Rong *et al.*
European Journal of Medical Research (2025) 30:345
https://doi.org/10.1186/s40001-025-02606-3

European Journal of Medical Research

RESEARCH

Open Access

Intelligent predictive risk assessment and management of sarcopenia in chronic disease patients using machine learning and a web-based tool

Ke Rong[1†], Gu li jiang Yi ke ran[2†], Changgui Zhou[1] and Xinglin Yi[3,4*]

Abstract

Background Individuals with chronic diseases are at higher risk of sarcopenia, and precise prediction is essential for its prevention. This study aims to develop a risk scoring model using longitudinal data to predict the probability of sarcopenia in this population over next 3–5 years, thereby enabling early warning and intervention.

Methods Using data from a nationwide survey initiated in 2011, we selected patient data records from wave 1 (2011–2012) and follow-up data from wave 3 (2015–2016) as the study cohort. Retrospective data collection included demographic information, health conditions, and biochemical markers. After excluding records with missing values, a total of 2891 adults with chronic conditions were enrolled. Sarcopenia was assessed based on the Asian Working Group for Sarcopenia (AWGS) 2019 guidelines. A generalized linear mixed model (GLMM) with random effects and diverse machine learning models were utilized to explore feature contributions to sarcopenia risk. The Recursive Feature Elimination (RFE) algorithm was employed to optimize the full Multilayer Perceptron (MLP) model and develop an online application tool.

Results Among total population, 580 (20.1%) individuals were diagnosed with sarcopenia in wave 1 (2011–2012), and 638 (22.1%) were diagnosed in wave 3 (2015–2016), while 2165 (74.9%) individuals were not diagnosed with sarcopenia across the study period. MLP model, performed better than other three classic machine learning models, demonstrated a ROC AUC of 0.912, a PR AUC of 0.401, a sensitivity of 0.875, a specificity of 0.844, a Kappa value of 0.376, and an F1 score of 0.44. According to MLP model-based SHapley Additive exPlanations (SHAP) scoring, weight, age, BMI, height, total cholesterol, PEF, and gender were identified as the most important features of chronic disease individuals for sarcopenia. Using the RFE algorithm, we selected six key variables—weight, age, BMI, height, total cholesterol, and gender—achieving an ROC AUC of about 0.9 for the online application tool.

Conclusion We developed an MLP machine learning model that incorporates only six easily accessible variables, enabling the prediction of sarcopenia risk in individuals with chronic diseases. Additionally, we created a practical online application tool to assist in decision-making and streamline clinical assessments.

Keywords Sarcopenia, Chronic disease, Machine learning, Predictive modeling, Longitudinal study

[†]Ke Rong and Gu li jiang·Yi ke ran contributed equally to this work.

*Correspondence:
Xinglin Yi
xinglinyi2024@163.com
Full list of author information is available at the end of the article

Introduction

Sarcopenia, characterized by the progressive loss of skeletal muscle mass and strength, is increasingly recognized as being closely associated with a range of chronic diseases. Numerous cohort studies have demonstrated that individuals with chronic diseases, particularly those over the age of 60, are at an elevated risk of developing sarcopenia. For instance, a systematic review of 11 population-based studies on patients with heart failure revealed that the prevalence of sarcopenia in individuals with cardiovascular disease was ~ 34%, significantly higher than in healthy controls [1]. Similarly, a cross-sectional study of patients with type 2 diabetes found that the rate of sarcopenia was nearly twice as high compared to age-matched healthy controls [2]. In patients with chronic kidney disease (CKD), global prevalence rates for sarcopenia and severe sarcopenia were reported to be 24.5% and 21%, respectively, with approximately half of dialysis-dependent CKD patients affected [3]. Furthermore, patients with chronic respiratory and digestive diseases also exhibit a higher prevalence of sarcopenia [4, 5].

The relationship between sarcopenia and chronic diseases is likely mediated through complex mechanisms, including persistent inflammation, oxidative stress, hormonal changes, and physical inactivity. Chronic comorbidities often exacerbate the inflammatory milieu, leading to increased muscle protein degradation and reduced muscle protein synthesis. In chronic conditions, persistent elevation of pro-inflammatory cytokines creates a catabolic environment in skeletal muscle, impairing protein synthesis while enhancing protein degradation. This inflammatory state is characterized by increased levels of TNF-α, IL-6, and IL-1β, which activate the NF-κB signaling pathway, subsequently triggering muscle atrophy [6–9]. The inflammatory cascade extends to activating the Akt/mTOR/FoxO3a and other signaling pathways, further promoting protein breakdown and muscle wasting [10, 11]. In this process, the gut microbiota–muscle axis [12], and cellular death mechanisms [13], and enhanced activity of the ubiquitin–proteasome [14] also play a role in the oxidative stress activation which further exacerbates mitochondrial damage and subsequent metabolic dysfunction [15]. In the context of diabetes, the presence of insulin resistance and altered growth factor signaling in chronic conditions such as diabetes further contributes to muscle wasting and weakness [16].

Conversely, sarcopenia has been shown to exacerbate the clinical course of several chronic diseases through its effects on muscle function, inflammation, and metabolic dysregulation. In cardiovascular disease (CVD), sarcopenia is associated with reduced physical capacity, increased frailty, and diminished ability to tolerate medical interventions, all of which lead to higher rates of hospitalization and mortality [17]. The bidirectional interaction between sarcopenia and chronic diseases underscores the importance of early detection and intervention.

Sarcopenia not only worsens the prognosis of chronic diseases but also significantly contributes to the increased healthcare costs and burden associated with these conditions. In this context, a multidisciplinary approach that combines physical assessments with laboratory biomarkers may be particularly effective for early identification and prediction of sarcopenia in patients with chronic disease.

To date, serum creatinine, cystatin C, uric acid, and sarcopenia index have been identified as potential predictors of sarcopenia in chronic respiratory diseases, and may also be relevant for other chronic conditions [18, 19]. However, most existing studies have primarily focused on exploring the correlation between these markers and sarcopenia, often through cross-sectional analyses. There remains a lack of effective and convenient predictive models that integrate independent predictors to forecast the risk of sarcopenia in individuals with chronic diseases. The increasing availability of large datasets and complex algorithms has led to the widespread use of machine learning for prediction in medicine. Notably, Guan et al. recently used XGBoost to predict new-onset atrial fibrillation (9.2%) across cohorts, significantly outperforming linear logistic regression [20]. Machine learning models excel at capturing non-linear relationships and managing high-dimensional data without prior assumptions about variable interactions, making them ideal for analyzing complex biological markers like serum creatinine and cystatin C. Explainability methods such as SHAP can further elucidate threshold effects and complex predictor interactions, enhancing model understanding, as shown by Qiu et al. [21]. Building on these strengths, the aim of our study is to predict the risk of developing sarcopenia over the next 3–5 years. We will achieve this through a prospective analysis of the large, longitudinal China Health and Retirement Longitudinal Study (CHARLS) cohort, incorporating explainable machine learning techniques. This approach will allow us to investigate physiological, biochemical, and inflammatory markers (e.g., total cholesterol, Cystatin C, C-reactive protein [CRP]) not only to develop robust predictive models but also to comprehensively identify and quantify the independent factors contributing to sarcopenia incidence.

Methods

Participants enrollment and variable collection

The CHARLS is a nationally representative, high-quality longitudinal survey targeting individuals aged 45 and older in China. It collects comprehensive data across various domains, including health status, retirement, income, wealth, family structure, and social support. To date, five waves of data collection have been completed: the baseline survey in 2011, followed by biennial surveys in 2013, 2015, 2018, and 2021.

For this study, data from waves 1 (2011–2012) and 3 (2015–2016) were analyzed. We focused on individuals with chronic diseases affecting the lungs, heart, liver, kidneys, digestive system, and psychiatric conditions, as well as those diagnosed with hypertension and diabetes. The dataset included demographic variables such as age, gender, height, weight, body mass index (BMI), education level, and smoking status. In addition, pulmonary function and laboratory test results were examined, including peak expiratory flow (PEF), total cholesterol, triglycerides, triglyceride-glucose index (TyG), calculated as ln(fasting triglycerides [mg/dL] × fasting glucose [mg/dL]/2), uric acid, blood urea nitrogen (BUN), cystatin C, CRP, glycated hemoglobin (HbA1c), high-density lipoprotein cholesterol (HDL-C), and low-density lipoprotein cholesterol (LDL-C). These variables were retrospectively obtained from the CHARLS database. CHARLS employs a multi-stage stratified probability sampling method to select tens of thousands of respondents. Data on individual demographics (e.g., smoking status, education level) were collected through face-to-face interviews and questionnaires. Physical performance data (e.g., timed walk measurements, blood pressure) were measured using physical performance scales. Blood test data (e.g., BUN, CRP) were collected from patient blood samples and sent to hospitals for examination. These data are typically followed up and updated every 2 years, providing high-quality longitudinal data suitable for tracking changes over time. These data are typically updated and followed up every 2 years in waves, providing high-quality longitudinal tracking data. Notably, considering that only the follow-up surveys from wave 1 (2011–2012) and wave 3 (2015–2016) included blood test indicators, we chose to incorporate data from these two waves. The rationale for selecting these variables was to include as many relevant variables as possible from those available in CHARLS, which aligns with approaches depicted in similar studies [22, 23].

Assessment of chronic disease

Chronic diseases were defined based on the CHARLS DA007 questionnaire, which asks: "Have you been diagnosed with [conditions listed below] by a doctor?" A positive response to any of the listed conditions was considered indicative of a chronic disease. These conditions include hypertension, dyslipidemia, diabetes or hyperglycemia, cancer or malignant tumors, chronic lung diseases such as chronic bronchitis or emphysema, liver disease, cardiac conditions including myocardial infarction, coronary heart disease, angina, and congestive heart failure, stroke, kidney disease, gastrointestinal diseases, psychiatric disorders, memory-related diseases such as Alzheimer's, Parkinson's, and cerebral atrophy, arthritis or rheumatism, and asthma.

Assessment of sarcopenia

In this study, sarcopenia was defined according to the 2019 criteria established by the Asian Working Group for Sarcopenia (AWGS) [24]. The assessment of sarcopenia comprised three components: muscle strength, appendicular skeletal muscle mass (ASM), and physical performance. Hand grip strength was measured using the Yuejian WL-1000 dynamometer, with the maximum value from two trials on both dominant and non-dominant hands recorded. Diagnostic thresholds were set at < 28 kg for men and < 18 kg for women, as specified by the AWGS 2019 guidelines.

Muscle mass was estimated using the ASM formula: ASM = 0.193 × weight (kg) + 0.107 × height (cm) − 4.157 × sex − 0.037 × age (years) − 2.631, where sex was coded as 1 for males and 2 for females [25]. Low muscle mass was defined as height-adjusted ASM (ASM/Height^2) below the 20 th percentile of the population, specifically < 5.45 kg/m^2 for women and < 7.15 kg/m^2 for men. Previous research has demonstrated a strong consistency between the ASM formula and dual-energy X-ray absorptiometry (DXA) in the Chinese population, highlighting its value in measuring ASM [26]. Physical performance was considered low if walking speed was less than 1.0 m/s or if participants required 12 s or more to complete the five-time chair stand test. Sarcopenia was diagnosed when low muscle mass was accompanied by either reduced physical performance or diminished muscle strength.

Inclusion and exclusion criterion

As illustrated in the flowchart (Fig. 1), we included patients diagnosed with chronic diseases as defined earlier. The exclusion criteria were as follows: patients with missing variable values, those who did not participate in interviews during both wave 1 and wave 3, and individuals with extreme outliers (BMI > 100 or height < 100 cm). Consequently, 22,982 individuals were excluded from the cohort, resulting in a final sample of 2891 individuals with chronic diseases.

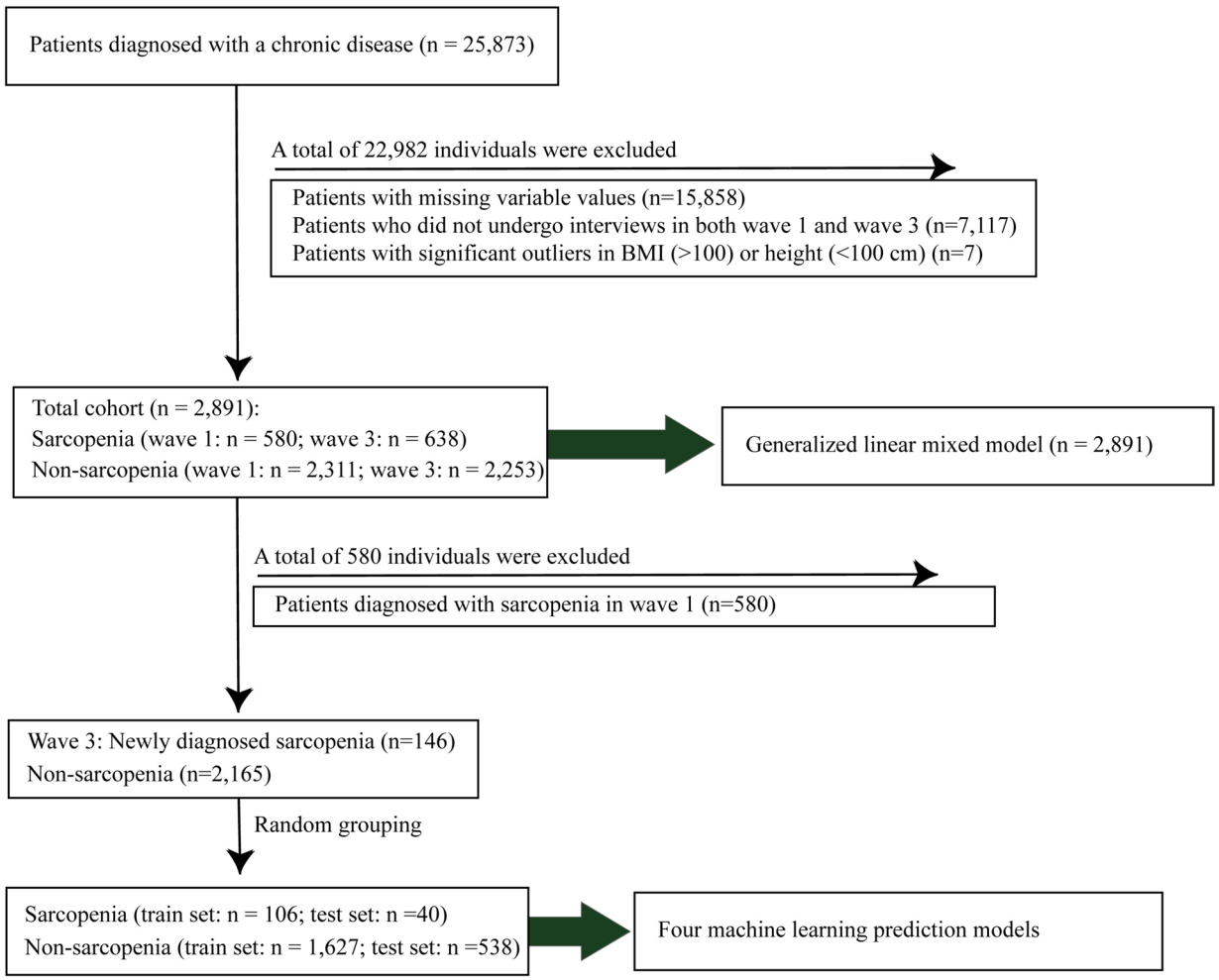

Fig. 1 Flowchart of inclusion and exclusion criteria. *BMI* body mass index

Among these, 580 individuals were diagnosed with sarcopenia in wave 1 (2011–2012), and 638 were diagnosed in wave 3 (2015–2016), while 2165 individuals were not diagnosed with sarcopenia across the study period. Since the data were obtained from an openly available consortium, no written consent was required for this study.

Statistical analysis

The chi-square test or Fisher's test was used for categorical data comparison, ANOVA or Kruskal–Wallis for multiple group comparisons, and *t*-test or Wilcoxon rank-sum test for two-group comparisons of continuous variables. The generalized linear mixed model (GLMM) with a random individual ID intercept was utilized to estimate the association between covariates and sarcopenia. Given that the features of each ID are repeated twice, the GLMM model, which accommodates repeated measures data, is deemed more suitable and is thus selected here over other linear models. The model can be expressed as following, where β_0 is the fixed intercept term representing the baseline log-odds of the sarcopenia when the covariate X_i is zero; β_1 is the fixed effect coefficient for the covariate X_i, quantifying its effect on the log-odds of the event; u_i is the random effect for individual i, accounting for individual-specific deviations from the overall effect; and $\text{logit}(p_i)$ is the logit link function defined as $\log\frac{p_i}{1-p_i}$.

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_i + u_i.$$

In our GLMM model, recognizing that the relationship between covariates and sarcopenia may not be strictly linear, we categorized continuous variables into three or four groups based on their quantiles (e.g., tertiles or quartiles), depending on the data distribution (Table S1). The practice of converting continuous variables into categorical ones can, to some extent, allow linear models

to capture potential non-linear associations by estimating distinct effects for different ranges of the variable, thereby relaxing the strict linearity assumption. This approach has been commonly adopted in previous epidemiological research [27, 28]. The model's fixed effects included time since baseline, each covariate, and the interaction terms between each covariate and time. The coefficient for each covariate within the fixed effects represents its effect on sarcopenia at baseline, while the coefficient for the interaction between each covariate and time reflects how the covariate's influence on sarcopenia evolves over time.

We developed four classical machine learning models—K-Nearest Neighbors (KNN), Random Forest (RF), XGBoost (XGB), and Multilayer Perceptron (MLP)—to prospectively predict the probability of sarcopenia onset within the next 3–5 years. Among these, RF is a tree-based model, while XGB is both a tree-based and boosting model. In contrast, KNN is a distance-based model that classifies data points based on the proximity to their nearest neighbors, and MLP is a neural network-based model designed to capture complex nonlinear relationships in the data. Specifically, an MLP consists of interconnected nodes (neurons) organized in layers: an input layer, one or more hidden layers, and an output layer. Each connection between neurons carries a weight, and neurons in the hidden and output layers typically apply a non-linear activation function (sigmoid in this analysis) to the weighted sum of their inputs. The performance of machine learning models is highly dependent on the configuration of key hyperparameters. For instance, in the case of the MLP, critical hyperparameters such as the number of units in the hidden layer (which determines the model's capacity to capture complex patterns), regularization strength (which controls overfitting by penalizing large weights), and number of epochs (which defines the number of training iterations) were identified and optimized through a subsequent Bayesian hyperparameter tuning process.

As shown in Fig. 1, individuals already diagnosed with sarcopenia at baseline were excluded from the analysis. The dataset was then randomly split into training and testing sets in a 75% to 25% ratio, ensuring similar distributions of both covariates and outcome variables across the two sets. To enhance model stability during training, five-fold cross-validation was applied to the training set.

Hyperparameter tuning for each machine learning model was performed using Bayesian optimization, starting with an initial random search of 10 evaluations and proceeding up to a maximum of 50 iterations. Early stopping was implemented if model performance did not improve after 10 consecutive iterations. Model comparison was conducted based on Receiver Operating Characteristic Area Under the Curve (ROC AUC), Precision-Recall Area Under the Curve (PR AUC), calibration, and five-fold cross-validation results. The AUC values of each pair of models were compared using the DeLong test method to statistically assess differences in performance. SHapley Additive exPlanations (SHAP) is a unified framework for interpreting the predictions of complex machine learning models by attributing each feature's contribution based on principles from cooperative game theory. In this analysis, numerical variables are represented in their continuous numerical form rather than being categorized into quantiles. Based on the magnitude of the SHAP values, the contributions of different variables to sarcopenia are ranked from highest to lowest. Subsequently, recursive feature elimination (RFE) is performed within the optimal model: features are iteratively removed one by one, starting with the least important according to the SHAP ranking from the model retrained at each step. This process aims to obtain the highest possible AUC value, thereby maintaining predictive accuracy while ensuring model simplicity. The performance is evaluated after each feature removal, and when the AUC value curve versus the number of remaining features begins to plateau, the smallest feature set achieving performance close to the maximum is considered optimal. The RFE algorithm was first proposed by Guyon et al. [29]. This algorithm, owing to its effectiveness in identifying impactful features, reducing model complexity, and potentially improving generalization by eliminating irrelevant or redundant predictors while often maintaining high predictive performance, was subsequently applied increasingly in the medical field [30, 31]. The application of RFE here aims to optimize the machine learning algorithm's feature set by addressing feature redundancy, striving to maintain predictive accuracy while enhancing model parsimony. This aligns with our main purpose in this research, which is to establish a convenient and precise model for medical utilization.

Sample size calculation was performed using the enrichment strategy proposed by Diggle in 2002, which is particularly suitable for longitudinal studies involving specific subpopulations or enriched cohorts, such as individuals with chronic diseases.

$$N = \frac{2\left(u_{1-\alpha/2} + u_{1-\beta}\right)^2 s^2 (1 + (n-1)\rho)}{n(\mu_1 - \mu_2)^2}.$$

Diggle's formula, as presented above, calculates the required sample size N based on several factors: the standard normal quantiles corresponding to the desired two-sided significance level $1 - \alpha/2$ and $1 - \beta$, the pooled

variance (s^2), the number of repeated measurements (n), the within-subject correlation (ρ), and the difference between groups or slopes ($\mu_1 - \mu_2$). In this context, s_1 and s_2 represent the standard deviations of the two groups, while n_1 and n_2 denote the sample sizes of the two groups in the formula. For this analysis, we computed the required sample size using the aforementioned formula as implemented in the "lmmpower" package in R. We specified a desired power of 0.80 (setting $1-\beta$) and a significance level of 0.05 (setting α) using the function's default values. We ultimately determined that at least 1849 samples are required for the GLMM model [32]. According to the guideline for the development of prediction models published in January 2024, at least 1034 samples are needed to construct a predictive model for binary outcomes [33]. The"Tidymodels"package in R software version 4.4.1 was used to establish machine learning models, while the"lme4"package was employed to develop the GLMM model.

Web tool development and prediction

By iteratively eliminating features using the RFE algorithm, we identified six key variables—weight, age, BMI, height, total cholesterol, and the commonly included gender variable—as inputs. This approach achieved an ROC AUC of approximately 0.9, culminating in the development of a practical online application tool. The "shiny" package was used to design the webpage layout, comprising three main components: an introduction section outlining the tool and its functionalities, an input section with fields for the five variables, and a prediction and visualization section. After entering the required data and clicking the yellow"Predict"button, the tool automatically displays the predicted probability of sarcopenia along with a SHAP force plot, offering insights into the risk of sarcopenia development in individuals with chronic diseases over the next 3–5 years.

Results

1Baseline characteristics of individuals

A total of 2,891 individuals with chronic diseases met the inclusion and exclusion criteria. Among them, 580 individuals (20.1%) were diagnosed with sarcopenia at wave 1 (2011–2012). By wave 3 (2015–2016), the total number of sarcopenia cases had risen to 638 (22.1%), including 146 new cases that developed in the interim, while 88 individuals previously diagnosed at wave 1 had recovered (Fig. 1). Compared to non-sarcopenic individuals, those with sarcopenia were more likely to be female, older, smokers, and have lower education levels. They also exhibited lower weight, height, BMI, PEF, and lipid levels (total cholesterol, triglycerides, LDL-C), but higher HDL-C levels. Additionally, BUN and cystatin C levels were elevated compared to non-sarcopenic individuals. These observations align with the expected characteristics exhibited by individuals with sarcopenia. For instance, age-related physiological changes and hormonal differences could contribute to the risk of muscle loss [34]. The observed lower weight, height, BMI, and PEF are also directly clinically significant, as they inherently reflect the reduced muscle mass and impaired physical function, including respiratory muscle strength, that characterize the sarcopenic phenotype [35]. Interestingly, a lower prevalence of diabetes and reduced TyG levels were observed in the sarcopenia group. Furthermore, sarcopenia patients were more likely to have chronic lung, liver, and digestive diseases, but unexpectedly, they had a lower prevalence of hypertension and other chronic heart diseases. The higher prevalence of chronic lung, liver, and digestive diseases among those with sarcopenia is clinically meaningful, aligning with the understanding that these conditions often promote systemic inflammation, malnutrition, and catabolic states conducive to muscle wasting (Table 1).

GLMM's fixed effects

The fixed effects of the GLMM, including the coefficients for each characteristic and their interaction terms with time, are presented in Table S2. As for main effect, the multivariable GLMM analysis revealed the clinical significance of several key factors associated with sarcopenia. Notably, older age and higher levels of Cystatin C, a marker potentially reflecting renal function and inflammation, were confirmed as significant risk factors associated with increased odds of sarcopenia. Conversely, protective factors were also evident: higher education levels, never smoking, greater height, and better PEF were significantly associated with lower odds of sarcopenia. These results align with the characteristics of sarcopenia individuals, which are generally observed in older adults with poor nutrition, a higher burden of chronic diseases, or those affected by unhealthy lifestyle habits. As for the interaction effects with time, higher education levels, the presence of diabetes and chronic heart disease, non-smoking status, higher BMI, elevated levels of total cholesterol, triglycerides, uric acid, BUN, and CRP, glycated hemoglobin, lower HDL-C, and higher LDL-C levels were associated with an increased risk of sarcopenia (Fig. 2, Table S2, Figure S1). These results indicated that individuals with these characteristics may experience an accelerated trajectory towards developing sarcopenia during the follow-up. Clinically, this suggests these factors predict a faster rate of muscle health decline, beyond their baseline association. Biologically, the interactions

Table 1 Baseline characteristics of individuals with chronic diseases at wave 1 and at wave 3

Variable	Wave 1					Wave 3			
	Feature	Without sarcopenia (*N* = 2311)	Sarcopenia (*N* = 580)	*p* value	Variable	Feature	Without sarcopenia (*N* = 2253)	Sarcopenia (*N* = 638)	*p* value
Gender	Female	1014 (43.9%)	282 (48.6%)	0.045	Gender	Female	995 (44.2%)	301 (47.2%)	0.191
	Male	1297 (56.1%)	298 (51.4%)			Male	1258 (55.8%)	337 (52.8%)	
Smoking status	Current	616 (26.7%)	205 (35.3%)	< 0.001	Smoking status	Current	543 (24.1%)	188 (29.5%)	0.021
	Former	223 (9.6%)	49 (8.4%)			Former	398 (17.7%)	109 (17.1%)	
	Never	1472 (63.7%)	326 (56.2%)			Never	1312 (58.2%)	341 (53.4%)	
Height, m	Mean ± SD	1.6 ± 0.1	1.5 ± 0.1	< 0.001	Height, m	Mean ± SD	1.6 ± 0.1	1.5 ± 0.1	< 0.001
Weight, kg	Mean ± SD	63.0 ± 9.8	45.7 ± 4.8	< 0.001	Weight, kg	Mean ± SD	63.7 ± 10.3	45.8 ± 5.0	< 0.001
BMI, kg/m^2	Mean ± SD	25.0 ± 3.5	19.4 ± 1.8	< 0.001	BMI, kg/m^2	Mean ± SD	25.3 ± 3.9	19.7 ± 2.0	< 0.001
PEF, L/min	Mean ± SD	294.5 ± 121.4	239.0 ± 108.4	< 0.001	PEF, L/min	Mean ± SD	314.5 ± 118.3	241.5 ± 110.0	< 0.001
Total cholesterol, mg/dL	Mean ± SD	194.9 ± 38.1	190.1 ± 38.7	0.007	Total cholesterol, mg/dL	Mean ± SD	186.8 ± 36.9	182.5 ± 38.5	0.01
Triglycerides, mg/dL	Mean ± SD	142.9 ± 102.2	106.2 ± 67.7	< 0.001	Triglycerides, mg/dL	Mean ± SD	155.0 ± 93.0	110.7 ± 65.4	< 0.001
TyG	Mean ± SD	8.8 ± 0.7	8.5 ± 0.6	< 0.001	TyG	Mean ± SD	8.8 ± 0.6	8.5 ± 0.6	< 0.001
Uric Acid, mg/dL	Mean ± SD	4.4 ± 1.2	4.4 ± 1.3	0.21	Uric Acid, mg/dL	Mean ± SD	5.0 ± 1.4	4.8 ± 1.4	< 0.001
BUN, mg/dL	Mean ± SD	15.6 ± 4.3	16.7 ± 5.2	< 0.001	BUN, mg/dL	Mean ± SD	15.8 ± 4.5	16.5 ± 5.5	< 0.001
Cystatin C, mg/L	Mean ± SD	1.0 ± 0.3	1.1 ± 0.3	< 0.001	Cystatin C, mg/L	Mean ± SD	0.9 ± 0.2	0.9 ± 0.3	< 0.001
CRP, mg/L	Mean ± SD	2.5 ± 5.5	2.7 ± 8.6	0.752	CRP, mg/L	Mean ± SD	3.0 ± 6.3	2.7 ± 7.8	0.484
Glycated Hemoglobin, %	Mean ± SD	5.3 ± 0.8	5.2 ± 0.6	< 0.001	Glycated Hemoglobin, %	Mean ± SD	6.1 ± 1.1	5.9 ± 1.0	< 0.001
HDL C, mg/dL	Mean ± SD	48.6 ± 13.9	57.6 ± 17.0	< 0.001	HDL C, mg/dl	Mean ± SD	49.7 ± 10.9	55.8 ± 13.3	< 0.001
LDL C, mg/dL	Mean ± SD	118.0 ± 35.6	112.6 ± 34.4	0.001	LDL C, mg/dL	Mean ± SD	104.4 ± 29.1	100.7 ± 28.2	0.004
Age, year	Mean ± SD	58.5 ± 8.7	65.6 ± 8.5	< 0.001	Age, year	Mean ± SD	62.2 ± 8.5	69.7 ± 8.6	< 0.001
Education status	Associate degree	553 (23.9%)	139 (24%)	< 0.001	Education status	Associate degree	543 (24.1%)	149 (23.4%)	< 0.001
	Bachelor's degree	491 (21.2%)	60 (10.3%)			Bachelor's degree	479 (21.3%)	72 (11.3%)	
	Others	207 (9%)	24 (4.1%)			Others	203 (9%)	28 (4.4%)	
	Vocational school	1060 (45.9%)	357 (61.6%)			Vocational school	1028 (45.6%)	389 (61%)	
Hypertension	No	1347 (58.3%)	429 (74%)	< 0.001	Hypertension	No	1113 (49.4%)	406 (63.6%)	< 0.001
	Yes	964 (41.7%)	151 (26%)			Yes	1140 (50.6%)	232 (36.4%)	
Diabetes	No	2063 (89.3%)	556 (95.9%)	< 0.001	Diabetes	No	1899 (84.3%)	580 (90.9%)	< 0.001
	Yes	248 (10.7%)	24 (4.1%)			Yes	354 (15.7%)	58 (9.1%)	
Chronic lung disease	No	2038 (88.2%)	469 (80.9%)	< 0.001	Chronic lung disease	No	1874 (83.2%)	462 (72.4%)	< 0.001
	Yes	273 (11.8%)	111 (19.1%)			Yes	379 (16.8%)	176 (27.6%)	
Chronic heart disease	No	1841 (79.7%)	508 (87.6%)	< 0.001	Chronic heart disease	No	1623 (72%)	495 (77.6%)	0.006
	Yes	470 (20.3%)	72 (12.4%)			Yes	630 (28%)	143 (22.4%)	
Chronic psychiatric disorders	No	2273 (98.4%)	572 (98.6%)	0.787	Chronic psychiatric disorders	No	2186 (97%)	620 (97.2%)	0.945
	Yes	38 (1.6%)	8 (1.4%)			Yes	67 (3%)	18 (2.8%)	
Chronic liver disease	No	2219 (96%)	543 (93.6%)	0.017	Chronic liver disease	No	2065 (91.7%)	589 (92.3%)	0.647

Table 1 (continued)

Variable	Wave 1				Variable	Wave 3			
	Feature	Without sarcopenia (*N* = 2311)	Sarcopenia (*N* = 580)	*p* value		Feature	Without sarcopenia (*N* = 2253)	Sarcopenia (*N* = 638)	*p* value
	Yes	92 (4%)	37 (6.4%)			Yes	188 (8.3%)	49 (7.7%)	
Chronic kidney disease	No	2111 (91.3%)	531 (91.6%)	0.94	Chronic kidney disease	No	1955 (86.8%)	542 (85%)	0.264
	Yes	200 (8.7%)	49 (8.4%)			Yes	298 (13.2%)	96 (15%)	
Chronic digest disease	No	1586 (68.6%)	366 (63.1%)	0.013	Chronic digest disease	No	1338 (59.4%)	328 (51.4%)	< 0.001
	Yes	725 (31.4%)	214 (36.9%)			Yes	915 (40.6%)	310 (48.6%)	

TyG triglyceride-glucose index

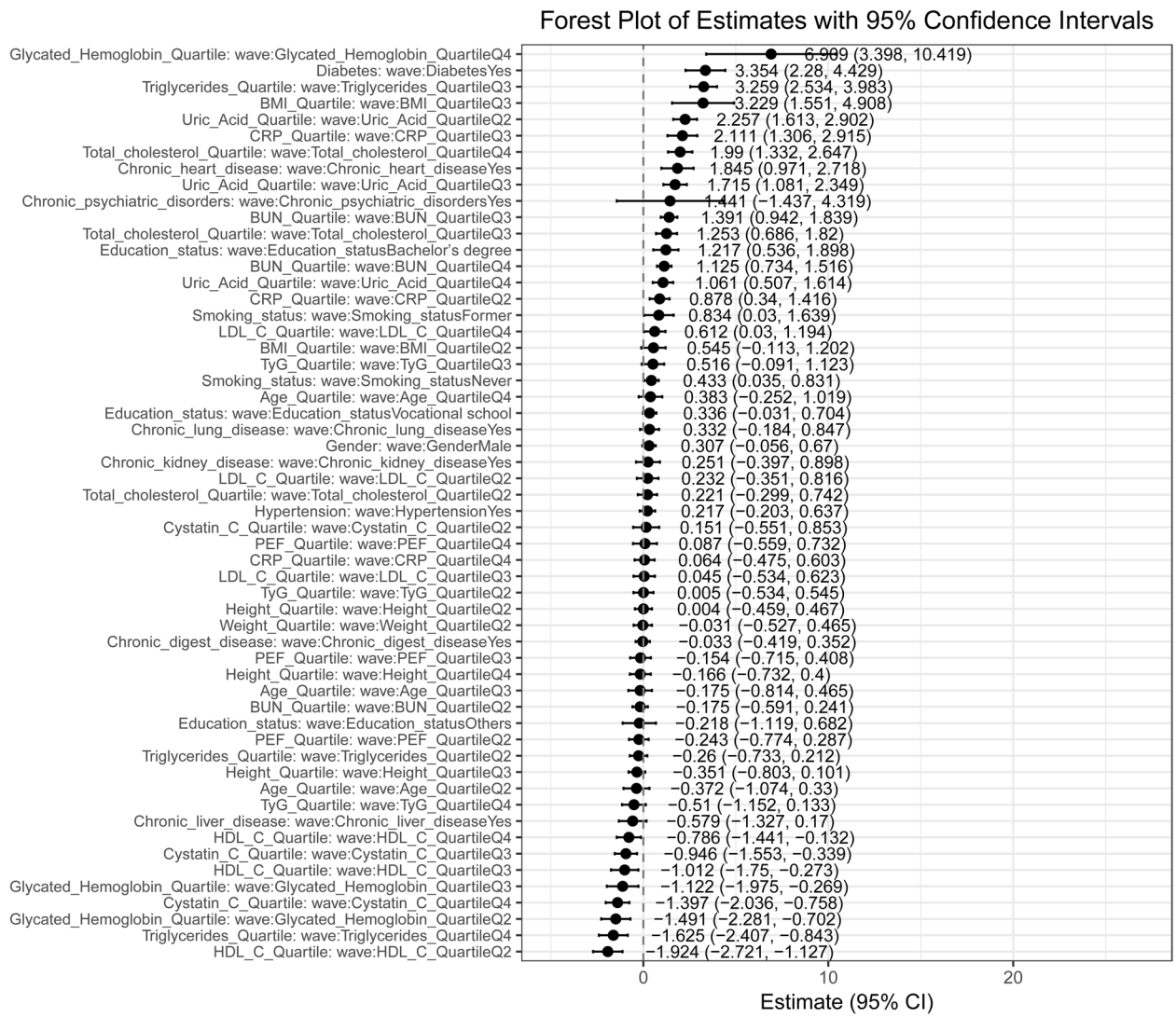

Fig. 2 Forest plot of estimates of interaction effects of different variable with time, including confidence intervals for GLMM. *GLMM* generalized linear mixed model, *CRP* C-reactive protein, *PEF* peak expiratory flow, *TyG* triglyceride-glucose index, *BUN* blood urea nitrogen, *HbA1c* glycated haemoglobin, *HDL-C* high-density lipoprotein cholesterol, *LDL-C* low-density lipoprotein cholesterol

involving conditions like diabetes and heart disease, alongside markers of inflammation (CRP) and metabolic stress (HbA1c, Uric Acid, BUN), likely reflect the ongoing detrimental impact of sustained inflammation, insulin resistance, and catabolism on muscle maintenance over time. All of which showed significant p-values (< 0.05).

Development and evaluation of machine learning models

As outlined in the flowchart in the Methods section, individuals with chronic diseases who had already developed sarcopenia at wave 1 were pre-excluded. This left a total of 2311 individuals for the longitudinal prediction process, with 1733 allocated to the training set and 578 to the test set (Table S3). Given that the time interval between wave 1 and wave 3 is approximately 3–5 years, the prediction model was designed to prospectively estimate the incidence of sarcopenia over this period.

After performing Bayesian optimization to determine the optimal hyperparameters for the four models, the primary hyperparameter results are as follows: KNN (number of neighbors = 20, weight function ="biweight"), RF (number of features to consider at each split = 6, number of trees = 457, minimum number of samples required to split a node = 48), XGB (number of features to consider at each split = 19, minimum number of samples required to split a node = 29, maximum tree depth = 15, learning rate = 0.00149, minimum loss reduction required for a split = 0.401, subsample ratio of the training instances = 0.908), and MLP (number of units in the hidden layer = 2, regularization strength = 0.0270, number of epochs = 405). All models' learning curves demonstrated good fitting (Figure S2), indicating that the models effectively captured the underlying patterns in the data without significant overfitting or underfitting. This suggests that the models generalize well to unseen data, maintaining both high accuracy and stability during the training and validation phases.

Several model evaluations were conducted as follows. In the training set, ROC AUC and PR AUC values indicated that KNN, RF, MLP, and XGB ranked from highest to lowest in terms of performance (Fig. 3A, C). However, in the test set, the MLP model demonstrated the best performance in both ROC AUC and PR AUC metrics (Fig. 3B, D). The DeLong test results for ROC AUC, presented in Table 2, confirmed that MLP outperformed other models, although the difference compared to the RF model in the test set was not statistically significant. The MLP model achieved its optimal cutoff value at 0.293, with an accuracy of 0.846 on the test set. It also demonstrated a ROC AUC of 0.912, a PR AUC of 0.401, a sensitivity of 0.875, a specificity of 0.844, a Kappa value of 0.376, and an F1 score of 0.44. The comprehensive evaluation metrics are presented in Table S4. The reasons for the relatively lower PR AUC, Kappa, and F1 scores observed here may be attributed to the inherent class imbalance within the dataset, which often leads to models predicting the majority class more frequently, resulting in fewer false positives and hence depressing these specific metrics that are sensitive to false positive rates. However, considering the clinical implications in this research, the primary focus is on ensuring that no potentially affected patients are missed, which justifies accepting a moderately higher false positive rate. This principle is particularly relevant in our study, considering the limited observation window during which individuals determined to be high-risk might not yet exhibit discernible symptoms. Nevertheless, there is potential for further improvement in our future research, in which we plan to explore models specifically designed for enhanced performance on tabular data and better handling of class imbalance, and utilize more advanced data pre-processing techniques such as Synthetic Minority Over-sampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), or related variants to better balance the class distribution before model training. Calibration plots revealed excellent consistency between predicted and observed outcomes for the MLP model, whereas the RF model showed poorer calibration (Figures S3, S4). Furthermore, based on the five-fold cross-validation results, the MLP model achieved the highest ROC AUC (Figure S5A) and PR AUC (Figure S5B), indicating superior stability and accuracy. Consequently, we concluded that the MLP model with two units in the hidden layer was the best-performing model in our study. Detailed hyperparameter tuning results for the MLP model are visualized in Figure S6, while the neural network structure is illustrated in Figure S7. Several potential reasons could explain the MLP's stronger performance in this context. Sarcopenia's multifactorial nature likely involves complex, non-linear relationships and high-order interactions between diverse predictors. MLPs excel at capturing such patterns through their interconnected layers and non-linear activation functions, potentially modeling smooth or global relationships more effectively than the axis-aligned partitions inherent in tree-based models like RF and XGB.

SHAP-based feature importance ranking and dependence plot

Based on the trained MLP model, we calculated and visualized the SHAP values using the"shapviz"package (https://cran.r-project.org/web/packages/shapviz/), as shown in Fig. 4. Weight, age, BMI, height, total cholesterol, PEF, and gender were identified as the most important features (Fig. 4A, B). The association between

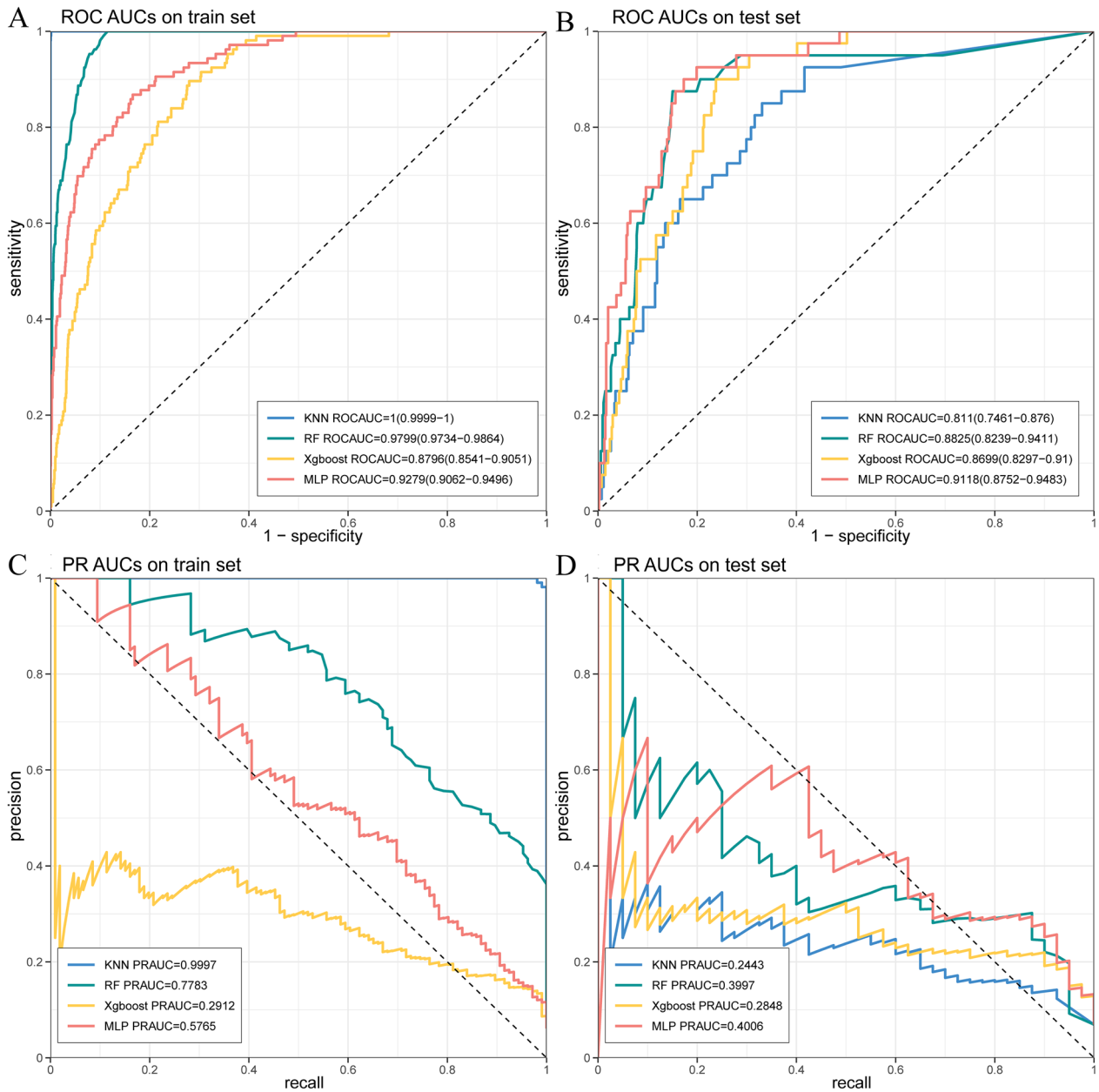

Fig. 3 ROC and PR AUC plots for machine learning models. ROC AUC plots for the training set (**A**) and test set (**B**), along with PR AUC plots for the training set (**C**) and test set (**D**), as generated by the machine learning models.*KNN* K-Nearest Neighbors, *RF* random forest, *XGB* XGBoost, *MLP* multilayer perceptron

characteristics and sarcopenia among individuals with chronic diseases was not a simple linear relationship. The probability of sarcopenia increased sharply when weight dropped below approximately 60 kg, while it remained stable for BMI values above 60. Similarly, sarcopenia risk rose significantly when age exceeded 60 years or BMI fell below 24 kg/m^2, but remained stable when age was below 60 years or BMI was above 24 kg/m^2.

Additionally, BUN levels above 30 mg/dL and HDL-C levels exceeding 50 mg/dL were associated with a notable increase in sarcopenia probability, with no significant decrease observed otherwise. Total cholesterol, cystatin C, triglycerides, and CRP demonstrated a positive linear relationship with sarcopenia risk, whereas height, LDL-C, and TyG showed a negative linear relationship (Figs. 4B, 5). Furthermore, characteristics such as male gender, lower education levels, smoking, chronic digestive

Table 2 DeLong test results for comparing ROC AUC between models

Train set			Test set		
Model comparison	***p* value**	**Z statistic**	**Model comparison**	***p* value**	**Z statistic**
KNN vs RF	**1.28E−09**	**6.06**	KNN vs RF	0.07	− 1.77
KNN vs XGB	**2.27E−20**	**9.24**	KNN vs XGB	0.12	− 1.53
KNN vs MLP	**7.29E−11**	**6.51**	KNN vs MLP	**0.003**	**− 2.92**
RF vs XGB	**4.80E−17**	**8.39**	RF vs XGB	0.577	0.55
RF vs MLP	**5.17E−07**	**5.01**	RF vs MLP	0.12	− 1.51
XGB vs MLP	**5.47E−06**	**− 4.54**	XGB vs MLP	**0.004**	**− 2.84**

KNN K-nearest neighbors, *RF* random forest, *XGB* XGBoost, *MLP* multilayer perceptron

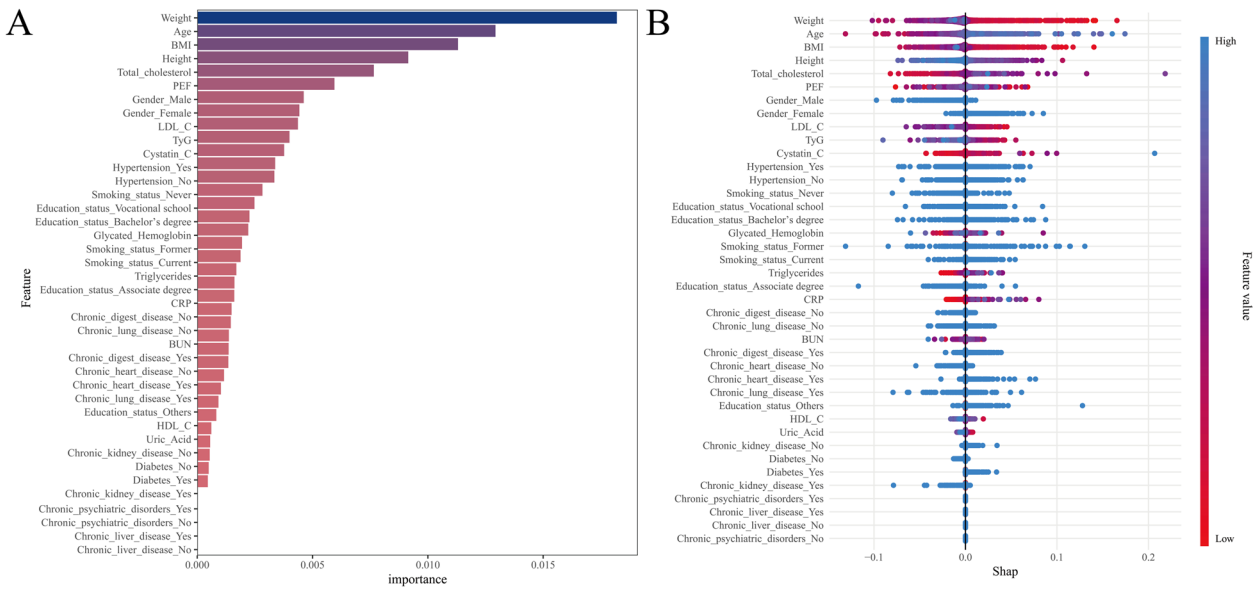

Fig. 4 SHAP results of the MLP model: feature importance bar plot (**A**) and beeswarm plot (**B**)

diseases, chronic heart disease, and diabetes were associated with an elevated risk of sarcopenia (Fig. 6).

RFE algorithm for simplifying the MLP full model

In this section, the RFE algorithm was applied to simplify the MLP full model by iteratively removing less important features based on their SHAP values. We recorded the ROC AUC value in the test set after each iteration to examine the trend of ROC AUC as the number of features decreased. Notably, three key feature count thresholds—5, 12, and 17—marked significant improvements in ROC AUC, as shown in Fig. 6, where the values reached 0.89, 0.91, and 0.924, respectively. After reaching 17 features, the ROC AUC plateaued, indicating no further performance gains with additional features. Based on this curve, we concluded that retaining only the top five features—weight, age, BMI, height, and total cholesterol—along with the common characteristic of gender, enables the MLP model to achieve a satisfactory level of accuracy while maintaining a simplified structure (Fig. 7).

Web tool development and prediction

We utilized an MLP model consisting of only five variables—weight, age, BMI, height, and total cholesterol—selected through the RFE algorithm based on the most important contributors identified by SHAP analysis. As shown in Fig. 8, a web-based tool was developed to easily and accurately predict the probability of sarcopenia occurrence over the next 3–5 years. The website is easily accessible to everyone at https://sasuki.shinyapps.io/wutiaowu2/. After entering the required data and clicking the yellow"Predict"button, the tool automatically displays the predicted probability of sarcopenia along with a SHAP force plot. The SHAP force plot provides a detailed explanation of how each input variable contributes to the predicted probability, helping users understand the

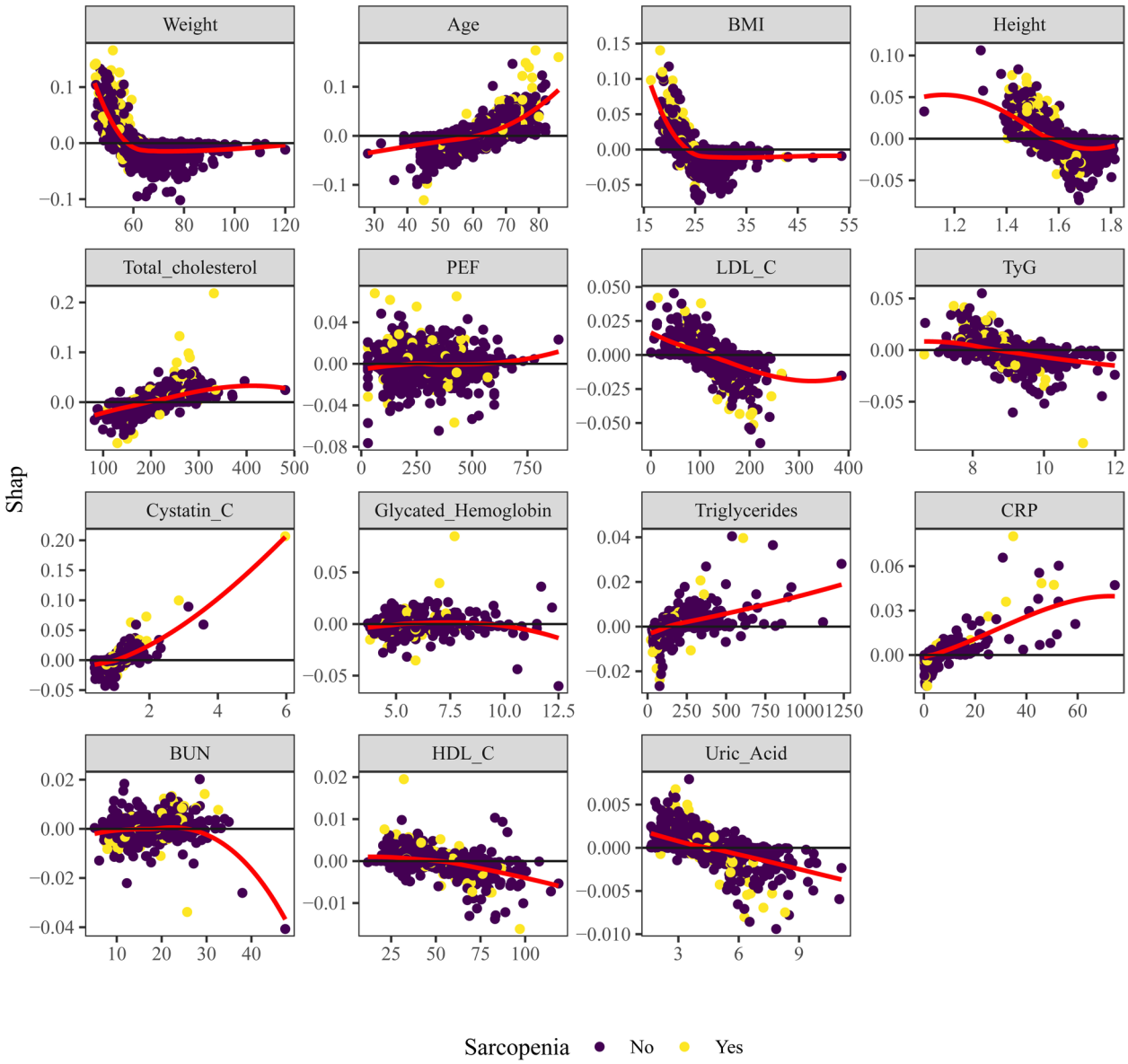

Fig. 5 SHAP results for continuous variables in the MLP model

specific factors driving the sarcopenia risk for an individual with chronic diseases. This offers both a predictive and interpretive framework for risk assessment and decision-making.

Discussion

Sarcopenia is increasingly recognized as a critical prognostic factor in patients with chronic diseases. This muscle-wasting syndrome amplifies the clinical burden and is associated with diminished functional capacity, higher hospitalization rates, and an elevated risk of complications. Due to the relative difficulty in obtaining longitudinal data, previous studies have predominantly relied on cross-sectional analyses, which are often descriptive and lack both systematic exploration and predictive value. This study conducted a comprehensive investigation into the factors influencing the future development of sarcopenia in individuals with chronic conditions, providing valuable insights into its predictive determinants.

Machine learning models demonstrate superior performance in sarcopenia prediction compared to conventional statistical approaches. Seok et al. utilized large-scale national health data from Korea,

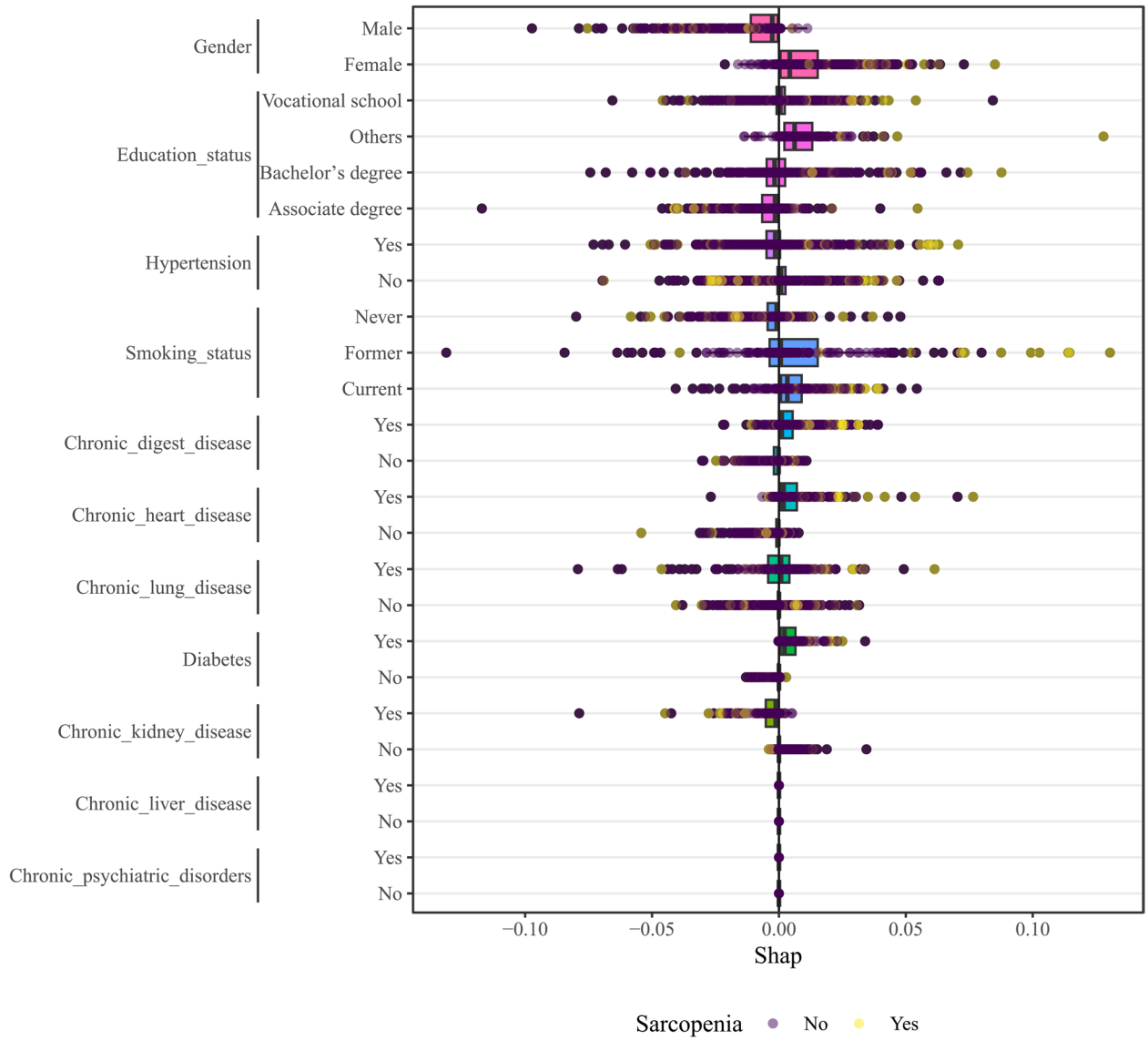

Fig. 6 SHAP results for categorical variables in the MLP model

incorporating comprehensive health characteristics and objectively measured physical activity intensity through validated questionnaires in older adults. Their analysis revealed limitations of logistic regression in handling complex feature interactions, achieving a lower predictive accuracy (AUC: 0.839) than a deep neural network (DNN) model (AUC: 0.869) [36]. Similarly, in a cross-sectional predictive analysis, Kim et al. utilized the same national database, which included 1597 individuals (29%) with sarcopenia. Their findings demonstrated that logistic regression was outperformed by more advanced machine learning models (AUC: 0.85 vs. 0.93) in diagnosing sarcopenia [37]. Although these studies achieved promising predictive performance, they were limited to cross-sectional analyses, which primarily reinforced the predictive value of BMI and physical strength while lacking longitudinal validation. In a recent longitudinal follow-up study, Yin and colleagues similarly utilized data from 11,661 Asian individuals in the openly accessible CHARLS dataset to develop a deep learning model based on common functional capacity (FC) variables. Their gradient boosting classifier (GBC) model, which incorporated 23 common FC features, achieved an AUC of 0.831 in a cross-sectional setting and 0.833 in a longitudinal setting [35]. Yin's study produced SHAP results similar to ours, indicating that weight, age, and height were among the most significant predictors of sarcopenia. By replacing FC variables such as jogging

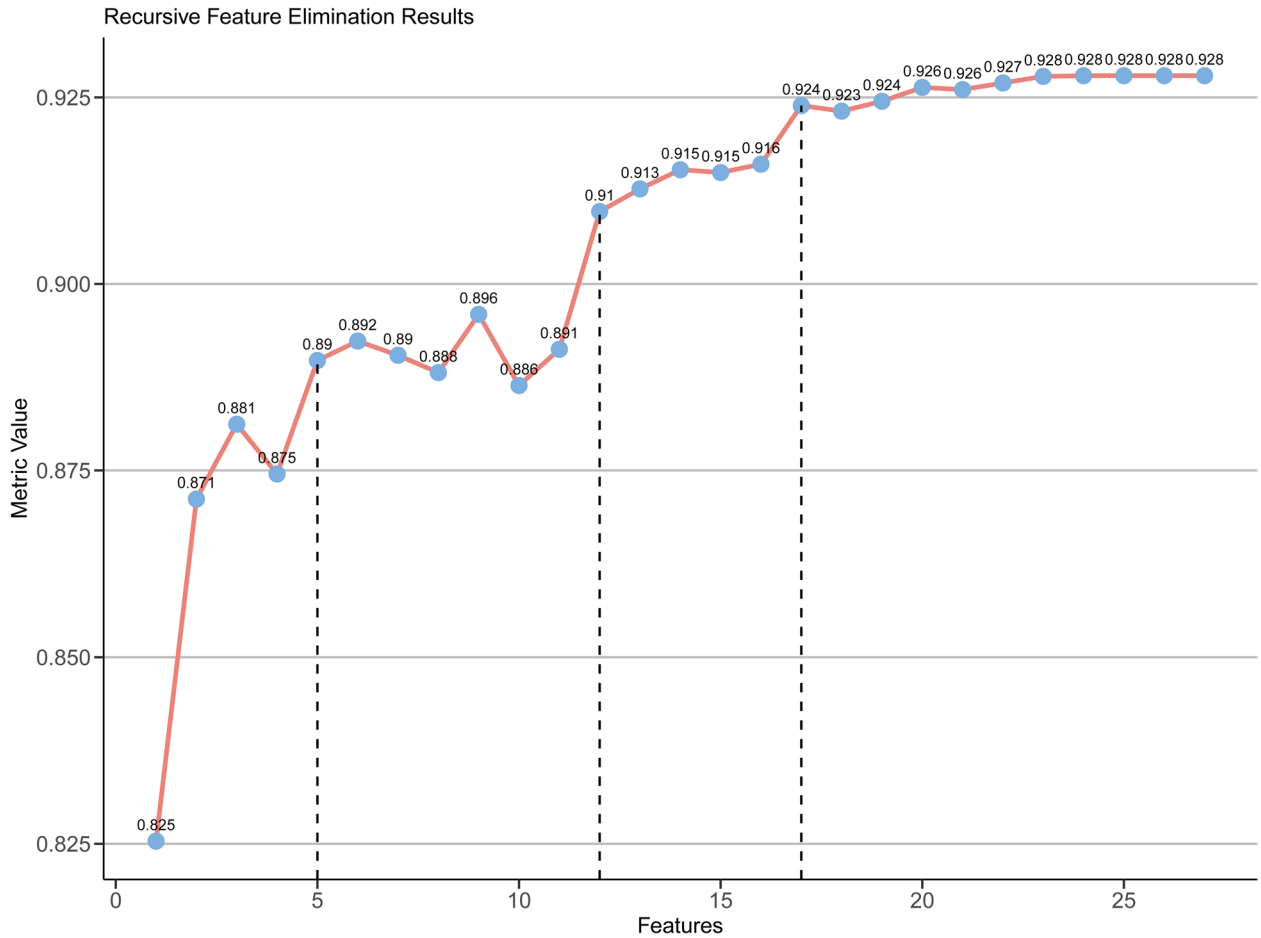

Fig. 7 RFE algorithm-based streamlined model using SHAP importance from the MLP model. *RFE* recursive feature elimination

1 km and lifting 5 kg with readily available serological test, such as lipid profiles and renal function, our model achieved an accuracy of 0.846, a ROC AUC of 0.912, a PR AUC of 0.401, a Kappa value of 0.376, and an F1 score of 0.44. The lower PR AUC, Kappa, and F1 scores are likely due to dataset class imbalance, causing models to favor the majority class and thus reduce false positives, affecting these FP-sensitive metrics. Clinically, however, the priority is minimizing false negatives (missing affected patients), which warrants accepting a higher false positive rate, particularly relevant due to the limited observation window for symptom onset. Future work plans to address this by employing models better suited for imbalanced tabular data and utilizing oversampling techniques like SMOTE/ADASYN. Overall, these metrics slightly outperform Yin's model and are particularly notable for using only six easily obtainable clinical features: weight, age, BMI, height, total cholesterol, and gender. The risk assessment tool developed in this study utilizes a limited set of easily obtainable variables to effectively identify high-risk individuals with chronic diseases who may develop sarcopenia. Accessible through a user-friendly web interface, this tool prompts users to adopt preventive interventions such as tailored exercise regimens, increased intake of amino acids and vitamin D, and other evidence-based measures to mitigate sarcopenia risk. Notably, while this tool prioritizes sensitivity in detecting at-risk populations (resulting in a higher false-positive rate), we recommend that all identified individuals be considered high-risk and encouraged to pursue proactive clinical interventions, including medical consultations.

Apart from height, weight, BMI, age, and gender, which demonstrated strong predictive value for sarcopenia—consistent with Yin's study and previous research [35, 38, 39]—our study also uncovered several additional intriguing findings, as outlined below. Through the main effects analysis of the GLMM, elevated cystatin C levels, along with lower uric acid and BUN levels, were identified as risk indicators for sarcopenia. This finding is consistent with several cross-sectional studies that have reported similar directional

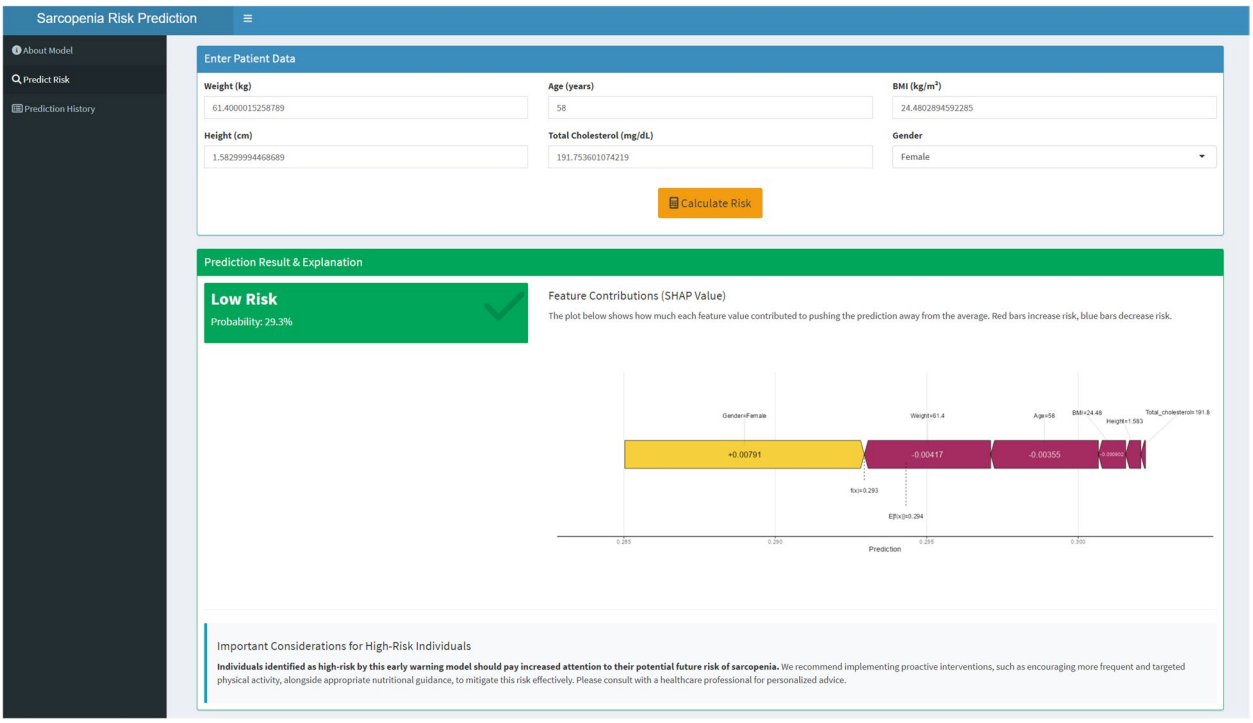

Fig. 8 Web calculator for the streamlined MLP model

effects of renal function markers on sarcopenia [40–43]. However, when examining the interaction effects of these characteristics with time (wave), we found that lower cystatin C levels, higher uric acid, and higher BUN levels were correlated with an increased risk of sarcopenia. The opposing directions of the main effects and the interaction effects of characteristics with time in the GLMM suggest a dual role of renal function-related indicators in sarcopenia. Elevated cystatin C levels could signify declining renal function, which correlates with muscle wasting and the progression of sarcopenia [44, 45]. Lower uric acid and BUN levels might reflect poor nutritional status or insufficient protein intake, leading to reduced muscle synthesis and mass, which are hallmarks of sarcopenia [46–48]. On the other hand, cystatin C, as a cysteine protease inhibitor, plays a role in regulating oxidative stress; thus, its reduction may indicate compromised antioxidant capacity and chronic inflammation, further contributing to sarcopenia development over time [49, 50]. Furthermore, chronic inflammation can elevate uric acid and BUN levels, exacerbating protein catabolism and muscle degradation. Additionally, the elevation of CRP, which is positively associated with sarcopenia, further supports this observation. By combining SHAP explanations derived from the MLP model, we are inclined to note that elevated levels of cystatin C and CRP, which indicate heightened inflammation, as well as lower uric acid and BUN levels, which reflect reduced serum protein metabolism, are more valuable in predicting the risk of sarcopenia.

A recent Genome-wide association study suggested that serum lipid levels, including LDL-C and HDL-C, may genetically reduce the risk of sarcopenia, a finding further supported by multivariate logistic regression analysis in the same study [51]. Using SHAP analysis, we observed a similar trend: when LDL-C and HDL-C levels exceed their respective thresholds of 100 mg/dL and 50 mg/dL, the risk of sarcopenia appears to decrease. Conversely, when total cholesterol and triglyceride levels surpass their normal values of 200 mg/dL and 150 mg/dL, respectively, the risk of sarcopenia shows a linear increase. The PKB/Akt and mTORC1 pathways play a crucial role in this process [52, 53]. Elevated levels of total cholesterol and triglycerides are typically associated with insulin resistance and chronic systemic inflammation, leading to metabolic disturbances, intramuscular fat infiltration, mitochondrial dysfunction, and increased reactive oxygen species (ROS) production, which accelerate muscle degradation and heighten the risk of sarcopenia [54]. Moreover, the triglyceride-to-HDL-C ratio (TG/HDL-C) has been identified as a potential marker of insulin resistance and sarcopenia, further underscoring

the distinct roles of different lipid components in muscle metabolism and health [55].

Among individuals with chronic diseases, the interaction effect of diabetes with time and the MLP-based SHAP analysis both demonstrated that diabetes is closely and positively associated with an increased risk of sarcopenia. In diabetic patients, the actions of GLP-1 and GIP on pancreatic β-cells are impaired, leading to reduced insulin secretion and complicating glycemic control. This reduction in insulin secretion further weakens the PI3 K/Akt/mTOR pathway, negatively impacting muscle protein synthesis [56, 57]. Moreover, diabetes is often associated with a chronic inflammatory state, marked by elevated levels of inflammatory cytokines such as TNF-α and IL-6. These cytokines activate the NF-κB pathway, enhancing the ubiquitin–proteasome system (UPS) activity and accelerating muscle protein degradation [58]. In such patients, medications like GLP-1 receptor agonists (GLP-1 RAs) and SGLT2 inhibitors (SGLT2i) may help counteract insulin resistance-induced muscle atrophy by inhibiting myostatin expression and activating the PI3 K pathway. Additionally, a healthy diet, adequate protein and energy intake, and regular exercise are strongly recommended to reduce the risk of sarcopenia [59]. Other chronic diseases, such as chronic digestive and heart diseases, also showed significantly higher SHAP values for sarcopenia compared to those without these conditions. Therefore, it is also essential to implement tailored nutritional planning for individuals with these conditions.

This study has several limitations. First, as it is based on data from a single center, the developed model may be subject to inevitable population biases. To address this, further studies with larger and more diverse sample sizes are necessary. Second, although the neural network model we developed demonstrates high accuracy, sensitivity, and specificity, other metrics such as positive predictive value, PR AUC, and F1 score are not yet optimal. Furthermore, compared to gold-standard methods such as Dual-energy X-ray Absorptiometry (DXA), the ASM formula's estimated values may exhibit certain biases. This formula derives muscle mass estimates through a regression model rather than directly measuring skeletal muscle mass, which may not fully capture individual variability or the specific characteristics of our study cohort—particularly in individuals with chronic diseases, which can independently alter body composition. Additionally, some potential confounders, such as nutritional intake and medication use, were not adequately accounted for. Patients with missing values for certain features were also excluded, inevitably introducing selection bias and confounding bias. Consequently, the generalizability of this study requires further validation through longitudinal studies in diverse ethnic and larger populations. Third, some discrepancies were observed between the GLMM model and the SHAP model for certain features, such as LDL-C. These discrepancies probably reflect the difference between SHAP capturing complex ML model behavior and the GLMM's linear framework. This highlights that predicting sarcopenia is not straightforward and suggests that future longitudinal studies providing richer, more detailed follow-up data would likely substantially improve machine learning prediction accuracy. Nevertheless, the web tool developed in this research indeed provides a longitudinal prediction scheme for sarcopenia in individuals with chronic diseases, which is consistent with biological and medical principles. We believe this tool holds significant value for early warning of high-risk populations and encouraging them in advance to pursue preventive measures, such as medication or exercise.

Conclusion

This study utilized data from the Asian Follow-up Database, comprising 2891 individuals with chronic diseases, all of whom had follow-up data available for both wave 1 and wave 3. Weight, age, BMI, height, and total cholesterol were identified as the most significant predictors of sarcopenia risk within the following 3–5 years in patients with chronic diseases. Specifically, individuals with a weight below 60 kg, age over 60 years, BMI less than 24 kg/m^2, height under 1.6 m, or total cholesterol exceeding 200 mg/dL were found to have an increased risk of developing sarcopenia. We recommend that individuals with these high-risk factors be prioritized for targeted health warnings and proactive interventions. The full model, developed using the best-performing MLP model, achieved a ROC AUC of 0.912, PR AUC of 0.401, sensitivity of 0.875, specificity of 0.844, a Kappa value of 0.376, and an F1 score of 0.44 on the test set. A streamlined model developed based on longitudinal follow-up data provides an accurate and convenient tool for the early prediction of sarcopenia. This model, created using the RFE algorithm with only weight, age, BMI, height, total cholesterol, and gender as inputs, achieved a ROC AUC of approximately 0.9. This streamlined tool was further developed into a web-based calculator for practical use (accessible at https://sasuki.shinyapps.io/wutiaowu2/). Individuals identified as high-risk by this early warning model should pay increased attention to their potential future risk of sarcopenia. We recommend implementing proactive interventions, such as encouraging more frequent and targeted physical activity, to mitigate this risk effectively.

Abbreviations

ADASYN	Adaptive synthetic sampling
Akt	Protein kinase B (Part of signaling pathway)
ANOVA	Analysis of variance
ASM	Appendicular skeletal muscle mass (kg/m^2 (when height-adjusted: ASM/Height^2))
AUC	Area under the curve
AWGS	Asian Working Group for Sarcopenia
BMI	Body Mass Index (kg/m^2)
BUN	Blood urea nitrogen (mg/dL)
CHARLS	China Health and Retirement Longitudinal Study
CKD	Chronic kidney disease
CRP	C-reactive protein (mg/L)
CVD	Cardiovascular disease
DNN	Deep neural network
DXA	Dual-energy X-ray absorptiometry
FC	Functional capacity
FoxO3a	Forkhead box protein O3a (part of signaling pathway)
GBC	Gradient boosting classifier
GIP	Glucose-dependent insulinotropic polypeptide
GLMM	Generalized linear mixed model
GLP-1	Glucagon-like peptide-1
GLP-1 RAs	Glucagon-like peptide-1 receptor agonists
HbA1c	Glycated hemoglobin (%)
HDL-C	High-density lipoprotein cholesterol (mg/dL)
IL-1β	Interleukin-1 beta
IL-6	Interleukin-6
KNN	K-nearest neighbors
LDL-C	Low-density lipoprotein cholesterol (mg/dL)
METS-IR	Metabolic score for insulin resistance
MLP	Multilayer perceptron
mTOR	Mammalian target of rapamycin (part of pathway)
NF-κB	Nuclear factor kappa-light-chain-enhancer of activated B cells
NHANES	National Health and Nutrition Examination Survey
PEF	Peak expiratory flow (L/min)
PI3 K	Phosphoinositide 3-kinase (part of signaling pathway)
PR AUC	Precision-recall area under the curve
RF	Random forest
RFE	Recursive feature elimination
ROC AUC	Receiver operating characteristic area under the curve
ROS	Reactive oxygen species
SGLT2i	Sodium-glucose cotransporter 2 inhibitors
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic minority over-sampling technique
TG/HDL-C	Triglyceride-to-HDL-C ratio (ratio)
TNF-α	Tumor necrosis factor-alpha
TyG	Triglyceride-Glucose Index (ln(fasting triglycerides [mg/dL] × fasting glucose [mg/dL]/2))
UPS	Ubiquitin-proteasome system
XGB	Extreme gradient boosting

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40001-025-02606-3.

Supplementary Material 1: Figure S1. Interaction effects of characteristics with time in GLMM. Figure S2. Learning curves of machine learning algorithms. Figure S3. Calibration plots of machine learning algorithms in the training set. Figure S4. Calibration plots of machine learning algorithms in the test set. Figure S5. Cross-validation ROC and PR AUC plots of machine learning models: ROC AUC plots, PR AUC plots. Figure S6. Visualization of hyperparameter tuning results for the MLP model. Figure S7. Illustration of the neural network structure. The numbers of input, hidden, and output layers are 1, 2, and 1, respectively

Supplementary Material 2: Table S1. Baseline characteristics of individuals with chronic diseases using quantile category. Table S2. Results of the generalized linear mixed model for fixed effects. Table S3. Baseline comparison between train set and test set. Table S4. Comprehensive model performance

Acknowledgements
We sincerely acknowledge all those who selflessly share their publicly available statistical data and open-source computational software.

Author contributions
X Y conceived and supervised the study. K R and G·Y collected the data and conducted the analysis. C Z completed visualization. X Y prepared the final draft and approved the final submission.

Funding
Not available.

Data availability
All data in current analysis can be requested from the corresponding author.

Declarations

Ethics approval and consent to participate
This study was conducted via openly available statistics and therefore did not require ethical approval.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Pulmonary and Critical Care Medicine, Yongchuan Hospital of Chongqing Medical University, Chongqing, China. [2]Kuitun Hospital of Ili Kazakh Autonomous Prefecture, Kuitun, Chongqing 833200, China. [3]The First Hospital Affiliated with Third Military Medical University, Chongqing, China. [4]Department of Respiratory Medicine, Southwest Hospital of Third Military Medical University, Chongqing, China.

Received: 5 March 2025 Accepted: 16 April 2025
Published online: 29 April 2025

References

1. Zhang Y, Zhang J, Ni W, Yuan X, Zhang H, Li P, et al. Sarcopenia in heart failure: a systematic review and meta-analysis. ESC Heart Fail. 2021;8:1007–17.
2. Koo BK, Roh E, Yang YS, Moon MK. Difference between old and young adults in contribution of β-cell function and sarcopenia in developing diabetes mellitus. J Diabetes Investig. 2016;7:233–40.
3. Duarte MP, Almeida LS, Neri SGR, Oliveira JS, Wilkinson TJ, Ribeiro HS, et al. Prevalence of sarcopenia in patients with chronic kidney disease: a global systematic review and meta-analysis. J Cachexia Sarcopenia Muscle. 2024;15:501–12.
4. Guañabens N, Parés A. Osteoporosis in chronic liver disease. Liver Int. 2018;38:776–85.
5. Sepúlveda-Loyola W, Osadnik C, Phu S, Morita AA, Duque G, Probst VS. Diagnosis, prevalence, and clinical impact of sarcopenia in COPD: a systematic review and meta-analysis. J Cachexia Sarcopenia Muscle. 2020;11:1164–76.
6. Bano G, Trevisan C, Carraro S, Solmi M, Luchini C, Stubbs B, et al. Inflammation and sarcopenia: a systematic review and meta-analysis. Maturitas. 2017;96:10–5.
7. Buchmann N, Fielitz J, Spira D, König M, Norman K, Pawelec G, et al. Muscle mass and inflammation in older adults: impact of the metabolic syndrome. Gerontology. 2022;68:989–98.
8. Gao J, Deng M, Li Y, Yin Y, Zhou X, Zhang Q, et al. Resistin as a systemic inflammation-related biomarker for sarcopenia in patients with chronic obstructive pulmonary disease. Front Nutr. 2022;9: 921399.
9. de Sire R, Rizzatti G, Ingravalle F, Pizzoferrato M, Petito V, Lopetuso L, et al. Skeletal muscle-gut axis: emerging mechanisms of sarcopenia for intestinal and extra intestinal diseases. Minerva Gastroenterol Dietol. 2018;64:351–62.
10. Oh H-J, Jin H, Lee J-Y, Lee B-Y. Silk peptide ameliorates sarcopenia through the regulation of Akt/mTOR/FoxO3a signaling pathways and

the inhibition of low-grade chronic inflammation in aged mice. Cells. 2023;12:2257.
11. Livshits G, Kalinkovich A. Restoration of epigenetic impairment in the skeletal muscle and chronic inflammation resolution as a therapeutic approach in sarcopenia. Ageing Res Rev. 2024;96: 102267.
12. Xu J, Pan Y, Zhang J, Dai S, Xu L. Sarcopenia in liver cirrhosis: perspectives from epigenetics and microbiota. Front Med (Lausanne). 2023;10:1264205.
13. Jin H, Xie W, He M, Li H, Xiao W, Li Y. Pyroptosis and sarcopenia: frontier perspective of disease mechanism. Cells. 2022;11:1078.
14. Haberecht-Müller S, Krüger E, Fielitz J. Out of control: the role of the ubiquitin proteasome system in skeletal muscle during inflammation. Biomolecules. 2021;11:1327.
15. Nishikawa H, Fukunishi S, Asai A, Yokohama K, Nishiguchi S, Higuchi K. Pathophysiology and mechanisms of primary sarcopenia (Review). Int J Mol Med. 2021;48:1–8.
16. Yang YJ, Kim DJ. An overview of the molecular mechanisms contributing to musculoskeletal disorders in chronic liver disease: osteoporosis, sarcopenia, and osteoporotic sarcopenia. Int J Mol Sci. 2021;22:2604.
17. Zuo X, Li X, Tang K, Zhao R, Wu M, Wang Y, et al. Sarcopenia and cardiovascular diseases: a systematic review and meta-analysis. J Cachexia Sarcopenia Muscle. 2023;14:1183–98.
18. Zhao X, Su R, Hu R, Chen Y, Xu X, Yuan Y, et al. Sarcopenia index as a predictor of clinical outcomes among older adult patients with acute exacerbation of chronic obstructive pulmonary disease: a cross-sectional study. BMC Geriatr. 2023;23:89.
19. Hsu W-H, Wang S-Y, Chao Y-M, Chang K-V, Han D-S, Lin Y-L. Novel metabolic and lipidomic biomarkers of sarcopenia. J Cachexia Sarcopenia Muscle. 2024;15:2175–86.
20. Guan C, Gong A, Zhao Y, Yin C, Geng L, Liu L, et al. Interpretable machine learning model for new-onset atrial fibrillation prediction in critically ill patients: a multi-center study. Crit Care. 2024;28:349.
21. Qiu W, Chen H, Dincer AB, Lundberg S, Kaeberlein M, Lee S-I. Interpretable machine learning prediction of all-cause mortality. Commun Med (Lond). 2022;2:125.
22. Jiang M, Ren X, Han L, Zheng X. Associations between sarcopenic obesity and risk of cardiovascular disease: a population-based cohort study among middle-aged and older adults using the CHARLS. Clin Nutr. 2024;43:796–802.
23. Liu Y, Cui J, Cao L, Stubbendorff A, Zhang S. Association of depression with incident sarcopenia and modified effect from healthy lifestyle: the first longitudinal evidence from the CHARLS. J Affect Disord. 2024;344:373–9.
24. Chen L-K, Woo J, Assantachai P, Auyeung T-W, Chou M-Y, Iijima K, et al. Asian working group for sarcopenia: 2019 consensus update on sarcopenia diagnosis and treatment. J Am Med Dir Assoc. 2020;21:300-307.e2.
25. Zhou S, Liu Y, Zhang Y, Luo N, Chen Q, Ge M, et al. Association between persistent musculoskeletal pain and incident sarcopenia in China: the mediating effect of depressive symptoms. Front Public Health. 2024;12:1416796.
26. Wen X, Wang M, Jiang C-M, Zhang Y-M. Anthropometric equation for estimation of appendicular skeletal muscle mass in Chinese adults. Asia Pac J Clin Nutr. 2011;20:551–6.
27. Duan M, Zhao X, Li S, Miao G, Bai L, Zhang Q, et al. Metabolic score for insulin resistance (METS-IR) predicts all-cause and cardiovascular mortality in the general population: evidence from NHANES 2001–2018. Cardiovasc Diabetol. 2024;23:243.
28. Zhang Y, Wang F, Tang J, Shen L, He J, Chen Y. Association of triglyceride glucose-related parameters with all-cause mortality and cardiovascular disease in NAFLD patients: NHANES 1999–2018. Cardiovasc Diabetol. 2024;23:262.
29. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn. 2002;46:389–422.
30. Wei H-L, Billings SA. Modeling COVID-19 Pandemic Dynamics Using Transparent, Interpretable, Parsimonious and Simulatable (TIPS) Machine Learning Models: A Case Study from Systems Thinking and System Identification Perspectives. In: Recent Advances in AI-enabled Automated Medical Diagnosis. CRC Press; 2022.
31. Deng F, Zhao L, Yu N, Lin Y, Zhang L. Union With Recursive Feature Elimination: A Feature Selection Framework to Improve the Classification Performance of Multicategory Causes of Death in Colorectal Cancer. Lab Invest. 2024;104: 100320.
32. Diggle PJ, Heagerty P, Liang KY, Zeger SL. Analysis of Longitudinal Data. 2nd ed. Oxford: Oxford University Press; 2002.
33. Collins GS, Dhiman P, Ma J, Schlussel MM, Archer L, Van Calster B, et al. Evaluation of clinical prediction models (part 1): from development to external validation. BMJ. 2024;384: e074819.
34. Hosoi T, Yakabe M, Hashimoto S, Akishita M, Ogawa S. The roles of sex hormones in the pathophysiology of age-related sarcopenia and frailty. Reprod Med Biol. 2024;23: e12569.
35. Yin L, Zhao J. An Artificial Intelligence Approach for Test-Free Identification of Sarcopenia. J Cachexia Sarcopenia Muscle. 2024;15:2765–80.
36. Seok M, Kim W. Sarcopenia Prediction for Elderly People Using Machine Learning: A Case Study on Physical Activity. Healthcare. 2023;11:1334.
37. Kim J. Machine-learning classifier models for predicting sarcopenia in the elderly based on physical factors. Geriatr Gerontol Int. 2024;24:595–602.
38. Zhang G, Fu J, Zhang H, Xu X, Cao Z. The impact of Life's Essentials 8 on sarcopenia prevalence among adults in the United States. Exp Gerontol. 2024;198: 112631.
39. Lin W-S, Hsu N-W, Yang S-H, Chen Y-T, Tsai C-C, Pan P-J. Predicting sarcopenia in community-dwelling older adults through comprehensive physical fitness tests. BMC Geriatr. 2024;24:932.
40. Shah LN, Leonard MB, Ziolkowski SL, Grimm P, Long J. Cystatin C and Creatinine Concentrations Are Uninformative Biomarkers of Sarcopenia: A Cross-Sectional NHANES Study. J Ren Nutr. 2023;33:538–45.
41. Matsuzawa R, Nagai K, Takahashi K, Mori T, Onishi M, Tsuji S, et al. Serum Creatinine-Cystatin C Based Screening of Sarcopenia in Community Dwelling Older Adults: A Cross-Sectional Analysis. J Frailty Aging. 2024;13:116–24.
42. Liu X, Chen X, Hu F, Xia X, Hou L, Zhang G, et al. Higher uric acid serum levels are associated with sarcopenia in west China: a cross-sectional study. BMC Geriatr. 2022;22:121.
43. Gao H, Wang J, Zou X, Zhang K, Zhou J, Chen M. High blood urea nitrogen to creatinine ratio is associated with increased risk of sarcopenia in patients with chronic obstructive pulmonary disease. Exp Gerontol. 2022;169: 111960.
44. Malmgren L, Grubb A. Muscle mass, creatinine, cystatin C and selective glomerular hypofiltration syndromes. Clin Kidney J. 2023;16:1206–10.
45. An JN, Kim J-K, Lee H-S, Kim SG, Kim HJ, Song YR. Serum cystatin C to creatinine ratio is associated with sarcopenia in non-dialysis-dependent chronic kidney disease. Kidney Res Clin Pract. 2022;41:580–90.
46. Nahas PC, Rossato LT, de Branco FMS, Azeredo CM, Rinaldi AEM, de Oliveira EP. Serum uric acid is positively associated with muscle strength in older men and women: Findings from NHANES 1999–2002. Clin Nutr. 2021;40:4386–93.
47. Floriano JP, Nahas PC, de Branco FMS, Dos Reis AS, Rossato LT, Santos HO, et al. Serum Uric Acid Is Positively Associated with Muscle Mass and Strength, but Not with Functional Capacity, in Kidney Transplant Patients. Nutrients. 2020;12:2390.
48. Wang XH, Mitch WE. Mechanisms of muscle wasting in chronic kidney disease. Nat Rev Nephrol. 2014;10:504–16.
49. Kang D-H, Nakagawa T, Feng L, Watanabe S, Han L, Mazzali M, et al. A role for uric acid in the progression of renal disease. J Am Soc Nephrol. 2002;13:2888–97.
50. Kitago M, Seino S, Shinkai S, Nofuji Y, Yokoyama Y, Toshiki H, et al. Cross-Sectional and Longitudinal Associations of Creatinine-to-Cystatin C Ratio with Sarcopenia Parameters in Older Adults. J Nutr Health Aging. 2023;27:946–52.
51. Huang H, Yu X, Jiang S, Wang C, Chen Z, Chen D, et al. The relationship between serum lipid with sarcopenia: Results from the NHANES 2011–2018 and bidirectional Mendelian randomization study. Exp Gerontol. 2024;196: 112560.
52. Shad BJ, Smeuninx B, Atherton PJ, Breen L. The mechanistic and ergogenic effects of phosphatidic acid in skeletal muscle. Appl Physiol Nutr Metab. 2015;40:1233–41.
53. Mahfouz R, Khoury R, Blachnio-Zabielska A, Turban S, Loiseau N, Lipina C, et al. Characterising the inhibitory actions of ceramide upon insulin signaling in different skeletal muscle cell models: a mechanistic insight. PLoS ONE. 2014;9: e101865.

54. Li C-W, 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;13:781–94.
55. Nur Zati Iwani AK, Jalaludin MY, Yahya A, Mansor F, Md Zain F, Hong JYH, et al. TG: HDL-C Ratio as Insulin Resistance Marker for Metabolic Syndrome in Children With Obesity. Front Endocrinol (Lausanne). 2022;13:852290.
56. Prasun P. Role of mitochondria in pathogenesis of type 2 diabetes mellitus. J Diabetes Metab Disord. 2020;19:2017–22.
57. Feng L, Gao Q, Hu K, Wu M, Wang Z, Chen F, et al. Prevalence and Risk Factors of Sarcopenia in Patients With Diabetes: A Meta-analysis. J Clin Endocrinol Metab. 2022;107:1470–83.
58. Kamiya M, Mizoguchi F, Yasuda S. Amelioration of inflammatory myopathies by glucagon-like peptide-1 receptor agonist via suppressing muscle fibre necroptosis. J Cachexia Sarcopenia Muscle. 2022;13:2118–31.
59. Hashimoto Y, Takahashi F, Okamura T, Hamaguchi M, Fukui M. Diet, exercise, and pharmacotherapy for sarcopenia in people with diabetes. Metabolism. 2023;144: 155585.

Publisher's Note