scientific reports

Check for updates

OPEN

Life's Crucial 9 and NAFLD from association to SHAP-interpreted machine learning predictions

Jianxin Xi[1], Yuguo Chen[1], Chen Jie[2], Jason Chi Shing Law[3], Zhongqi Fan[1] & Guoyue Lv[1]✉

Non-alcoholic fatty liver disease (NAFLD) is the most prevalent chronic liver disease worldwide. Cardiovascular disease (CVD) and NAFLD share multiple common risk factors. Life's Crucial 9 (LC9), a novel indicator for comprehensive assessment of cardiovascular health (CVH), has not yet been studied in terms of its association with or predictive value for NAFLD. This study analyzed data from 10,197 participants in the National Health and Nutrition Examination Survey (NHANES) from 2007 to 2018. The association between LC9 and NAFLD was assessed using weighted logistic regression, while weighted Cox proportional hazards models were applied to evaluate the relationship between LC9 and all-cause mortality among NAFLD patients. Restricted cubic spline (RCS) analysis was conducted to explore dose-response relationships, and Kaplan-Meier survival curves were utilized to examine differences in survival outcomes. Machine learning (ML) approaches were employed to construct predictive models, with the optimal model further interpreted using SHapley Additive exPlanations (SHAP). An increase of 10 points in LC9 was negatively associated with the risk of NAFLD (model 3: OR = 0.39, 95% CI = 0.36 – 0.42, *P* < 0.001) and all-cause mortality in NAFLD patients (model 3: HR = 0.78, 95% CI = 0.67 – 0.91, *P* < 0.001). A non-linear relationship was observed between LC9 and NAFLD (*P* < 0.0001 for nonlinearity). Among the eight ML models, the Support Vector Machine (SVM) demonstrated the best predictive performance (AUC = 0.873). SHAP analysis indicated that LC9 was the most significant predictor in the model. LC9 demonstrated a nonlinear negative association with NAFLD and a linear negative association with all-cause mortality in NAFLD patients. Maintaining a higher LC9 score may reduce the risk of NAFLD and all-cause mortality among NAFLD patients. The predictive model developed using Support Vector Machine (SVM) exhibited strong clinical predictive value, with LC9 being the most critical factor in the model, facilitating self-risk assessment and targeted intervention.

Keywords NAFLD, Life's Crucial 9, Machine learning, SHAP, NHANES

Abbreviations

NAFLD	Non-alcoholic fatty liver disease
CVD	Cardiovascular disease
LS7	Life's Simple 7
AHA	American Heart Association
CVH	Cardiovascular health
LE8	Life's Essential 8
LC9	Life's Crucial 9
NHANES	National Health and Nutrition Examination Surveys
PA	Physical activity
BMI	Body mass index
Non-HDL-C	Non-high-density lipoprotein cholesterol
DASH	Dietary Approaches to Stop Hypertension
PHQ-9	Patient Health Questionnaire-9
TG	Triglycerides
GGT	γ-Glutamyltransferase

[1]Department of Hepatobiliary and Pancreatic Surgery, General Surgery Center, The First Hospital of Jilin University, No.1 Xinmin Street, Chaoyang District, Changchun City, Jilin Province, China. [2]Department of Radiology, The First Hospital of Jilin University, Jilin, China. [3]Faculty of Medicine, The University of Hong Kong, Pok Fu Lam, Hong Kong SAR, China. ✉email: lvgy@jlu.edu.cn

CAP	Controlled attenuation parameter
PIR	Ratio of family income to poverty
Alb	Albumin
ALP	Alkaline phosphatase
ALT	Alanine aminotransferase
AST	Aspartate aminotransferase
TB	Total bilirubin
SCr	Serum creatinine
UA	Uric acid
OR	Odds ratios
CI	Confidence interval
XGBoost	Extreme Gradient Boosting
DT	Decision tree
GLM	Generalized linear model
NNET	Neural network
NB	Naive Bayes
KNN	K-Nearest Neighbors
RF	Random Forest
SVM	Support vector machines
LR	Linear regression
ML	Machine learning
AUROC	Area under the receiver operating characteristic curve
DCA	Decision curve analysis

Non-alcoholic fatty liver disease (NAFLD), characterized by the accumulation of fat in the liver, is the most common chronic liver disease worldwide. Its prevalence continues to rise globally, posing a significant threat to public health[1]. NAFLD is associated with complex metabolic disorders and is closely linked to the incidence and mortality of cardiovascular disease (CVD)[2]. Obesity, atherosclerotic dyslipidemia, insulin resistance, and type 2 diabetes (T2D) are common risk factors shared by NAFLD and CVD[2]. Therefore, the management of CVD risk factors is equally important in the routine care of NAFLD patients[3].

Life's Simple 7 (LS7), which includes three health behaviors (diet, physical activity (PA) and nicotine exposure), four health factors (body mass index (BMI), non-high-density lipoprotein cholesterol (non-HDL-C), blood glucose, and blood pressure), was initially proposed by the American Heart Association (AHA) and is used to assess cardiovascular health (CVH)[4]. With the growing research on factors influencing CVH, sleep health was added to LS7, resulting in the formation of Life's Essential 8 (LE8)[5]. Recently, recognizing the importance of mental health in the prevention of CVD, Life's Crucial 9 (LC9) was created as a new indicator, incorporating mental health into LE8[6]. Therefore, The LC9 score includes four health behaviors (diet, physical activity (PA), nicotine exposure, and sleep health), four health factors (body mass index (BMI), non-high-density lipoprotein cholesterol (non-HDL-C), blood glucose, and blood pressure), as well as mental health components. Currently, depression has become a global mental health issue that cannot be ignored[7]. Moreover, depression has been confirmed as an independent risk factor for CVD[8]. Similarly, depression has been shown to increase the risk of NAFLD, and its causal relationship with NAFLD has been reported in several studies[9,10]. Therefore, the inclusion of mental health in Life's Crucial 9 (LC9) may be more sensitive to individual differences. However, as of now, no studies have explored its association with NAFLD or its predictive performance for NAFLD outcomes.

Therefore, in the present study, we utilized data from the National Health and Nutrition Examination Surveys (NHANES) to investigate the association between LC9 and the risk of NAFLD, as well as its correlation with all-cause mortality in NAFLD patients. A NAFLD predictive model based on LC9 scores was developed using various machine learning techniques, thereby demonstrating the significance and predictive value of LC9 in forecasting NAFLD outcomes.

Methods

Study population

NHANES is a comprehensive, long-term epidemiological study focusing on non-institutionalized residents of the United States. This study employs a stratified, multistage probability sampling technique to gather baseline information and evaluate health status (https://wwwn.cdc.gov/nchs/nhanes/default.aspx). Given that the NHANES program has received approval from the Ethics Review Board of the National Center for Health Statistics (NCHS), and the dataset is publicly available, there was no necessity for additional ethical clearance.

This study analyzed publicly available data from the 2007–2018 survey cycles of the National Health and Nutrition Examination Survey (NHANES), which included 59,842 participants. Participants were sequentially excluded based on the following criteria:[1] incomplete VCTE or FLI data;[2] age < 20 years;[3] positive for hepatitis B surface antigen, hepatitis C antibody, or hepatitis C RNA;[4] diagnosis of autoimmune hepatitis;[5] significant alcohol consumption;[6] incomplete LC9 data;[7] missing covariate data. Ultimately, a total of 10,197 participants were included for analysis (Fig. 1). Additionally, data on follow-up and mortality status were obtained from the National Death Index records through December 31, 2019 (https://www.cdc.gov/nchs/data-linkage/mortality.htm), which were used to analyze the impact of LC9 on all-cause mortality in 4,678 NAFLD patients.

Measurement of LC9 score

LC9 builds upon LE8 by incorporating a depression score. In previous studies, the detailed algorithms for each component of the LE8 scoring system have been publicly documented (Supplemental Table 1)[11]. The LE8 score

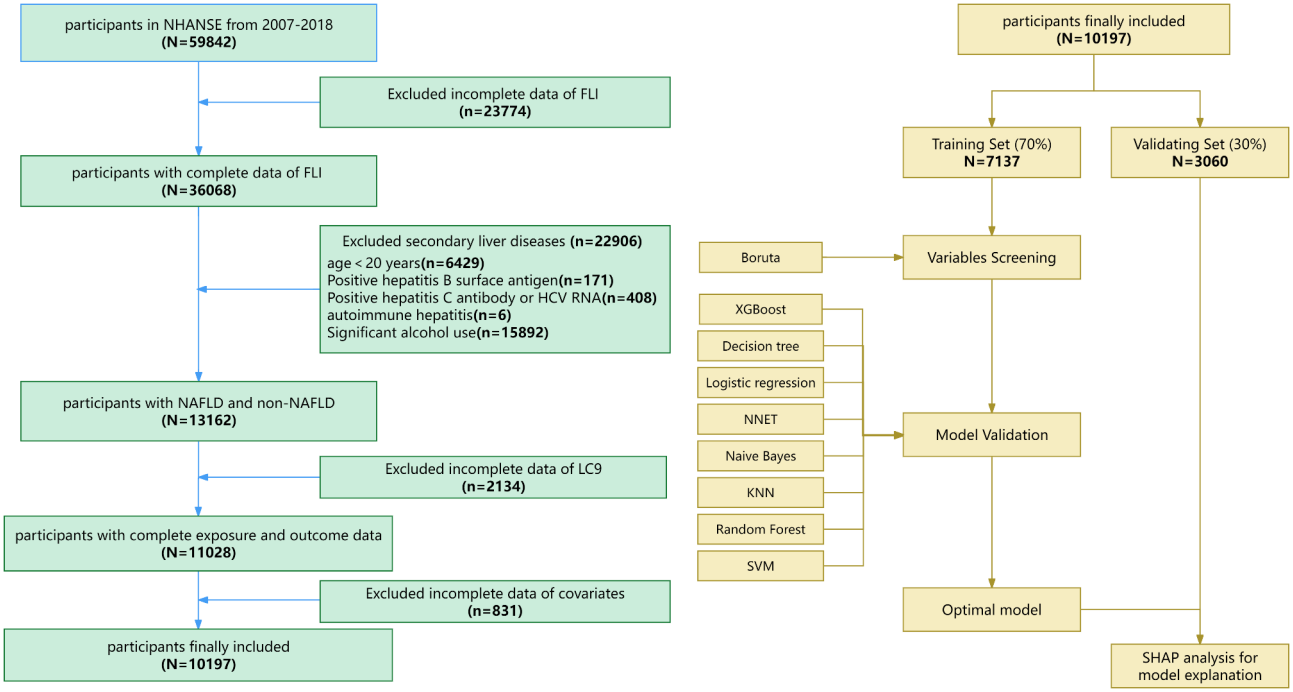

Fig. 1. Flowchart illustrating selection of the study population and analysis in NHANES from 2007 to 2018.

includes four health behaviors (diet, physical activity (PA), nicotine exposure, and sleep health) and four health factors (body mass index (BMI), non-high-density lipoprotein cholesterol (non-HDL-C), blood glucose, and blood pressure). The score for each CVH indicator is based on the association between health outcomes and risk, with allocations made using a modified Delphi method, ranging from 0 to 100 points. The final overall LE8 score is derived by calculating the average of the eight individual indicators.

In the calculation of individual indicators, the diet metric is evaluated using the Dietary Approaches to Stop Hypertension (DASH) Index. Dietary information was obtained from two 24-hour dietary recalls and integrated with food pattern equivalents data from the U.S. Department of Agriculture. The final DASH score was calculated utilizing the 'dietaryindex' R package[12]. Physical activity, nicotine exposure, sleeping information, diabetes history, and medication history were collected through self-report questionnaires. BMI and blood pressure were measured during physical examinations. Blood lipid levels, fasting blood glucose, and glycated hemoglobin levels were obtained through measurements of blood samples sent to central laboratories. Additionally, the depression score was calculated using the Patient Health Questionnaire-9 (PHQ-9) score. The depression score was assigned as follows: 100 for a PHQ-9 score of 0 to 4, 75 for a score of 5 to 9, 50 for a score of 10 to 14, 25 for a score of 15 to 19, and 0 for a score of 20 to 27[13]. Finally, the LC9 score was derived by calculating the average of the depression score and the eight indicators included in the LE8 score.

Definition of NAFLD

For participants from 2007 to 2018, FLI was utilized to evaluate the presence of hepatic steatosis. The FLI was calculated as follows:

$$\begin{aligned} \mathrm{L} = {} & 0.953 \times \ln(\mathrm{TG}) + 0.139 \times \mathrm{BMI} + 0.718 \times \ln(\mathrm{GGT}) \\ & + 0.053 \times \text{Waist Circumference} - 15.745 \end{aligned}$$

$$\begin{aligned} \mathrm{L} = {} & 0.953 \times \ln(\mathrm{TG}) + 0.139 \times \mathrm{BMI} + 0.718 \times \ln(\mathrm{GGT}) \\ & + 0.053 \times \text{Waist Circumference} - 15.745 \end{aligned}$$

In this formula, triglycerides (TG) were measured in mg/dL, BMI in kg/m^2, γ-glutamyltransferase (GGT) in U/L, and waist circumference in centimeters. According to previous studies, an FLI ≥ 60 was used to define hepatic steatosis[14]. Building upon this, NAFLD was defined as an FLI score ≥ 60, with exclusions made for individuals with viral hepatitis, autoimmune hepatitis, and significant alcohol consumption (defined as > 3 drinks per day for men or > 2 drinks per day for women, where a "drink" is equivalent to 14 g of pure alcohol[15]).

Assessment of covariates

Based on existing literature and clinical relevance, this study included covariates potentially associated with NAFLD. Demographic variables incorporated were age, gender, race and ethnicity, the ratio of family income to poverty (PIR), educational level, and marital status. Age was categorized into three groups: ≤ 39 years, 40 – 60 years, and > 60 years. Race and ethnicity were classified as Mexican American, Other Hispanic, Non-Hispanic

White, Non-Hispanic Black, and Other Race. The ratio of family income to poverty (PIR) was divided into three categories: <1.0, 1.0 – 3.0, and > 3.0. Educational level was classified into three categories: < high school, = high school, and > high school. Marital status was classified as "Married" for individuals who were married or living with a partner, and "Unmarried" for those who were widowed, divorced, separated, or never married. Additionally, anthropometric and laboratory covariates included albumin (Alb), alkaline phosphatase (ALP), GGT, alanine aminotransferase (ALT), aspartate aminotransferase (AST), total bilirubin (TB), serum creatinine (SCr), and uric acid (UA).

Statistical analysis

Baseline analysis and correlation analysis

The overall population was divided into NAFLD and non-NAFLD groups, and LC9 was categorized into quartiles (Q1-Q4). Categorical variables are presented as numbers (percentages) and were compared between groups using the weighted χ^2 test with the Rao and Scott second-order correction. Continuous variables are presented as medians (IQRs) and were compared between groups using the weighted Kruskal-Wallis test. Weighted binary logistic regression models were used to calculate odds ratios (ORs) and 95% confidence intervals (CIs) to examine the association between LC9 and NAFLD. Weighted Cox proportional hazards regression models were employed to calculate hazard ratios (HRs) and 95% CIs to explore the relationship between LC9 and all-cause mortality in the NAFLD population. Person-time was measured from the date of the NHANES interview to either the date of death or the end of follow-up (December 31, 2019), whichever came first. Survival rates across groups were compared using Kaplan-Meier curves. Three models were developed in this study: model 1, which was unadjusted for any covariates; model 2, which adjusted for age, gender, race and ethnicity, educational level, PIR, and marital status; and model 3, which further adjusted for Alb, ALP, GGT, ALT, AST, TB, SCr, and UA. The dose-response relationship between exposure and outcome was determined using a four-node restricted cubic spline (RCS). In the presence of a non-linear relationship, segmented regression using the 'segment' R package was applied to identify the existence of inflection points. Finally, subgroup analysis was performed as a sensitivity analysis to assess the stability of the association between LC9 and NAFLD, as well as the association between LC9 and all-cause mortality in NAFLD patients across different subgroups. Furthermore, to further validate the robustness of our results, we conducted two additional sets of sensitivity analyses. First, we defined MASLD, a new concept alternative to NAFLD, using the USFLI index and performed a correlation analysis with data from 2007 to 2018. Second, we conducted a further correlation analysis using VCTE data from 2017 to 2020.

Model development and validation

Variables selection and machine learning model construction and validation Among the final 10,197 participants, the data were randomly split into a training set ($n = 7,137$) and a testing set ($n = 3,060$) in a 7:3 ratio. To mitigate the impact of high-dimensional data on the performance of the machine learning (ML) models, variable selection was performed in the training set using the Boruta algorithm. This approach facilitated the identification of the most significant predictors, which were subsequently incorporated into the model development process[16]. Subsequently, eight ML algorithms, namely Extreme Gradient Boosting (XGBoost), Decision Tree (DT), Logistic Regression (GLM), Neural Networks (NNET), Naive Bayes (NB), K-Nearest Neighbors (KNN), Random Forest (RF), and Support Vector Machines (SVM), were utilized to train and construct predictive models to predict the occurrence of NAFLD. To ensure optimal performance and robustness, all eight models underwent 10-fold cross-validation. Additionally, traditional linear regression (LR) methods were also used to construct the models in order to evaluate the advantages and disadvantages of using machine learning compared to linear regression for model construction. Model accuracy was assessed using the area under the receiver operating characteristic curve (AUROC), and the reliability and clinical applicability of the models were further evaluated through calibration curves and decision curve analysis (DCA) to determine their net benefit.

Model interpretability Based on the selection of the optimal model, the Shapley Additive Explanations (SHAP) algorithm was utilized to enhance the interpretability of the model[17]. This algorithm assigns a corresponding attribution value (SHAP value) to each predictor variable, facilitating a quantitative assessment of the contribution of each variable to the model's predictive performance[18]. By visualizing the mean absolute SHAP values, the relative importance of each variable within the model can be ranked, providing a comprehensive understanding of their individual impact on model predictions[19]. The SHAP honeycomb diagram provides an intuitive visualization of the contributions of all variables to the model's predictions. The SHAP waterfall plot illustrates the direction and magnitude of the impact of each feature on the final predicted value for an individual sample. SHAP dependence plots allow for the examination of the dependency relationship between a given variable and its corresponding SHAP values, as well as the interaction effects between different variables.

All statistical analyses in this study were conducted using R (version 4.4.1). Two-tailed tests were performed, and a p-value of < 0.05 considered statistically significant. The overall study workflow is illustrated in Fig. 1.

Results

Baseline characteristics

A total of 4,681 NAFLD patients and 5516 non-NAFLD controls were included in the baseline analysis, with baseline characteristics presented in Table 1. Notably, in terms of demographic characteristics, NAFLD patients tended to be older, predominantly male, Mexican American, with lower educational level, and were more likely to be unmarried. Laboratory indicators revealed that NAFLD patients generally had lower levels of Alb and TB, but higher levels of ALP, GGT, ALT, AST, SCr, and UA. No significant differences in PIR were observed between the NAFLD and non-NAFLD groups. Furthermore, the LC9 score was lower in NAFLD patients, with the majority of these patients falling into the first quartile (Q1) following quartile stratification of the LC9 score.

Characteristic[2]	*N*	Overall *N* = 96,671,625[1]	Non-NAFLD *N* = 56,501,048[1]	NAFLD *N* = 40,170,578[1]	*p*-value
Age (years)	10,197				<0.001
20-39		3276 (34%)	2122 (40%)	1154 (25%)	
40-60		3529 (39%)	1728 (36%)	1801 (43%)	
> 60		3392 (27%)	1666 (24%)	1726 (32%)	
Gender	10,197				<0.001
Male		5108 (49%)	2468 (42%)	2640 (57%)	
Female		5089 (51%)	3048 (58%)	2041 (43%)	
Race and ethnicity	10,197				<0.001
Mexican American		1106 (5.6%)	483 (4.6%)	623 (7.0%)	
Other Hispanic		872 (4.5%)	450 (4.3%)	422 (4.6%)	
Non-Hispanic White		5001 (73%)	2713 (73%)	2288 (73%)	
Non-Hispanic Black		2085 (9.8%)	1096 (9.3%)	989 (10%)	
Other Race		1133 (7.2%)	774 (8.6%)	359 (5.3%)	
PIR	10,197				0.11
< 1.0		1430 (8.9%)	763 (8.8%)	667 (8.9%)	
1.0–3.0		3814 (30%)	1935 (29%)	1879 (32%)	
> 3.0		4953 (61%)	2818 (62%)	2135 (59%)	
Educational level	10,197				<0.001
< High school		1406 (8.5%)	687 (7.4%)	719 (9.9%)	
= High school		2045 (19%)	1007 (17%)	1038 (21%)	
> High school		6746 (73%)	3822 (76%)	2924 (69%)	
Marital status	10,197				0.001
Married		3693 (32%)	2082 (34%)	1611 (30%)	
Unmaried		6504 (68%)	3434 (66%)	3070 (70%)	
Alb (g/dl)	10,197	4.30 (4.10, 4.50)	4.40 (4.10, 4.60)	4.20 (4.00, 4.50)	<0.001
ALP (IU/L)	10,197	63 (51, 76)	59 (48, 71)	68 (56, 82)	<0.001
GGT (IU/L)	10,197	19 (13, 28)	15 (12, 21)	25 (18, 38)	<0.001
ALT (U/L)	10,197	21 (16, 28)	19 (15, 23)	25 (19, 34)	<0.001
AST (U/L)	10,197	22 (19, 27)	22 (19, 26)	24 (20, 29)	<0.001
TB (mg/dl)	10,197	0.60 (0.50, 0.80)	0.60 (0.50, 0.80)	0.60 (0.40, 0.80)	<0.001
SCr (umol/L)	10,197	76 (65, 88)	73 (64, 86)	79 (66, 90)	<0.001
UA (mg/dl)	10,197	5.30 (4.40, 6.30)	4.90 (4.10, 5.70)	5.90 (5.00, 6.80)	<0.001
DASH diet score	10,197	100 (75, 100)	100 (75, 100)	100 (75, 100)	<0.001
Physical activity score	10,197	100 (70, 100)	100 (70, 100)	100 (70, 100)	<0.001
Nicotine exposure score	10,197	70 (30, 100)	100 (70, 100)	30 (15, 70)	<0.001
Sleep health score	10,197	60 (40, 100)	80 (60, 100)	60 (40, 80)	<0.001
Body mass index score	10,197	100 (60, 100)	100 (100, 100)	100 (60, 100)	<0.001
Blood lipid score	10,197	80 (50, 100)	100 (50, 100)	55 (30, 100)	<0.001
Glucose score	10,197	80 (50, 100)	80 (50, 100)	80 (50, 100)	<0.001
Blood pressure score	10,197	100 (80, 100)	100 (90, 100)	100 (40, 100)	<0.001
Depression score	10,197	100 (100, 100)	100 (100, 100)	100 (100, 100)	<0.001
LC9 score	10,197	77 (67, 86)	83 (76, 91)	69 (61, 77)	<0.001
LC9 score quartile[3]	10,197				<0.001
Q1		3176 (25%)	828 (11%)	2348 (44%)	
Q2		2721 (26%)	1280 (21%)	1441 (32%)	
Q3		2319 (25%)	1592 (29%)	727 (20%)	
Q4		1981 (24%)	1816 (40%)	165 (3.8%)	

Table 1. Weighted characteristics of the study population based on NAFLD. PIR, Ratio of family income to poverty; Alb, Albumin; ALP, Alkaline Phosphatase; GGT, Gamma Glutamyl Transferase; ALT, Alanine Aminotransferase; AST, Aspartate Aminotransferase; TB, Total Bilirubin; SCr, Serum Creatinine; UA, Uric acid; LC9, Life's Crucial 9. [1]N: Weighted population. [2]Medians (IQRs) for continuous variables: P value was calculated by design-based KruskalWallis test. % for categorical variables: P value was calculated by Pearson's X^2: Rao & Scott adjustment. [3]Q1–Q4 indicates quartile 1–quartile 4.

Impact of LC9 score on NAFLD and all-cause mortality in NAFLD patients

The results of the weighted logistic regression analysis for the three models demonstrated a significant negative association between LC9 score and the risk of NAFLD, as summarized in Table 2. In model 3 (fully adjusted model), each 10-point increase in the LC9 score was associated with a 61% reduction in the risk of NAFLD (OR = 0.39, 95% CI = 0.36–0.42, $P < 0.001$). When the LC9 score was categorized into quartiles, compared to the lowest quartile, the adjusted ORs for the second, third, and highest quartiles were 0.38 (95% CI = 0.31–0.45, $P < 0.001$), 0.21 (95% CI = 0.17–0.26, $P < 0.001$), and 0.04 (95% CI = 0.03–0.05, $P < 0.001$), respectively. Additionally, a trend test of the LC9 score quartiles demonstrated a significant decline in the risk of NAFLD as the LC9 score increased (P for trend < 0.001).

Additionally, among the 4,379 NAFLD patients, after a median follow-up of 6.70 years (interquartile range: 3.67 to 9.92 years), the relationship between the LC9 score and all-cause mortality was assessed using Weighted Cox proportional hazards regression models. The results are also presented in Table 2. The results from model 3 suggest that for each 10-point increase in the LC9 score, the risk of all-cause mortality in NAFLD patients is reduced by 22% (HR = 0.78, 95% CI = 0.67–0.91, $P < 0.001$). Compared to the lowest quartile of LC9 score, the adjusted HR for the second quartile was 0.72 (95% CI: 0.49–1.04, $P = 0.076$), for the third quartile was 0.48 (95% CI: 0.27–0.85, $P = 0.012$), and for the highest quartile was 0.55 (95% CI: 0.18–1.70, $P = 0.302$). Regrettably, no significant trend effect in all-cause mortality risk was observed among NAFLD patients with increasing LC9 score quartiles (P for trend = 0.196). When LC9 score was dichotomized into low and high LC9 score groups, the high LC9 score group demonstrated significantly lower all-cause mortality compared to the low LC9 score group (Supplemental Fig. 1).

Dose–response relationship and sensitivity analyses

The dose-response relationships between LC9 and NAFLD, as well as between LC9 and all-cause mortality among NAFLD patients, were examined using RCS curves. The results are presented in Fig. 2. The RCS curve analysis indicated a significant non-linear relationship between LC9 score and NAFLD across model 1, model 2, and model 3 (all $P < 0.0001$ for nonlinearity). The positions of the four knots on the RCS curve were identified at 50.0, 68.9, 80.0, and 94.4.

In contrast, no significant nonlinear relationship was observed between the LC9 score and all-cause mortality among NAFLD patients (all $P > 0.05$ for nonlinearity). The four knot locations for the RCS curve were at 33.0, 54.1, 75.3, and 96.5. As the LC9 score increased, both the risk of NAFLD and the all-cause mortality rate among NAFLD patients decreased significantly. To further analyze the relationship between the LC9 score and NAFLD, segmented regression was performed using the 'segmented' R package. The result identified a breakpoint at 60.11. When the LC9 score exceeded this threshold, its incremental increases had a more pronounced effect on reducing the risk of NAFLD (Supplemental Fig. 2).

Subgroup analyses were conducted across demographic variables including age, gender, race and ethnicity, PIR, educational level, and marital status. The results are presented in Fig. 3. The negative association between LC9 and NAFLD remained consistent across all analyzed subgroups. Notably, a significant interaction was identified between LC9 and PIR with the risk of NAFLD ($p = 0.016$). Additionally, a significant interaction was also observed between LC9 and race and ethnicity ($p = 0.007$). This association was more pronounced among participants with PIR > 3.0 (OR per 10-point increase: 0.37, 95% CI: 0.34–0.40) and other race (OR per 10-point increase: 0.39, 95% CI: 0.33–0.45). Conversely, no significant interactions were observed between LC9 and subgroup variables with the risk of all-cause mortality among individuals with NAFLD. Moreover, whether using the 2007–2018 data to define MASLD through the USFLI index or utilizing the 2017–2020 data to define NAFLD through VCTE data, the association between LC9 and NAFLD remained consistently stable (Supplementary Tables 2 and 3).

	NAFLD				All-cause mortality		
	Model 1 OR (95%CI) *P*-value	**Model 2 OR (95%CI) *P*-value**	**Model 3 OR (95%CI) *P*-value**		**Model 1 HR (95%CI) *P*-value**	**Model 2 HR (95%CI) *P*-value**	**Model 3 HR (95%CI) *P*-value**
LC9 Per 10-score increase	0.35 (0.33, 0.37) < 0.001	0.32 (0.29, 0.34) < 0.001	0.39 (0.36, 0.42) < 0.001	LC9 Per 10-score increase	0.72 (0.63, 0.81) < 0.001	0.74 (0.64, 0.86) < 0.001	0.78 (0.67, 0.91) 0.001
LC9 quartile				LC9 quartile			
Q1 (24.4–67.2)	Ref.	Ref.	Ref.	Q1 (24.4–60.6)	Ref.	Ref.	Ref.
Q2 (67.2–77.2)	0.38 (0.32, 0.45) < 0.001	0.32 (0.27, 0.38) < 0.001	0.38 (0.31, 0.45) < 0.001	Q2 (60.6–69.4)	0.60 (0.41, 0.89) 0.011	0.64 (0.44, 0.94) 0.021	0.72 (0.49, 1.04) 0.076
Q3 (77.2–86.1)	0.17 (0.14, 0.20) < 0.001	0.14 (0.11, 0.17) < 0.001	0.21 (0.17, 0.26) < 0.001	Q3 (69.4–77.2)	0.35 (0.20, 0.61) < 0.001	0.44 (0.25, 0.79) 0.006	0.48 (0.27, 0.85) 0.012
Q4 (86.1–100)	0.02 (0.02, 0.03) < 0.001	0.02 (0.01, 0.03) < 0.001	0.04 (0.03, 0.05) < 0.001	Q4 (77.2–96.7)	0.38 (0.13, 1.10) 0.075	0.52 (0.17, 1.58) 0.249	0.55 (0.18, 1.70) 0.302
P for trend	< 0.001	< 0.001	< 0.001	*P* for trend	0.031	0.162	0.196

Table 2. Association between LC9 and NAFLD, and association between LC9 and all-cause mortality in NAFLD patients. Model 1 was unadjusted. Model 2 was adjusted for age, gender, race and ethnicity, educational level, PIR, marital status. Model 3 was adjusted for age, gender, race and ethnicity, educational level, PIR, marital status, Alb, ALP, GGT, ALT, AST, TB, SCr, UA.

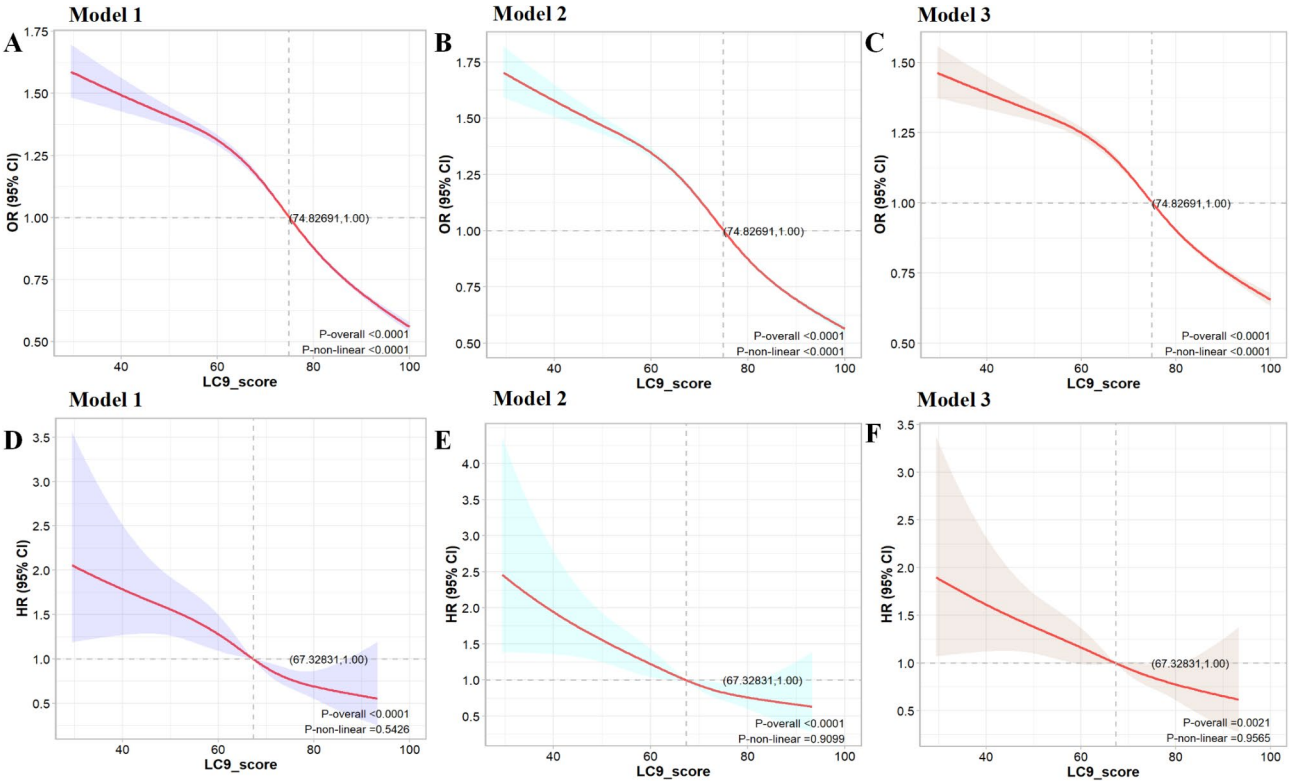

Fig. 2. Association of LC9 with NAFLD and all-cause mortality among individuals with NAFLD. In Model 1, the odds ratio (OR) and hazard ratio is presented without adjusting for any variables. In Model 2, the OR and HR are adjusted for age, gender, race and ethnicity, educational level, PIR, and marital status. In Model 3, the OR and HR are additionally adjusted for Alb, ALP, GGT, ALT, AST, TB, SCr, and UA. Shaded areas represent 95% CIs. The LC9 value corresponding to an OR of 1 is 74.83 and an HR of 1 is 67.33.

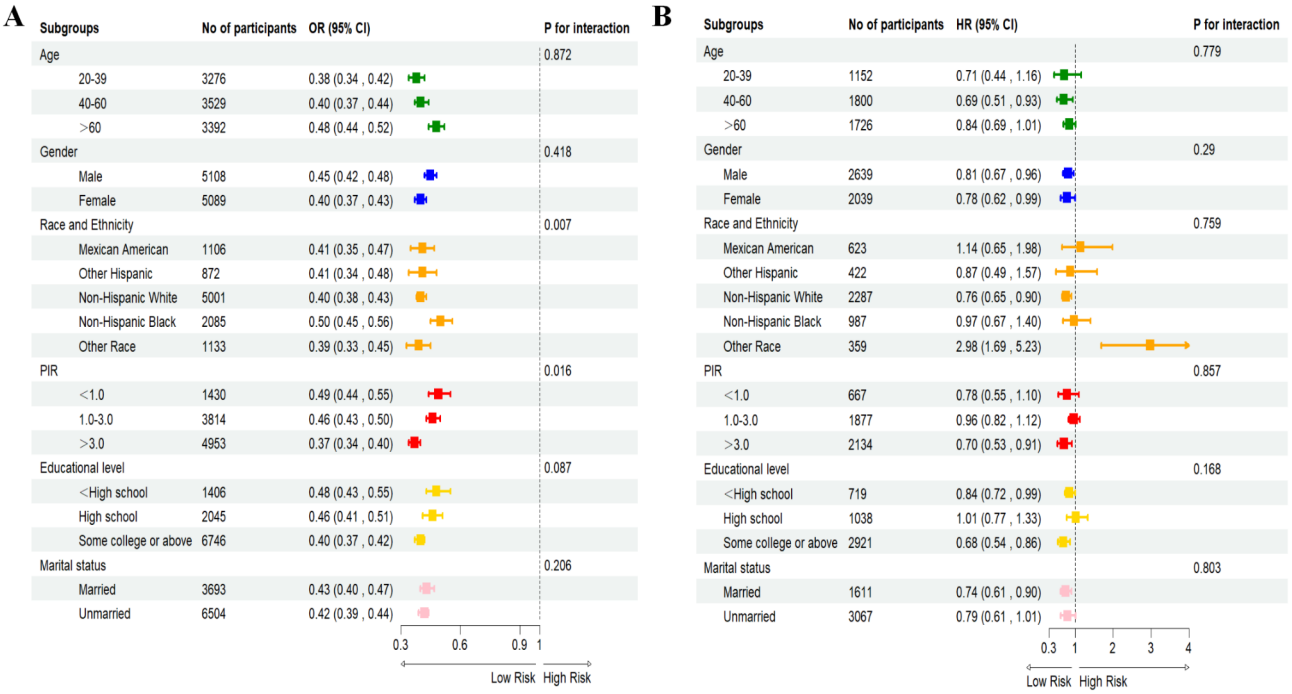

Fig. 3. Subgroup analysis of the association of LC9 with NAFLD and all-cause mortality among individuals with NAFLD.

Variable selection and model development and evaluation

In this study, 15 potential effective predictive variables were selected using the Boruta algorithm with shadow features (Fig. 4). After filtering, 15 valid variables were identified for subsequent model prediction, including LC9, Age, Gender, Race and ethnicity, PIR, Educational level, Alb, ALP, GGT, ALT, AST, TB, SCr, UA and marital status.

Figure 5A and B show the ROC curves of eight ML models, including XGBoost, DT, GLM, NNET, NB, KNN, RF, and SVM, in both the training and testing sets. Moreover, calibration curves and decision curve analysis (DCA) curves for the eight ML models were analyzed to evaluate the predictive accuracy and clinical value of the models (Supplemental Fig. 3). It is apparent that, both in the training set (AUROC = 0.873) and the testing set (AUROC = 0.868), SVM model demonstrated superior predictive performance and optimal fit. Furthermore, compared to the LR model (AUROC of the training set = 0.868 and AUROC of the testing set = 0.863), the SVM model demonstrated superior predictive performance (Supplementary Fig. 4).Consequently, in the subsequent analyses, the best-performing model (SVM) was further examined for interpretability.

Visualization of feature importance

The importance of each feature variable in the SVM model and its contribution to the model's predictions were assessed using the SHAP algorithm (Fig. 6A and B). Among the 15 predictor variables included in the model, LC9 had the highest average absolute SHAP value, which was consistent with expectations, indicating that LC9 is the most important risk factor for NAFLD. The SHAP waterfall plot was employed to visualize the direction and magnitude of impact of different variable features on the final predicted value for the second participant within the study population, thereby improving the transparency and interpretability of the model (Fig. 6C). In Fig. 6D, the relationship between the variables in the model and their corresponding SHAP values is illustrated. A notable linear relationship is observed, where an increase in the LC9 score is associated with a decrease in the SHAP values. Considering that the AUROC and SHAP results might vary across different quartiles, we performed model predictions based on LC9 quartiles. The AUROC results remained largely consistent across different LC9 quartiles, while the SHAP results showed variability (Supplementary Figs. 5 and 6). Additionally, we also incorporated LC9 as a categorical variable into the model, and the results were generally consistent with those obtained when LC9 was treated as a continuous variable (Supplementary Fig. 7).

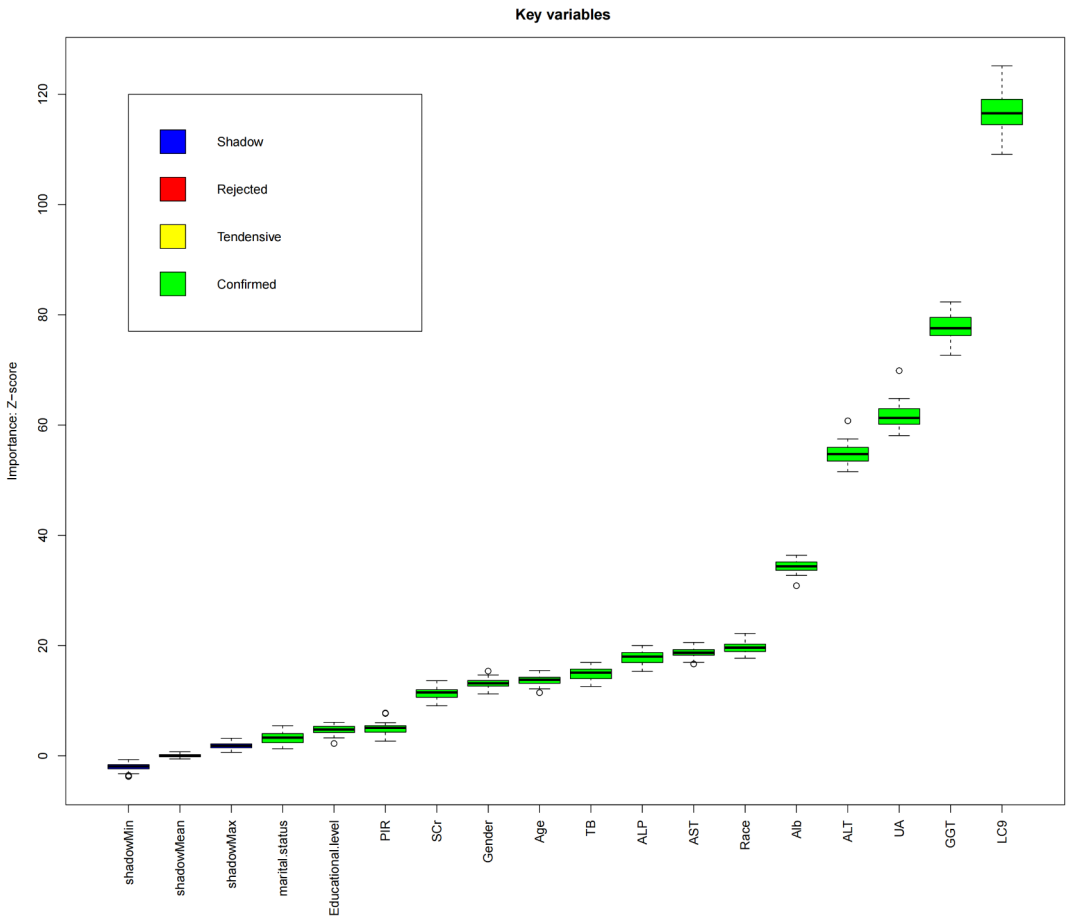

Fig. 4. Image of Boruta algorithm for selecting ML model variables.

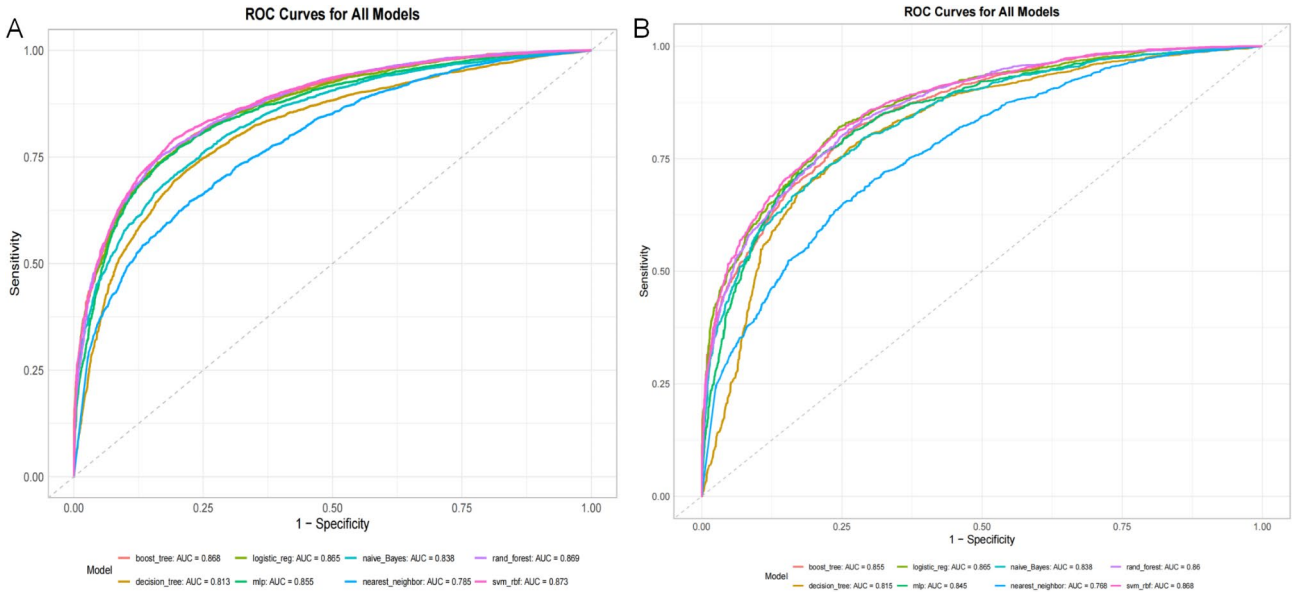

Fig. 5. (A) ROC curves of the training set. (B) ROC curves of the testing set.

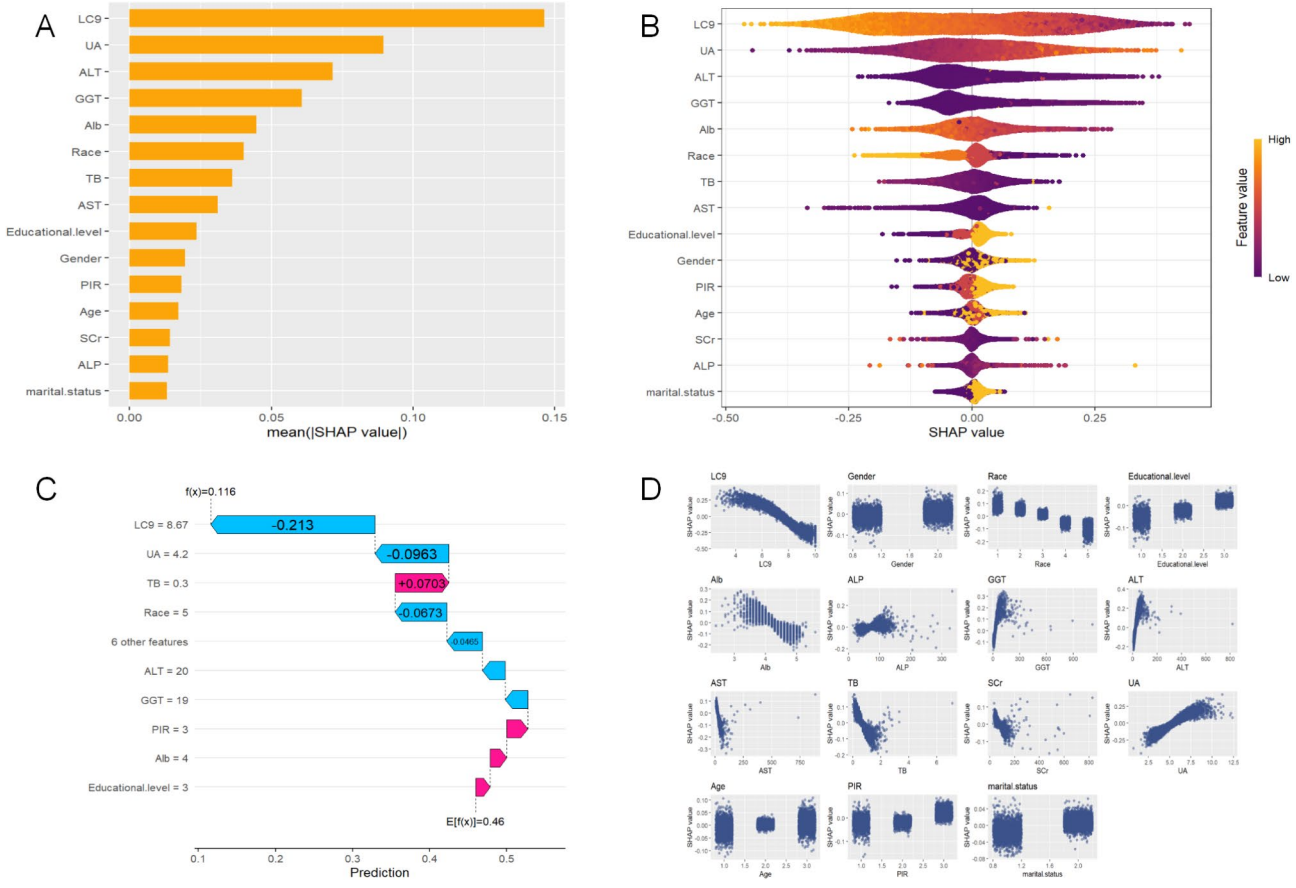

Fig. 6. SHAP diagram of SVM model. (A) SHAP value ranking of the variables in the model. (B) SHAP honeycomb diagram of the SVM model. (C) SHAP waterfall plot for the second sample in the study population. (D) the dependency relationships between SHAP values and different variables.

Discussion

This study, utilizing data from NHANES 2007–2018, investigated the association between the LC9 score and non-alcoholic fatty liver disease (NAFLD), as well as its relationship with all-cause mortality in individuals with NAFLD. In addition, eight ML models were developed to predict the risk of NAFLD. After adjusting for multiple covariates, the results consistently demonstrated a significant negative correlation between LC9 and both NAFLD and all-cause mortality in NAFLD patients, with a pronounced nonlinear relationship between LC9 and NAFLD risk. Subgroup analyses and sensitivity analyses further supported the robustness of these findings. Among the eight ML models, SVM model exhibited the best predictive performance, with an AUROC of 0.856 following cross-validation, highlighting its superior accuracy and reliability. The SHAP algorithm revealed that LC9 was the most influential variable in the best-performing model, underscoring its critical clinical value in predicting NAFLD risk.

Currently, the associations between LS7 and LE8 with NAFLD have been widely reported. A cross-sectional analysis from the Multi-Ethnic Study of Atherosclerosis, involving 3,901 participants, demonstrated a negative correlation between LS7 and the prevalence of NAFLD[20]. Another cross-sectional study analyzing 2,679 participants from the NHANES database for 2017–2018 revealed that lower LS7 scores were associated with a significantly increased risk of NAFLD, advanced liver fibrosis, and cirrhosis, compared to higher LS7 score groups[21]. Furthermore, these results were corroborated by a cohort study in Korean adults, where higher LS7 scores were not only linked to a reduced incidence of NAFLD but also promoted its regression[22]. LE8, as an improved version of LS7, has also been extensively studied. A cross-sectional study based on the NHANES dataset revealed a non-linear negative association between LE8 and NAFLD[23]. He et al. utilized the UK Biobank cohort, including 266,645 participants without a history of liver disease, and through a median follow-up period of 11.9 years, demonstrated that higher LE8 scores significantly reduce the risk of incident severe NAFLD, with this effect remaining unaffected by genetic risk[24]. Additionally, higher LE8 scores were found to mitigate the impact of air pollution on the risk of severe NAFLD in individuals with type 2 diabetes[25]. Furthermore, a large cohort study in the Chinese population, involving 21,844 participants with a median follow-up of 2.3 years, highlighted the significant role of favorable LE8 scores in reducing the incidence of NAFLD and promoting its regression[26]. Recent studies have explored the impact of LE8 on all-cause mortality and CVD mortality in patients with NAFLD/MASLD. Maintaining favorable LE8 scores has been shown to alleviate the risk of both all-cause mortality and CVD mortality in these patients[27,28]. LC9, a recently proposed new index, has garnered limited research attention. However, its associations with CVD mortality and all-cause mortality have been established[29]. Currently, no studies have investigated the relationship between LC9 and NAFLD or its predictive value for NAFLD.

Although the underlying biological mechanisms linking various CVH assessment indicators to the development and progression of NAFLD, as well as mortality in NAFLD patients, remain unclear, research into the mechanisms by which CVH components are associated with NAFLD is ongoing. The pathogenesis of NAFLD is multifactorial, involving a complex interplay of obesity, insulin resistance, chronic inflammation, oxidative stress, and dysregulation of lipid metabolism[26]. The DASH diet, by reducing energy density and increasing dietary fiber content, can significantly reduce body weight and obesity index[30]. Additionally, the DASH diet adjusts trace elements, with high calcium and magnesium promoting fat breakdown and fatty acid saponification, while low sodium helps regulate leptin levels to inhibit fat accumulation[31,32]. Physical activity can improve insulin resistance in skeletal muscles and the liver and suppress the production of inflammatory factors[33,34]. Similarly, smoking can promote the production of inflammatory factors, reduce the levels of anti-inflammatory factors, exacerbate oxidative stress, and induce hepatocyte apoptosis, thereby worsening NAFLD[35,36]. Sleep disturbances can lead to inter-organ imbalance of sympathetic-parasympathetic branches, with an overactive sympathetic nervous system increasing the risk of obesity and insulin resistance[37,38]. In individuals with depression, increased activity of monoamine oxidase-A enzyme can promote cellular oxidative stress and exert detrimental effects on NAFLD by mediating the development of obesity and diabetes[39,40]. In conclusion, the association between LC9 score and NAFLD is both scientifically plausible and supported by existing evidence.

To date, there are no specific pharmacological treatments for NAFLD, making a healthy lifestyle the cornerstone of its management, with its importance being self-evident. In contrast to methods that focus solely on a single risk factor for NAFLD, LC9 integrates multiple indicators, providing a comprehensive evaluation of health behaviors and factors in NAFLD patients, while also considering the impact of mental health on disease outcomes. Additionally, the LC9 score provides a "visualization" of CVH, facilitating self-assessment by patients and allowing them to make adjustments based on the evaluation results, thereby offering significant clinical value. To the best of our knowledge, this study is the first to establish a relationship between LC9 and NAFLD, as well as to investigate the impact of LC9 on all-cause mortality in NAFLD patients. In addition, this study is the first to apply ML techniques to develop a predictive model for NAFLD based on the LC9 score. Utilizing the NHANES dataset, which is designed with a stratified, multistage probability sampling approach, we implemented weight adjustments to ensure that the results are nationally representative of the U.S. population. Furthermore, potential confounding factors were addressed through covariate adjustments, and subgroup analyses were conducted to reinforce the robustness of the findings. Nevertheless, this study has several limitations. This study includes both cross-sectional and cohort components. The relationship between LC9 and NAFLD was investigated through a cross-sectional analysis, and thus, it cannot establish a causal relationship between LC9 and NAFLD. Although multiple covariates were considered, we were still unable to exclude the influence of all potential confounders. Additionally, health behavior indicators based on self-reported questionnaires are prone to measurement errors. Lastly, the results of this study are specific to U.S. adults, and the ML model developed here has not been externally validated, limiting its applicability to other populations. Future research should involve large-scale, prospective, multi-center cohort studies, or Mendelian randomization approaches to validate

causal relationships. Additionally, external validation and refinement of the machine learning models across diverse populations are needed to enhance their predictive accuracy and clinical utility.

Conclusion

In summary, the findings of our study demonstrate that in a nationally representative cohort of U.S. adults, LC9 is negatively associated with both the risk of NAFLD and all-cause mortality among NAFLD patients. Moreover, the NAFLD prediction model based on LC9, developed using multiple ML techniques, exhibits strong accuracy and holds considerable clinical value for individual NAFLD risk assessment. This model provides a foundation for timely self-intervention through lifestyle changes in NAFLD patients, potentially mitigating disease progression and preventing the advancement to more severe stages.

Data availability

Data utilized in this investigation are available for public access via the NHANES website (https://www.cdc.gov/nchs/nhanes/index.htm). The datasets analyzed during this study can be found in the NHANES repository, and all pertinent data are accessible to the public in accordance with the guidelines established by the NHANES program.

Received: 30 December 2024; Accepted: 3 March 2025
Published online: 18 March 2025

References

1. Riazi, K. et al. The prevalence and incidence of NAFLD worldwide: a systematic review and meta-analysis. *Lancet Gastroenterol. Hepatol.* **7**(9), 851–861 (2022).
2. Deprince, A., Haas, J. T. & Staels, B. Dysregulated lipid metabolism links NAFLD to cardiovascular disease. *Mol. Metab.* **42**, 101092 (2020).
3. Adams, L. A., Anstee, Q. M., Tilg, H. & Targher, G. Non-alcoholic fatty liver disease and its relationship with cardiovascular disease and other extrahepatic diseases. *Gut* **66**(6), 1138–1153 (2017).
4. Lloyd-Jones, D. M. et al. Defining and setting National goals for cardiovascular health promotion and disease reduction: the American heart association's strategic impact goal through 2020 and beyond. *Circulation* **121**(4), 586–613 (2010).
5. Lloyd-Jones, D. M. et al. Life's essential 8: updating and enhancing the American heart association's construct of cardiovascular health: a presidential advisory from the American heart association. *Circulation* **146**(5), e18–e43 (2022).
6. Gaffey, A. E., Rollman, B. L. & Burg, M. M. Strengthening the pillars of cardiovascular health: psychological health is a crucial component. *Circulation* **149**(9), 641–643 (2024).
7. Global burden. Of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. *Lancet* **396**(10258), 1204–1222 (2020).
8. Levine, G. N. et al. Psychological health, well-being, and the mind-heart-body connection: a scientific statement from the American heart association. *Circulation* **143**(10), e763–e83 (2021).
9. Li, S., Li, S., Duan, F. & Lu, B. Depression and NAFLD risk: a meta-analysis and Mendelian randomization study. *J. Affect. Disord* **352**, 379–385 (2024).
10. Ruan, X. et al. Depression and 24 Gastrointestinal diseases: a Mendelian randomization study. *Transl Psychiatry* **13**(1), 146 (2023).
11. Ma, H. et al. Cardiovascular health and life expectancy among adults in the United States. *Circulation* **147**(15), 1137–1146 (2023).
12. Zhan, J. J. et al. Dietaryindex: a user-friendly and versatile R package for standardizing dietary pattern analysis in epidemiological and clinical studies. *Am. J. Clin. Nutr.* **120**(5), 1165–1174 (2024).
13. Zhang, Z., Jackson, S. L., Gillespie, C., Merritt, R. & Yang, Q. Depressive symptoms and mortality among US adults. *JAMA Netw. Open.* **6**(10), e2337011 (2023).
14. Wu, S. et al. Non-alcoholic fatty liver is associated with increased risk of irritable bowel syndrome: a prospective cohort study. *BMC Med.* **20**(1), 262 (2022).
15. Zhang, Y. et al. Association of triglyceride glucose-related parameters with all-cause mortality and cardiovascular disease in NAFLD patients: NHANES 1999–2018. *Cardiovasc. Diabetol.* **23**(1), 262 (2024).
16. Speiser, J. L., Miller, M. E., Tooze, J. & Ip, E. A comparison of random forest variable selection methods for classification prediction modeling. *Expert Syst. Appl.* **134**, 93–101 (2019).
17. Wang, H., Shao, Y., Zhou, S., Zhang, C. & Xiu, N. Support vector machine classifier via L(0/1) Soft-Margin loss. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(10), 7253–7265 (2022).
18. Alabi, R. O., Elmusrati, M., Leivo, I., Almangush, A. & Mäkitie, A. A. Machine learning explainability in nasopharyngeal cancer survival using LIME and SHAP. *Sci. Rep.* **13**(1), 8984 (2023).
19. Parsa, A. B., Movahedi, A., Taghipour, H., Derrible, S. & Mohammadian, A. K. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accid. Anal. Prev.* **136**, 105405 (2020).
20. Oni, E. et al. Life's simple 7 and nonalcoholic fatty liver disease: the multiethnic study of atherosclerosis. *Am. J. Med.* **134**(4), 519–525 (2021).
21. Fan, H. et al. Ideal cardiovascular health metrics are associated with reduced severity of hepatic steatosis and liver fibrosis detected by transient elastography. *Nutrients* **14**, 24 (2022).
22. Jang, E. H. et al. Cardiovascular health metrics in the development and regression of nonalcoholic fatty liver disease: A cohort study. *J. Clin. Med.* **8**, 5 (2019).
23. Wang, L., Yi, J., Guo, X. & Ren, X. Associations between life's essential 8 and non-alcoholic fatty liver disease among US adults. *J. Transl. Med.* **20**(1), 616 (2022).
24. He, P. et al. A healthy lifestyle, life's essential 8 scores and new-onset severe NAFLD: A prospective analysis in UK biobank. *Metabolism* **146**, 155643 (2023).
25. Aimuzi, R., Xie, Z., Qu, Y. & Jiang, Y. Air pollution, life's essential 8, and risk of severe non-alcoholic fatty liver disease among individuals with type 2 diabetes. *BMC Public. Health* **24**(1), 1350 (2024).
26. Yaqin, W. et al. Cumulative exposure to AHA life's essential 8 is associated with nonalcoholic fatty liver disease: a large cohort study. *Nutr. Metab. (Lond.)* **21**(1), 38 (2024).
27. Zhang, Y., Wang, P., Tu, F., Kang, H. & Fu, C. Life's essential 8 and mortality in US adults with metabolic dysfunction-associated steatotic liver disease. *BMC Public. Health* **24**(1), 3411 (2024).
28. Sun, M., Qiu, Y., Zhang, L. & Chen, G. The correlation between life's essential 8 and cardiovascular disease and mortality in individuals with nonalcoholic fatty liver disease: a cross-sectional study. *Sci. Rep.* **14**(1), 23999 (2024).
29. Ge, J., Peng, W. & Lu, J. Predictive value of life's crucial 9 for cardiovascular and all-cause mortality: a prospective cohort study from the NHANES 2007 to 2018. *J. Am. Heart Assoc.* **13** (20), e036669 (2024).

30. Rooholahzadegan, F. et al. The effect of DASH diet on glycemic response, meta-inflammation and serum LPS in obese patients with NAFLD: a double-blind controlled randomized clinical trial. *Nutr. Metab. (Lond.)* **20**(1), 11 (2023).
31. Conlin, P. R. et al. The effect of dietary patterns on blood pressure control in hypertensive patients: results from the dietary approaches to stop hypertension (DASH) trial. *Am. J. Hypertens.* **13**(9), 949–955 (2000).
32. Soltani, S., Chitsazi, M. J. & Salehi-Abargouei, A. The effect of dietary approaches to stop hypertension (DASH) on serum inflammatory markers: a systematic review and meta-analysis of randomized trials. *Clin. Nutr.* **37**(2), 542–550 (2018).
33. von Loeffelholz, C., Roth, J., Coldewey, S. M. & Birkenfeld, A. L. The role of physical activity in nonalcoholic and metabolic dysfunction associated fatty liver disease. *Biomedicines* **9**, 12 (2021).
34. Gleeson, M. et al. The anti-inflammatory effects of exercise: mechanisms and implications for the prevention and treatment of disease. *Nat. Rev. Immunol.* **11**(9), 607–615 (2011).
35. Arnson, Y., Shoenfeld, Y. & Amital, H. Effects of tobacco smoke on immunity, inflammation and autoimmunity. *J. Autoimmun.* **34**(3), J258–J265 (2010).
36. Azzalini, L. et al. Cigarette smoking exacerbates nonalcoholic fatty liver disease in obese rats. *Hepatology* **51**(5), 1567–1576 (2010).
37. Sabath, E., Báez-Ruiz, A. & Buijs, R. M. Non-alcoholic fatty liver disease as a consequence of autonomic imbalance and circadian desynchronization. *Obes. Rev.* **16**(10), 871–882 (2015).
38. Thorp, A. A. & Schlaich, M. P. Relevance of sympathetic nervous system activation in obesity and metabolic syndrome. *J. Diabetes Res.* **2015**, 341583 (2015).
39. Youssef, N. A. et al. Associations of depression, anxiety and antidepressants with histological severity of nonalcoholic fatty liver disease. *Liver Int.* **33**(7), 1062–1070 (2013).
40. Bica, T., Castelló, R., Toussaint, L. L. & Montesó-Curto, P. Depression as a risk factor of organic diseases:an international integrative review. *J. Nurs. Scholarsh.* **49**(4), 389–399 (2017).

Acknowledgements

We would like to acknowledge the National Health and Nutrition Examination Survey (NHANES) for providing the publicly available data used in this study. We also thank the researchers and participants who contributed to the data collection process. Special thanks to the team members who assisted in data analysis and interpretation. Finally, we would like to express our gratitude to the institutions and organizations that support research in this field, enabling studies like ours to advance scientific knowledge.

Author contributions

Formal analysis & writing: Jianxin Xi and Jason Chi Shing Law; Statistical analysis: Chen Yuguo; Figure production: Chen Jie; Review & editing: Lv Guoyue and Fan Zhongqi.

Funding

This research is funded by the Natural Science Foundation of China (82241223, U20A20360), the Natural Science Foundation of Jilin Province (No. YDZJ202201ZYTS014), and the Financial Department of Jilin Province (Grant NO. JLSWSRCZX2020-045).

Competing interests

The authors declare no competing interests.

Ethics approval

The present study utilizes publicly accessible data from the National Health and Nutrition Examination Survey (NHANES). Given that the utilized data are anonymized and publicly available, ethical approval was not deemed necessary for this research.

Consent to participate

This study used publicly available data from the National Health and Nutrition Examination Survey (NHANES). NHANES is a program conducted by the Centers for Disease Control and Prevention (CDC) that collects health and nutritional data from a representative sample of the U.S. population. The NHANES data used in this research are de-identified and publicly available, so no individual consent was required for this study. The study was conducted in accordance with ethical standards and institutional guidelines for research using publicly available data.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-92777-0.

Correspondence and requests for materials should be addressed to G.L.