Application of Interpretable Machine Learning Algorithms to Predict Sarcopenia in Patients with Chronic Liver Disease: Findings from CHARLS

Bin Liang, Xue Qiu, Jiehua Deng, Jiansheng Huang, Yequan Lu, Yongyu Chen

Submitted to: Journal of Medical Internet Research
on: December 03, 2024

Table of Contents

Original Manuscript **5**
Supplementary Files **20**
Figures **21**
Figure 1 ..... **22**
Figure 2 ..... **23**

Application of Interpretable Machine Learning Algorithms to Predict Sarcopenia in Patients with Chronic Liver Disease: Findings from CHARLS

Bin Liang[1*] PhD; Xue Qiu[2*] PhD; Jiehua Deng[3] PhD; Jiansheng Huang[4] PhD; Yequan Lu[5] PhD; Yongyu Chen[1] PhD

[1]Minzu Hospital of Guangxi Medical University Nanning CN
[2]Guangxi Medical University Nanning CN
[3]Eighth Affiliated Hospital of Sun Yat-sen University Shenzhen CN
[4]First Affiliated Hospital of GuangXi Medical University Nanning CN
[5]Nanxishan Hospital of Guangxi Zhuang Autonomous Region Guilin CN
[*]these authors contributed equally

Corresponding Author:
Yongyu Chen PhD
Minzu Hospital of Guangxi Medical University
No.232, Minxiu Dong Road, Nanning, Guangxi
Nanning
CN

Abstract

Background: Diagnosing sarcopenia is challenging due to limited access to bio-impedance analysis (BIA) and dual-energy X-ray absorptiometry (DXA).

Objective: This study sought to establish and evaluate machine learning (ML) models for predicting sarcopenia in CLD patients.

Methods: This study retrospectively analyzed data from the China Health and Retirement Longitudinal Study (CHARLS). Feature selection was conducted via univariate logistic regression, least absolute shrinkage and selection operator (LASSO), and variance inflation factor (VIF). Eight ML algorithms were applied to build predictive models, with performance assessed using area under the receiver operating characteristic curve (AUC), confusion matrix-derived index, calibration curves, and decision curve analysis (DCA). Feature importance was evaluated with SHapley Additive exPlanations (SHAP).

Results: 712 participants from CHARLS 2015 were divided into a training set and internal validation set, with 365 patients from CHARLS 2011 used for external validation. Eight significant predictors of sarcopenia were identified: age, education level, episodic memory, waist circumference, body mass index (BMI), heart disease, hemoglobin, and mean corpuscular volume (MCV). Among the ML models, the support vector machine (SVM) performed best, with an AUC of 0.91 (95% CI 0.88–0.93) in the training set, and 0.88 (95% CI 0.85–0.91) and 0.82 (95% CI 0.79–0.85) in internal and external validation sets, respectively. Calibration curves, Brier scores, and DCA confirmed the SVM's robust performance. SHAP analysis revealed that BMI, age, and education level were the most important features.

Conclusions: The SVM model, integrating socio-demographic and clinical factors, showed excellent performance in predicting sarcopenia risk in CLD patients. It holds promise for guiding early diagnosis and treatment to improve patient outcomes. Clinical Trial: Not applicable.

(JMIR Preprints 03/12/2024:69585)
DOI: https://doi.org/10.2196/preprints.69585

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Application of Interpretable Machine Learning Algorithms to Predict Sarcopenia in Patients with Chronic Liver Disease: Findings from CHARLS

Authors and Affiliations

Bin Liang[a,1], Xue Qiu[b,1], Jiehua Deng[c], Jiansheng Huang[d], Yequan Lu[e], Yongyu Chen[f]*

a Department of Gastroenterology & Hepatology, Minzu Hospital of Guangxi Medical University, Nanning, Guangxi, China
b Guangxi Medical University, Nanning, Guangxi, China
c Department of Respiratory, The Eighth Affiliated Hospital of Sun Yat-sen University, Shenzhen, Guangdong, China
d Department of Cardiology, The first Affiliated Hospital of Guangxi Medical University, Nanning, Guangxi, China
e Department of Respiratory, Nanxishan Hospital of Guangxi Zhuang Autonomous Region, Guilin, Guangxi, China
f Department of Gastroenterology & Hematology, Minzu Hospital of Guangxi Medical University, Nanning, Guangxi, China

First author Bin Liang (liangbindoctor@163.com)
Co-First author Xue Qiu (xue1010qiu@163.com)
Co-author Jiehua Deng (2459828789@qq.com)
Co-author Jiansheng Huang (3319069747@qq.com)
Co-author Yequan Lu (2587279378@qq.com)

* Correspondence to:

Yongyu Chen, Department of Hematology, Minzu Hospital of Guangxi Medical University, No.232, Mingxiu Dong Road, Nanning, China
E-mail address: Chenyongyu@sr.gxmu.edu.cn
Telephone: +86-15678351189

1These authors contributed equally to this work.

Text word count:3119;
Abstract word count:262;
Tables: 2
Figures: 4

Abstract

Background

Diagnosing sarcopenia is challenging due to limited access to bio-impedance analysis (BIA) and dual-energy X-ray absorptiometry (DXA). This study sought to establish and evaluate machine learning (ML) models for predicting sarcopenia in CLD patients.

Methods

This study retrospectively analyzed data from the China Health and Retirement Longitudinal Study (CHARLS). Feature selection was conducted via univariate logistic regression, least absolute shrinkage and selection operator (LASSO), and variance inflation factor (VIF). Eight ML algorithms were applied to build predictive models, with performance assessed using area under the receiver operating characteristic curve (AUC), confusion matrix-derived index, calibration curves, and decision curve analysis (DCA). Feature importance was evaluated with SHapley Additive exPlanations (SHAP).

Results

712 participants from CHARLS 2015 were divided into a training set and internal validation set, with 365 patients from CHARLS 2011 used for external validation. Eight significant predictors of sarcopenia were identified: age, education level, episodic memory, waist circumference, body mass index (BMI), heart disease, hemoglobin, and mean corpuscular volume (MCV). Among the ML models, the support vector machine (SVM) performed best, with an AUC of 0.91 (95%

CI 0.88–0.93) in the training set, and 0.88 (95% CI 0.85–0.91) and 0.82 (95% CI 0.79–0.85) in internal and external validation sets, respectively. Calibration curves, Brier scores, and DCA confirmed the SVM's robust performance. SHAP analysis revealed that BMI, age, and education level were the most important features.

Conclusions
The SVM model, integrating socio-demographic and clinical factors, showed excellent performance in predicting sarcopenia risk in CLD patients. It holds promise for guiding early diagnosis and treatment to improve patient outcomes.

Keywords: Sarcopenia, Chronic liver disease, Machine learning, Predictive model

1 Introduction

Chronic liver disease (CLD) is a prevalent health issue globally, associated with significant morbidity and mortality (1). According to the World Health Organization, CLD is among the leading causes of morbidity and mortality globally, with an estimated 2 million deaths annually attributable to liver cirrhosis and liver cancer (2). In China, the burden is particularly pronounced, driven by lifestyle changes and the large number of individuals with hepatitis B virus (HBV) and hepatitis C virus (HCV) infections (3). Beyond liver-related complications like cirrhosis and hepatocellular carcinoma, CLD is increasingly associated with sarcopenia—a progressive loss of muscle mass and function that further worsens patient outcomes (4).

Sarcopenia, defined by decreased muscle mass, strength, and functionality, is now recognized as a significant health risk, particularly for those with CLD (5). Sarcopenia is linked to multiple adverse outcomes, increased mortality, longer hospital stays, and a higher incidence of complications (6, 7). Among CLD patients, sarcopenia affects approximately 20-70% of those with advanced disease stages, contributing to increased morbidity and mortality (8, 9). This impact highlights the necessity of early identification and intervention to mitigate sarcopenia's adverse effects on CLD progression and overall patient health. Despite its importance, diagnosing sarcopenia in CLD patients remains challenging. Conventional diagnostic methods based on appendicular skeletal muscle mass, such as bio-impedance analysis (BIA) and dual-energy X-ray absorptiometry (DXA), require specialized equipment often unavailable in resource-limited settings (10). Additionally, these techniques may be less accurate for CLD patients due to fluid retention and metabolic alterations affecting muscle mass measurements (11). Thus, there is a critical need for accessible, reliable tools to assess sarcopenia in CLD populations.

Machine learning (ML) has emerged as a promising tool in the realm of predictive modeling, offering innovative solutions to overcome the limitations of traditional diagnostic methods. Recent studies have demonstrated the efficacy of ML models in predicting sarcopenia using simple clinical parameters. For instance, a study by Wu et al. highlighted the potential of a ML model that utilized basic clinical information to predict sarcopenia in patients undergoing peritoneal dialysis, emphasizing the feasibility of such models in clinical settings (12). Kang et al. developed a predictive model using data from the Korea National Health and Nutrition Examination Surveys, achieving notable accuracy in identifying sarcopenia risk factors among older adults (13). This aligns with findings from Zhang et al., who emphasized the role of clinical characteristics and laboratory markers in developing effective ML-based diagnostic tools for sarcopenia (14). These findings suggest that ML can leverage existing clinical data to provide actionable insights into sarcopenia risk, particularly in populations affected by chronic conditions.

This study aims to create a ML-based predictive model for sarcopenia risk assessment in CLD patients, utilizing data from the China Health and Retirement Longitudinal Study (CHARLS). By focusing on accessible socio-demographic and clinical indicators, this model aspires to provide a practical tool for early intervention, improving patient outcomes and addressing sarcopenia's health burden among those with CLD.

2 Methods

2.1 Study population

This study investigated retrospective data from the China Health and Retirement Longitudinal Study (CHARLS), which is dedicated to examining various aspects of health, economic status, and retirement among people 45 of age and older across China. CHARLS encompasses a rich dataset collected from a diverse cohort, involving families from 450 communities in 28 provinces, making it a robust resource for understanding aging in the Chinese population. The study protocol underwent thorough evaluation and received approval from the Biomedical Ethics Committee of Peking University (Approval number: IRB00001052-11015). In accordance with ethical norms, signed informed consent was secured from all participants before their inclusion in the study (15). The dataset is publicly accessible for further

research at http://charls.pku.edu.cn. Patients with CLD were identified through self-reported data in the CHARLS questionnaire. A total of 712 CLD patients were included from the CHARLS 2015 dataset, and 365 patients were drawn from the CHARLS 2011 dataset, with 94 (13.18%) and 49 (13.42%) patients diagnosed with sarcopenia in the respective datasets.

2.2 Outcomes

Sarcopenia was defined using the 2019 criteria from the Asian Working Group for Sarcopenia (AWGS), focusing on muscle strength, appendicular skeletal muscle mass (ASM), and physical performance (16). Hand grip strength was measured using the YuejianTM WL-1000 dynamometer, where each participant's dominant and non-dominant hand was tested twice, resulting in four total measurements. The two highest values from these trials were selected, and their average was calculated to determine the participant's grip strength. The thresholds for low grip strength were set at < 28 kg for men and < 18 kg for women (15). The equation for ASM computation was as followed:

$ASM = 0.193 \times weight(kg) + 0.107 \times height(cm) - 4.157 \times gender - 0.037 \times age(years) - 2.631$, where male is coded as 1 and female as 2. The skeletal muscle mass index (SMI), calculated as ASM divided by height squared (17), has the determination threshold for low muscle mass were established at < 7.28 kg/m² for men and < 5.55 kg/m² for women. The Short Physical Performance Battery (SPPB) was applied to evaluate physical performance, encompassing assessments of standing balance (held for 10 seconds in semi-tandem, side-by-side, and tandem positions), walking speed (timed over a 2.5 m distance with an average of two trials recorded), and a 5-repetition chair stand test (duration to complete five consecutive stands) (16). A score of ≤ 9 on the SPPB, a chair stand time ≥ 12 seconds, or a walking speed < 1.0 m/s indicated low physical performance. Individuals who met the criteria for low muscle mass in conjunction with either low muscle strength or low physical performance were classified as having sarcopenia (16).

2.3 Extraction of the potential predictors

Potential predictive variables, including socio-demographic factors and clinical factors, were screened for analysis based on clinical relevance and insights derived from earlier research (18-22). Socio-demographic factors included age, gender, education level, marital status and residence. Clinical factors encompassed the outcomes of diverse medical scale assessments, existing medical conditions, anthropometric measurements, and laboratory parameters.

Mental health was gauged via the Center for Epidemiologic Studies Depression 10 (CESD10) scale, with scores ranging from 0 to 30. Smoking and drinking status were recorded as binary variables (no/yes). Life satisfaction was graded using a five-point scale, from “completely satisfied” to “not at all satisfied”. Episodic memory was assessed through participants’ recall of a list of 10 words, yielding scores from 0 to 10. Mental status was evaluated through a cognitive assessment with a maximum score of 11. The Instrumental Activities of Daily Living (IADL) score was derived from a specific scale, with scores ranging from 0 to 5, indicating varying levels of dependency.

Anthropometric measurements comprised body mass index (BMI, in kg/m²) and waist circumference (in cm). Hypertension was characterized by self-reporting physician diagnosis, using antihypertensive medicines, and having a blood pressure of 140/90 mmHg or higher. Diabetes was defined by self-reporting physician diagnosis, usage of hypoglycemic medications, fasting blood glucose ≥ 126 mg/dL, and/or HbA1c level ≥ 6.5%. Other medical conditions and chronic diseases such as hypertension, diabetes, disability, chronic lung disease, heart disease, stroke, arthritis or rheumatism, dyslipidemia, kidney disease, digestive disease, and asthma were classified based on self-reported diagnoses. Laboratory parameters measured included total cholesterol (TC), triglycerides (TG), high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), uric acid, hematocrit, hemoglobin, white blood cell count (WBC), mean corpuscular volume (MCV), platelet count, serum creatinine, and cystatin C, each reported in their respective units.

2.4 Establishment and evaluation of machine learning models

First, we conducted a feature selection procedure to identify significant predictors using univariate logistic regression and least absolute shrinkage and selection operator (LASSO) regression techniques. Subsequently, we employed the variance inflation factor (VIF) to identify and exclude predictors with strong multicollinearity (VIF□10).

Second, the remaining predictive factors were input into eight ML algorithms: decision tree (DT), k-nearest neighbors (KNN), light gradient boosting machine (LGBM), logistic regression (LR), naive Bayes (NB), random forest (RF), support vector machine (SVM), and extreme gradient boosting (XGB). The receiver operating characteristic (ROC) curves, calibration curves and decision curve analysis (DCA) were performed, and the area under the receiver operating characteristic curve (AUC) was the primary metric used for comparing model performance. Confusion matrix-derived index, including accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score, were calculated to assess the predictive performance of these machine learning models across the training set, internal validation set, and external validation set.

Finally, we utilized SHapley Additive exPlanations (SHAP) approach to clarify the contributing roles of participants' features to the predictions. The SHAP feature importance plot demonstrated the global significance of each feature, where larger mean absolute SHAP values indicated greater relevance to model predictions. Additionally, the SHAP summary plot illustrated the impact of each feature on the model, with each point representing a patient's feature SHAP value, color-coded from orange to purple to reflect the range of feature values.

2.5 Statistical analyses

Data were analyzed based on the distribution characteristics of the variables. Continuous variables were expressed as mean ± standard deviation or median with interquartile range (IQR), based on their adherence to a normal distribution. Counts and percentages were utilized to report categorical variables. Group comparisons were performed using the Chi-square test for categorical data and the independent samples t-test for continuous data. All statistical analyses were executed using EmpowerStat (version 6.0) and DCPM (version 5.3.8, Jingding Medical Technology Co., Ltd). A two-tailed approach was adopted for all tests, with a significance threshold set at $P < 0.05$.

3 Results

3.1 Participants' characteristics

The flowchart of the study is outlined in ***Figure 1***. After applying rigorous inclusion and exclusion criteria and handling missing data through multiple imputation, we initially identified 712 individuals with chronic liver disease from the CHARLS 2015 dataset. The sample was partitioned into a training set comprising 497 patients and an internal validation set consisting of 215 patients, utilizing a 7:3 random allocation method. Additionally, we included an external validation cohort of 365 chronic liver disease patients from the CHARLS 2011 dataset to mitigate the overfitting effect.

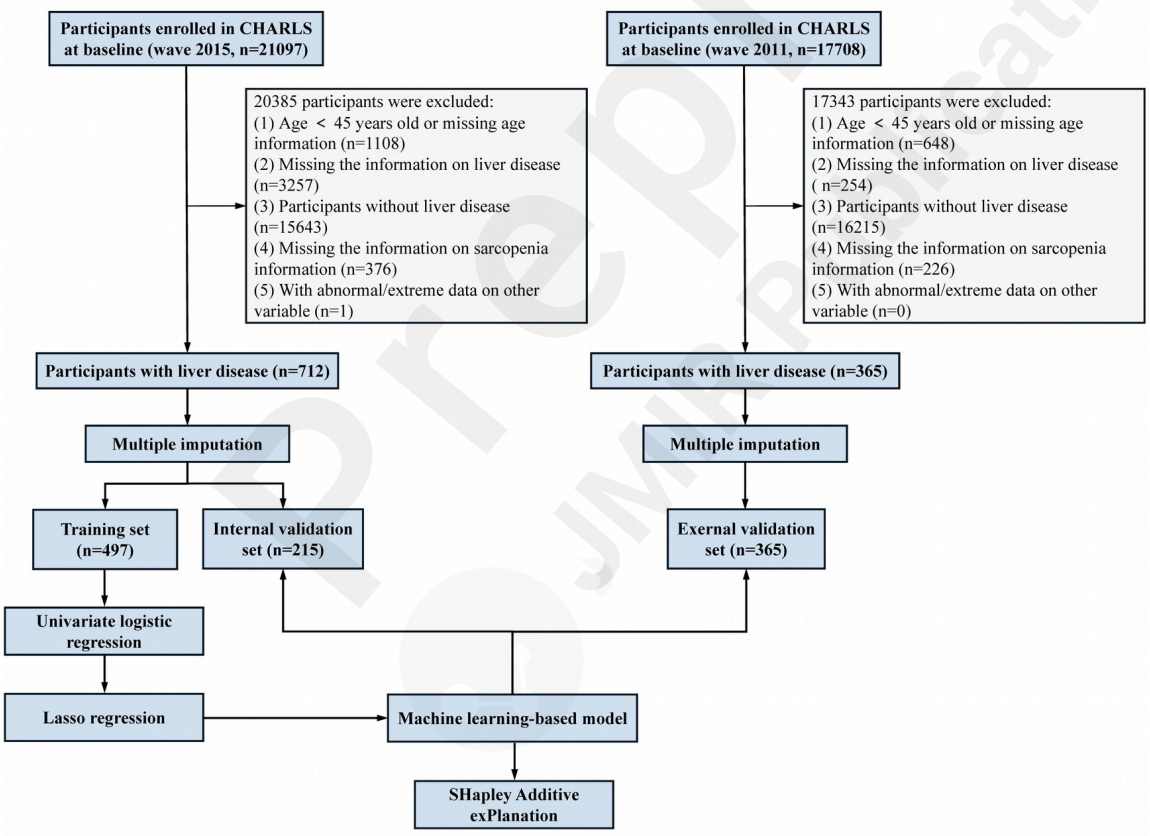


Figure 1. Flowchart of the study population.
Note: CHARLS, China Health and Retirement Longitudinal Study.

The sample included 255 men (51.31%) and 242 women (48.69%) in the training set, while the internal validation set had 103 men (47.91%) and 112 women (52.09%). In the external validation set, there were 198 men (54.25%) and 167 women (45.75%). The median age was consistent across the training and internal validation sets at 61 years, whereas the external validation set had a slightly younger median age of 58 years. The participants' socio-demographic and clinical characteristics were comprehensively outlined in ***Table 1***.

Table 1. Baseline characteristics of the study samples

Characteristics	Training set (n = 497)	Internal validation set (n = 215)	External validation set (n = 365)	P-value
Age, years	61.00 (54.00,67.00)	61.00 (55.00,66.00)	58.00 (52.00,64.00)	<0.001
Gender (%)				0.330
Female	242 (48.69)	112 (52.09)	167 (45.75)	
Male	255 (51.31)	103 (47.91)	198 (54.25)	
Education level, n (%)				0.969
Illiterate	211 (42.45)	90 (41.86)	153 (41.92)	
Primary school	108 (21.73)	46 (21.40)	72 (19.73)	
Middle school	107 (21.53)	44 (20.47)	80 (21.92)	
High school and above	71 (14.29)	35 (16.28)	60 (16.44)	
Marital status (%)				0.576
Unmarried, n (%)	63 (12.68)	24 (11.16)	38 (10.41)	
Married, n (%)	434 (87.32)	191 (88.84)	327 (89.59)	
Residence, n (%)				0.171
Urban	207 (41.65)	74 (34.42)	150 (41.10)	
Rural	290 (58.35)	141 (65.58)	215 (58.90)	
CESD-10	9.00 (4.00,14.00)	8.00 (3.00,13.00)	9.00 (4.00,15.00)	0.385
Smoking status, n (%)				0.072
No	366 (73.64)	166 (77.21)	251 (68.77)	
Yes	131 (26.36)	49 (22.79)	114 (31.23)	
Drinking status, n (%)				0.074
No	318 (63.98)	152 (70.70)	257 (70.41)	
Yes	179 (36.02)	63 (29.30)	108 (29.59)	
Life satisfaction, n (%)				<0.001
Completely satisfied	10 (2.01)	5 (2.33)	19 (5.21)	
Very satisfied	49 (9.86)	16 (7.44)	55 (15.07)	
Somewhat satisfied	263 (52.92)	112 (52.09)	246 (67.40)	
Not very satisfied	150 (30.18)	71 (33.02)	41 (11.23)	
Not at all satisfied	25 (5.03)	11 (5.12)	4 (1.10)	
Episodic memory	3.50 (2.00,5.00)	3.50 (2.00,5.00)	3.50 (2.50,4.50)	0.270
Mental status	9.00 (6.00,10.00)	9.00 (7.00,10.00)	9.00 (6.00,10.00)	0.407
IADL score	0.00 (0.00,1.00)	0.00 (0.00,1.00)	0.00 (0.00,1.00)	0.896
Body mass index, kg/m^2	24.57 (22.43,27.30)	24.30 (22.17,27.41)	24.12 (22.05,26.51)	0.173
Waist, cm	88.20 (81.90,96.30)	89.00 (81.00,96.10)	87.40 (80.00,94.00)	0.057
Hypertension, n (%)				0.005
No	205 (41.25)	90 (41.86)	189 (51.78)	
Yes	292 (58.75)	125 (58.14)	176 (48.22)	
Diabetes, n (%)				0.021
No	363 (73.04)	161 (74.88)	296 (81.10)	

Yes	134 (26.96)	54 (25.12)	69 (18.90)	
Disability, n (%)				<0.001
No	288 (57.95)	125 (58.14)	287 (78.63)	
Yes	209 (42.05)	90 (41.86)	78 (21.37)	
Chronic lung disease, n (%)				0.004
No	376 (75.65)	160 (74.42)	307 (84.11)	
Yes	121 (24.35)	55 (25.58)	58 (15.89)	
Heart disease, n (%)				<0.001
No	318 (63.98)	143 (66.51)	277 (75.89)	
Yes	179 (36.02)	72 (33.49)	88 (24.11)	
Stroke, n (%)				0.665
No	471 (94.77)	203 (94.42)	350 (95.89)	
Yes	26 (5.23)	12 (5.58)	15 (4.11)	
Arthritis or rheumatism, n (%)				0.002
No	205 (41.25)	88 (40.93)	192 (52.60)	
Yes	292 (58.75)	127 (59.07)	173 (47.40)	
Dyslipidemia, n (%)				<0.001
No	329 (66.20)	149 (69.30)	306 (83.84)	
Yes	168 (33.80)	66 (30.70)	59 (16.16)	
Kidney disease, n (%)				<0.001
No	362 (72.84)	162 (75.35)	314 (86.03)	
Yes	135 (27.16)	53 (24.65)	51 (13.97)	
Digestive disease, n (%)				<0.001
No	230 (46.28)	101 (46.98)	230 (63.01)	
Yes	267 (53.72)	114 (53.02)	135 (36.99)	
Asthma, n (%)				0.453
No	451 (90.74)	192 (89.30)	337 (92.33)	
Yes	46 (9.26)	23 (10.70)	28 (7.67)	
TC, mg/dL	177.61 (158.69,198.84)	177.22 (154.83,199.61)	188.08 (169.33,203.44)	<0.001
TG, mg/dL	123.01 (92.04,168.14)	130.97 (91.15,175.22)	112.50 (87.61,149.62)	0.028
HDL-C, mg/dL	49.42 (43.63,55.68)	48.17 (42.47,54.08)	47.17 (40.98,56.06)	0.081
LDL-C, mg/dL	94.59 (81.47,114.29)	93.82 (81.85,112.93)	111.34 (94.72,126.36)	<0.001
Uric acid, mg/dL	4.96 (4.20,5.70)	5.10 (4.30,5.70)	4.56 (3.95,5.08)	<0.001
Hematocrit, %	41.95 (39.30,44.40)	42.11 (38.70,44.70)	42.24 (39.50,44.70)	0.585
Hemoglobin, g/dL	13.90 (13.06,14.90)	13.80 (12.90,15.00)	14.90 (13.90,15.60)	<0.001
White blood cell, 10^9/L	5.71 (5.00,6.60)	5.79 (5.00,6.50)	5.82 (5.10,6.47)	0.729
MCV, fl	92.00 (88.90,94.85)	92.70 (89.75,96.45)	91.19 (88.64,94.40)	0.014
Platelet, 10^9/L	190.07 (154.00,223.00)	190.00 (166.50,220.50)	197.30 (165.00,224.08)	0.080
Serum creatinine, mg/dL	0.77 (0.67,0.89)	0.78 (0.68,0.90)	0.78 (0.70,0.88)	0.847
Cystatin C, mg/L	0.84 (0.75,0.94)	0.84 (0.77,0.95)	1.01 (0.94,1.09)	<0.001

Note: CESD-10, Center for epidemiologic studies depression scale-10; IADL; Instrumental activities of daily living; TC, total cholesterol; TG, triglycerides; HDL-C, high density lipoprotein cholesterol; LDL-C, low density lipoprotein

cholesterol; MCV, mean corpuscular volume. Values presented as median (interquartile range), n (%).

3.2 Selection of predictors

Table 2 illustrated the potential factors for predicting sarcopenia in patients with CLD. For predictors selection, 8 predictors (age, education level, episodic memory, waist, BMI, heart disease, hemoglobin, MCV) emerged as the most significant features for establishing a prediction model through LASSO regularization, utilizing the tenfold cross-validation method. ***Figure 2*** showed the coefficients associated with each variable in the LASSO regression model. Notably, the variance inflation factors for these 8 variables were all less than 10 (***Table S1***), indicating the absence of multicollinearity and further substantiating the validity of the selected predictors.

Table 2. Results of univariate logistic regression analysis in the training set

Variables	β	Standard Error	Odds Ratio (95%Confidence)	Z	*P-value*
Age	0.064	0.015	1.067(1.035-1.100)	4.208	<0.001
Education level	-0.472	0.143	0.624(0.465-0.816)	-3.299	0.001
Rural residence	0.981	0.310	2.667(1.487-5.050)	3.165	0.002
Episodic memory	-0.467	0.084	0.627(0.528-0.736)	-5.531	<0.001
Mental status	-0.199	0.047	0.820(0.747-0.899)	-4.206	<0.001
Waist	-0.071	0.012	0.932(0.909-0.951)	-6.105	<0.001
Body mass index	-1.474	0.189	0.229(0.152-0.319)	-7.813	<0.001
Hypertension	-0.587	0.268	0.556(0.328-0.939)	-2.193	0.028
Diabetes	-1.421	0.441	0.242(0.091-0.531)	-3.220	0.001
Heart disease	-1.024	0.335	0.359(0.179-0.670)	-3.059	0.002
Dyslipidemia	-1.154	0.358	0.315(0.148-0.610)	-3.224	0.001
TC	-0.013	0.004	0.987(0.979-0.995)	-3.174	0.002
TG	-0.019	0.004	0.981(0.974-0.988)	-5.084	<0.001
HDL-C	0.048	0.012	1.049(1.025-1.073)	4.091	<0.001
LDL-C	-0.017	0.006	0.983(0.972-0.994)	-3.025	0.002
Hematocrit	-0.088	0.028	0.916(0.865-0.967)	-3.095	0.002
Hemoglobin	-0.333	0.082	0.717(0.608-0.838)	-4.081	<0.001
White blood cell	-0.207	0.098	0.813(0.666-0.980)	-2.105	0.035
MCV	0.049	0.023	1.050(1.006-1.099)	2.139	0.032
Cystatin C	0.861	0.419	2.367(1.075-5.848)	2.059	0.040

Note: TC, total cholesterol; TG, triglycerides; HDL-C, high density lipoprotein cholesterol; LDL-C, low density lipoprotein cholesterol; MCV, mean corpuscular volume.

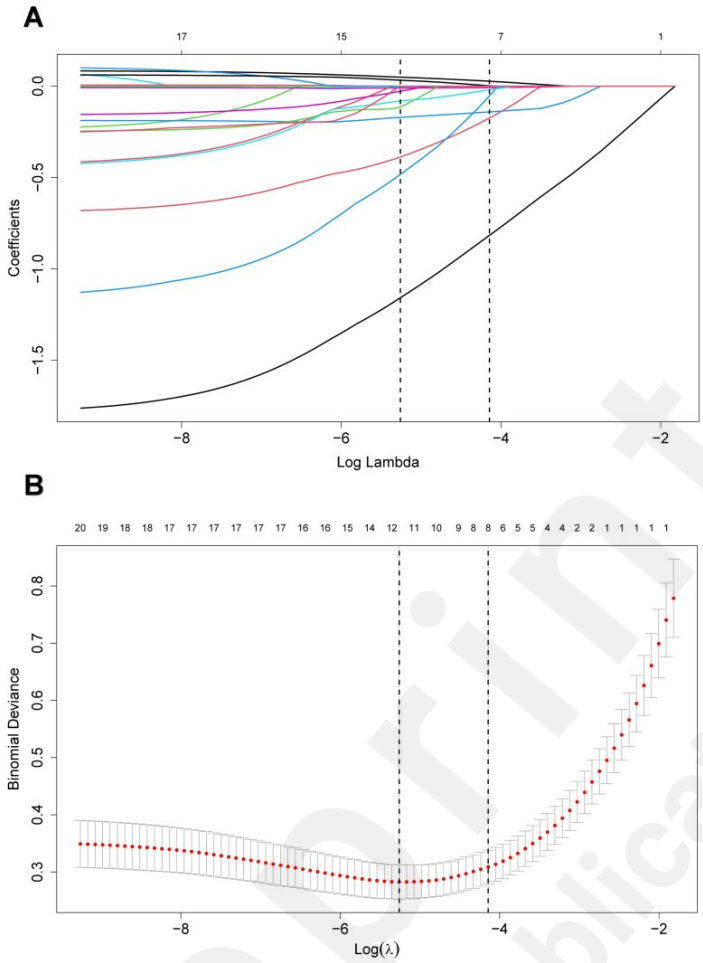


Figure 2. Identification of key predictors for sarcopenia in patients with liver disease by the LASSO regression model. (A) Coefficient profile based on the logarithmic (lambda) sequence, with non-zero coefficients resulting from the chosen optimal lambda. (B) Optimal lambda parameter determined through tenfold cross-validation, accompanied by a graph displaying the partial likelihood deviation curve as a function of log(lambda), with a vertical line indicating the optimal value based on the 1-standard error of minimum criterion.

3.3 Comparison and interpretation of ML-based model

All these 8 predictors were inputted into DT, KNN, LGBM, LR, NB, RF, SVM, and XGB models. The evaluation of the eight machine learning models aimed at predicting sarcopenia risk across training, internal validation, and external validation sets is detailed in ***Table 3*** and illustrated in ***Figure 3***. RF, XGB, KNN, and LGBN model demonstrated a propensity for overfitting, as evidenced by their superior performance on the training dataset compared to a significant decrease in accuracy on both internal and external validation sets. Figure 3 demonstrated the comparison if ROC curves for different ML models addressing sarcopenia across training, internal validation and external validation sets. After synthesizing multiple evaluation metrics, it was concluded that the SVM model demonstrated the utmost superior performance in predicting sarcopenia for CLD patients, particularly due to its robust predictive accuracy across training, internal validation, and external validation sets, with the AUC values remaining the highest among all models with good stability.

Table 3. Diagnostic performance of the machine learning models for sarcopenia in participants with liver disease

Models	Accuracy	Sensitivity	Specificity	PPV	NPV	F1 score
Training set						
SVM	0.899	1.000	0.884	0.565	1.000	0.722
LR	0.928	0.969	0.921	0.649	0.995	0.778
DT	0.960	0.909	0.966	0.769	0.988	0.833
RF	1.000	1.000	1.000	1.000	1.000	1.000
XGB	0.972	0.923	0.979	0.870	0.988	0.896

KNN	1.000	1.000	1.000	1.000	1.000	1.000
LGBN	0.960	0.938	0.963	0.792	0.990	0.859
NBM	0.914	0.969	0.905	0.606	0.995	0.746
Internal validation set						
SVM	0.884	0.966	0.871	0.538	0.994	0.691
LR	0.902	0.931	0.898	0.587	0.988	0.720
DT	0.921	0.688	0.962	0.759	0.946	0.721
RF	0.907	0.724	0.935	0.636	0.956	0.677
XGB	0.893	0.759	0.914	0.579	0.960	0.657
KNN	0.893	0.621	0.935	0.600	0.941	0.610
LGBN	0.893	0.828	0.903	0.571	0.971	0.676
NBM	0.856	0.759	0.871	0.478	0.959	0.587
External validation set						
SVM	0.901	1.000	0.886	0.576	1.000	0.731
LR	0.921	0.959	0.915	0.635	0.993	0.764
DT	0.929	0.709	0.968	0.796	0.949	0.750
RF	0.940	0.898	0.946	0.721	0.984	0.800
XGB	0.923	0.939	0.921	0.648	0.990	0.767
KNN	0.890	0.592	0.937	0.592	0.937	0.592
LGBN	0.901	0.959	0.892	0.580	0.993	0.723
NBM	0.874	0.816	0.883	0.519	0.969	0.635

Note: SVM, Support Vector Machine; LR, Logistic Regression; DT, Decision Tree; RF, Random Forest; XGB, Extreme Gradient Boosting; KNN, K-Nearest Neighbors; LGBM, LightGBM; NBM, Naive Bayes Model.

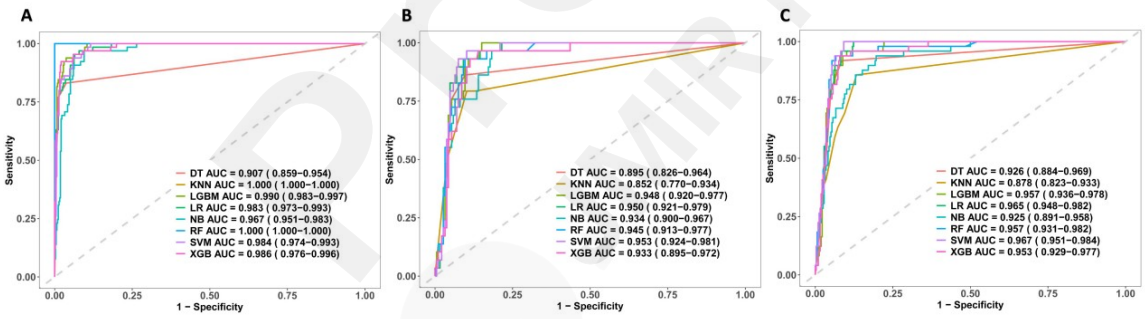


Figure 3. ROC analyses of applied machine learning models. (A) Training set. (B) Internal validation set. (C) External validation set. Note: DT, decision tree; KNN, K-nearest neighbors; LGBM, LightGBM; LR, logistic regression; NB, naive bayes model; RF, random forest; SVM, support vector machine; XGB, eXtreme Gradient Boosting.

Also, we constructed calibration curves, Brier score calculation and DCA to investigate these models' calibration and clinical utility (***Figure S1***, ***Figure S2***, and ***Table S2***). Among all these models, the SVM demonstrated notable performance in both calibration and clinical applicability. Decision curve analysis showed that it yielded a greater net benefit compared to a treat all-or-none strategy with a risk threshold of < 70%.

To elucidate how the SVM model was predictive of sarcopenia in the CLD population, we conducted an additional analysis using SHAP (***Figure 4***), a widely recognized method for assessing feature importance. As illustrated in Figure 4A, the bar graph showcased the importance of each feature as evaluated by SHAP values, with mean absolute SHAP values used to rank the eight most significant variables contributing to the predictive model. In order, the top three most important variables were BMI, age and education level. Additionally, Figure 4B revealed the correlation between feature

values and their respective SHAP values within the dataset, where each point represents the feature value and SHAP value for individual patients. We can see that the increases in BMI and education level have a positive impact and push the prediction toward reduced sarcopenia prevalence, whereas the increase in age has a negative impact and pushes the prediction toward elevated sarcopenia odds.

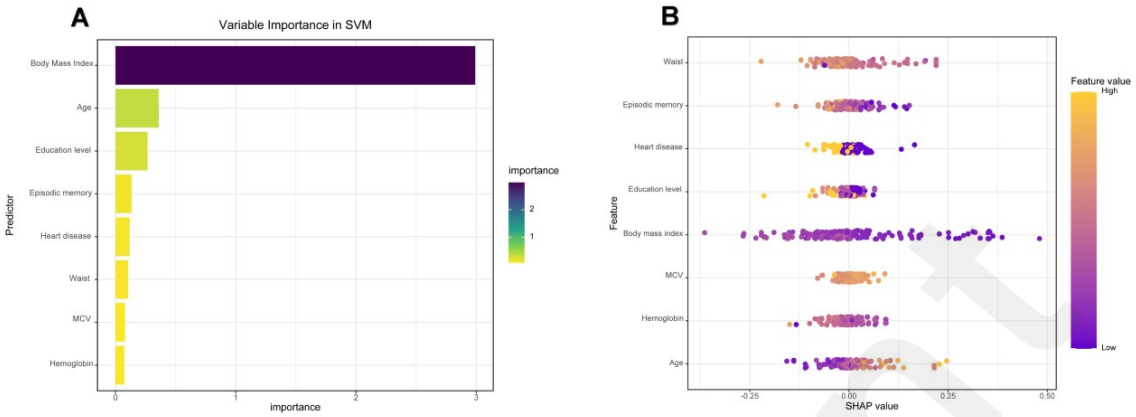


Figure 4. SHAP analysis of SVM model. (A) Feature importance ranking based on SHAP values, represented in a matrix diagram to illustrate the significance of each covariate in SVM model. (B) Characteristic attributes visualized by SHAP, with the abscissa representing SHAP values; each line denotes a feature, with orange dots indicating higher eigenvalues and purple dots indicating lower eigenvalues. SHAP, Shapley additive explanations; SVM, support vector machine; MCV, mean corpuscular volume.

4 Discussion

In this study, we developed and validated ML-based predictive models by utilizing socio-demographic and clinical data from the China Health and Retirement Longitudinal Study (CHARLS) 2011 and 2015 datasets. Among the eight ML models evaluated, the support vector machine (SVM) model emerged as the most effective, demonstrating superior performance metrics including AUROC. This model's optimization through automated parameter tuning and rigorous cross-validation ensured its robustness and reliability. Additionally, SHAP noted that, in order, the top three most important variables were BMI, age and education level.

Traditional logistic regression (LR) models are limited in their ability to handle numerous variables and often fail to capture complex, nonlinear relationships within data, resulting in suboptimal predictive accuracy, especially when dealing with multifactorial conditions (23). In contrast, ML approaches can process a greater number of predictors and uncover novel patterns in the data, offering a more nuanced understanding of disease risk (24, 25). As a classical ML algorithm, the SVM model demonstrated remarkable predictive capabilities, outperforming conventional risk assessment tools such as logistic regression (26).

Previous studies have employed logistic regression and diverse ML algorithms to distinguish between individuals at elevated risk and those without concerning sarcopenia. In a study including older patients who had patellar fractures, for example, Chen et al. established a novel predictive nomogram for postoperative sarcopenia, with a low body mass index and advanced age being the primary risk variables (27). Zou and Shao developed a predictive model for diabetic patients using the CHARLS dataset (21). Li et al. focused on elderly Chinese, also utilizing CHARLS data (22). These studies often overlooked the unique challenges faced by patients with CLD, while our study stands out by applying ML techniques specifically to CLD population, adding diversity to existing prediction models. Additionally, we diverge from Wu et al. (12), who predicted sarcopenia in peritoneal dialysis patients using basic clinical information, and Kang et al. (13), who used random forest models for postmenopausal women. Our research not only employed advanced ML algorithms but also comprehensively compared them to traditional logistic regression models. Although Zhang et al. developed an efficient ML model with clinical and laboratory data (14), our study included both internal and external validation using CHARLS 2011 and 2015 datasets. Our rigorous cross-validation and parameter optimization enhance the model's robustness, a step not always emphasized in previous studies. Finally, our SHAP value analysis provides a novel interpretation of feature contributions, offering deeper insights into the interplay of clinical and socio-demographic variables related to sarcopenia. These aspects position our research as a significant contribution to the field, improving predictive modeling and early identification of sarcopenia.

In our study, the SHAP results indicated that age emerged as the primary risk factor, with older individuals exhibiting a higher odds of developing sarcopenia, aligning with previous findings (28， 29). Previous studies have emphasized that older individuals are particularly vulnerable to sarcopenia due to age-related hormonal imbalances, inflammation, muscle

regenerative decline and functional deterioration (30-32). Furthermore, our results indicated that lower educational attainment is linked to a heightened risk of sarcopenia, which is consistent with previous study (33, 34). This negative relationship may be attributed to socio-economic factors including occupation, economic conditions, and living environment, which influence access to healthcare services, exposure to risk factors, and psychosocial well-being (35). What's more, our study demonstrated that body mass index (BMI) and waist circumference were both inversely related with sarcopenia prevalence, suggesting that lower body mass and central adiposity may contribute to muscle loss in patients with CLD. This relationship can be found in other cross-sectional or cohort studies (36-38). However, some studies indicate the opposite, revealing a possible connection between obesity and the onset of sarcopenia (39). One possible explanation for these divergent findings could lie in the role of adipose inflammation, which may result in the redistribution of intra-abdominal fat and fat infiltration into muscle tissue. This, in turn, could exacerbate muscle loss, contributing to the pathogenesis of sarcopenic obesity (40).

Although our study yields encouraging findings, it is crucial to recognize a number of limitations. First, the SVM model was developed using data primarily from Chinese patients, which limits its generalizability to other racial and ethnic groups. Therefore, further validation in diverse populations is necessary to assess the model's broader applicability. Second, this study employed a cross-sectional design, which, although useful for predicting the risk of sarcopenia, can only establish associations and cannot infer causal relationships between predictors and outcomes. Moreover, the retrospective approach to data collection across various centers led to occurrences of incomplete data. While this limitation was addressed through stringent inclusion and exclusion criteria and a reasonably large sample size, the sample still has some constraints due to the exclusion of participants with incomplete data. Future studies with larger sample sizes and prospective international multicenter designs are essential to enhance the validation of the SVM model's performance and its applicability across different populations.

In conclusion, this study successfully established a SVM model for assessing sarcopenia risk in CLD patients, utilizing accessible socio-demographic and clinical data to achieve high predictive accuracy. Future research should aim to integrate additional data sources, such as genetic and lifestyle factors, to further enhance model accuracy. Broader validation studies are also necessary to confirm these findings and ensure the model's robustness across diverse patient populations. By continuing to refine and validate these predictive models, we can improve early detection and intervention strategies for sarcopenia in patients with CLD, ultimately enhancing patient outcomes and quality of life.

Abbreviations

BIA	Bio-impedance analysis
DXA	Dual-energy X-ray absorptiometry
CLD	Chronic liver disease
ML	Machine learning
CHARLS	China Health and Retirement Longitudinal Study
LASSO	Least absolute shrinkage and selection operator
VIF	Variance inflation factor
AUC	Area under the receiver operating characteristic curve
DCA	Decision curve analysis
SHAP	SHapley Additive exPlanations
BMI	Body mass index
MCV	Mean corpuscular volume
SVM	Support vector machine
ASM	Appendicular skeletal muscle mass
SPPB	Short Physical Performance Battery
CESD10	Center for Epidemiologic Studies Depression 10
IADL score	Instrumental Activities of Daily Living score
BMI	Body mass index
TC	Total cholesterol
TG	Triglycerides
HDL-C	High-density lipoprotein cholesterol
LDL-C	Low-density lipoprotein cholesterol
WBC	White blood cell count
MCV	Mean corpuscular volume
DT	Decision tree
KNN	k-nearest neighbors
LGBM	Light gradient boosting machine
NB	Naive Bayes

RF	Random forest
XGB	Extreme gradient boosting
ROC curve	Receiver operating characteristic curve
PPV	Positive predictive value
NPV	Negative predictive value
IQR	Interquartile range

Acknowledgments
The present study used data from the CHARLS. The authors thank the CHARLS research team and individuals included in the study. CHARLS is supported by Peking University, the National Natural Science Foundation of China, the National Institute on Aging, and the World Bank. The authors thank the creators of EmpowerStats and DCPM for their pivotal support and advanced analytical tools.

Ethic Statement
The authors confirm adherence to the ethical guidelines for publication set forth by the Journal of Cachexia, Sarcopenia and Muscle: update 2019. Guidelines for ethical conduct at both national and international levels were adhered to, including the Deontological Code of Ethics and the 1964 Declaration of Helsinki along with its subsequent amendments. All patients provided written consent for the use of their data, and the study protocol of the CHARLS received approval from the Ethical Review Committee of Peking University (approval number: IRB00001052-11015).

Data Availability Statement
The data from the CHARLS study are freely accessible via the URL: https://charls.pku.edu.cn/en/.

Conflicts of Interests
All authors have no conflict of interest related to this publication.

References

1. Devarbhavi H, Asrani SK, Arab JP, Nartey YA, Pose E, Kamath PS. Global burden of liver disease: 2023 update. J Hepatol. 2023;79(2):516-37.
2. Global burden of 288 causes of death and life expectancy decomposition in 204 countries and territories and 811 subnational locations, 1990-2021: a systematic analysis for the Global Burden of Disease Study 2021. Lancet. 2024;403(10440):2100-32.
3. Xiao J, Wang F, Wong NK, He J, Zhang R, Sun R, et al. Global liver disease burdens and research trends: Analysis from a Chinese perspective. J Hepatol. 2019;71(1):212-21.
4. Bunchorntavakul C. Sarcopenia and Frailty in Cirrhosis: Assessment and Management. Med Clin North Am. 2023;107(3):589-604.
5. Tandon P, Montano-Loza AJ, Lai JC, Dasarathy S, Merli M. Sarcopenia and frailty in decompensated cirrhosis. J Hepatol. 2021;75 Suppl 1(Suppl 1):S147-s62.
6. Ooi PH, Hager A, Mazurak VC, Dajani K, Bhargava R, Gilmour SM, et al. Sarcopenia in Chronic Liver Disease: Impact on Outcomes. Liver Transpl. 2019;25(9):1422-38.
7. Zeng X, Shi ZW, Yu JJ, Wang LF, Luo YY, Jin SM, et al. Sarcopenia as a prognostic predictor of liver cirrhosis: a multicentre study in China. J Cachexia Sarcopenia Muscle. 2021;12(6):1948-58.
8. Carey EJ, Lai JC, Wang CW, Dasarathy S, Lobach I, Montano-Loza AJ, et al. A multicenter study to define sarcopenia in patients with end-stage liver disease. Liver Transpl. 2017;23(5):625-33.
9. Kim G, Kang SH, Kim MY, Baik SK. Prognostic value of sarcopenia in patients with liver cirrhosis: A systematic review and meta-analysis. PLoS One. 2017;12(10):e0186990.
10. Montano-Loza AJ, Duarte-Rojo A, Meza-Junco J, Baracos VE, Sawyer MB, Pang JX, et al. Inclusion of Sarcopenia Within MELD (MELD-Sarcopenia) and the Prediction of Mortality in Patients With Cirrhosis. Clin Transl Gastroenterol. 2015;6(7):e102.
11. Traub J, Bergheim I, Eibisberger M, Stadlbauer V. Sarcopenia and Liver Cirrhosis-Comparison of the European Working Group on Sarcopenia Criteria 2010 and 2019. Nutrients. 2020;12(2).
12. Wu J, Lin S, Guan J, Wu X, Ding M, Shen S. Prediction of the sarcopenia in peritoneal dialysis using simple clinical information: A machine learning-based model. Semin Dial. 2023;36(5):390-8.
13. Kang YJ, Yoo JI, Ha YC. Sarcopenia feature selection and risk prediction using machine learning: A cross-sectional study. Medicine (Baltimore). 2019;98(43):e17699.
14. Zhang H, Yin M, Liu Q, Ding F, Hou L, Deng Y, et al. Machine and deep learning-based clinical characteristics and laboratory markers for the prediction of sarcopenia. Chin Med J (Engl). 2023;136(8):967-73.
15. Zhao Y, Hu Y, Smith JP, Strauss J, Yang G. Cohort profile: the China Health and Retirement Longitudinal Study

(CHARLS). Int J Epidemiol. 2014;43(1):61-8.
16. Chen LK, Woo J, Assantachai P, Auyeung TW, Chou MY, Iijima K, et al. Asian Working Group for Sarcopenia: 2019 Consensus Update on Sarcopenia Diagnosis and Treatment. J Am Med Dir Assoc. 2020;21(3):300-7.e2.
17. Wen X, Wang M, Jiang CM, Zhang YM. Anthropometric equation for estimation of appendicular skeletal muscle mass in Chinese adults. Asia Pac J Clin Nutr. 2011;20(4):551-6.
18. Tuo S, Yeo YH, Chang R, Wen Z, Ran Q, Yang L, et al. Prevalence of and associated factors for sarcopenia in patients with liver cirrhosis: A systematic review and meta-analysis. Clin Nutr. 2024;43(1):84-94.
19. Shafiee G, Ostovar A, Maleki Birjandi S, Nabipour I, Larijani B, Heshmat R. Development of a Simple and Practical Screening Tool for Detection of Sarcopenia in Older People: The Bushehr Elderly Health Program. Front Med (Lausanne). 2021;8:655759.
20. Cai G, Ying J, Pan M, Lang X, Yu W, Zhang Q. Development of a risk prediction nomogram for sarcopenia in hemodialysis patients. BMC Nephrol. 2022;23(1):319.
21. Zou M, Shao Z. Construction and evaluation of sarcopenia risk prediction model for patients with diabetes: a study based on the China health and retirement longitudinal study (CHARLS). Diabetol Metab Syndr. 2024;16(1):230.
22. Li Q, Cheng H, Cen W, Yang T, Tao S. Development and validation of a predictive model for the risk of sarcopenia in the older adults in China. Eur J Med Res. 2024;29(1):278.
23. Lei D, Zhang H, Liu H, Li Z, Wu Y. Maximal uncorrelated multinomial logistic regression. IEEE Access. 2019;7:89924-35.
24. Ming C, Viassolo V, Probst-Hensch N, Chappuis PO, Dinov ID, Katapodi MC. Machine learning techniques for personalized breast cancer risk prediction: comparison with the BCRAT and BOADICEA models. Breast Cancer Res. 2019;21(1):75.
25. Ratul IJ, Wani UH, Nishat MM, Al-Monsur A, Ar-Rafi AM, Faisal F, et al. Survival Prediction of Children Undergoing Hematopoietic Stem Cell Transplantation Using Different Machine Learning Classifiers by Performing Chi-Square Test and Hyperparameter Optimization: A Retrospective Analysis. Comput Math Methods Med. 2022;2022:9391136.
26. Hamet P, Tremblay J. Artificial intelligence in medicine. metabolism. 2017;69:S36-S40.
27. Chen YS, Cai YX, Kang XR, Zhou ZH, Qi X, Ying CT, et al. Predicting the risk of sarcopenia in elderly patients with patellar fracture: development and assessment of a new predictive nomogram. PeerJ. 2020;8:e8793.
28. Wu X, Li X, Xu M, Zhang Z, He L, Li Y. Sarcopenia prevalence and associated factors among older Chinese population: Findings from the China Health and Retirement Longitudinal Study. PLoS One. 2021;16(3):e0247617.
29. Altaf S, Malmir K, Mir SM, Olyaei GR, Aftab A, Rajput TA. Prevalence and associated risk factors of sarcopenia in community-dwelling older adults in Pakistan: a cross-sectional study. BMC Geriatr. 2024;24(1):497.
30. Alway SE, Myers MJ, Mohamed JS. Regulation of satellite cell function in sarcopenia. Front Aging Neurosci. 2014;6:246.
31. van Atteveld VA, Van Ancum JM, Reijnierse EM, Trappenburg MC, Meskers CGM, Maier AB. Erythrocyte sedimentation rate and albumin as markers of inflammation are associated with measures of sarcopenia: a cross-sectional study. BMC Geriatr. 2019;19(1):233.
32. Roubenoff R, Castaneda C. Sarcopenia-understanding the dynamics of aging muscle. Jama. 2001;286(10):1230-1.
33. Zhang Y, Wang R, Chen Z, Zhou F, Su S. Educational attainment, brain cortical structure, and sarcopenia: a Mendelian randomization study. Front Public Health. 2024;12:1415343.
34. Dai S, Shu D, Meng F, Chen Y, Wang J, Liu X, et al. Higher Risk of Sarcopenia in Older Adults with Type 2 Diabetes: NHANES 1999-2018. Obes Facts. 2023;16(3):237-48.
35. Swan L, Warters A, O'Sullivan M. Socioeconomic Inequality and Risk of Sarcopenia in Community-Dwelling Older Adults. Clin Interv Aging. 2021;16:1119-29.
36. Blanco-Reina E, Ocaña-Riola R, Ariza-Zafra G, García-Merino MR, Aguilar-Cano L, Valdellós J, et al. Prevalence and Clinical Conditions Related to Sarcopaenia among Older Persons Living in the Community. J Clin Med. 2022;11(13).
37. Bernabeu-Wittel M, González-Molina Á, Fernández-Ojeda R, Díez-Manglano J, Salgado F, Soto-Martín M, et al. Impact of Sarcopenia and Frailty in a Multicenter Cohort of Polypathological Patients. J Clin Med. 2019;8(4).
38. Yoo MC, Won CW, Soh Y. Association of high body mass index, waist circumference, and body fat percentage with sarcopenia in older women. BMC Geriatr. 2022;22(1):937.
39. Curtis M, Swan L, Fox R, Warters A, O'Sullivan M. Associations between Body Mass Index and Probable Sarcopenia in Community-Dwelling Older Adults. Nutrients. 2023;15(6).
40. Li CW, Yu K, Shyh-Chang N, Jiang Z, Liu T, Ma S, et al. Pathogenesis of sarcopenia and the relationship with fat mass: descriptive review. J Cachexia Sarcopenia Muscle. 2022;13(2):781-94.

Tables and Figure legends

Table 1. Baseline characteristics of the study samples

Note: CESD-10, Center for epidemiologic studies depression scale-10; IADL; Instrumental activities of daily living; TC, total cholesterol; TG, triglycerides; HDL-C, high density lipoprotein cholesterol; LDL-C, low density lipoprotein cholesterol; MCV, mean corpuscular volume. Values presented as median (interquartile range), n (%).

Table 2. Results of univariate logistic regression analysis in the training set

Note: TC, total cholesterol; TG, triglycerides; HDL-C, high density lipoprotein cholesterol; LDL-C, low density lipoprotein cholesterol; MCV, mean corpuscular volume.

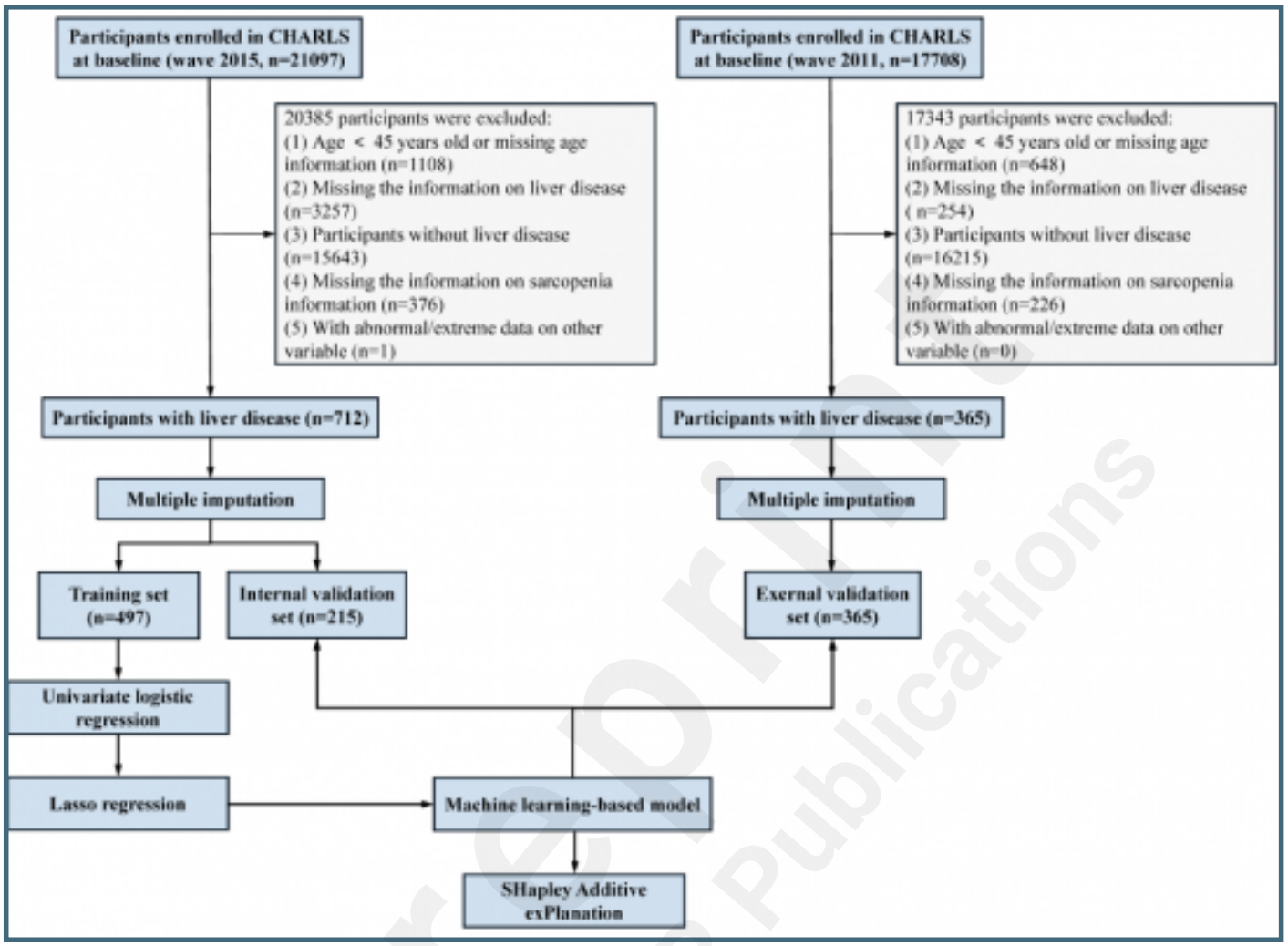
Table 5. Diagnostic performance of the machine learning models for sarcopenia in participants with liver disease

Note: SVM, Support Vector Machine; LR, Logistic Regression; DT, Decision Tree; RF, Random Forest; XGB, Extreme Gradient Boosting; KNN, K-Nearest Neighbors; LGBM, LightGBM; NBM, Naive Bayes Model.

Supplementary Files

Figures

Flowchart of the study population.



Identification of key predictors for sarcopenia in patients with liver disease by the LASSO regression model.