference having the most substantial impact on NAFLD identification.

In Figure 6, we present the SHAP summary plot, which illustrates the contribution of each variable to the prediction of NAFLD for each instance. This plot provides a detailed view of how individual features influence the model's predictions. Notably, the plot shows that higher values of certain features are positively correlated with an increased likelihood of NAFLD. Specifically, waist circumference, insulin levels, ALT, TG, and HbA1c are prominently associated with the presence of NAFLD. Each point on the plot represents a SHAP value for a feature, with the color gradient indicating the normalized value of the feature (ranging from low in blue to high in red). The SHAP values (positioned on the x-axis) reflect the impact of each feature on the model's output. Features with positive SHAP values push the prediction towards higher risk, while negative values push it towards lower risk. For instance, higher waist circumference and insulin levels significantly increase the likelihood of NAFLD, as indicated by their clustering on the right side of the plot with predominantly red points.

To interpret the influence of individual features on predicted outcomes, we utilized the PDP technique for interpreting our machine learning model. As presented in Figure 7, waist circumference, insulin, TG, ALT, FPG, and HbA1c exhibit positive correlations with the likelihood of NAFLD. This indicates that higher values of these

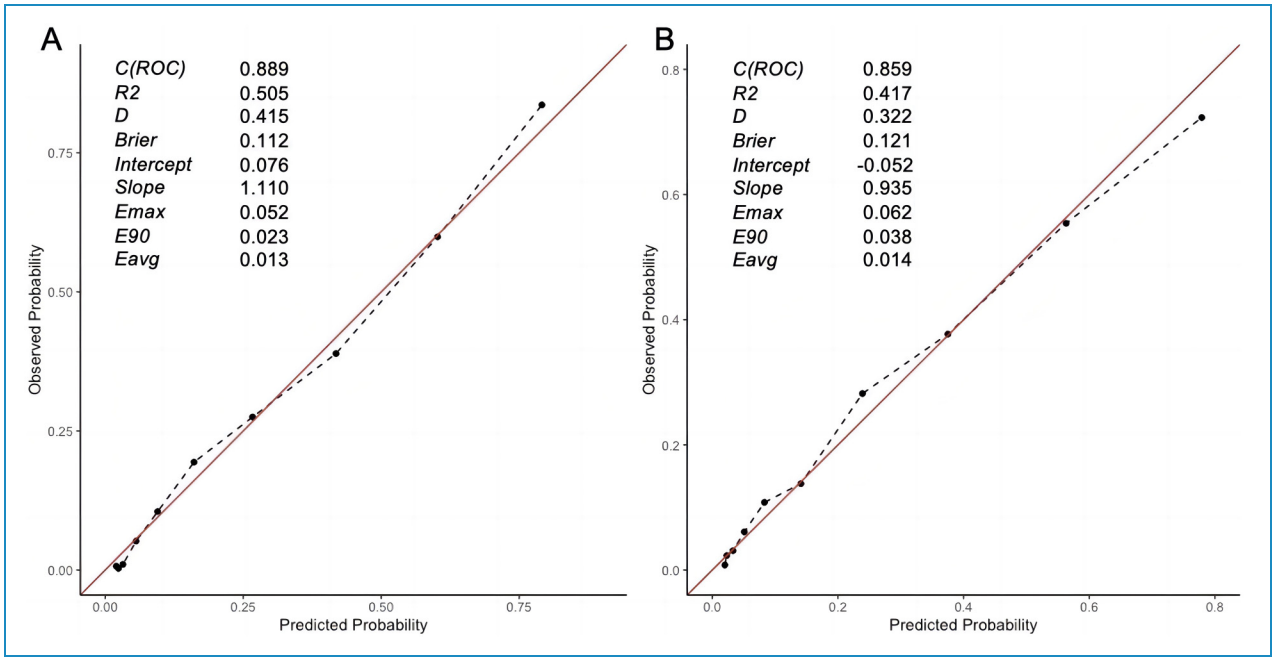

Figure 3. Calibration curves of the XGBoost model on the training set (A) and the validation set (B). The calibration curves demonstrated a high degree of reliability by showing that the predicted probability was close to the observed probability.

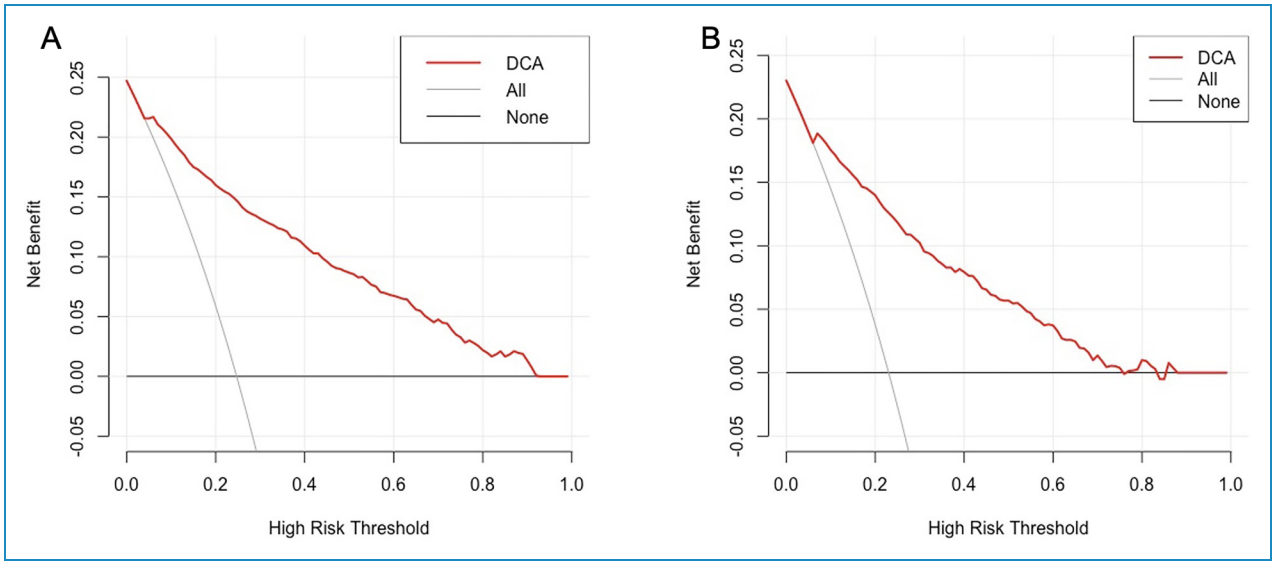

Figure 4. Decision curve analysis of the XGBoost model on the training set (A) and validation set (B). Decision curve analysis of the XGBoost model on the training and validation set, indicating clinical net benefits of approximately 25%. The threshold probability was represented on the *x*-axis, while the clinical net benefits were displayed on the *y*-axis. The grey line indicated the strategy of the assumption that all patients have received the assessment of the XGBoost model, while the horizontal black line demonstrated the strategy of the assumption that no patient has received the evaluation of the XGBoost model.

features are associated with an increased risk of NAFLD. Conversely, HDL-C levels show a negative association with NAFLD, as higher HDL-C levels are associated with a reduced risk.

Figure 8 provides an insightful illustration of the XGBoost model's predictions using the LIME technique. This figure showcases the model's explanations for four randomly selected cases from the validation set, highlighting the contributions of various features to each prediction. The color-coded bars indicate whether each feature supports (blue) or contradicts (red) the predicted outcome. The length of the bars represents the weight of each

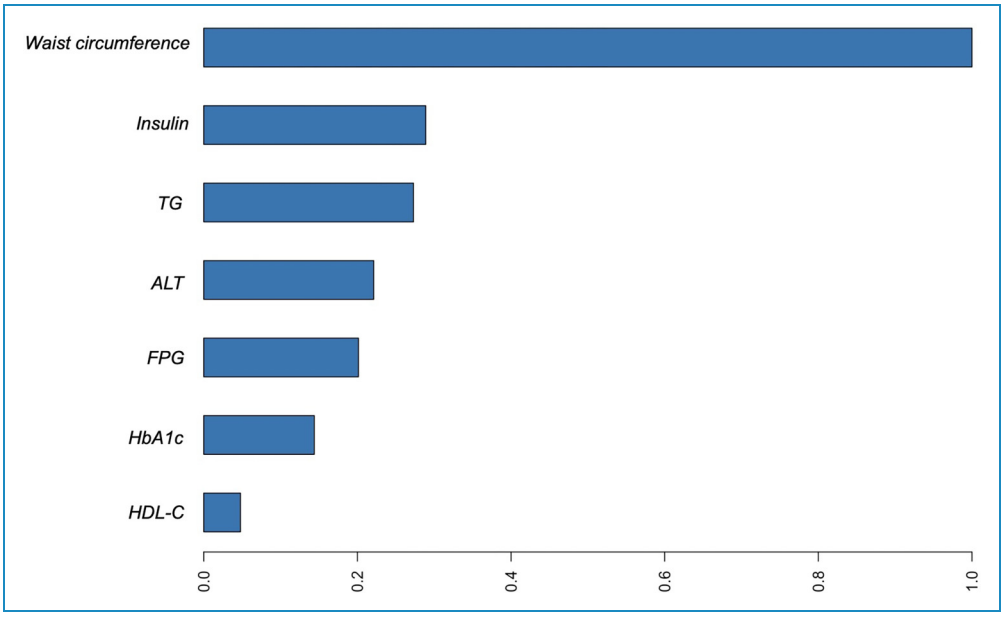

Figure 5. Variable importance of the XGBoost model on the training set.
TG: triglyceride; ALT: alanine aminotransferase; FPG: fasting plasma glucose; HbA1c: glycohemoglobin; HDL-C: high-density lipoprotein cholesterol.

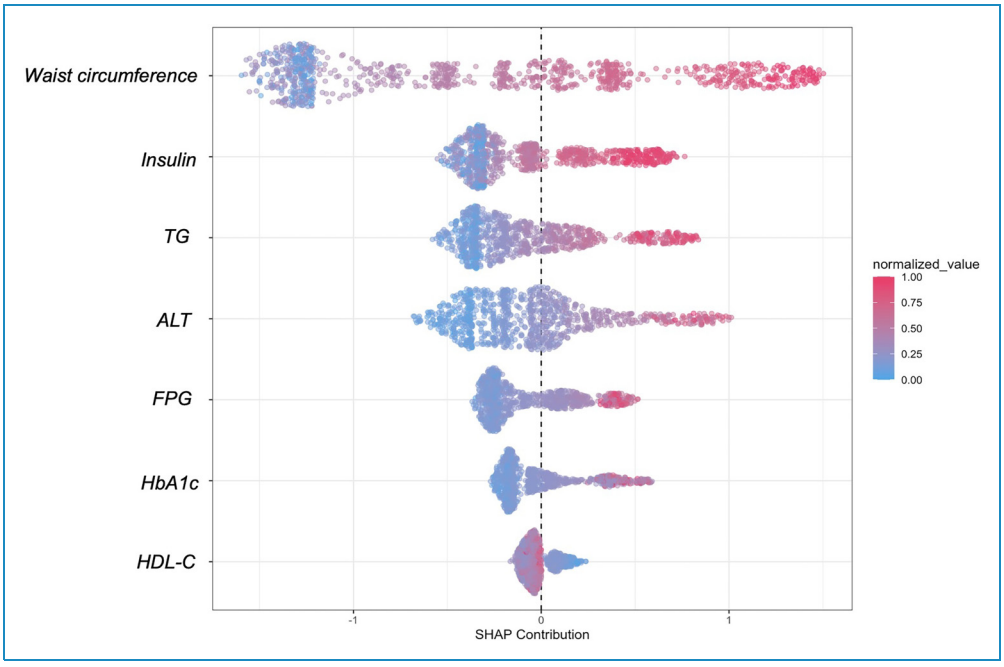

Figure 6. SHAP plot of the XGBoost model. The closer the variable values were to 1, the greater the likelihood that NAFLD would be identified.
SHAP: SHapley additive explanation; ALT:alanine aminotransferase; TG: triglyceride; FPG: fasting plasma glucose; HbA1c: glycohemoglobin; HDL-C: high-density lipoprotein cholesterol.

feature's contribution to the prediction. For instance, in case #7 (label 1, indicating the presence of NAFLD), the XGBoost model predicted a high probability of 0.71 for NAFLD. The most significant variable contributing to the prediction was TG, followed by HbA1c, HDL-C, FPG, waist circumference, ALT, and insulin levels. Similarly, case #3 (label 0, indicating the absence of NAFLD) had a predicted probability of 0.87 for not having NAFLD. The

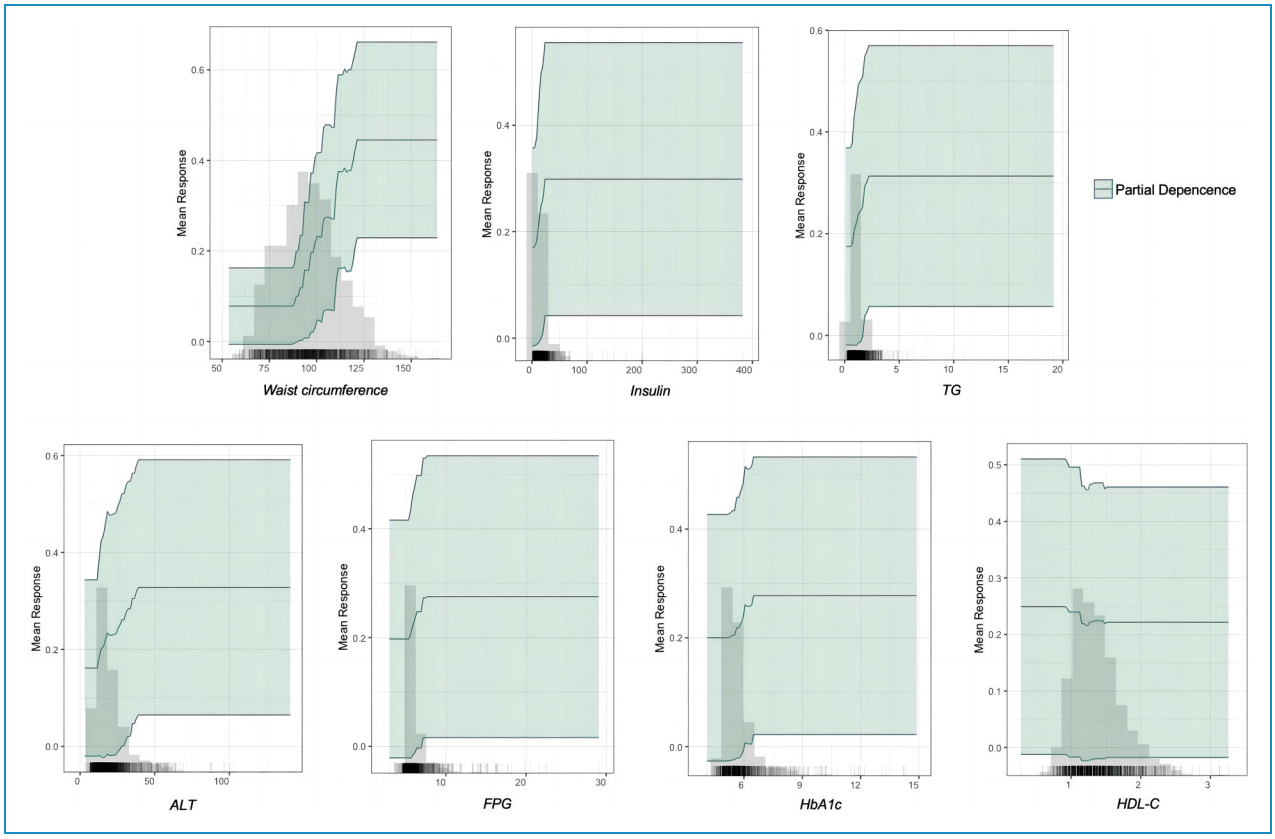

Figure 7. PDP for the important variables in the XGBoost model.
PDP: partial dependence plot; TG: triglyceride; ALT: alanine aminotransferase; FPG, fasting plasma glucose; HbA1c: glycohemoglobin; HDL-C: high density lipoprotein cholesterol.

primary contributors to this prediction were TG and HbA1c levels, which strongly contradicted the presence of NAFLD, while HDL-C supported the prediction.

The code for automated machine learning modeling is available on the following website: https://osf.io/nvwek, while the code for model deployment can be accessed at https://osf.io/6cxmb. Clinical practitioners are invited to further refine and optimize the model by utilizing these open-source resources in their future work.

Discussion

In this investigation, we have developed a series of AutoML models to detect NAFLD utilizing the NHANES database. These models were all superior to existing scoring systems such as FLI, LAP, HSI, NAFLD index, ZJU index, FSI, and FLD index. Notably, the model employing the XGBoost algorithm emerged as the superior performer within the ensemble of AutoML models on the validation set. To augment the interpretability of the model, we implemented an array of explanatory tools including variable importance, SHAP, PDP, and LIME. Additionally, we introduced 'Shiny NAFLD', a newly developed R Shiny tool based on the XGBoost model for identifying NAFLD using demographic data and clinical data. 'Shiny NAFLD' stands as a testament to the practical utility of our research, offering an accessible platform for healthcare professionals to identify NAFLD.

The significant demographic differences observed in our study suggest that higher CAP values are associated with specific population characteristics. The higher proportion of males and the older median age in the CAP $\geq$ 302 dB/m group indicate that gender and age may influence liver fat accumulation, as noted previously.[36] The significant racial differences, with a higher proportion of non-Hispanic whites in the CAP $\geq$ 302 dB/m group, point to potential genetic or lifestyle factors that warrant further investigation.[37]

In addition, the biochemical differences highlight several health risks associated with higher CAP values. Consistent with Ayada et al.'s findings, elevated levels of ALT, AST, and GGT suggest that individuals in the CAP $\geq$ 302 dB/m group may have a higher risk of liver dysfunction.[38] The higher levels of HbA1c and FPG in the CAP $\geq$ 302 dB/m group underscore a greater prevalence of impaired glucose metabolism or diabetes, which aligns with studies showing a strong association between NAFLD and metabolic disorders such as type 2 diabetes.[39] Additionally, the abnormal lipid

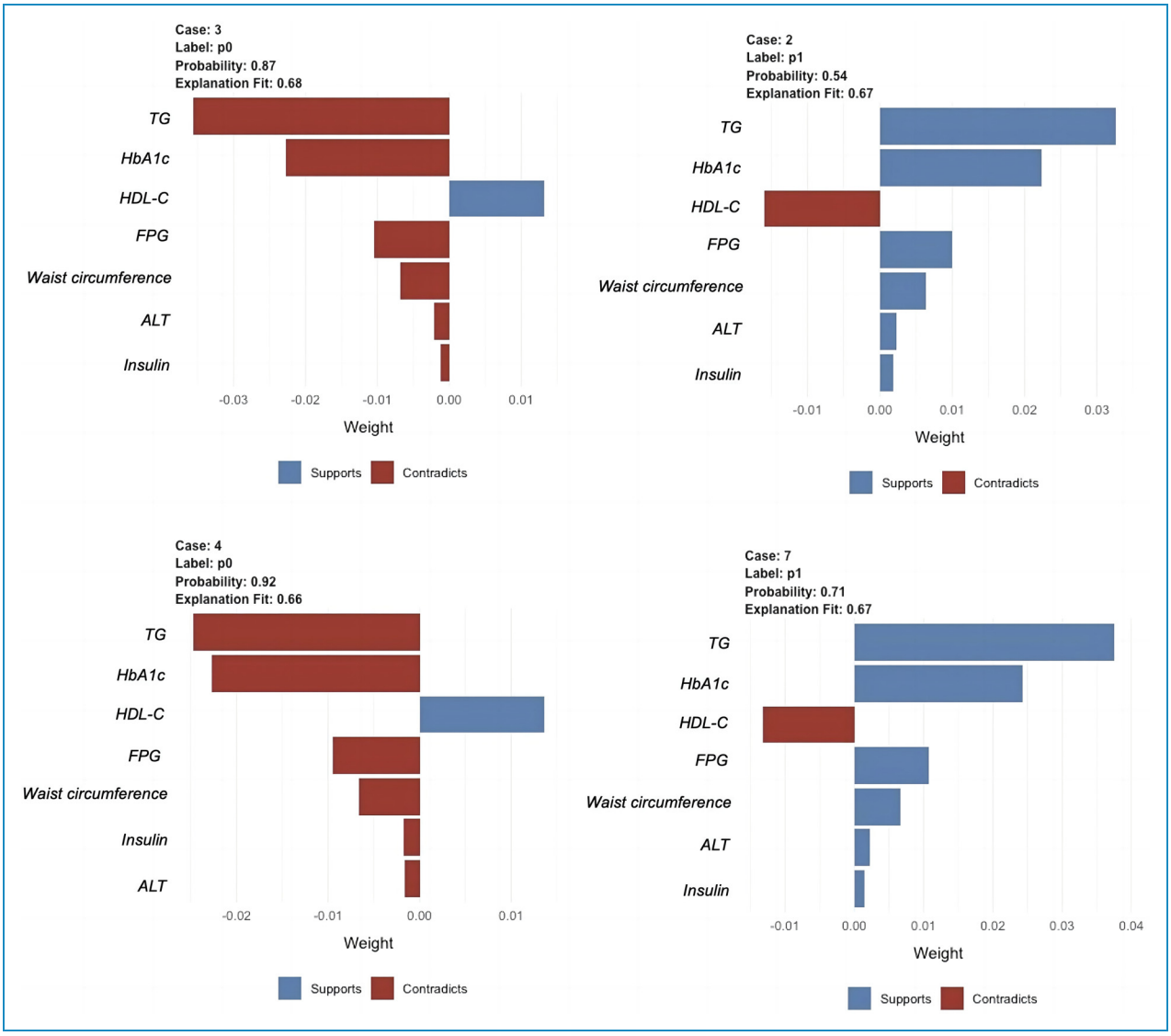

Figure 8. LIME plots of the XGBoost model. Four individuals were selected randomly from the validation set to see the impact of the main features on the outcome.
LIME: local interpretable model agnostic explanation; TG: triglyceride; HbA1c: glycohemoglobin; HDL-C: high-density lipoprotein cholesterol; FPG: fasting plasma glucose; ALT: alanine aminotransferase.

profiles, including elevated TC, TG, and LDL-C, suggest an increased risk of cardiovascular diseases, corroborating previous research that highlights the link between NAFLD and cardiovascular risk.[40] Furthermore, the anthropometric measurements further emphasize the health risks, with significantly higher weight, BMI, arm circumference, waist circumference, and hip circumference in the CAP $\geq$ 302 dB/m group, indicating a strong association between NAFLD and obesity, which is a known risk factor for various metabolic disorders.[41,42] Overall, our findings highlight the importance of regular monitoring and targeted interventions for individuals with NAFLD.[43] These individuals are at an increased risk for liver dysfunction, metabolic disorders such as diabetes, cardiovascular issues, and obesity-related complications.[38–41] Comprehensive health assessments and personalized treatment plans are crucial for managing these risks and improving health outcomes.

Considering the adverse effects associated with invasive liver biopsies, the subjectivity inherent in ultrasound examinations, and the prohibitive costs associated with FibroScan diagnostics, our AutoML models stand out as high-quality tools for NAFLD diagnosis. Traditional scoring systems, including FLI,[11] LAP,[12] HSI,[13] NAFLD index,[15] ZJU index,[16] FSI,[17] and FLD index,[14] demonstrated inferior accuracy and AUC compared to the models we built using AutoML. This aligns with recent literature emphasizing the need for more accurate and non-invasive diagnostic tools for NAFLD.[5,8] In this study, we evaluated a suite of models constructed using AutoML's

six algorithms (GBM, DRF, XRT, GLM, XGBoost, and DL) to assess the presence of NAFLD. The XGBoost model demonstrated superior performance with an AUC of 0.859, an accuracy of 0.795, a sensitivity of 0.773, a specificity of 0.802, a PPV of 0.538, an NPV of 0.922, an LR + of 3.898, and an LR− of 0.283. The AUC metric proved particularly useful in addressing the challenges of unbalanced data, as it inherently weights both classes equally, unlike accuracy.[44] Additionally, a high AUC indicates that the model has a strong capability to distinguish between patients with and without NAFLD, making it a reliable measure for evaluating the effectiveness of our diagnostic approach.

Given that early-stage NAFLD is often asymptomatic and, if left untreated, may progress to more severe conditions such as cirrhosis and hepatocellular carcinoma, our research prioritizes the early detection of NAFLD patients, making the sensitivity metric particularly critical. Sensitivity, defined as the proportion of true positives among all confirmed cases, measures our model's ability to correctly identify actual cases of NAFLD. The XGBoost model's high sensitivity of 0.773 underscores its success in accurately detecting NAFLD among those afflicted, which is particularly significant given the asymptomatic nature of early-stage NAFLD, where early detection is crucial for preventing disease progression.

Furthermore, specificity indicates the model's ability to correctly identify patients without NAFLD, thus reducing the rate of false positives. The XGBoost model achieved a specificity of 0.802, minimizing unnecessary further testing and patient anxiety. PPV measures the proportion of true positive results among all positive predictions, with the XGBoost model demonstrating a PPV of 0.538, indicating the reliability of a positive NAFLD diagnosis. NPV represents the proportion of true negative results among all negative predictions, and the XGBoost model showed an impressive NPV of 0.922, highlighting its effectiveness in ruling out NAFLD when the prediction is negative. Moreover, LR + and LR− provide additional insights into the model's diagnostic utility, with a high LR + indicating a strong association between a positive test result and the presence of NAFLD, and a low LR− suggesting that a negative test result is effective in ruling out the disease. Consequently, given its optimal balance of these metrics, the XGBoost model emerged as the most effective tool in our analysis for NAFLD detection.

The XGBoost model demonstrated robust calibration, with Brier scores of 0.112 for the training set and 0.121 for the validation set, and calibration slopes of 1.110 and 0.935, respectively. The DCA showed that the XGBoost model provides a higher net benefit for NAFLD identification compared to "no assessment" or "all assessment" strategies. Combining the results of the calibration curves and DCA, we can assert that our XGBoost model not only performs well in terms of predictive accuracy but also offers tangible benefits in clinical settings. The robust calibration ensures that the predicted risks are reliable, while the decision curve analysis highlights the practical advantages of implementing the model in patient care.

To enhance the interpretability of our XGBoost model, we employed several techniques: variable importance analysis, SHAP, PDP, and LIME. These techniques highlight the most critical predictors, such as waist circumference and insulin levels, and provide detailed insights into how individual features influence the model's predictions. Variable importance analysis ranks features based on their contribution to the model's predictions, thereby identifying the most influential factors. Additionally, SHAP values quantify the contribution of each feature to individual predictions, demonstrating the positive correlation of features like waist circumference and insulin with NAFLD risk. Furthermore, PDP illustrates the relationship between a feature and the predicted outcome while keeping all other features constant, which helps in understanding the marginal effect of a feature. Finally, LIME offers local explanations for specific predictions, providing case-specific insights that clarify the model's decision-making process. By using these methods, we ensure that our model's predictions are understandable, trustworthy, and clinically relevant. These interpretability techniques improve the transparency and applicability of the model, enhancing trust in its use for clinical decision-making.

NAFLD exhibits a complex, bidirectional association with metabolic syndrome components, acting both as a contributing factor and a result of the metabolic syndrome (MetS).[1,45] In our study, waist circumference was found to be the most crucial variable in the variable importance plot, followed by insulin, TG, ALT, FPG, HbA1c, and HDL-C. It is noteworthy that the majority of these critical variables (waist circumference, insulin, TG, FPG, and HbA1c) are core components of MetS, which demonstrates the feasibility of metabolic syndrome markers in identifying NAFLD.

Previous research has established waist circumference as an independent risk factor and a potent predictor of NAFLD.[46,47] A substantial retrospective study that analyzed physical examination data from Chinese adults yielded predictive models for NAFLD, accentuating the value of waist circumference as an indicator.[20] Several studies have demonstrated a strong correlation between obesity and NAFLD.[48–50] However, emerging research suggests that this relationship might be impacted by fat distribution.[51] A condition termed 'lean NAFLD' occurs in individuals who develop NAFLD despite having a normal BMI (<25 kg/m^2 in non-Asian, <23 kg/m^2 in Asian).[52] Such patients usually have central obesity or possess additional metabolic risk factors.[53] A meta-analysis from Pang et al. investigated the independent associations of central and general obesity with NAFLD.[42] Their findings highlighted that the pooled OR for waist circumference

(3.14 [2.07–4.77]) was higher compared to BMI (2.85 [1.60–5.08]), with binary variables and using the nonobese cohort as a reference. This indicates that abdominal obesity may pose a higher risk for NAFLD than general obesity.

Insulin has been proven as a strong predictor of NAFLD according to previous studies.[54,55] Bril et al.[56] demonstrated that intact molecules of insulin by mass spectrometry hold a high AUC of 0.90 for identifying NAFLD in individuals without diabetes. They also suggested measuring fasting intact insulin levels as a simple, non-invasive method of identifying the presence of NAFLD.

TG and FPG are two key metabolic variables that are modified in fatty liver and have a strong correlation with insulin resistance.[57] Insulin resistance, which is defined as the ability of insulin to inhibit glucose generation, lipid synthesis, and lipolysis, is the underlying cause of hyperglycemia and an increase in TG in NAFLD. Insulin resistance also limits the receptor-mediated entry of insulin into the liver, which reduces insulin clearance.[58,59] Furthermore, there will be hyperinsulinemia, leading to hepatic steatosis.[58] Tomizawa et al. reported that TG was the strongest predictor of NAFLD among markers of hyperlipidemia and diabetes.[60] Another study by Liu et al. proposed that elevated levels of circulating TG and high glucose levels are likely to increase the risk for NAFLD.[61] In addition, Zhang et al. observed that individuals with NAFLD had significantly higher FPG and TG levels than those with non-NAFLD.[62] Recently, the triglyceride and glucose index (TyG), derived from the product of TG and FPG, has been validated as a reliable biomarker for identifying NAFLD with a sensitivity of 72.2% and a specificity of 70.5%.[62] Lee et al. have also reported the predictive powers of the homeostatic model assessment for insulin resistance (HOMA-IR), a hepatic insulin resistance index consisting of FPG and Insulin, for predicting NAFLD.[63] Their findings reveal that HOMA-IR exhibits excellent predictive capacity for NAFLD with an AUC of 0.831.

HbA1c serves as a valuable marker for chronic glycemic control by reflecting average blood glucose levels over the preceding 8–12 weeks.[64] Ma et al. and Bae et al. discovered serum HbA1c level was significantly and independently associated with the risk for NAFLD.[65,66] In clinical practice, a mild to moderate elevation in ALT is frequently the sole laboratory abnormality in NAFLD patients and is considered an early, surrogate biomarker for the disease.[67] Furthermore, Zhang et al. conducted research on ALT as a diagnostic tool for NAFLD, with an AUC of 0.715.[62] Additionally, a Mendelian randomization analysis has demonstrated that genetically predicted elevated serum liver enzymes could increase NAFLD risk, whereas HDL-C was linked to a decreased risk of NAFLD.[68]

Logistic regression, a linear model commonly employed for binary classification tasks, struggled to tackle complex learning tasks. With the development of machine learning, leveraging algorithms, such as XGBoost, GBM, support vector machine (SVM), deep learning, and DRF, have significantly improved the efficacy of handling more intricate problems. Recent studies[69,70] provide robust evidence of the exceptional potential of machine learning in disease diagnosis and the anticipation of risk factors. Qin et al.[23] have demonstrated the efficacy of the SVM model for NAFLD screening achieving an impressive accuracy of 0.801 and an AUC of 0.850, using data from annual health examinations. Additionally, using data from NHANES 1988–1994, Atsawarungruangkit et al.[29] developed a random undersampling boosted trees model to predict NAFLD with an accuracy of 0.711. Due to the use of different databases and diagnostic methods, comparing the results is inappropriate. Noureddin et al.[28] have developed six different machine-learning models to identify NAFLD by leveraging demographic and clinical data from the NHANES 2017–2018 cohort, with participants identified through transient elastography. The performance metrics of the tested models exhibit an AUC in the range of 0.79 to 0.84, an accuracy spanning from 0.75 to 0.79, and a sensitivity varying between 0.53 and 0.71. These figures are comparatively lower than those of our proposed model, which has an AUC of 0.859 and an accuracy of 0.795, underscoring its superior diagnostic efficacy. The use of AutoML automates the selection and tuning of machine learning models, thereby simplifying the process for clinical practitioners and enhancing diagnostic efficacy.

Innovative imaging techniques leverage sophisticated algorithms and neural networks to significantly improve the precision of diagnostics.[71] However, such models require parameter adjustment and feature engineering that rely on human machine learning experts, hence constraining their widespread implementation. Our study addresses this challenge by using the H2O AutoML platform, which automates these processes, making machine learning more accessible to individuals lacking expertise in this domain and enhancing the efficiency of machine learning processes. Moreover, our development of the Shiny NAFLD web application provides a practical and user-friendly tool for healthcare professionals, offering a non-invasive method to enhance diagnostic accuracy. This is particularly significant in the context of NAFLD, where early and accurate diagnosis is crucial for preventing disease progression and implementing effective treatment strategies.

The integration of AutoML into NAFLD diagnostics represents a shift towards a more streamlined, individualized, and objective approach. This innovation is poised to markedly enhance patient outcomes by enabling the early detection of conditions, thereby improving prognoses and reducing dependence on diagnostic methods that are often costly and limited in availability. Particularly advantageous in primary care settings and regions with limited resources, AutoML emerges as a cost-effective and scalable solution, optimizing the distribution of medical resources. The incorporation of AutoML into clinical practices may

signify an advancement towards an accessible and efficient strategy for managing NAFLD.

This study highlights several key features: firstly, the application of AutoML streamlines the process of algorithm selection, hyperparameter adjustment, and optimal model output, offering a user-friendly tool for clinicians of varied statistical expertise. Based on the optimal model in AutoML, we present an R Shiny web-based application, Shiny NAFLD, to facilitate the identification of NAFLD in real clinical practice. Secondly, it leverages demographic and clinical information from NHANES, an open-accessed cross-sectional study in the United States, to develop and validate models to identify the presence of NAFLD in study participants. Furthermore, the diagnosis of NAFLD via ultrasound examination is often hampered by numerous factors, particularly the subjectivity of the examiner. This research utilizes the CAP measured at a threshold of ≥302 dB/m through LUTE, which has been shown to yield greater accuracy and sensitivity compared to conventional ultrasound methodologies. Lastly, the study enhances the interpretability of complex 'black-box' models through various visualization techniques, such as variable importance, SHAP, PDP, and LIME.

Some limitations of this study should be noted. Firstly, the same database was used for both the training and validation sets, which may constrain the model's generalizability. Using only the NHANES database for modeling may limit the generalizability of our findings, as the dataset may not fully represent diverse populations and clinical settings. To ensure broader applicability, it is essential to test the model with external datasets from various demographic and geographic backgrounds. Additionally, as a retrospective study, there is an inherent risk of selection bias and other potential biases that could affect the findings. Secondly, due to the limitations and restrictions of the NHANES database, secondary causes of hepatic fat accumulation, such as Wilson disease and inborn errors of metabolism (e.g., lecithin-cholesterol acyltransferase deficiency, cholesterol ester storage disease, Wolman disease), could not be ruled out. Additionally, the scope of the data, including any temporal limitations, should be considered, as data collected over a specific time may not capture all relevant trends and patterns. Lastly, there is ongoing debate surrounding the established cutoff values for steatosis when using CAP, suggesting that further research is required to reach a consensus on these diagnostic thresholds. Future research should focus on validating and refining the model across different populations and clinical environments. Additionally, extending this model to incorporate real-time data integration will enhance its usability and adaptability in diverse clinical scenarios.

Conclusion

This study demonstrated the effectiveness of using machine learning algorithms on the H2O AutoML platform to identify NAFLD by analyzing key variables. XGBoost emerged as the best performer, highlighting its potential for clinical diagnosis. We developed Shiny NAFLD, an R Shiny web application (http://39.101.122.171:3838/App2/), providing healthcare professionals with a non-invasive tool to enhance NAFLD diagnostic accuracy and support personalized treatment strategies. Future research should validate the models on external datasets, explore interpretability techniques, and investigate applications to other diseases.

Acknowledgements: We would like to thank AW and JZ for their assistance and guidance in this research.

Contributorship: AW and JZ researched the literature and conceived the study. LL (Lihe Liu), JL, LL (Lu Liu), JG, GX, MY, and XL were involved in protocol development, gaining ethical approval, and data analysis. LL (Lihe Liu), JL and LL (Lu Liu) wrote the first draft of the manuscript. All authors reviewed and edited the manuscript and approved the final version of the manuscript.

Data sharing statement: All NHANES data for this study are publicly available and can be found in: https://wwwn.cdc.gov/nchs/nhanes.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: We used publicly available datasets from the official NHANES website (https://wwwn.cdc.gov/nchs/nhanes). NHANES was conducted in accordance with the Declaration of Helsinki and approved by the NCHS Research Ethics Review Board (Continuation of Protocol #2011–17 and Protocol #2018-01, 26 October 2017). More details can be found at https://www.cdc.gov/nchs/nhanes/irba98.htm. Therefore, the need for ethical approval was not applicable.

Informed consent statement: Informed consent was obtained from all subjects involved in the study. More details can be found at the following links: https://www.cdc.gov/nchs/nhanes/participant/participant-confidentiality.htm, https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/documents.aspx?BeginYear=2019, and https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/documents.aspx?BeginYear=2017.

Guarantor: Airong Wu and Jinzhou Zhu.

Funding: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by the Youth Program of Suzhou Health Committee (grant number KJXW2019001) and the Suzhou Clinical Center of Digestive Diseases (grant number Szlcyxzx202101).

ORCID iD: Lihe Liu https://orcid.org/0009-0007-5269-8438

Supplemental material: Supplemental material for this article is available online.

References

1. Powell EE, Wong VW and Rinella M. Non-alcoholic fatty liver disease. *Lancet* 2021; 397: 2212–2224.
2. Angulo P, Kleiner DE, Dam-Larsen S, et al. Liver fibrosis, but no other histologic features, is associated with long-term outcomes of patients with nonalcoholic fatty liver disease. *Gastroenterology* 2015; 149: 389–397.e10.
3. Alba LM and Lindor K. Non-alcoholic fatty liver disease. *Aliment Pharmacol Ther* 2003; 17: 977–986.
4. Younossi ZM. Non-alcoholic fatty liver disease - A global public health perspective. *J Hepatol* 2019; 70: 531–544.
5. Li Y, Wang X, Zhang J, et al. Applications of artificial intelligence (AI) in researches on non-alcoholic fatty liver disease(NAFLD): a systematic review. *Rev Endocr Metab Disord* 2022; 23: 387–400.
6. Castera L, Friedrich-Rust M and Loomba R. Noninvasive assessment of liver disease in patients with nonalcoholic fatty liver disease. *Gastroenterology* 2019; 156: 1264–1281.e4.
7. Younossi ZM, Koenig AB, Abdelatif D, et al. Global epidemiology of nonalcoholic fatty liver disease-meta-analytic assessment of prevalence, incidence, and outcomes. *Hepatology* 2016; 64: 73–84.
8. Eddowes PJ, Sasso M, Allison M, et al. Accuracy of FibroScan controlled attenuation parameter and liver stiffness measurement in assessing steatosis and fibrosis in patients with nonalcoholic fatty liver disease. *Gastroenterology* 2019; 156: 1717–1730.
9. Sasso M, Audiere S, Kemgang A, et al. Liver steatosis assessed by controlled attenuation parameter (CAP) measured with the XL probe of the FibroScan: a pilot study assessing diagnostic accuracy. *Ultrasound Med Biol* 2016; 42: 92–103.
10. Sasso M, Beaugrand M, de Ledinghen V, et al. Controlled attenuation parameter (CAP): a novel VCTE guided ultrasonic attenuation measurement for the evaluation of hepatic steatosis: preliminary study and validation in a cohort of patients with chronic liver disease from various causes. *Ultrasound Med Biol* 2010; 36: 1825–1835.
11. Lee JH, Kim D, Kim HJ, et al. Hepatic steatosis index: a simple screening tool reflecting nonalcoholic fatty liver disease. *Dig Liver Dis* 2010; 42: 503–508.
12. Cuthbertson DJ, Weickert MO, Lythgoe D, et al. External validation of the fatty liver index and lipid accumulation product indices, using 1H-magnetic resonance spectroscopy, to identify hepatic steatosis in healthy controls and obese, insulin-resistant individuals. *Eur J Endocrinol* 2014; 171: 561–569.
13. Bedogni G, Kahn HS, Bellentani S, et al. A simple index of lipid overaccumulation is a good marker of liver steatosis. *BMC Gastroenterol* 2010; 10: 98.
14. Koehler EM, Schouten JN, Hansen BE, et al. External validation of the fatty liver index for identifying nonalcoholic fatty liver disease in a population-based study. *Clin Gastroenterol Hepatol* 2013; 11: 1201–1204.
15. Ichino N, Osakabe K, Sugimoto K, et al. The NAFLD Index: a simple and accurate screening tool for the prediction of non-alcoholic fatty liver disease. *Rinsho Byori* 2015; 63: 32–43.
16. Fuyan S, Jing L, Wenjun C, et al. Fatty liver disease index: a simple screening tool to facilitate diagnosis of nonalcoholic fatty liver disease in the Chinese population. *Dig Dis Sci* 2013; 58: 3326–3334.
17. Park YJ, Lim JH, Kwon ER, et al. Development and validation of a simple index system to predict nonalcoholic fatty liver disease. *Korean J Hepatol* 2011; 17: 19–26.
18. Deo RC. Machine learning in medicine. *Circulation* 2015; 132: 1920–1930.
19. Handelman GS, Kok HK, Chandra RV, et al. Edoctor: machine learning and the future of medicine. *J Intern Med* 2018; 284: 603–619.
20. Ji W, Xue M, Zhang Y, et al. A machine learning based framework to identify and classify non-alcoholic fatty liver disease in a large-scale population. *Front Public Health* 2022; 10: 846118.
21. Corey KE, Kartoun U, Zheng H, et al. Development and validation of an algorithm to identify nonalcoholic fatty liver disease in the electronic medical record. *Dig Dis Sci* 2016; 61: 913–919.
22. Wu CC, Yeh WC, Hsu WD, et al. Prediction of fatty liver disease using machine learning algorithms. *Comput Methods Programs Biomed* 2019; 170: 23–29.
23. Qin S, Hou X, Wen Y, et al. Machine learning classifiers for screening nonalcoholic fatty liver disease in general adults. *Sci Rep* 2023; 13: 3638.
24. Ma H, Xu CF, 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.
25. Yip TC, Ma AJ, Wong VW, 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: 447–456.
26. Chen YS, Chen D, Shen C, et al. A novel model for predicting fatty liver disease by means of an artificial neural network. *Gastroenterol Rep (Oxf)* 2021; 9: 31–37.
27. Peng HY, Duan SJ, Pan L, et al. Development and validation of machine learning models for nonalcoholic fatty liver disease. *Hepatobiliary Pancreat Dis Int* 2023; 22: 615–621.
28. Noureddin M, Ntanios F, Malhotra D, et al. Predicting NAFLD prevalence in the United States using national health and nutrition examination survey 2017–2018 transient elastography data and application of machine learning. *Hepatol Commun* 2022; 6: 1537–1548.
29. Atsawarungruangkit A, Laoveeravat P and 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: 1417–1427.
30. Liu C, Zhou SH, Su H, et al. An artificial neural network model combined with dietary retinol intake from different sources to predict the risk of nonalcoholic fatty liver disease. *Biomed Environ Sci* 2023; 36: 1123–1135.
31. Zhang H, Ji J, Liu Z, et al. Artificial intelligence for the diagnosis of clinically significant prostate cancer based on multimodal data: a multicenter study. *BMC Med* 2023; 21: 270.

32. Wang Y, Hong Y, Wang Y, et al. Automated multimodal machine learning for esophageal variceal bleeding prediction based on endoscopy and structured data. *J Digit Imaging* 2023; 36: 326–338.
33. Liu L, Zhang R, Shi Y, et al. Automated machine learning for predicting liver metastasis in patients with gastrointestinal stromal tumor: a SEER-based analysis. *Sci Rep* 2024; 14: 12415.
34. Blazek K, van Zwieten A, Saglimbene V, et al. A practical guide to multiple imputation of missing data in nephrology. *Kidney Int* 2021; 99: 68–74.
35. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng* 2018; 2: 749–760.
36. Kim D, Cholankeril G, Loomba R, et al. Prevalence of fatty liver disease and fibrosis detected by transient elastography in adults in the United States, 2017-2018. *Clin Gastroenterol Hepatol* 2021; 19: 1499–1501.e2.
37. Riazi K, Swain MG, Congly SE, et al. Race and ethnicity in non-alcoholic fatty liver disease (NAFLD): a narrative review. *Nutrients* 2022; 14: 4556–4574.
38. Ayada I, van Kleef LA, Alferink LJM, et al. Systematically comparing epidemiological and clinical features of MAFLD and NAFLD by meta-analysis: focusing on the non-overlap groups. *Liver Int* 2022; 42: 277–287.
39. Targher G, Corey KE, Byrne CD, et al. The complex link between NAFLD and type 2 diabetes mellitus - mechanisms and treatments. *Nat Rev Gastroenterol Hepatol* 2021; 18: 599–612.
40. Targher G, Byrne CD and Tilg H. NAFLD and increased risk of cardiovascular disease: clinical associations, pathophysiological mechanisms and pharmacological implications. *Gut* 2020; 69: 1691–1705.
41. Polyzos SA, Kountouras J and Mantzoros CS. Obesity and nonalcoholic fatty liver disease: from pathophysiology to therapeutics. *Metabolism* 2019; 92: 82–97.
42. Pang Q, Zhang JY, Song SD, et al. Central obesity and nonalcoholic fatty liver disease risk after adjusting for body mass index. *World J Gastroenterol* 2015; 21: 1650–1662.
43. Tincopa MA and Loomba R. Non-invasive diagnosis and monitoring of non-alcoholic fatty liver disease and non-alcoholic steatohepatitis. *Lancet Gastroenterol Hepatol* 2023; 8: 660–670.
44. Janitza S, Strobl C and Boulesteix AL. An AUC-based permutation variable importance measure for random forests. *BMC Bioinformatics* 2013; 14: 119.
45. Yki-Jarvinen H. Non-alcoholic fatty liver disease as a cause and a consequence of metabolic syndrome. *Lancet Diabetes Endocrinol* 2014; 2: 901–910.
46. Mansour-Ghanaei R, Mansour-Ghanaei F, Naghipour M, et al. The role of anthropometric indices in the prediction of non-alcoholic fatty liver disease in the PERSIAN Guilan cohort study (PGCS). *J Med Life* 2018; 11: 194–202.
47. Khamseh ME, Malek M, Abbasi R, et al. Triglyceride glucose Index and related parameters (triglyceride glucose-body mass Index and triglyceride glucose-Waist circumference) identify nonalcoholic fatty liver and liver fibrosis in individuals with overweight/obesity. *Metab Syndr Relat Disord* 2021; 19: 167–173.
48. Sarwar R, Pierce N and Koppe S. Obesity and nonalcoholic fatty liver disease: current perspectives. *Diabetes Metab Syndr Obes* 2018; 11: 533–542.
49. Li L, Liu DW, Yan HY, et al. Obesity is an independent risk factor for non-alcoholic fatty liver disease: evidence from a meta-analysis of 21 cohort studies. *Obes Rev* 2016; 17: 510–519.
50. Fabbrini E, Sullivan S and Klein S. Obesity and nonalcoholic fatty liver disease: biochemical, metabolic, and clinical implications. *Hepatology* 2010; 51: 679–689.
51. Ramírez-Vélez R, Izquierdo M, Correa-Bautista JE, et al. Liver fat content and body fat distribution in youths with excess adiposity. *J Clin Med* 2018; 7: 528–539.
52. Ye Q, Zou B, Yeo YH, et al. Global prevalence, incidence, and outcomes of non-obese or lean non-alcoholic fatty liver disease: a systematic review and meta-analysis. *Lancet Gastroenterol Hepatol* 2020; 5: 739–752.
53. Wei JL, Leung JC, Loong TC, et al. Prevalence and severity of nonalcoholic fatty liver disease in non-obese patients: a population study using proton-magnetic resonance spectroscopy. *Am J Gastroenterol* 2015; 110: 1306–1314. quiz 1315.
54. Kühn T, Nonnenmacher T, Sookthai D, et al. Anthropometric and blood parameters for the prediction of NAFLD among overweight and obese adults. *BMC Gastroenterol* 2018; 18: 113.
55. Kotronen A, Peltonen M, Hakkarainen A, et al. Prediction of non-alcoholic fatty liver disease and liver fat using metabolic and genetic factors. *Gastroenterology* 2009; 137: 865–872.
56. Bril F, McPhaul MJ, Kalavalapalli S, et al. Intact fasting insulin identifies nonalcoholic fatty liver disease in patients without diabetes. *J Clin Endocrinol Metab* 2021; 106: e4360–e4371.
57. Guerrero-Romero F, Simental-Mendía LE, González-Ortiz M, et al. The product of triglycerides and glucose, a simple measure of insulin sensitivity. Comparison with the euglycemic-hyperinsulinemic clamp. *J Clin Endocrinol Metab* 2010; 95: 3347–3351.
58. Najjar SM and Perdomo G. Hepatic insulin clearance: mechanism and physiology. *Physiology (Bethesda)* 2019; 34: 198–215.
59. Santoleri D and Titchenell PM. Resolving the paradox of hepatic insulin resistance. *Cell Mol Gastroenterol Hepatol* 2019; 7: 447–456.
60. Tomizawa M, Kawanabe Y, Shinozaki F, et al. Triglyceride is strongly associated with nonalcoholic fatty liver disease among markers of hyperlipidemia and diabetes. *Biomed Rep* 2014; 2: 633–636.
61. Liu Z, Zhang Y, Graham S, et al. Causal relationships between NAFLD, T2D and obesity have implications for disease subphenotyping. *J Hepatol* 2020; 73: 263–276.
62. Zhang S, Du T, Zhang J, et al. The triglyceride and glucose index (TyG) is an effective biomarker to identify nonalcoholic fatty liver disease. *Lipids Health Dis* 2017; 16: 15.
63. Lee JH, Park K, Lee HS, et al. The usefulness of metabolic score for insulin resistance for the prediction of incident nonalcoholic fatty liver disease in Korean adults. *Clin Mol Hepatol* 2022; 28: 814–826.
64. Diagnosis and classification of diabetes mellitus. *Diabetes Care* 2011; 34: S62–S69.
65. Ma H, Xu C, Xu L, et al. Independent association of HbA1c and nonalcoholic fatty liver disease in an elderly Chinese population. *BMC Gastroenterol* 2013; 13: 3.
66. Bae JC, Cho YK, Lee WY, et al. Impact of nonalcoholic fatty liver disease on insulin resistance in relation to HbA1c levels in nondiabetic subjects. *Am J Gastroenterol* 2010; 105: 2389–2395.
67. Patel DA, Srinivasan SR, Xu JH, et al. Persistent elevation of liver function enzymes within the reference range is

associated with increased cardiovascular risk in young adults: the Bogalusa heart study. *Metabolism* 2007; 56: 792–798.
68. Xie J, Huang H, Liu Z, et al. The associations between modifiable risk factors and nonalcoholic fatty liver disease: a comprehensive Mendelian randomization study. *Hepatology* 2023; 77: 949–964.
69. Chen K, Pan Y, Xiang X, et al. The nonalcoholic fatty liver risk in prediction of unfavorable outcome after stroke: a nationwide registry analysis. *Comput Biol Med* 2023; 157: 106692.
70. Zhang Z, Wang S, Zhu Z, et al. Identification of potential feature genes in non-alcoholic fatty liver disease using bioinformatics analysis and machine learning strategies. *Comput Biol Med* 2023; 157: 106724.
71. Zhao H, Qiu X, Lu W, et al. High-quality retinal vessel segmentation using generative adversarial network with a large receptive field. *Int J Imaging Syst Technol* 2020; 30: 828–842.