

RESEARCH

Open Access



Development and validation of a machine learning-based framework for assessing metabolic-associated fatty liver disease risk

Jiale Deng^{1†}, Weidong Ji^{1†}, Hongze Liu^{1†}, Lin Li¹, Zhe Wang¹, Yurong Hu³, Yushan Wang^{2*} and Yi Zhou^{1*}

Abstract

Background The existing predictive models for metabolic-associated fatty liver disease (MAFLD) possess certain limitations that render them unsuitable for extensive population-wide screening. This study is founded upon population health examination data and employs a comparison of eight distinct machine learning (ML) algorithms to construct the optimal screening model for identifying high-risk individuals with MAFLD in China.

Methods We collected physical examination data from 5,171,392 adults residing in the northwestern region of China, during the year 2021. Feature selection was conducted through the utilization of the Least Absolute Shrinkage and Selection Operator (LASSO) regression. Additionally, class balancing parameters were incorporated into the models, accompanied by hyperparameter tuning, to effectively address the challenges posed by imbalanced datasets. This study encompassed the development of both tree-based ML models (including Classification and Regression Trees, Random Forest, Adaptive Boosting, Light Gradient Boosting Machine, Extreme Gradient Boosting, and Categorical Boosting) and alternative ML models (specifically, k-Nearest Neighbors and Artificial Neural Network) for the purpose of identifying individuals with MAFLD. Furthermore, we visualized the importance scores of each feature on the selected model.

Results The average age (standard deviation) of the 5,171,392 participants was 51.12 (15.00) years, with 52.47% of the participants being females. MAFLD was diagnosed by specialized physicians. 20 variables were finally included for analyses after LASSO regression model. Following ten rounds of cross-validation and parameter optimization for each algorithm, the CatBoost algorithm exhibited the best performance, achieving an Area Under the Receiver Operating Characteristic Curve (AUC) of 0.862. The ranking of feature importance indicates that age, BMI, triglyceride, fasting plasma glucose, waist circumference, occupation, high density lipoprotein cholesterol, low density lipoprotein cholesterol, total cholesterol, systolic blood pressure, diastolic blood pressure, ethnicity and cardiovascular diseases are the top 13 crucial factors for MAFLD screening.

[†]Jiale Deng, Weidong Ji and Hongze Liu contributed equally to this work.

*Correspondence:

Yushan Wang
34160869@qq.com

Yi Zhou
zhouyi@mail.sysu.edu.cn

Full list of author information is available at the end of the article



Conclusion This study utilized a large-scale, multi-ethnic physical examination data from the northwestern region of China to establish a more accurate and effective MAFLD risk screening model, offering a new perspective for the prediction and prevention of MAFLD.

Keywords Metabolic-associated fatty liver disease, Machine learning, Screening model, Prediction model

Introduction

Metabolic-associated fatty liver disease (MAFLD), previously recognized as nonalcoholic fatty liver disease (NAFLD), is intricately linked with metabolic perturbations and has emerged as the predominant chronic liver ailment on a global scale in the 21st century [1, 2]. MAFLD harbors the potential to engender a spectrum of grave afflictions, including liver hepatocellular carcinoma, cirrhosis, and fibrosis [3, 4], alongside maladies associated with metabolic syndrome, notably diabetes, hypertension, cardiovascular disease, chronic kidney disease, and systemic inflammation [5–8]. Timely screening and vigilant management of MAFLD hold utmost significance in averting and ameliorating the potential gravity of its complications and ensuing consequences.

During routine screening, accurately identifying MAFLD can prove challenging, given that patients might not manifest overt signs or symptoms during the initial phases. Moreover, traditional biochemical assessments fall short of directly quantifying hepatic steatosis and inflammatory conditions. Consequently, the susceptible population at risk of MAFLD could be considerable, warranting screening across a substantial cohort. Initial diagnosis and staging of MAFLD commonly entail a fusion of clinical background, laboratory analyses, and imaging assessments. Attaining more precise outcomes necessitates the utilization of liver tissue biopsy, which presently stands as the widely acknowledged benchmark, delivering intricate insights into hepatic histology, fibrosis progression, and inflammation intensity [9]. Nevertheless, liver biopsy involves invasiveness, accompanied by inherent risks and discomfort, rendering it potentially restricted within clinical application. In recent times, numerous researchers have harnessed non-invasive indicators and machine learning (ML) algorithms to discern MAFLD susceptibility, yielding promising outcomes [10–12]. Notwithstanding, these investigations are not without limitations. Certain predictive models demonstrate specificity toward particular populations [10], while others lean on intricate data collection such as omics data and biomarkers [11, 12]. Furthermore, a subset of studies grapples with diminutive sample sizes or disregards ethnic nuances [13, 14]. Consequently, the current MAFLD predictive models fall short in catering to the requisites of extensive population health screening.

Xinjiang, located in China, is an expansive and culturally diverse region. Owing to the distinct dietary preferences and genetic variances among its various ethnic

groups, the prevalence of overweight and obesity in this area is notably high [15–17], resulting in substantial accumulation of hepatic fat content. Moreover, the region's distinctive climatic extremities, desertification, air quality concerns, and relatively secluded geographic positioning profoundly influence the lifestyles and nutritional intake of its inhabitants [18], indirectly elevating the vulnerability to MAFLD. Hence, embarking upon MAFLD prediction research encompassing a wide-ranging population in Xinjiang is attuned to the region's ethnically diverse and substantial sample characteristics. This approach is instrumental in advancing our comprehension of the risk factors inherent to MAFLD within this geographical context. Given the background described above, this study aims to develop the optimal ML model for identifying MAFLD patients from a large-scale health examination population in Western China. Two categories of models, including tree-based models and other ML models, were constructed and compared for their predictive performance. Furthermore, the study analyzed the important predictive factors to facilitate large-scale MAFLD screening and gain comprehensive insights into MAFLD risk factors, providing novel research perspectives.

Methods

Study population

The data employed in this study originated from the China Xinjiang National Health Examination Program, carried out in the year 2021. To acquire a thorough grasp of the research framework and the criteria for participant selection adopted within this initiative, we direct interested readers to our preceding study [19]. This investigation collated health examination data from a substantial cohort, comprising a total of 9,382,225 individuals. The exclusion criteria encompassed: (i) Missing values for important variables related to the diagnosis of MAFLD, such as plasma triglycerides and high-density lipoprotein cholesterol ($n=3,752,890$). (ii) Age less than 18 years or older than 100 years ($n=444,569$). (iii) Participants with liver cirrhosis, liver tumors, and liver cancer ($n=13,374$). After implementing meticulous screening protocols, a cohort of 5,171,392 participants, drawn from 14 diverse regions, was deemed suitable for subsequent analysis (Fig. 1). For the data included in this study, we utilized the random forest algorithm for data imputation. The detailed demographic distribution across each of these regions is as follows: Hotan (735,022), Ili (723,546), Aksu (729,658), Changji (373,288), Tacheng (323,568), Bayingol

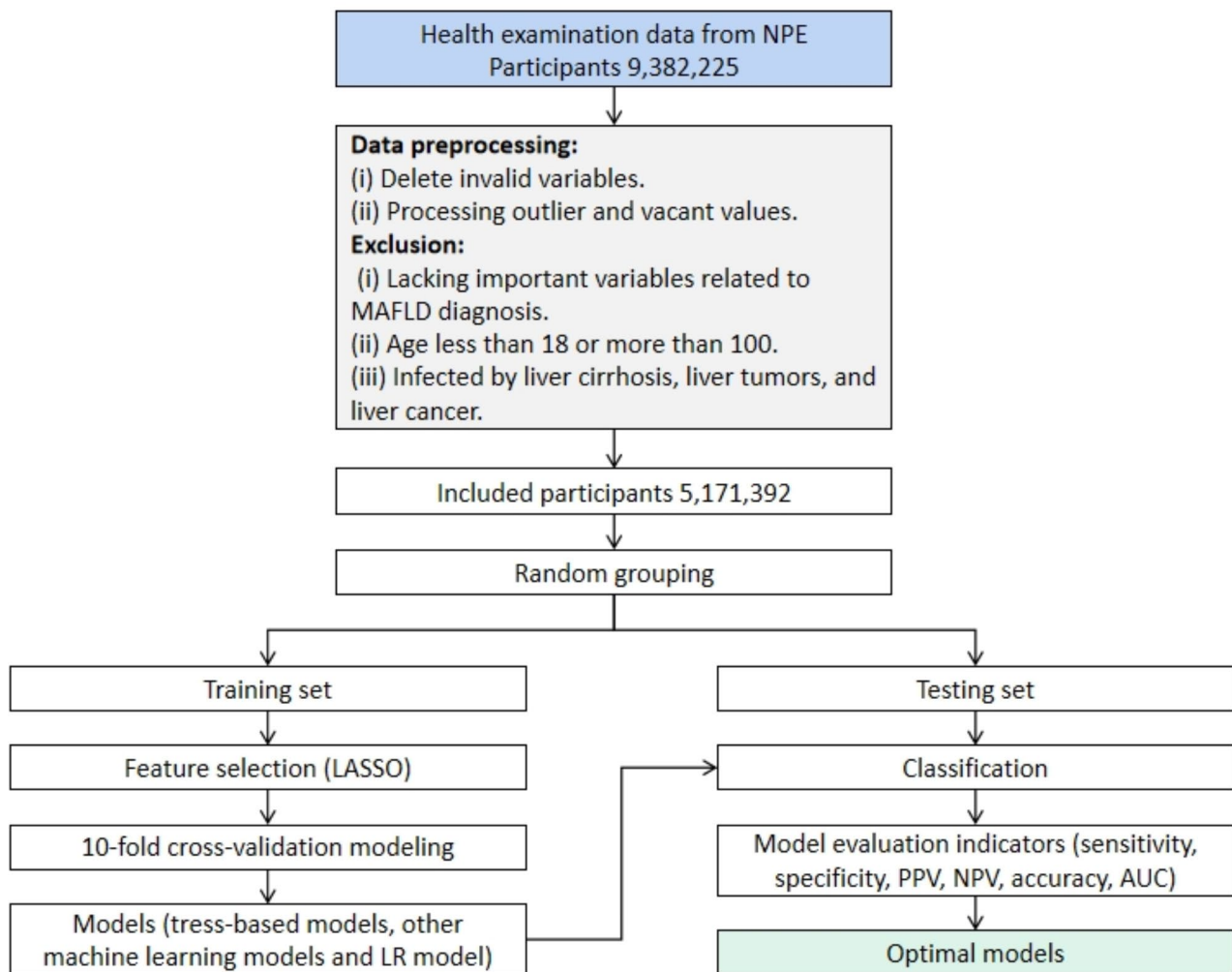


Fig. 1 Analysis process of this study. LASSO, least absolute shrinkage and selection operator; PPV, positive predictive value; NPV, negative predictive value; AUC, area under the receiver operating characteristic curve

Mongolian (319,160), Altay (182,453), Turpan (174,962), Bortala Mongolian (105,143), Hami (166,406), Kizilsu Kirgiz (136,787), Karamay (75,276), Kashgar (924,608), and Urumqi (201,515). The study was approved by the Ethics Committee and Institutional Review Committee of First Affiliated Hospital of Xinjiang Medical University(K202101-20).

MAFLD diagnosis

MAFLD was diagnosed using any of the following three criteria: excessive overweight or obesity, type II diabetes mellitus, or metabolic dysregulation, in addition to radiological imaging-confirmed hepatic steatosis, according to the current assessment criteria [20]. Hepatic steatosis was determined by a diagnostic abdominal ultrasound and a physician through a questionnaire asking participants about their disease history (e.g., whether they have ever been diagnosed by a doctor with fatty liver,

fatty accumulation, or degeneration of the liver). A body mass index (BMI) of $\geq 23 \text{ kg/m}^2$ is defined as overweight or obesity (Asian cut-off value). Type II diabetes mellitus was defined by self-reported medical diagnosis, a history of type II diabetes, or a fasting glucose value $\geq 7.0 \text{ mmol/L}$. Metabolic dysregulation was defined by meeting two or more of the following criteria: (1) waist circumference (WC) $\geq 90/80 \text{ cm}$ (Asian cut-off value) in men/women, (2) blood pressure $\geq 130/85 \text{ mmHg}$ or specific medication, (3) plasma triglycerides $\geq 1.70 \text{ mmol/L}$ or specific medication, (4) high-density lipoprotein cholesterol (HDL-C) $< 1.0 \text{ mmol/L}$ for males and $< 1.3 \text{ mmol/L}$ for females, and (5) pre-diabetes status (fasting blood glucose level from 5.6 to 6.9 mmol/L or HbA1c from 39 to 47%).

Predictors considered

The research data underwent a meticulous preprocessing procedure, including normalization and standardization for all data. When considering which variables to be included as predictors, we refer to relevant clinical studies of MAFLD [21–24] or effective factors that have been used for machine learning prediction [25]. Subsequently, we meticulously chose 20 pertinent attributes from questionnaire surveys and customary medical examination components for the purpose of shaping the predictive model. These attributes encompassed sex, age, ethnicity, education, occupation, marital history (MS), exercise frequency (EF), eating habits (EH), smoking status (SS), drinking frequency (DF), cardiovascular diseases (CVD), waist circumference (WC), body mass index (BMI), systolic blood pressure (SBP), diastolic blood pressure (DBP), fasting blood glucose (FBG); total cholesterol (TC), triglyceride (TG), low-density lipoprotein cholesterol (LDLC) and high-density lipoprotein cholesterol (HDLC). For a comprehensive elucidation of these attributes, we kindly direct readers to consult Table 1.

Data preprocessing

Data preprocessing was performed using the sklearn library in Python. Specifically, the sklearn.preprocessing

module was utilized to transform categorical data into numerical labels using LabelEncoder, encode ordinal variables using OrdinalEncoder, and create dummy variables for nominal variables using OneHotEncoder. Additionally, we utilized the MinMaxScaler function for normalization. The principle of this function is to determine the minimum and maximum values for each feature, and then scale all values in the feature so that the minimum value corresponds to 0 and the maximum value corresponds to 1, effectively normalizing the entire range of values to fall within the interval [0, 1].

Grouping and feature selection

The participants were randomly divided into two sets: a training set comprising 4,137,133 individuals and a testing set comprising 1,034,279 individuals, in an 8:2 ratio. Subsequently, we employed Least Absolute Shrinkage and Selection Operator (LASSO) regression for variable selection on the training set. We used the glmnet package in R to perform feature extraction using LASSO regression, specifically employing a binomial logistic regression model type. LASSO regression aims to optimize the coefficient estimates of the model by introducing L1 regularization, which in turn promotes sparse solutions,

Table 1 Information description of included variables

Variable	Variable type
Sex	Categorical variable ("male" or "female")
Age	Continuous variable
Ethnicity	Categorical variable ("Han", "Uyghur", "Kazak", "Hui", "Kirgiz", "Mongol", "Tajik", "other")
Education	Categorical variable ("elementary school or below", "middle or high school", "college or above")
Occupation	Categorical variable ("Occupation1": "leaders of government agencies, party organizations, enterprises, and institutions, administrative and clerical personnel or military personnel", "Occupation2": "technical professionals or production and transportation personnel", "Occupation3": "those involved in agriculture, forestry, animal husbandry, fishing, and water conservancy", "Occupation4": "commercial and service sector personnel or others whose category could not be determined", "Occupation5": "unemployed individuals")
MS	Categorical variable ("single", "married", "divorced", "widowed")
EF	Categorical variable ("not exercising", "occasionally", "more than once a week", "daily")
EH	Categorical variable ("meat and vegetable balance", "meat based", "vegetarian based")
SS	Categorical variable ("never smoked", "smoking", "quit smoking")
DF	Categorical variable ("never", "occasionally", "often", "every day")
CVD	Categorical variable ("yes" or "no")
WC	Continuous variable
BMI	Continuous variable
SBP	Continuous variable
DBP	Continuous variable
FPG	Continuous variable
TC	Continuous variable
TG	Continuous variable
LDLC	Continuous variable
HDLC	Continuous variable

Note: MS, marital status; EF, exercise frequency; EH, eating habits; SS, smoking status; DF, drinking frequency; CVD, cardiovascular diseases; WC, waist circumference; BMI, body mass index; SBP, systolic blood pressure; DBP, diastolic blood pressure; FBG, fasting blood glucose; TC, total cholesterol; TG, triglyceride; LDLC, low-density lipoprotein cholesterol; HDLC, high-density lipoprotein cholesterol

i.e., forcing many regression coefficients to shrink to zero, thus enabling feature selection.

Prediction models

This study devised and compared two classes of MAFLD screening models: tree-based ML models (including Classification and Regression Trees (CART), Random Forest, Adaptive Boosting (ADABOOST), Light Gradient Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost), and Categorical Boosting (CatBoost), as well as other ML models (namely, k-Nearest Neighbors (KNN) and Artificial Neural Network (ANN)). Furthermore, to achieve the objective of easily accessible and convenient screening, we solely employed data gathered from questionnaire surveys as predictive factors for the models.

CART, Random Forest, ADABOOST, LightGBM, XGBoost, and CatBoost are all tree-based ML algorithms. They are capable of handling both classification and regression problems by constructing decision trees or optimizing gradient boosting decision trees to improve prediction performance. CART is a tree-based classification and regression algorithm that builds a decision tree model by recursively partitioning the dataset [26]. Random Forest improves prediction accuracy by constructing multiple independent decision trees and aggregating their results [27]. ADABOOST builds a strong classifier by training a series of weak classifiers and combining them with weighted voting. It gradually improves overall classification performance by adjusting sample weights to focus on misclassified samples [28]. LightGBM is an efficient gradient boosting framework developed by Microsoft. It accelerates model training speed using a histogram-based decision tree algorithm and has lower memory usage [29]. XGBoost is a classic gradient boosting framework that enhances model accuracy and robustness by using second-order Taylor expansion to approximate the loss function and regularization terms [30]. CatBoost, developed by Yandex, is a gradient boosting framework with automatic handling capability for categorical features. It can directly utilize statistical information from categorical features [31].

The KNN algorithm is based on the fundamental idea of finding the nearest neighbors to a new input instance in the training set [32], and then using the majority vote of these K nearest neighbors for classification prediction. The ANN algorithm simulates the structure of neurons in the human brain and consists of an input layer, hidden layer(s), and an output layer [33]. The input layer receives data and converts it into a suitable format. The hidden layer(s) contains multiple neurons that are used to extract features and transform the input data. The output layer performs classification or prediction based on the output from the hidden layer.

Model evaluation

For the purpose of refining model performance, parameter adjustments were meticulously executed across each model, hinging on the learning curve, with the aim of unearthing the most optimal parameter configurations. Furthermore, an appraisal of model efficacy entailed the utilization of a confusion matrix, through which sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, and the area under the receiver operating characteristic curve (AUC) were systematically evaluated. In particular, we generated 95% confidence intervals for the AUC. Additionally, to achieve better performance, we applied a threshold adjustment to the model predictions on the training set to find the threshold that maximized the Youden index (the sum of sensitivity and specificity minus 1).

Feature importance evaluation

The process of evaluating the importance of features through machine learning models involves a systematic quantification of the impact of individual features on the predictive performance of the model [34]. This is achieved by assessing how much the model's performance is compromised when a specific feature is either permuted or excluded. Machine learning models, including Random Forest, Gradient Boosting, and Neural Networks, exhibit distinct advantages in this realm due to their capacity to grasp intricate relationships and interactions among features. These models offer insights into both linear and nonlinear correlations, thereby facilitating the identification of pivotal features that substantially contribute to the model's accuracy. The integration of machine learning-based assessments of feature importance not only enhances the model's interpretability but also aids in the selection of relevant features and provides guidance for domain-specific insights.

Statistical analysis

In this study, continuous variables were represented using mean (standard deviation), while categorical variables were presented using counts (percentages). For each variable, t-tests or Mann-Whitney tests were used for continuous variables, and chi-square tests or Fisher's exact tests were used for categorical variables. Two-tailed p-values less than 0.05 were considered statistically significant.

Results

Basic characteristics

This study included 5,171,392 participants, with a mean age (standard deviation) of 51.12 (15.00) years, of which approximately 52.47% were females (Table 2). The majority of participants were of Uighur ethnicity (54.73%), 43.54% had completed only primary education or lower,

Table 2 Baseline characteristics of participants in this study

Characteristics	Total N= 5,171,392	Non-MAFLD N= 4,646,808	MAFLD N= 524,584	P-value
Sex				< 0.001
male	2,457,777 (47.53%)	2,209,878 (47.56%)	247,899 (47.26%)	
female	2,713,615 (52.47%)	2,436,930 (52.44%)	276,685 (52.74%)	
Age	51.12 (15.00)	50.49 (15.29)	56.70 (10.59)	0.000
Ethnicity				0.000
Han	1,500,321 (29.01%)	1,289,429 (27.75%)	210,892 (40.20%)	
Uyghur	2,830,454 (54.73%)	2,607,099 (56.11%)	223,355 (42.58%)	
Kazak	451,386 (8.73%)	405,064 (8.72%)	46,322 (8.83%)	
Hui	217,779 (4.21%)	192,264 (4.14%)	25,515 (4.86%)	
Kirgiz	171,452 (3.32%)	152,952 (3.29%)	18,500 (3.53%)	
Education				0.000
Elementary school or below	2,251,377 (43.54%)	2,042,455 (43.95%)	208,922 (39.83%)	
Middle or high school	2,461,380 (47.60%)	2,206,676 (47.49%)	254,704 (48.55%)	
College or above	458,635 (8.87%)	397,677 (8.56%)	60,958 (11.62%)	
Occupation				0.000
Occupation1	325,629 (6.30%)	272,291 (5.86%)	53,338 (10.17%)	
Occupation2	313,443 (6.06%)	264,069 (5.68%)	49,374 (9.41%)	
Occupation3	3,389,941 (65.55%)	3,136,958 (67.51%)	252,983 (48.23%)	
Occupation4	855,347 (16.54%)	729,141 (15.69%)	126,206 (24.06%)	
Occupation5	287,032 (5.55%)	244,349 (5.26%)	42,683 (8.14%)	
MS				0.000
spinsterhood	502,280 (9.71%)	489,050 (10.52%)	13,230 (2.52%)	
married	4,331,240 (83.75%)	3,857,570 (83.02%)	473,670 (90.29%)	
divorced	96,800 (1.87%)	87,235 (1.88%)	9,565 (1.82%)	
widowed	241,072 (4.66%)	212,953 (4.58%)	28,119 (5.36%)	
EF				0.000
never	4,512,045 (87.25%)	4,108,865 (88.42%)	403,180 (76.86%)	
occasionally	204,521 (3.95%)	168,795 (3.63%)	35,726 (6.81%)	
more than once a week	109,839 (2.12%)	91,514 (1.97%)	18,325 (3.49%)	
everyday	344,987 (6.67%)	277,634 (5.97%)	67,353 (12.84%)	
EH				0.000
meat and vegetable balance	5,011,848 (96.91%)	4,507,626 (97.00%)	504,222 (96.12%)	
meat based	80,525 (1.56%)	68,217 (1.47%)	12,308 (2.35%)	
vegetarian based	79,019 (1.53%)	70,965 (1.53%)	8,054 (1.54%)	
SS				0.000
never	4,667,913 (90.26%)	4,202,173 (90.43%)	465,740 (88.78%)	
smoking	474,768 (9.18%)	420,615 (9.05%)	54,153 (10.32%)	
quit smoking	28,711 (0.56%)	24,020 (0.52%)	4,691 (0.89%)	
DF				0.000
never	4,827,021 (93.34%)	4,348,617 (93.58%)	478,404 (91.20%)	
occasionally	310,453 (6.00%)	270,561 (5.82%)	39,892 (7.60%)	
often	28,579 (0.55%)	23,285 (0.50%)	5,294 (1.01%)	
everyday	5,339 (0.10%)	4,345 (0.09%)	994 (0.19%)	
CVD:				0.000
No	4,613,430 (89.21%)	4,188,959 (90.15%)	424,471 (80.92%)	
Yes	557,962 (10.79%)	457,849 (9.85%)	100,113 (19.08%)	
WC	86.99 (12.37)	86.00 (11.89)	95.80 (13.06)	0.000
BMI	25.57 (4.21)	25.15 (3.99)	29.33 (4.21)	0.000
SBP	123.25 (17.83)	122.31 (17.53)	131.58 (18.38)	0.000
DBP	74.47 (11.02)	73.93 (10.84)	79.23 (11.47)	0.000
FPG	5.55 (1.60)	5.46 (1.49)	6.31 (2.24)	0.000
TC	4.44 (1.18)	4.40 (1.18)	4.86 (1.15)	0.000

Table 2 (continued)

Characteristics	Total N=5,171,392	Non-MAFLD N=4,646,808	MAFLD N=524,584	P-value
TG	1.49 (1.04)	1.43 (0.98)	2.00 (1.38)	0.000
LDLC	2.48 (1.01)	2.44 (1.00)	2.81 (1.01)	0.000
HDLC	1.39 (0.66)	1.40 (0.66)	1.33 (0.57)	0.000

Continuous variables were characterized as mean (standard deviation), while categorical variables were described in terms of frequency (percentage)

MS, marital status; EH, eating habits; Occupation 1: leaders of government agencies, party organizations, enterprises, and institutions, administrative and clerical personnel or military personnel; Occupation 2: technical professionals or production and transportation personnel; Occupation 3: those involved in agriculture, forestry, animal husbandry, fishing, and water conservancy; Occupation 4: commercial and service sector personnel or others whose category could not be determined; Occupation 5: unemployed individuals; EF, exercise frequency; DF, drinking frequency; SS, smoking status; CVD, cardiovascular diseases; WC, waist circumference; BMI, body mass index; SBP, systolic blood pressure; DBP, diastolic blood pressure; FBG, fasting blood glucose; TC, total cholesterol; TG, triglyceride; LDLC, low-density lipoprotein cholesterol; HDLC, high-density lipoprotein cholesterol.

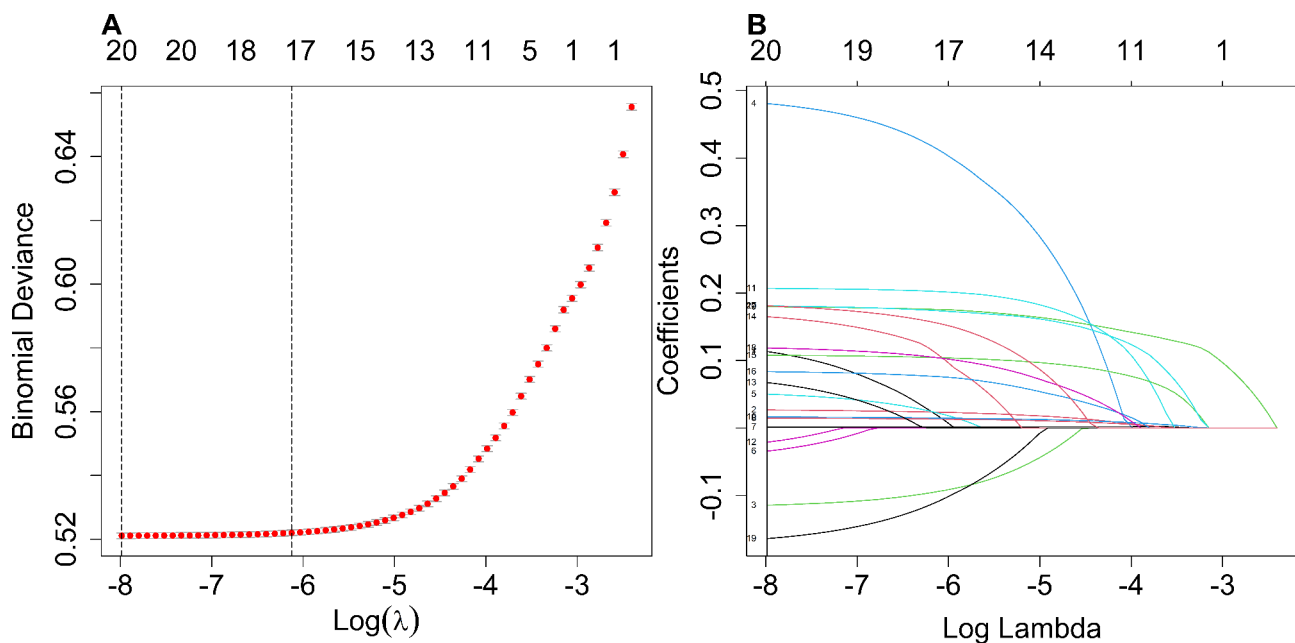


Fig. 2 Feature selection using LASSO regression in the training set. **(A)** Cross-validation was performed 10 times to select the optimal parameters (lambda) of the LASSO model. **(B)** LASSO coefficient profile of 20 characteristics. In the LASSO regression algorithm, as lambda is tuned, the shrinkage and variable selection process leads to a corresponding change in the trajectory of the coefficients of each characteristic related to MAFLD, which can be visualized in the LASSO coefficient profile. MAFLD, metabolic dysfunction-associated fatty liver disease; LASSO, least absolute shrinkage and selection operator

and 65.55% were engaged in agricultural, forestry, animal husbandry, fishing, and water management occupations.

Notably, the dataset revealed a conspicuous imbalance concerning the prevalence of MAFLD, with 524,584 individuals diagnosed with the condition against a significantly larger group of 4,646,808 without it, highlighting a ratio of roughly 1:9.

Statistical analysis revealed significant differences between participants with and without MAFLD in terms of age, sex, ethnicity, education, occupation, MS, EH, EF, DE, SS, CVD, WC, BMI, SBP, DBP, FBG, TC, TG, LDLC and HDLC.

Features extraction

In this study, we utilized LASSO regression to perform feature selection on the training dataset. As depicted

in Fig. 2, the outcomes of the feature selection process through LASSO regression unveiled that the model encompasses 20 non-zero coefficient variables. These variables include sex, age, ethnicity, education, occupation, MS, EF, EH, SS, DE, CVD, WC, BMI, SBP, DBP, FBG, TC, TG, LDLC and HDLC. These 20 variables were subsequently integrated as input features for the screening model developed within the framework of this research.

Tuning of parameters

In this study, a 10-fold cross-validation approach was implemented to fine-tune and optimize the parameters of the six tree-based models using the training dataset. The performance of these models, as indicated by their AUC values, was visualized across various parameter configurations.

More specifically, in the process of hyperparameter tuning, max_depth was evaluated from a range of 1 to 20, recording the AUC value for each configuration, and then selecting the max_depth corresponding to the highest AUC value for each model. The optimal hyperparameters for each model are presented in Table 3. The optimal values for max_depth were identified as 11 for CART, 16 for Random Forest, 5 for ADABOOST, 11 for LightGBM, 8 for XGBoost, and 6 for CatBoost, respectively. Default values are used for the remaining hyperparameters. As a result, we successfully trained and established six classification tree models that exhibited a noteworthy predicting performance.

Comparison of model performance

Tables 4 and 5 show the evaluation metrics corresponding to the performance of each model on the training and test data sets, respectively. Our observations reveal that the tree-based machine learning models outshine alternative machine learning methods in the context of conducting large-scale MAFLD screening within populations. Notably, the CatBoost algorithm emerges with remarkable prowess, attaining Sensitivity, Specificity, and AUC values of 0.814, 0.753 and 0.862, respectively. In contrast, the artificial neural network (ANN) displays relatively modest performance in this task. Figure 3

delineates the comprehensive receiver operating characteristic (ROC) curve on the training set and the test set, encapsulating all classifiers under scrutiny.

Importance of features

Within the confines of this study, we embarked on an evaluation and prioritization of feature importance specifically pertaining to the CatBoost model, which showcased the most superior performance and attained the highest AUC value. Based on the empirical findings delineated in Fig. 4, we discerned that BMI, age, TG, WC, FPG, occupation3 (pertaining to agriculture, forestry, animal husbandry, fishing, and water conservancy roles), HDLC, LDLC, TC, ethnicity (Uyghur), DBP, SBP and CVD emerged as the foremost 13 pivotal predictive factors. These factors were identified through the utilization of the CatBoost model to anticipate MAFLD within a sizeable population, grounded in the insights gleaned from questionnaire data.

Discussion

Given the swiftly increasing prevalence of MAFLD, the task of identifying prospective MAFLD patients and implementing suitable therapeutic interventions has become an exigent priority. Within the confines of this study, a comprehensive cohort of 5,171,392 adults aged

Table 3 The optimal hyperparameters of each algorithm

Algorithms	Hyperparameters
CART	{'ccp_alpha': 0.0, 'class_weight': {0: 0.5565471129755745, 1: 4.9210921980753906}, 'criterion': 'gini', 'max_depth': 11, 'max_features': None, 'max_leaf_nodes': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'random_state': 0, 'splitter': 'best'}
RF	{'bootstrap': True, 'ccp_alpha': 0.0, 'class_weight': {0: 0.5565471129755745, 1: 4.9210921980753906}, 'criterion': 'gini', 'max_depth': 15, 'max_features': 'auto', 'max_leaf_nodes': None, 'max_samples': None, 'min_impurity_decrease': 0.0, 'min_impurity_split': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 100, 'n_jobs': 30, 'oob_score': False, 'random_state': 0, 'verbose': 0, 'warm_start': False}
ADABOOST	{'algorithm': 'SAMME.R', 'base_estimator': 'deprecated', 'estimator__ccp_alpha': 0.0, 'estimator__class_weight': None, 'estimator__criterion': 'gini', 'estimator__max_depth': 5, 'estimator__max_features': None, 'estimator__max_leaf_nodes': None, 'estimator__min_impurity_decrease': 0.0, 'estimator__min_samples_leaf': 1, 'estimator__min_samples_split': 2, 'estimator__min_weight_fraction_leaf': 0.0, 'estimator__random_state': 0, 'estimator__splitter': 'best', 'estimator': DecisionTreeClassifier(max_depth=5, random_state=0), 'learning_rate': 1.0, 'n_estimators': 50, 'random_state': None}
LightGBM	{'boosting_type': 'gbdt', 'class_weight': None, 'colsample_bytree': 1.0, 'importance_type': 'split', 'learning_rate': 0.1, 'max_depth': 11, 'min_child_samples': 20, 'min_child_weight': 0.001, 'min_split_gain': 0.0, 'n_estimators': 100, 'n_jobs': 48, 'num_leaves': 31, 'objective': None, 'random_state': 0, 'reg_alpha': 0.0, 'reg_lambda': 0.0, 'subsample': 1.0, 'subsample_for_bin': 200000, 'subsample_freq': 0, 'verbosity': -1, 'scale_pos_weight': 9.842184396150781}
XGBoost	{'objective': 'binary:logistic', 'use_label_encoder': False, 'base_score': 0.5, 'booster': 'gbtree', 'colsample_bylevel': 1, 'colsample_bynode': 1, 'colsample_bytree': 1, 'enable_categorical': False, 'gamma': 0, 'gpu_id': -1, 'importance_type': None, 'interaction_constraints': '', 'learning_rate': 0.300000012, 'max_delta_step': 0, 'max_depth': 8, 'min_child_weight': 1, 'missing': nan, 'monotone_constraints': '()', 'n_estimators': 100, 'n_jobs': 60, 'num_parallel_tree': 1, 'predictor': 'auto', 'random_state': 0, 'reg_alpha': 0, 'reg_lambda': 1, 'scale_pos_weight': 9.842184396150781, 'subsample': 1, 'tree_method': 'exact', 'validate_parameters': 1, 'verbosity': None, 'eval_metric': 'error'}
CatBoost	{'loss_function': 'Logloss', 'eval_metric': 'AUC', 'iterations': 500, 'learning_rate': 0.03, 'random_seed': 0, 'l2_leaf_reg': 0, 'random_state': 0, 'depth': 6, 'scale_pos_weight': 9.842184396150781, 'boosting_type': 'Ordered'}
KNN	{'algorithm': 'auto', 'leaf_size': 30, 'metric': 'minkowski', 'metric_params': None, 'n_jobs': None, 'n_neighbors': 100, 'p': 2, 'weights': 'uniform'}
ANN	{'activation': 'relu', 'alpha': 0.0001, 'batch_size': 'auto', 'beta_1': 0.9, 'beta_2': 0.999, 'early_stopping': False, 'epsilon': 1e-08, 'hidden_layer_sizes': (20), 'learning_rate': 'constant', 'learning_rate_init': 0.001, 'max_fun': 15000, 'max_iter': 200, 'momentum': 0.9, 'n_iter_no_change': 10, 'nesterovs_momentum': True, 'power_t': 0.5, 'random_state': 0, 'shuffle': True, 'solver': 'adam', 'tol': 0.0001, 'validation_fraction': 0.1, 'verbose': False, 'warm_start': False}

Table 4 Performance of each algorithm in the training set

Algorithms	Sensitivity	Specificity	PPV	NPV	Accuracy	AUC (95% CI)	Threshold
CART	0.833	0.721	0.253	0.974	0.733	0.855 (0.8542–0.8553)	0.501
RF	0.845	0.758	0.283	0.977	0.767	0.881 (0.8803–0.8813)	0.498
ADABOOST	0.831	0.736	0.262	0.975	0.746	0.862 (0.8612–0.8622)	0.489
LightGBM	0.828	0.743	0.267	0.974	0.752	0.864 (0.8631–0.8641)	0.538
XGBoost	0.851	0.747	0.276	0.978	0.758	0.878 (0.8777–0.8786)	0.517
CatBoost	0.842	0.757	0.282	0.977	0.766	0.878 (0.8771–0.8780)	0.530
KNN	0.833	0.723	0.254	0.975	0.734	0.857 (0.8565–0.8575)	0.110
ANN	0.829	0.729	0.257	0.974	0.739	0.857 (0.8568–0.8578)	0.095

Note: PPV, positive predictive value; NPV, negative predictive value; AUC, area under the receiver operating characteristic curve; KNN, K-Nearest Neighbor; ANN, Artificial Neural Network; CART, Classification and Regression Tree; RF, Random Forest; ADABOOST, Adaptive Boosting; LightGBM, Light Gradient Boosting Machine; XGBoost, Extreme Gradient Boosting; CatBoost, Categorical Boosting

Table 5 Performance of each algorithm in the test set

Algorithms	Sensitivity	Specificity	PPV	NPV	Accuracy	AUC (95% CI)	Threshold (determined by the training set)
CART	0.824	0.713	0.243	0.973	0.724	0.845 (0.8438–0.8461)	0.501
RF	0.806	0.749	0.265	0.972	0.755	0.856 (0.8553–0.8574)	0.498
ADABOOST	0.823	0.731	0.256	0.974	0.741	0.855 (0.8535–0.8557)	0.489
LightGBM	0.823	0.740	0.262	0.974	0.748	0.859 (0.8583–0.8603)	0.538
XGBoost	0.823	0.742	0.264	0.974	0.750	0.861 (0.8600–0.8621)	0.517
CatBoost	0.814	0.753	0.270	0.973	0.759	0.862 (0.8612–0.8632)	0.530
KNN	0.815	0.715	0.243	0.972	0.725	0.841 (0.8397–0.8419)	0.110
ANN	0.838	0.716	0.249	0.975	0.728	0.856 (0.8547–0.8568)	0.095

Note: PPV, positive predictive value; NPV, negative predictive value; AUC, area under the receiver operating characteristic curve; KNN, K-Nearest Neighbor; ANN, Artificial Neural Network; CART, Classification and Regression Tree; RF, Random Forest; ADABOOST, Adaptive Boosting; LightGBM, Light Gradient Boosting Machine; XGBoost, Extreme Gradient Boosting; CatBoost, Categorical Boosting

18 and above was enlisted. Leveraging their physical examination data, we endeavored to develop and juxtapose AI algorithms intended for the large-scale population screening of MAFLD. Our comprehensive inquiry sheds light on the exemplary performance exhibited by the CatBoost algorithm within the domain of MAFLD screening. Remarkably, BMI, age, TG, WC, FPG, occupation3, HDLC, LDLC, TC, ethnicity (Uyghur), DBP, SBP and CVD have emerged as the pivotal predictive factors of significance.

In contrast to conventional statistical models, ML models offer substantially enhanced data analysis and predictive capacities in disease prognostication. These models

possess the capability to manage extensive, high-dimensional, and intricate medical datasets, effectively discerning latent patterns and predictive principles to elevate the precision of predictions. Presently, an array of investigations has been undertaken to predict MAFLD employing laboratory indicators. Several scholars have integrated lipidomics, metabolomics, genomics, transcriptomics, and biomarkers as predictive variables for the formulation of models [35–37]. While these models have exhibited commendable outcomes in predicting MAFLD, acquiring such data through extensive health screenings proves impractical and hampers endeavors aimed at the broad-scale screening of diseases. Additionally, certain

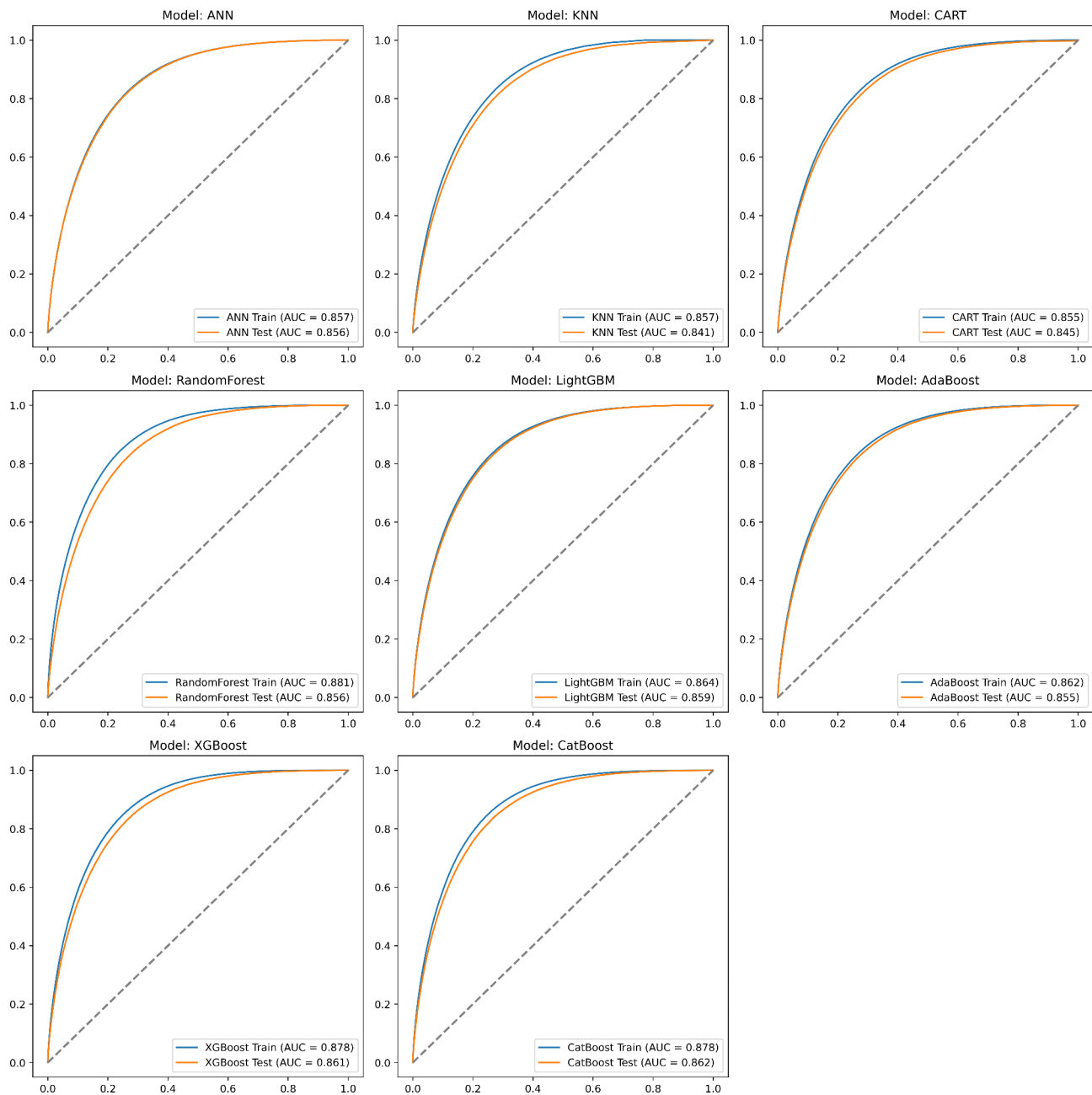


Fig. 3 ROC curves on the training set and the test set for KNN, ANN, CART, RF, ADABOOST, LightGBM, XGBoost and CatBoost respectively. ROC, receiver operating characteristic; KNN, K-Nearest Neighbor; ANN, Artificial Neural Network; CART, Classification and Regression Tree; RF, Random Forest; ADABOOST, Adaptive Boosting; LightGBM, Light Gradient Boosting Machine; XGBoost, Extreme Gradient Boosting; CatBoost, Categorical Boosting

investigations center on the interplay between distinct ailments and MAFLD, encompassing cardiovascular diseases, diabetes, and liver fibrosis, while others pertain solely to particular demographics, such as adolescents or individuals with obesity [35, 38–40]. Some studies suffer from the limitation of small sample sizes [35, 38, 39], or do not consider ethnic-specific factors [13, 14]. Studies constrained by these limitations might encounter challenges in extrapolating research conclusions and prediction models to the broader populace. In stark contrast,

our investigation encompasses a dataset of 5,171,392 participants hailing from the Xinjiang region, distinguished by its expansive sample size and ethnically diverse population. Consequently, the implications of our study hold promise for advancing MAFLD screening and prognostication within a sizable Chinese demographic.

The aforementioned studies on MAFLD prediction models have all achieved notably high AUCs, with some models reaching above 0.8 [13, 35, 36, 40] and others exceeding 0.9 [35] in their test sets, demonstrating good

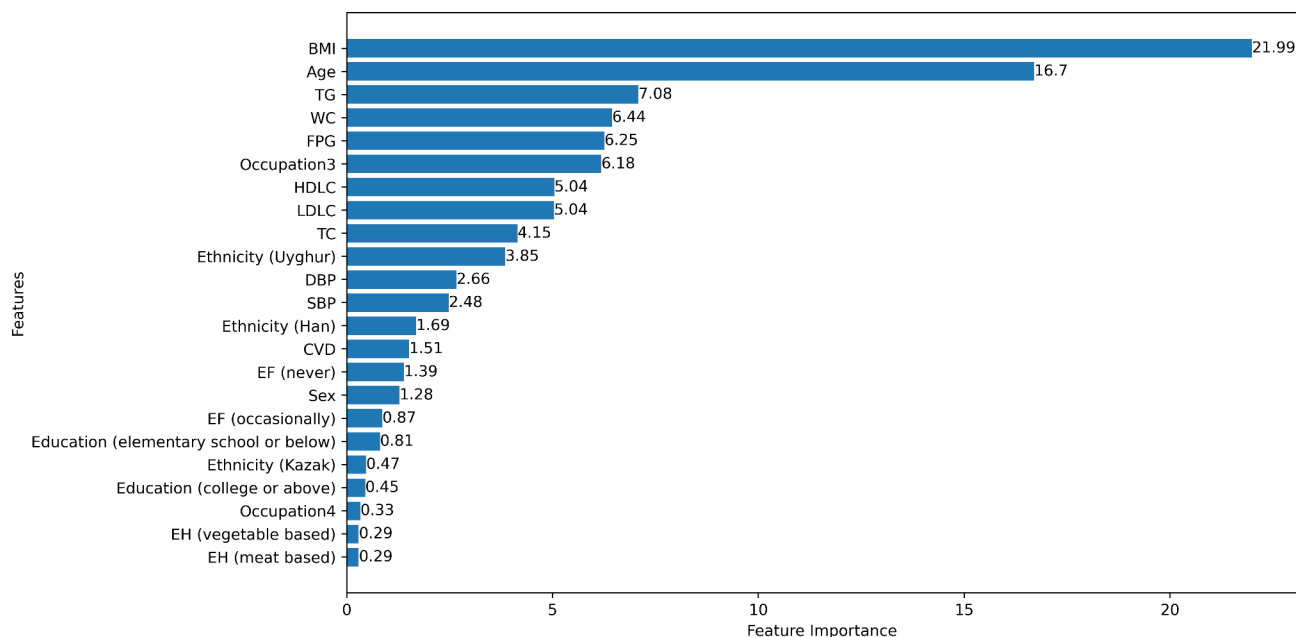


Fig. 4 Feature importance of CatBoost algorithm. Ethnicity (Hui), Ethnicity (Kirgiz), Ethnicity (Kazak), Ethnicity (Uyghur) and Ethnicity (Han) are dummy variables of Ethnicity. MS, marital status; EF, exercise frequency; EH, eating habits; SS, smoking status; DF, drinking frequency; CVD, cardiovascular diseases; WC, waist circumference; BMI, body mass index; SBP, systolic blood pressure; DBP, diastolic blood pressure; FBG, fasting blood glucose; TC, total cholesterol; TG, triglyceride; LDLC, low-density lipoprotein cholesterol; HDLC, high-density lipoprotein cholesterol

predictive performance under their respective study conditions. All models constructed and optimized in this study achieved AUCs above 0.8 in the test set, among which the best-performing CatBoost model reached an AUC of 0.862. This, to a certain extent, suggests that the variables selected in this study hold predictive value under the context of large-scale population disease screening, and the constructed models exhibit commendable performance in MAFLD screening.

In this study, we have identified that BMI, age, TG, FPG, WC, occupation, HDLC, LDLC, TC, SBP, DBP and CVD are pivotal factors for MAFLD screening. Obesity has been widely confirmed to be highly associated with MAFLD; therefore, BMI and WC are also important predictive factors for MAFLD. Age is highly correlated with MAFLD, with a higher age being associated with an increased risk of MAFLD, a factor that has been considered in many MAFLD prediction studies [41]. TG, HDLC, LDLC, and TC are components of lipid profiles in blood tests, and these parameters have been confirmed to be highly correlated with MAFLD. This is because elevated lipid levels are prone to causing the accumulation of fat in the liver [42]. Studies have shown a significant association between fasting blood glucose levels and the incidence and severity of MAFLD. Elevated fasting blood glucose levels serve as markers of diabetes and insulin resistance, both of which are notably linked to MAFLD. Insulin resistance leads to inadequate utilization of insulin, promoting fat accumulation in the liver

and facilitating the development of MAFLD. Moreover, elevated fasting blood glucose itself may directly harm the liver, causing inflammation and fibrosis in liver cells, further exacerbating the condition of MAFLD [43]. Individuals immersed in activities such as agriculture, forestry, animal husbandry, fishing, and water conservancy are predominantly engaged in physically demanding labor. Prolonged exposure to physical labor or regular exercise exerts a favorable influence in averting the onset of MAFLD, a notion substantiated by an array of scholarly investigation [44]. Research indicates that elevated systolic and diastolic blood pressure are associated with an increased risk of MAFLD. Epidemiological investigations reveal that the prevalence of MAFLD among hypertensive patients is approximately 49.5%, significantly higher than that in the general population [42]. Furthermore, MAFLD appears to be closely linked to hypertension and endothelial dysfunction, seemingly serving as an independent risk factor for prehypertension and hypertension [45]. Our research findings highlight a robust correlation between MAFLD and cardiovascular disease. On one hand, cardiovascular disease risk factors (including hypertension, hyperlipidemia, and diabetes) can culminate in anomalous hepatic fat buildup, giving rise to MAFLD. Moreover, inflammatory and vascular injury elements prompted by cardiovascular disease can permeate the systemic circulation, fostering the progression of MAFLD. On the other hand, individuals with MAFLD frequently manifest obesity and metabolic irregularities,

such as insulin resistance and elevated cholesterol levels. These factors compromise endothelial function and catalyze the advancement of atherosclerosis, thereby exacerbating the vulnerability to cardiovascular disease [38]. Ethnicity is also an important predictor. Ethnic group may be associated with regional, dietary habits, climate, genes and other factors. Previous studies have reported that there are differences in the incidence of MAFLD among different ethnic groups or regions [46].

This study boasts several notable strengths. Foremost, the MAFLD prediction model devised herein leverages variables garnered from physical examination data. Consequently, when juxtaposed with conventional MAFLD diagnostic methodologies, this model emerges as swifter, cost-efficient, and conducive to preliminary MAFLD screening within extensive populations. Additionally, our investigation encompasses a substantial and diverse Chinese demographic, encompassing an array of ethnic backgrounds. This inclusivity adeptly captures the influence of ethnicity-specific elements on the disease, markedly bolstering the applicability of our model. Furthermore, the meticulous sample selection adhering to scientifically grounded inclusion and exclusion criteria characterizes this study. This approach adeptly retains comprehensive data hailing from extensive health screening questionnaires, thus seamlessly aligning with the tenets of epidemiological research on MAFLD in real-world scenarios.

However, this study also has certain limitations. Firstly, our model is based on health check questionnaire data derived from a large-scale Chinese population, excluding data from other countries. The peculiarities of Xinjiang's ethnic structure and geographical environment may impact the generalizability of this model to other populations, despite it being the first model established based on a multi-ethnic population comprising millions of samples. Secondly, this study is cross-sectional in design, which restricts our ability to ascertain causal relationships between certain factors and MAFLD, as exemplified by the relationship between exercise frequency and MAFLD. Follow-up cohort studies are needed to address this limitation. Thirdly, the quantification of alcohol consumption did not strictly adhere to the exclusion criteria for MAFLD (males > 30 g/day, females > 20 g/day), which could potentially impact the predictive ability of our model. The self-reported component of predictive factors may introduce bias due to inaccuracies or incomplete reporting. Furthermore, the exclusion of a large number of participants with indeterminate disease outcomes may increase potential selection bias.

Conclusions

The severity and prevalence of MAFLD have garnered heightened recognition from the public, propelling the dire necessity for the formulation of a large-scale, population-oriented early screening model. Grounded in a multi-ethnic and expansive sample populace, this study exclusively harnessed questionnaire surveys and customary medical examination components to meticulously establish and juxtapose tree-based MAFLD predictive models against alternative ML methodologies. We identified the optimal MAFLD predictive model and extensively analyzed the interactions between various risk factors and MAFLD. The study results demonstrated that our MAFLD screening model achieved satisfactory predictive performance, providing a new and more economical and efficient approach for the prevention and screening of MAFLD.

Acknowledgements

Thanks to the health commission of Xinjiang Uygur Autonomous Region and the health management institute of Xinjiang medical university for data support. Thanks to all the participants for their help.

Author contributions

ZY, WYS conceived of the study and designed the study. DJL, JWD and LHZ did the main analyses. LL, WZ and HYR contributed to refining the ideas, carrying out additional analyses. All authors discussed the results and revised the manuscript. All authors read and approved the final manuscript.

Funding

This work is supported by the Key Research and Development Program of China (2022YFC3601600, 2021YFC2009400) and the National Natural Science Foundation of China (NSFC) (61876194), the Province Natural Science Foundation of Guangdong (2021A1515011897), the Science and Technology Innovation Special Project of Guangdong Province, China (202011020004) and the Province Natural Science Foundation of Xinjiang, China (2016D01C330). The corresponding author had final responsibility for the decision to submit the manuscript for publication.

Data availability

The data underlying this article will be shared on reasonable request to the corresponding author.

Declarations

Ethical approval and consent to participate

The study was conducted in accordance with the principles outlined in the Declaration of Helsinki and was approved by the Medical Ethics Committee of the First Affiliated Hospital of Xinjiang Medical University, and the ethical review approval number is K202101-20. Informed consent was obtained from all subjects.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹Zhongshan School of Medicine, Sun Yat-sen University, 74 Zhongshan 2nd Road, Yuexiu District, Guangzhou 510080, Guangdong, China

²People's Hospital of Xinjiang Uygur Autonomous Region, 91 Tianchi Road, Urumqi 830054, Xinjiang, China

³School of Computer Science, China University of Geosciences, Wuhan, Beihe 430074, China

Received: 26 October 2023 / Accepted: 26 August 2024

Published online: 18 September 2024

References

- Eslam M, Sanyal AJ, George J. MAFLD: A Consensus-Driven proposed nomenclature for metabolic Associated fatty liver disease. *Gastroenterology*. 2020;158(7):1999–2014.
- Powell EE, Wong VW, Rinella M. Non-alcoholic fatty liver disease. *Lancet*. 2021;397(10290):2212–24.
- Vitale A, Svegliati-Baroni G, Ortolani A, Cucco M, Dalla RG, Giannini EG, Piscaglia F, Rapaccini G, Di Marco M, Caturelli E, et al. Epidemiological trends and trajectories of MAFLD-associated hepatocellular carcinoma 2002–2033: the ITA.LI.CA database. *Gut*. 2023;72(1):141–52.
- van Kleef LA, Ayada I, Alferink L, Pan Q, de Knegt RJ. Metabolic dysfunction-associated fatty liver disease improves detection of high liver stiffness: the Rotterdam Study. *Hepatology*. 2022;75(2):419–29.
- Wang TY, Wang RF, Bu ZY, Targher G, Byrne CD, Sun DQ, Zheng MH. Association of metabolic dysfunction-associated fatty liver disease with kidney disease. *NAT REV NEPHROL*. 2022;18(4):259–68.
- Lee H, Lee YH, Kim SU, Kim HC. Metabolic dysfunction-Associated fatty liver Disease and Incident Cardiovascular Disease Risk: a Nationwide Cohort Study. *Clin Gastroenterol Hepatol*. 2021;19(10):2138–47.
- Uehara T, Wakui H, Tamura K. Metabolic dysfunction-associated fatty liver disease reflects a significantly higher risk of hypertension than non-alcoholic fatty liver disease. *HYPERTENS RES*. 2023;46(5):1165–7.
- Lim S, Kim JW, Targher G. Links between metabolic syndrome and metabolic dysfunction-associated fatty liver disease. *Trends Endocrinol Metab*. 2021;32(7):500–14.
- Hashimoto E, Taniai M, Tokushige K. Characteristics and diagnosis of NAFLD/NASH. *J Gastroenterol Hepatol*. 2013;28(Suppl 4):64–70.
- Zhang D, Zhang L, Chen S, Chen R, Zhang X, Bai F. Prevalence and risk factors of metabolic-Associated fatty liver Disease among Hospital Staff. *Diabetes Metab Syndr Obes*. 2023;16:1221–34.
- Lischka J, Lieb K, Schanzer A, Hojreh A, Ba-Ssalamah A, de Gier C, Walleczek NK, Zeyda M, Greber-Platzer S. Risk scores for metabolic dysfunction-Associated fatty liver disease in Pediatric obesity. *ANN NUTR METAB*. 2022;78(5):247–54.
- Xu J, Liu X, Wu S, Zhang D, Liu X, Xia P, Ling J, Zheng K, Xu M, Shen Y, et al. RNA-binding proteins in metabolic-associated fatty liver disease (MAFLD): from mechanism to therapy. *BIOSCI TRENDS*. 2023;17(1):21–37.
- Zhan H, Chen S, Gao F, Wang G, Chen SD, Xi G, Yuan HY, Li X, Liu WY, Byrne CD et al. AutoFibroNet: a deep learning and multi-photon microscopy-derived automated network for liver fibrosis quantification in MAFLD. *Aliment Pharmacol Ther* 2023.
- Perakakis N, Polyzos SA, Yazdani A, Sala-Vila A, Kountouras J, Anastasilakis AD, Mantzoros CS. Non-invasive diagnosis of non-alcoholic steatohepatitis and fibrosis with the use of omics and supervised learning: a proof of concept study. *METABOLISM*. 2019;101:154005.
- He J, Guo H, Ding YS, Liu JM, Zhang M, Ma RL, Zhang JY, Li SG, Sun F, Xu SZ, et al. [Epidemiological study on overweight and obesity among rural adult residents in Hazakh, Uyghur and Han populations in Xinjiang]. *Zhonghua Liu Xing Bing Xue Za Zhi*. 2013;34(12):1164–8.
- He J, Guo S, Liu J, Zhang M, Ding Y, Zhang J, Li S, Xu S, Niu Q, Guo H, et al. Ethnic differences in prevalence of general obesity and abdominal obesity among low-income rural Kazakh and Uyghur adults in far western China and implications in preventive public health. *PLoS ONE*. 2014;9(9):e106723.
- Fan JG, Kim SU, Wong VW. New trends on obesity and NAFLD in Asia. *J HEPA-TOL*. 2017;67(4):862–73.
- Zhai F, He Y, Wang Z, Hu Y. [Status and characteristic of dietary intake of 12 minority nationalities in China]. *Wei Sheng Yan Jiu*. 2007;36(5):539–41.
- Ji W, Zhang Y, Cheng Y, Wang Y, Zhou Y. Development and validation of prediction models for hypertension risks: a cross-sectional study based on 4,287,407 participants. *Front Cardiovasc Med*. 2022;9:928948.
- Eslam M, Newsome PN, Sarin SK, Anstee QM, Targher G, Romero-Gomez M, Zelber-Sagi S, Wai-Sun WW, Dufour JF, Schattenberg JM, et al. A new definition for metabolic dysfunction-associated fatty liver disease: an international expert consensus statement. *J HEPATOL*. 2020;73(1):202–9.
- Vilar-Gomez E, Nephew LD, Vuppalanchi R, Gawrieh S, Mladenovic A, Pike F, Samala N, Chalasani N. High-quality diet, physical activity, and college education are associated with low risk of NAFLD among the US population. *Hepatology*. 2022;75(6):1491–506.
- Yuan S, Chen J, Li X, Fan R, Arsenault B, Gill D, Giovannucci EL, Zheng JS, Larsson SC. Lifestyle and metabolic factors for nonalcoholic fatty liver disease: mendelian randomization study. *EUR J EPIDEMIOL*. 2022;37(7):723–33.
- Byrne CD, Targher G. NAFLD: a multisystem disease. *J HEPATOL*. 2015;62(1 Suppl):S47–64.
- Li H, Yu XH, Ou X, Ouyang XP, Tang CK. Hepatic cholesterol transport and its role in non-alcoholic fatty liver disease and atherosclerosis. *PROG LIPID RES*. 2021;83:101109.
- Xue Y, Xu J, Li M, Gao Y. Potential screening indicators for early diagnosis of NAFLD/MAFLD and liver fibrosis: triglyceride glucose index-related parameters. *Front Endocrinol (Lausanne)*. 2022;13:951689.
- Loh WY. Classification and regression trees. *WIREs Data Min Knowl Discov*. 2011;1(1):14–23.
- Breiman L. Random forests. *MACH LEARN*. 2001;45(1):5–32.
- Feng D, Liu Z, Wang X, Chen Y, Chang J, Wei D, Jiang Z. Machine learning-based compressive strength prediction for concrete: an adaptive boosting approach. *CONSTR BUILD MATER*. 2020;230:117000.
- Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T. LightGBM: a highly efficient gradient boosting decision Tree. *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 30 (NIPS 2017)* 2017, 30.
- Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: 2016/1/1 2016; Ithaca. ACM; 2016:785–794.
- Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 31(NIPS)* (2018) 2018, 31.
- Chen L, Yap KH. A fuzzy K-nearest-neighbor algorithm to blind image deconvolution. 2003 IEEE INTERNATIONAL CONFERENCE ON SYSTEMS, MAN AND CYBERNETICS, VOLS 1–5, CONFERENCE PROCEEDINGS 2003:2049–2054.
- Jain AK, Mao J, Mohiuddin KM. Artificial neural networks: a tutorial. *Comput (Long Beach Calif)*. 1996;29(3):31–44.
- Huynh-Thu VA, Saeys Y, Wehenkel L, Geurts P. Statistical interpretation of machine learning-based feature importance scores for biomarker discovery. *Bioinformatics*. 2012;28(13):1766–74.
- Jimenez-Rivera C, Hadjiyannakis S, Davila J, Hurteau J, Aglipay M, Barrowman N, Adamo KB. Prevalence and risk factors for non-alcoholic fatty liver in children and youth with obesity. *BMC PEDIATR*. 2017;17(1):113.
- Oh TG, Kim SM, Caussy C, Fu T, Guo J, Bassirian S, Singh S, Madamba EV, Bettencourt R, Richards L, et al. A Universal gut-microbiome-derived signature predicts cirrhosis. *CELL METAB*. 2020;32(5):878–88.
- Han N, He J, Shi L, Zhang M, Zheng J, Fan Y. Identification of biomarkers in nonalcoholic fatty liver disease: a machine learning method and experimental study. *FRONT GENET*. 2022;13:1020899.
- Ciardullo S, Cannistraci R, Mazzetti S, Mortara A, Perseghin G. Nonalcoholic fatty liver Disease, Liver Fibrosis and Cardiovascular Disease in the adult US Population. *FRONT ENDOCRINOL*. 2021;12:71484.
- Riekkki H, Aitokari L, Kivelä L, Lahti S, Hiltunen P, Vuorela N, Huhtala H, Lakka TA, Kurppa K. Prevalence and associated factors of metabolic-associated fatty liver disease in overweight Finnish children and adolescents. *FRONT ENDOCRINOL*. 2023;14:1090344.
- Oses M, Cadenas-Sanchez C, Medrano M, Galbete A, Miranda-Ferrua E, Ruiz JR, Sanchez-Valverde F, Ortega FB, Cabeza R, Villanueva A, et al. Development of a prediction protocol for the screening of metabolic associated fatty liver disease in children with overweight or obesity. *PEDIATR OBES*. 2022;17(9):e12917.
- Peng H, Pan L, Ran S, Wang M, Huang S, Zhao M, Cao Z, Yao Z, Xu L, Yang Q, et al. Prediction of MAFLD and NAFLD using different screening indexes: a cross-sectional study in U.S. adults. *Front Endocrinol (Lausanne)*. 2023;14:1083032.
- Kyhl LK, Nordestgaard BG, Tybjaerg-Hansen A, Nielsen SF. High fat in blood and body and increased risk of clinically diagnosed non-alcoholic fatty liver disease in 105,981 individuals. *ATHEROSCLEROSIS*. 2023;376:1–10.
- Zou Y, Yu M, Sheng G. Association between fasting plasma glucose and non-alcoholic fatty liver disease in a nonobese Chinese population with normal blood lipid levels: a prospective cohort study. *LIPIDS HEALTH DIS* 2020, 19(1).
- Nath P, Panigrahi MK, Sahu MK, Narayan J, Sahoo RK, Patra AA, Jena S, Patnaik AK, Jena A, Singh SP. Effect of Exercise on NAFLD and its risk factors: comparison of moderate versus low intensity Exercise. *J Clin Translational Hepatol*. 2020;8(2):120–6.
- Cai J, Zhang XJ, Li H. Progress and challenges in the prevention and control of nonalcoholic fatty liver disease. *MED RES REV*. 2019;39(1):328–48.

46. Wu Y, Zheng Q, Zou B, Yeo YH, Li X, Li J, Xie X, Feng Y, Stave CD, Zhu Q, et al. The epidemiology of NAFLD in Mainland China with analysis by adjusted gross regional domestic product: a meta-analysis. *HEPATOL INT.* 2020;14(2):259–69.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.