BMC Geriatrics

RESEARCH

Open Access

A prediction study on the occurrence risk of heart disease in older hypertensive patients based on machine learning

Fei Si[1], Qian Liu[1] and Jing Yu[1*]

Abstract

Objective Constructing a predictive model for the occurrence of heart disease in elderly hypertensive individuals, aiming to provide early risk identification.

Methods A total of 934 participants aged 60 and above from the China Health and Retirement Longitudinal Study with a 7-year follow-up (2011–2018) were included. Machine learning methods (logistic regression, XGBoost, DNN) were employed to build a model predicting heart disease risk in hypertensive patients. Model performance was comprehensively assessed using discrimination, calibration, and clinical decision curves.

Results After a 7-year follow-up of 934 older hypertensive patients, 243 individuals (26.03%) developed heart disease. Older hypertensive patients with baseline comorbid dyslipidemia, chronic pulmonary diseases, arthritis or rheumatic diseases faced a higher risk of future heart disease. Feature selection significantly improved predictive performance compared to the original variable set. The ROC-AUC for logistic regression, XGBoost, and DNN were 0.60 (95% CI: 0.53–0.68), 0.64 (95% CI: 0.57–0.71), and 0.67 (95% CI: 0.60–0.73), respectively, with logistic regression achieving optimal calibration. XGBoost demonstrated the most noticeable clinical benefit as the threshold increased.

Conclusion Machine learning effectively identifies the risk of heart disease in older hypertensive patients based on data from the CHARLS cohort. The results suggest that older hypertensive patients with comorbid dyslipidemia, chronic pulmonary diseases, and arthritis or rheumatic diseases have a higher risk of developing heart disease. This information could facilitate early risk identification for future heart disease in older hypertensive patients.

Keywords Hypertension, Heart Disease, Machine Learning, Risk Prediction, Older Patients

Introduction

Hypertension remains a prominent global public health issue, posing a significant threat to human health. It is estimated that in 2019, the global prevalence of hypertension in adults reached 1.3 billion, with a continuous upward trend [1]. In China alone, approximately 330 million individuals suffer from cardiovascular diseases, with a staggering 245 million cases attributed to hypertension [2]. Hypertension stands as a major risk factor for heart disease, contributing to nearly 10.8 million deaths annually [3]. As a modifiable risk factor, hypertension continues to be a primary contributor to cardiovascular diseases. With the accelerated aging of the global population, the prevalence of hypertension is steadily increasing, particularly among older hypertensive patients. Notably, the prognosis for heart disease in older hypertensive patients tends to be worse, leading to a heavier disease burden. Given the substantial population of older hypertensive patients, early identification of those

*Correspondence:
Jing Yu
ery_jyu@lzu.edu.cn
[1] Department of Cardiology, The Second Hospital & Clinical Medical School, Lanzhou University, No. 82 Cuiyingmen, Lanzhou 730000, China

at high risk of developing heart disease through epidemiological data could facilitate personalized interventions, thereby improving the prognosis for this vulnerable population. Therefore, the construction of a predictive model for heart disease risk in elderly hypertensive individuals is of paramount importance, contributing significantly to the early identification of high-risk individuals and subsequent implementation of personalized interventions for improved outcomes. In disease prediction research, regression analysis is a commonly used predictive method, such as Cox regression [4, 5]. However, these methods may struggle with nonlinear problems and, when dealing with numerous predictor variables, might fail to discern complex relationships between predictors and outcomes. Machine learning, as a category of computer-dependent algorithms, has gained widespread application in disease risk prediction in recent years by iteratively learning from input data to efficiently predict on new datasets [6].

Logistic regression serves as the most fundamental machine learning model, known for its simple model setup and frequent use as a benchmark against other machine learning models. XGBoost (Extreme Gradient Boosting) is an ensemble algorithm belonging to the boosting family. It improves upon the Gradient Boosting Decision Tree (GBDT) algorithm by incorporating algorithmic and engineering enhancements, thereby further enhancing model computation speed and efficiency [7, 8]. Deep learning is also a common model in the field of machine learning. Deep Neural Networks (DNN) is a neural network comprising multiple hidden layers, serving as a foundational deep learning model with numerous successful applications in classification problems [9, 10].

In this study, leveraging epidemiological information (demographics, lifestyle behaviors, and medical history), XGBoost and DNN were applied to predict heart disease in older hypertensive patients. Feature engineering was employed to optimize the predictive models, and comparisons were made with the traditional logistic regression (LR) model. The evaluation of predictive models was comprehensively conducted using discrimination, calibration, and decision curves. This study aims to provide insights for selecting the optimal model for early screening of heart disease in older hypertensive patients.

Materials and methods

Data source

The data for this study were obtained from the China Health and Retirement Longitudinal Study (CHARLS), a high-quality longitudinal cohort study representative of the Chinese population aged 45 and above. In this study, we utilized data from the year 2011 as the baseline and the most recent follow-up in 2018 as the outcome assessment. Inclusion criteria for the study participants were: (1) baseline age ≥ 60 years, and (2) baseline diagnosis of hypertension (systolic blood pressure ≥ 140 mmHg and/or diastolic blood pressure ≥ 90 mmHg.). Exclusion criteria were: (1) missing data on baseline hypertension status and heart disease (including heart attack, coronary heart disease, angina, congestive heart failure, or other heart problems), (2) baseline presence of heart disease, and (3) missing data on heart disease status in 2018. A total of 934 eligible study participants were ultimately identified, with 243 experiencing a heart disease event in 2018.

Variable selection and measurement

Based on a combination of literature review and the CHARLS database, this study incorporated four major categories of variables at baseline. These include demographic variables such as age, gender, and waist-to-height ratio; behavioral factors including smoking, alcohol consumption, and physical activity; medical history encompassing dyslipidemia (elevated low-density lipoprotein, triglycerides, and total cholesterol, or decreased high-density lipoprotein levels), diabetes, malignant tumors, chronic lung diseases (e.g., chronic bronchitis, emphysema, excluding tumors or cancer), heart and liver diseases, stroke, kidney diseases, digestive system disorders, emotional and mental health issues, memory-related diseases, arthritis or rheumatic diseases, and asthma. The predicted outcome was the occurrence of heart disease, as self-reported by elderly individuals or reported by their family members. Specifically, the study followed individuals with hypertension at baseline in 2011 to ascertain whether they developed heart disease during the continuous follow-up until 2018.

Data preprocessing

Firstly, an analysis of outliers was conducted. Since the variables did not follow a normal distribution, outliers were identified using Tukey's test. Specifically, data points below Q1-1.5*QR or above Q3+1.5*QR (where Q1 and Q3 are the lower and upper quartiles, and IQR is the interquartile range) were considered outliers and treated as missing values. Detected outliers were handled as missing values in this study.

Furthermore, various variables exhibited different levels of missingness, as shown in Table 1, and the heatmap of missing values is depicted in Fig. 1. Except for the waist-to-height ratio, the missing proportions for other variables were all less than 4.00% (with the waist-to-height ratio having the highest missing proportion at 17.77%). For missing values, imputation was performed using the random forest algorithm. For continuous variables, imputation was done using a weighted average of

Table 1 Sample Missing Data Overview

Variables	Valid Samples	Missing Samples	Missing Ratios
Sociodemographic			
Age	934	0	0.00%
Gender	934	0	0.00%
Waist-to-Height Ratio	768	166	17.77%
Lifestyle Behaviors			
Sleep Duration	900	34	3.64%
Alcohol Consumption	933	1	0.11%
Smoking	934	0	0.00%
Medical History			
Dyslipidemia	911	23	2.46%
Diabetes	927	7	0.75%
Tumors	929	5	0.54%
Chronic Respiratory Diseases	931	3	0.32%
Liver Diseases	932	2	0.21%
Kidney Diseases	929	5	0.54%
Stroke	930	4	0.43%
Gastrointestinal Diseases	933	1	0.11%
Psychological Issues	930	4	0.43%
Memory-Related Disorders	931	3	0.32%
Arthritis or Rheumatic Disease	931	3	0.32%
Asthma	932	2	0.21%

non-missing samples, while for categorical variables, imputation was based on the category with the highest average proximity. To address differences in scale among variables, the Min–Max normalization method was applied, ensuring that all variable values fell within the range of [0,1].

Feature Selection

The selection of predictor variables has a decisive impact on the performance of the final model. In this study, feature selection was conducted using the method of information gain combined with stepwise forward/backward selection. Specifically, the process involved calculating the information gain for each variable and arranging them in descending order. The variable with the highest information gain (V1) was then incorporated into the random forest (RF) model, and its F1 score was computed. Subsequently, the variable with the second-highest information gain (V2) was added, and the F1 score for the model containing both V1 and V2 variables (denoted as F1') was calculated. If F1' > F1, V2 was included; otherwise, V2 was excluded. This process was iteratively repeated for all variables until all variables were explored.

Model construction and validation

In this study, predictive models were constructed using logistic regression, XGBoost from ensemble learning, and DNN from deep learning. The dataset was divided into training and testing sets in a 70:30 ratio, with the training set used for model training and the testing set for model performance evaluation. During the training process, grid search was employed to fine-tune model hyperparameters. In the testing phase, model discrimination was assessed using accuracy, precision, recall, and F1 score. The Area Under the Curve (AUC) metric with a 95% confidence interval (CI) was estimated using the bootstrap method. Model calibration was evaluated using the Hosmer and Lem show goodness-of-fit test. Additionally, decision curve analysis (DCA) was utilized to assess the clinical utility of the models. The entire model development process is illustrated in Fig. 2.

Statistical analysis

Continuous variables were described using median and interquartile range, while categorical variables were described using percentages. All analyses in this study were conducted using R version 3.6.0. Specifically, the XGBoost model was implemented using the XGBoost package, and the construction of the deep neural network utilized the h2o package. Model AUC calculation and pairwise comparisons of AUC were performed using the pROC package. Model calibration analysis was conducted using the PredictABEL package, with the random seed set to 123. A two-sided p-value < 0.05 was considered statistically significant.

Results

Baseline characteristics of study participants

After a 7-year follow-up of 934 hypertensive patients at baseline, 243 individuals experienced heart disease, resulting in a 7-year incidence rate of 26.03%. A comparison of baseline characteristics between the heart disease and non-heart disease groups is presented in Table 2. The average age of the study population was 62.83 years, with males comprising 47.20%. The waist-to-height ratio was 0.52, and the average sleep duration was 5 h. The smoking rate was higher than the alcohol consumption rate (38.12% vs. 27.62%). Among chronic diseases, arthritis or rheumatic diseases, digestive system disorders, dyslipidemia, chronic lung diseases, and diabetes exhibited a higher prevalence compared to other conditions. By comparing baseline characteristics between hypertensive patients with and without accompanying heart disease, it was observed that hypertensive patients with baseline dyslipidemia,

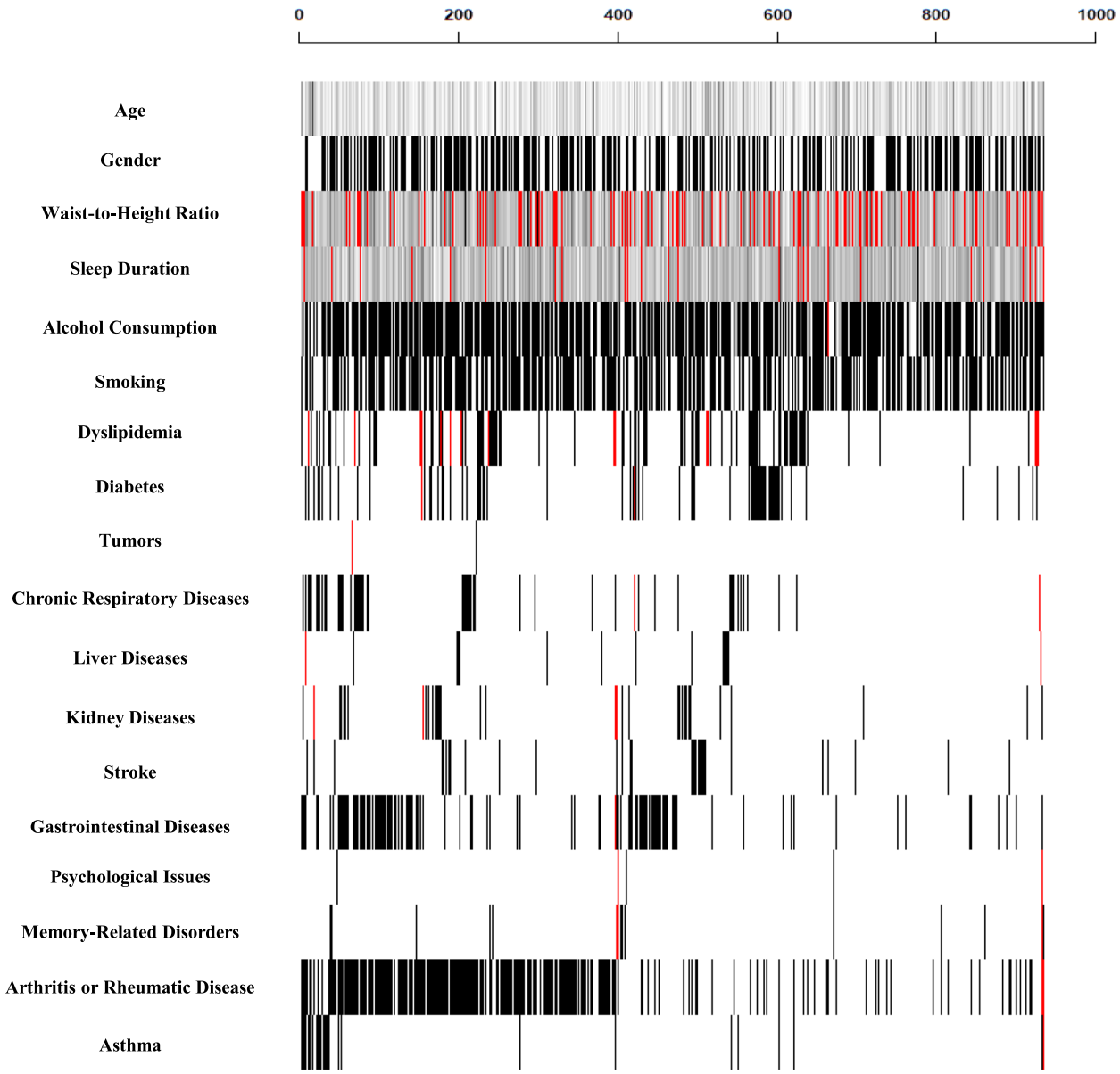

Fig. 1 Heatmap of Missing Values

chronic lung diseases, and arthritis or rheumatic diseases had a higher risk of developing heart disease.

Feature selection

Feature selection was performed across the entire dataset through feature engineering, and the results are illustrated in Fig. 3. Figure 3-A displays the importance of each variable, with the top five being waist-to-height ratio, age, sleep duration, alcohol consumption, and arthritis or rheumatic diseases. Figure 3-B presents the outcomes of feature selection, identifying a total of 7 variables. These include demographic variables age and waist-to-height ratio; lifestyle factor alcohol consumption; and medical history variables diabetes, chronic lung diseases, dyslipidemia, and arthritis or rheumatic diseases.

Model performance

The models underwent parameter optimization through tenfold cross-validation and grid search on the training set. Logistic regression used default parameters; XGBoost was tuned for learning rate (eta), feature

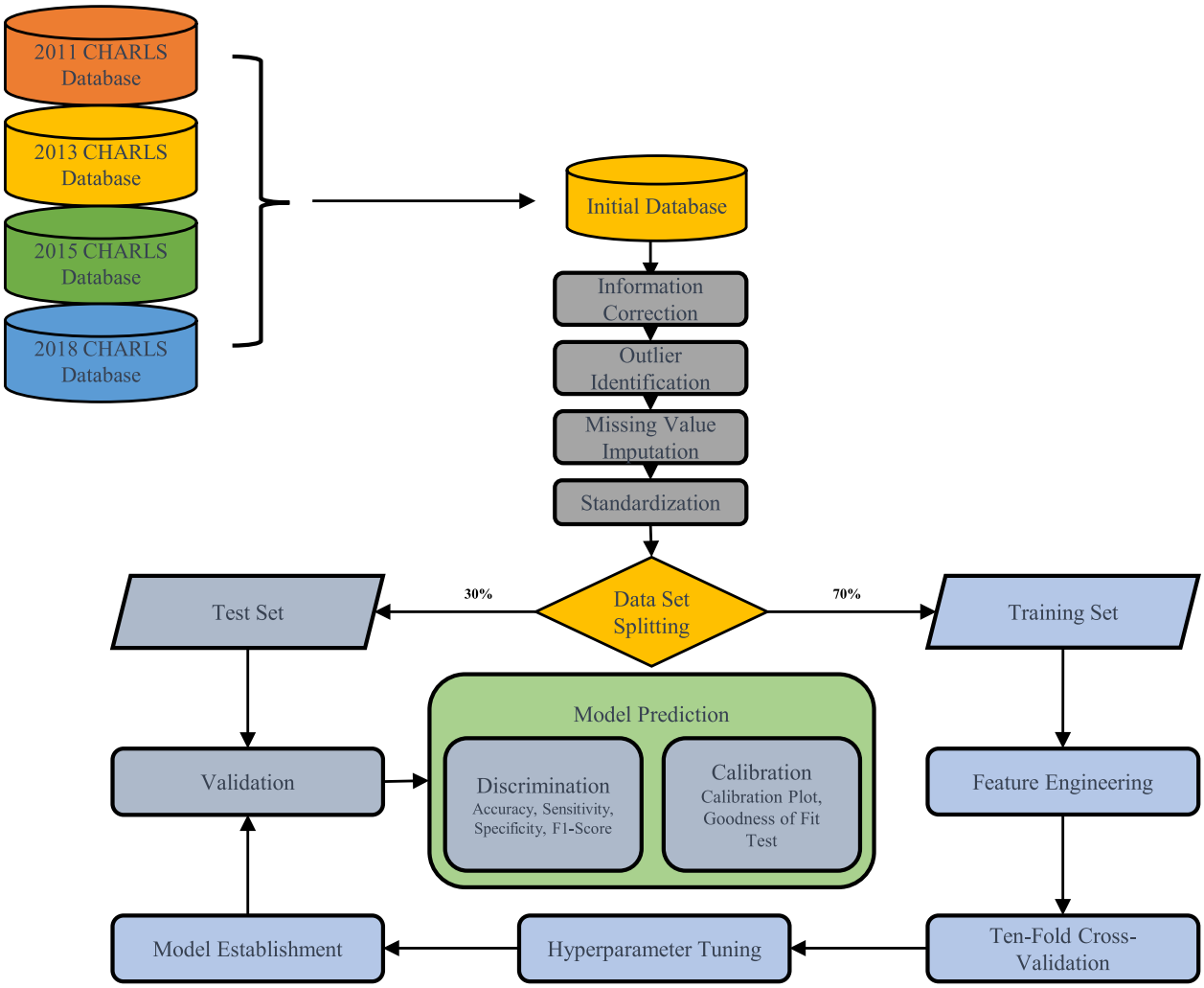

Fig. 2 Model Development Workflow

sampling rate, sample rate, maximum depth, minimum child weight, and minimum loss reduction for node splitting (gamma); for DNN, the study configured a network with 4 hidden layers and iterated through 1–20 neurons in each hidden layer. Results of hyperparameter tuning are presented in Table 3. Table 4 displays the predictive performance of the models on the full variable set and the feature-selected variable set. On the full variable set, LR exhibited the highest accuracy (0.67) and specificity (0.81), followed by DNN, while XGBoost had relatively lower accuracy (0.57) and specificity (0.62). Sensitivity ranked from high to low as DNN (0.50), XGBoost (0.46), and LR (0.32), with a similar trend observed in F1 values, where DNN had the highest (0.44) and LR the lowest (0.35). AUC values for LR, XGBoost, and DNN were 0.60 (0.53–0.67), 0.55 (0.48–0.63), and 0.57 (0.49–0.67), respectively, with no statistically significant differences ($P > 0.05$, Table 5).

After feature selection, there was a noticeable decrease in accuracy and specificity but a substantial increase in sensitivity and F1 values, especially for XGBoost and DNN, with sensitivities reaching 0.74 and 0.73, and F1 values rising to 0.48. AUC values for LR, XGBoost, and DNN on the feature-selected set were 0.60 (0.53–0.68), 0.64 (0.57–0.71), and 0.67 (0.60–0.73), respectively. Notably, DNN's AUC on the feature-selected set was significantly higher than LR ($P = 0.037$). Additionally, comparing the models' AUC before and after feature selection revealed a significant improvement for XGBoost and DNN (P values of 0.022 and 0.004, respectively).

Calibration results for the models are presented in Fig. 4, where LR exhibited the best calibration ($\chi 2 = 10.378$, $P = 0.239$), followed by DNN ($\chi 2 = 18.082$, $P = 0.021$) and XGBoost ($\chi 2 = 26.206$, $P = 0.001$). Further analysis using DCA is detailed in Fig. 5. Setting

Table 2 Baseline Characteristics of the Study Population

	Entire Population ($N=924$)	Heart Disease ($n=243$)	Non-Heart Disease ($n=691$)	*P-value*
Sociodemographic				
Age	62.83(66.25,71.75)	62.5(66.08,71.33)	63(66.25,71.92)	0.167
Gender (Male)	441(47.2%)	102(41.98%)	339(49.06%)	0.062
Waist-to-Height Ratio	0.52(0.52,0.61)	0.51(0.51,0.63)	0.52(0.52,0.61)	0.080
Lifestyle Behaviors				
Sleep Duration (Hours)	5(5,8)	5(5,8)	5(5,8)	0.952
Alcohol Consumption (Yes)	258(27.62%)	58(23.87%)	200(28.94%)	0.134
Smoking (Yes)	356(38.12%)	83(34.16%)	273(39.51%)	0.145
Medical History				
Dyslipidemia (Yes)	147(16.14%)	57(23.46%)	90(13.02%)	< 0.001
Diabetes (Yes)	101(10.09%)	33(13.58%)	68(9.84%)	0.119
Tumors (Yes)	3(0.32%)	1(0.41%)	2(0.29%)	1.000
Chronic Respiratory Diseases (Yes)	98(10.53%)	35(14.4%)	63(9.12%)	0.028
Liver Diseases (Yes)	21(2.25%)	7(2.88%)	14(2.03%)	0.451
Kidney Diseases (Yes)	53(5.71%)	14(5.76%)	39(5.64%)	1.000
Stroke (Yes)	47(5.04%)	12(4.94%)	35(5.06%)	1.000
Gastrointestinal Diseases (Yes)	191(20.47%)	52(21.4%)	139(20.12%)	0.644
Psychological Issues (Yes)	6(0.65%)	0(0.00%)	6(0.87%)	0.349
Memory-Related Disorders (Yes)	22(2.36%)	5(2.06%)	17(2.46%)	1.000
Arthritis or Rheumatic Disease(Yes)	389(41.78%)	115(47.33%)	274(39.65%)	0.041
Asthma (Yes)	46(4.94%)	18(7.41%)	28(4.05%)	0.056

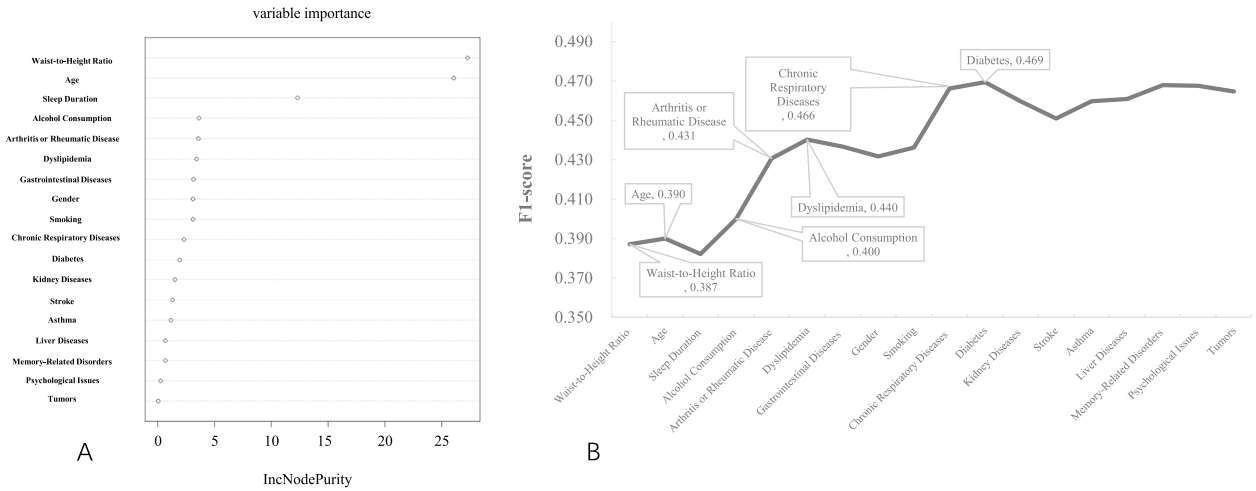

Fig. 3 Feature Selection. Panel **A** depicts the relative importance of each variable, while Panel **B** illustrates the results of feature selection combining random forest with forward and backward selection

the threshold at ≤ 0.21, DNN and LR outperformed XGBoost; for thresholds between 0.21–0.34, XGBoost's utility surpassed LR but remained below DNN; and for thresholds exceeding 0.34, XGBoost consistently outperformed DNN and LR.

SHAP Interpretability analysis

To further elucidate the black-box nature of the XGBoost model, we utilized the SHAP method to assess the interpretability of the XGBoost model. The SHAP results reveal that the top three important predictive factors are Age, WHtR, and Alcohol Consumption, as shown in

Table 3 Model Hyperparameter Tuning

Model	Search Space	Optimal Parameters
XGBoost	eta = c(0.05,0.1,0.2,0.5,1)	eta = 0.05
	colsample_bytree = (1/3,2/3,1)	colsample_bytree = 2/3
	ubsample = c(0.25,0.5,0.75,1)	subsample = 0.75
	max_depth = c(4,5,6)	max_depth = 5
	min_child_weigth = c(1:12)	min_child_weigth = 2
	gamma = c(0.1,0.2,0.3,0.4,0.5)	gamma = 0.1
DNN	activation = c("Tanh", "TanhWithDropout", "Rectifier", "RectifierWithDropout", "Maxout", "MaxoutWithDropout")	activation = "Rectifier"
	hidden = c(1,1,1,1): c(20,20,20,20)	hidden = c(8,9,8,7)

Table 4 Performance Metrics of Predictive Models

		Confusion Matrix			Accuracy	Sensitivity	Specificity	F1 Score	AUC
			Yes	No					
All Variables	LR	Yes	26	42	0.67	0.32	0.81	0.35	0.60(0.53–0.67)
		No	56	174					
	XGboost	Yes	38	83	0.57	0.46	0.62	0.37	0.55(0.48–0.63)
		No	44	133					
	DNN	Yes	41	65	0.64	0.50	0.70	0.44	0.57(0.49–0.67)
		No	41	151					
Feature Selection	LR	Yes	38	83	0.57	0.46	0.62	0.37	0.60(0.53–0.68)
		No	44	133					
	XGboost	Yes	61	111	0.56	0.74	0.49	0.48	0.64(0.57–0.71)
		No	21	105					
	DNN	Yes	60	106	0.57	0.73	0.51	0.48	0.67(0.60–0.73)
		No	22	110					

Table 5 Pairwise Comparison of AUC for Model Performance

Model	*Z-value*	*P-value*
LR vs XGB	0.949	0.342
LR vs DNN	0.561	0.574
XGB vs DNN	−0.368	0.712
LR_FET vs XGB_FET	−0.880	0.378
LR_FET vs DNN_FET	−2.081	0.037
XGB_FET vs DNN_FET	−0.644	0.519
LR vs LR_FET	−0.449	0.653
XGB vs XGB_FET	−2.284	0.022
DNN vs DNN_FET	−2.886	0.004

LR, XGBoost, and DNN represent models constructed on the entire variable set, while LR_FET, XGBoost_FET, and DNN_FET represent models built on variable sets after feature selection. Model AUC comparisons were conducted using the Delong test from the pROC package

Fig. 6. The y-axis represents the variable values arranged from high to low, while the x-axis represents the model's contribution to the prediction outcome. The prediction performance of variables for the occurrence probability of heart disease in older hypertensive individuals exhibits complexity. For instance, among the top three key variables, the effects of Age and WHtR on the outcome are relatively random, while in the case of Alcohol Consumption, an increase in alcohol consumption would increase the likelihood of heart disease occurrence in the model prediction for older hypertensive individuals.

Discussion

In recent years, machine learning models have been extensively applied to predict the short-term or long-term risk of various diseases, the risk of complications, analysis of risk factors, and mortality analysis [11–15]. This demonstrates the advantages of machine learning models in predicting various diseases. Particularly, there is an increasing application of these models in predicting the risk of heart disease, which is crucial for primary prevention. They play a foundational role in heart disease risk prevention and clinical decision-making. Numerous heart disease risk prediction models based on machine learning have shown good performance, but the focus

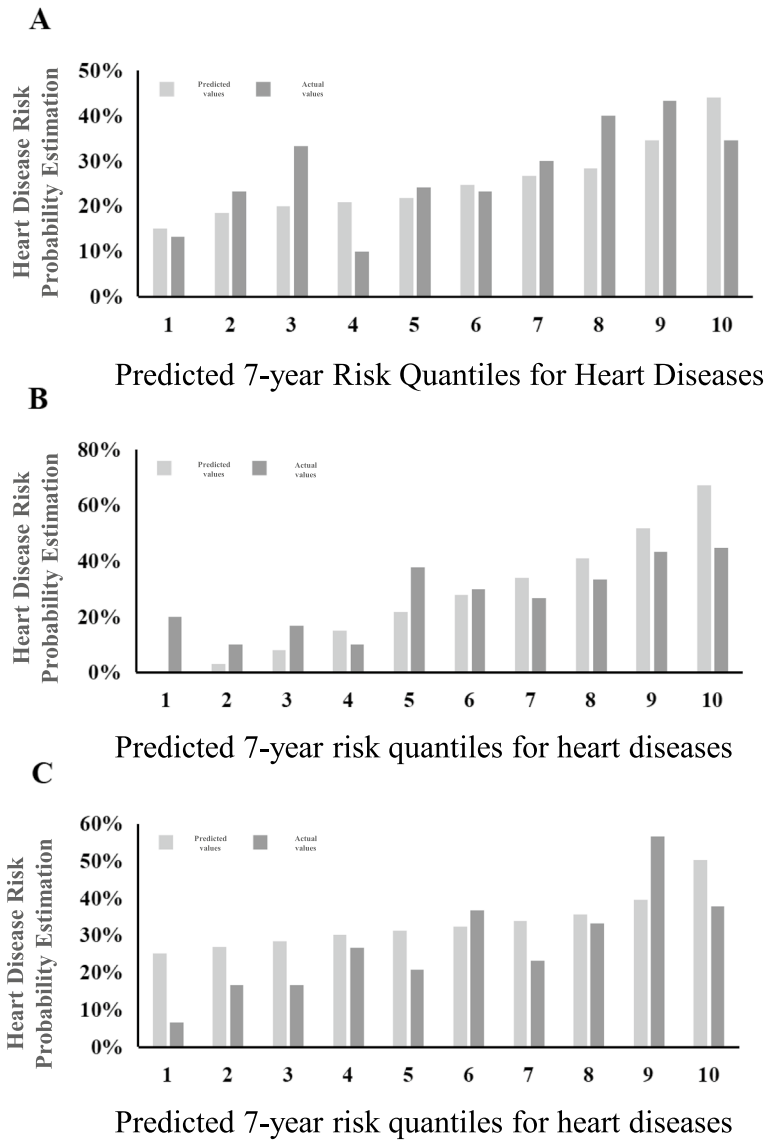

Fig. 4 Calibration Plots for Predictive Models. Panel A represents logistic regression, Panel B represents XGBoost, and Panel C represents DNN (Deep Neural Network)

of each study varies and each has its own limitations [16, 17]. Multiple studies on heart disease risk prediction models indicate that hypertension is one of the significant factors in predicting the risk of heart disease and cardiovascular events [18–20]. However, there is a scarcity of machine learning prediction models specifically targeting the risk of heart disease in hypertensive patients, which is an aspect worthy of attention. While some prediction models in this domain have demonstrated promising outcomes, they are predominantly based on cross-sectional studies. Consequently, these models are confined to short-term risk prediction for current heart disease occurrence in hypertensive patients, lacking the longitudinal depth necessary for long-term risk forecasting and thereby presenting significant limitations [17, 21]. Our study, conducted over a 7-year follow-up period, specifically targeted hypertensive patients, thereby contributing to the prediction of long-term cardiovascular disease risk and providing a basis for early risk identification, thus offering distinct advantages. Moreover, while many current machine learning-based prediction models primarily focus on cardiovascular diseases such as atherosclerosis and coronary heart disease, they often overlook other types of cardiac conditions (such as heart failure, rheumatic heart disease, etc.) and their complications. However, it is evident that hypertension is not

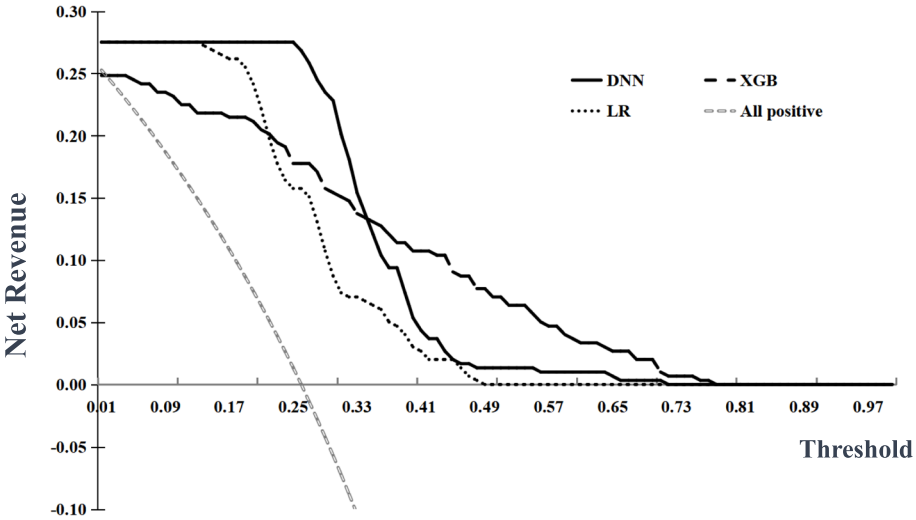

Fig. 5 Decision Curves Corresponding to Predictive Models

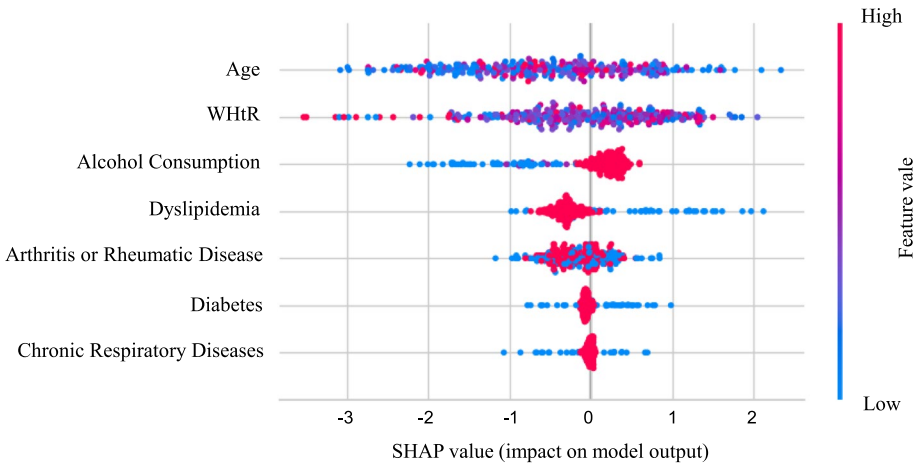

Fig. 6 SHAP Summary Plot

solely a significant risk factor for coronary heart disease. Therefore, our study simultaneously addresses various cardiac diseases, including coronary heart disease, heart failure, rheumatic heart disease, among others, resulting in a broader scope of prediction. Finally, our focus is on the older population. In an era marked by the continuous aging of society, the application of machine learning in geriatric diseases is becoming increasingly widespread [22]. However, there is a notable scarcity of research on the prediction of long-term cardiac risk in older hypertensive patients, particularly those in underserved communities. This study addresses precisely these focal issues.

Dyslipidemia has become a prevalent condition in the elderly, posing a significant risk for the occurrence and progression of cardiovascular diseases. Elevated levels of total cholesterol (TC), triglycerides (TG), and low-density lipoprotein cholesterol (LDL-C) are characteristic features of these diseases [2, 23]. The predictive results of our study suggest that hypertensive elderly patients with dyslipidemia have a greater risk of developing heart disease. Factors such as hypertension, diabetes, and dyslipidemia can mediate abnormal platelet activation, favoring pathological thrombus formation and cardiovascular diseases, making them common risk factors for the development of cardiovascular diseases [24]. Moreover, lifestyle factors and dietary habits in hypertensive patients often resemble those in individuals with dyslipidemia [25]. Recent research indicates that triglycerides are the main lipid indicator most likely to increase systolic and

diastolic blood pressure, with a particularly strong correlation between elevated triglycerides in small high-density lipoprotein (HDL) and increased blood pressure [26]. As blood pressure, lipid, and glucose levels are often latent, vascular diseases, myocardial infarctions, strokes, and other severe events have likely occurred by the time they are detected in this population. Therefore, for older hypertensive patients, dyslipidemia is a robust factor for predicting a higher risk of future heart disease. Timely detection and management of lipid levels can help reduce the risk of heart disease occurrence and prevent serious cardiovascular events.

Patients with chronic lung diseases may experience increased pulmonary circulation resistance, leading to pulmonary arterial hypertension and affecting the right heart system. Hypertension itself can contribute to cardiac remodeling, which further promotes the occurrence of diseases such as heart failure. Therefore, for older hypertensive patients, timely prevention, identification, and treatment of chronic lung diseases can help reduce the risk of future heart diseases. Rheumatic heart disease (RHD) is an acquired valvular disease that begins with untreated streptococcal pharyngeal infection, characterized by valve regurgitation and/or stenosis, often associated with complications such as arrhythmias, systemic embolism, infective endocarditis, pulmonary hypertension, heart failure, and death [27]. It remains a significant cause of cardiovascular disease-related deaths in developing countries [28]. Acute rheumatic fever (ARF) and RHD are important determinants of global cardiovascular disease incidence and mortality [27]. Additionally, in rheumatic diseases characterized by severe systemic inflammation, there is often an increased incidence and mortality of atherosclerosis and cardiovascular diseases [29].

Elderly patients with baseline hypertension will experience long-term damage to the target organ, the heart. This makes the occurrence of various heart diseases more likely, not limited to rheumatic heart disease. According to our predictive results, elderly individuals with hypertension who later develop arthritis or rheumatic diseases are more likely to develop heart disease. Therefore, in older hypertensive patients, particularly those with arthritis or rheumatic diseases, it is essential to strengthen prevention, identification, and management to reduce the risk of heart disease occurrence.

In the prediction using the full variable set, the predictive abilities of all three machine learning methods were relatively weak, with the highest AUC being 0.60 for LR. After feature engineering, a smaller set of more valuable variables was selected, and at this point, the performance of both XGBoost and DNN models showed significant improvement, reaching 0.64 and 0.67, respectively. This suggests that feature engineering not only reduces the dimensionality of the predictive data but also helps enhance the predictive performance of the models. It is noteworthy that during the feature selection process, there were inconsistencies between the selected features and the identified risk factors, such as gender, smoking, and drinking. These factors are known risk factors for heart disease but did not enter the feature selection results. Some studies suggest that factors influencing a disease may not necessarily contribute significantly to its prediction [30]. In other words, factors with a substantial impact on the disease may have a minimal contribution to prediction, highlighting an important difference between predictive research and causal inference.

This study also conducted discrimination analysis, calibration analysis, and clinical utility analysis on the predictive models. A good discrimination model should be able to differentiate between individuals with high and low future disease risks, often evaluated using AUC. Calibration reflects the consistency between predicted risk and actual risk. A systematic review of cardiovascular disease prediction models in 2015 found that 63% of studies reported the discrimination of predictive models, but only 36% reported model calibration, leading to variations in the quality of predictive models [31]. Currently, more and more predictive research recommends reporting both discrimination and calibration to achieve a more scientific and objective evaluation. Additionally, decision curve analysis can guide clinical application by determining the appropriate model threshold based on gains [32]. This study found differences in predictive model gains at different threshold ranges.

Through feature selection, this study identified seven important predictive variables: age, waist-to-height ratio, drinking, diabetes, dyslipidemia, lung diseases, and arthritis. Except for age, which is an immutable factor, the other factors can be improved through changes in diet, increased physical exercise, and weight control. Specifically, leveraging epidemiological data can aid in the early identification of heart disease risk in older hypertensive patients. For low-risk individuals, maintaining healthy lifestyle habits is recommended. For high-risk individuals, changing unhealthy habits, undergoing regular check-ups, and seeking medical examination and treatment for severe cases are advised.

The strengths of this study may include the large-scale community cohort survey, which provides representative and cost-effective data. The adoption of a high-risk population strategy, combined with epidemiological data, facilitates early screening and improves intervention efficiency. Moreover, the use of feature engineering techniques helps reduce the dimensionality of predictive variables, select better predictive models, and facilitates practical

application. Finally, a comprehensive evaluation of models using discrimination, calibration, and clinical decision curve analysis guides model selection from multiple perspectives. However, the study only included demographic, lifestyle, and disease history characteristics, and the sample size was relatively small, unable to fully demonstrate the advantages of machine learning algorithms in handling big data and multidimensional features. Furthermore, the study only conducted internal validation and did not perform external validation in a more extensive population, limiting the generalizability of the research. Overall, the results of the study provide valuable insights into predicting the future risk of heart disease in the elderly hypertensive population and early risk identification.

Authors' contributions

conceptualization: F.S. and J.Y. methodology: F.S., Q.L.and J.Y. formal analysis and investigation: F.S., Q.L.and J.Y. writing — review and editing: F.S., Q. L.and J.Yu. Figures 1, 2, 3, 4 and 5: F.S. Tables 1, 2, 3, 4 and 5: F.S.

Funding

This study was supported by the National Natural Science Foundation of China (NSFC 81960086,82160089), Gansu province health research project (GSWSKY2017-02), and the Cuiying Scientific and Technological Innovation Program of Lanzhou University Second Hospital (CY2021-MS-A13). This study was also supported by Special Fund Project for Doctoral Training of the Lanzhou University Second Hospital (YJS-BD-24) and International science and technology cooperation base (PR0124002).

Data availability

This study uses publicly available data from the China Health and Retirement Longitudinal Study. The raw data can be accessed through the following link: https://charls.pku.edu.cn/.

Declarations

Ethics approval and consent to participate

The study was approved by the Biomedical Ethics Committee of Peking University (Approval No: IRB00001052-11015; IRB00001052-11014) and strictly adhered to the ethical principles outlined in the Declaration of Helsinki. All participants provided written informed consent, ensuring the ethical compliance of the data collection process and the protection of participants' rights.

Consent for publication

Not Applicable.

Competing interests

The authors declare no competing interests.

Clinical Trial Number

This study was completed based on the publicly available longitudinal survey database, the China Health and Retirement Longitudinal Study (CHARLS), and is not a clinical trial, therefore, no clinical trial registration number was obtained.

Received: 26 February 2024 Accepted: 2 January 2025
Published online: 11 January 2025

References

1. Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019: a pooled analysis of 1201 population-representative studies with 104 million participants[J]. Lancet (London, England), 2021, 398(10304):957–980. https://doi.org/10.1016/s0140-6736(21)01330-1.
2. Report on Cardiovascular Health and Diseases in China 2021: An Updated Summary[J]. Biomedical and environmental sciences : BES, 2022, 35(7):573–603. https://doi.org/10.3967/bes2022.079.
3. Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019[J]. Lancet (London, England), 2020, 396(10258):1223–1249. https://doi.org/10.1016/s0140-6736(20)30752-2.
4. Chai X, Chen Y, Li Y,et al. Lower geriatric nutritional risk index is associated with a higher risk of all-cause mortality in patients with chronic obstructive pulmonary disease: a cohort study from the National Health and Nutrition Examination Survey 2013–2018[J]. BMJ open respiratory research, 2023, 10(1). https://doi.org/10.1136/bmjresp-2022-001518.
5. Paljärvi T, Tiihonen J, Lähteenvuo M, et al. Psychotic depression and deaths due to suicide[J]. J Affect Disord. 2023;321:28–32. https://doi.org/10.1016/j.jad.2022.10.035.
6. Greener JG, Kandathil SM, Moffat L, et al. A guide to machine learning for biologists[J]. Nat Rev Mol Cell Biol. 2022;23(1):40–55. https://doi.org/10.1038/s41580-021-00407-0.
7. Fan T, Wang J, Li L, et al. Predicting the risk factors of diabetic ketoacidosis-associated acute kidney injury: A machine learning approach using XGBoost[J]. Front Public Health. 2023;11:1087297. https://doi.org/10.3389/fpubh.2023.1087297.
8. Wang R, Zhang J, Shan B, et al. XGBoost Machine Learning Algorithm for Prediction of Outcome in Aneurysmal Subarachnoid Hemorrhage[J]. Neuropsychiatr Dis Treat. 2022;18:659–67. https://doi.org/10.2147/ndt.s349956.
9. Egger J, Gsaxner C, Pepe A, et al. Medical deep learning-A systematic meta-review[J]. Comput Methods Programs Biomed. 2022;221: 106874. https://doi.org/10.1016/j.cmpb.2022.106874.
10. Stahlschmidt SR, Ulfenborg B, Synnergren J. Multimodal deep learning for biomedical data fusion: a review[J]. Briefings in bioinformatics, 2022, 23(2). https://doi.org/10.1093/bib/bbab569.
11. Dong J, Feng T, Thapa-Chhetry B, et al. Machine learning model for early prediction of acute kidney injury (AKI) in pediatric critical care[J]. Critical care (London, England). 2021;25(1):288. https://doi.org/10.1186/s13054-021-03724-0.
12. Huang G, Jin Q, Mao Y. Predicting the 5-Year Risk of Nonalcoholic Fatty Liver Disease Using Machine Learning Models: Prospective Cohort Study[J]. J Med Internet Res. 2023;25: e46891. https://doi.org/10.2196/46891.
13. Cho SY, Kim SH, Kang SH, et al. Pre-existing and machine learning-based models for cardiovascular risk prediction[J]. Sci Rep. 2021;11(1):8886. https://doi.org/10.1038/s41598-021-88257-w.
14. Zafar A, Attia Z, Tesfaye M, et al. Machine learning-based risk factor analysis and prevalence prediction of intestinal parasitic infections using epidemiological survey data[J]. PLoS Negl Trop Dis. 2022;16(6): e0010517. https://doi.org/10.1371/journal.pntd.0010517.
15. Ke J, Chen Y, Wang X, et al. Machine learning-based in-hospital mortality prediction models for patients with acute coronary syndrome[J]. Am J Emerg Med. 2022;53:127–34. https://doi.org/10.1016/j.ajem.2021.12.070.
16. Alaa AM, Bolton T, Di Angelantonio E, et al. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants[J]. PLoS ONE. 2019;14(5): e0213653. https://doi.org/10.1371/journal.pone.0213653.
17. Xi Y, Wang H, Sun N. Machine learning outperforms traditional logistic regression and offers new possibilities for cardiovascular risk prediction: A study involving 143,043 Chinese patients with hypertension[J]. Frontiers in cardiovascular medicine. 2022;9:1025705. https://doi.org/10.3389/fcvm.2022.1025705.
18. Konstantonis G, Singh KV, Sfikakis PP, et al. Cardiovascular disease detection using machine learning and carotid/femoral arterial imaging frameworks in rheumatoid arthritis patients[J]. Rheumatol Int. 2022;42(2):215–39. https://doi.org/10.1007/s00296-021-05062-4.
19. Zhuang XD, Tian T, Liao LZ, et al. Deep Phenotyping and Prediction of Long-term Cardiovascular Disease: Optimized by Machine Learning[J]. Can J Cardiol. 2022;38(6):774–82. https://doi.org/10.1016/j.cjca.2022.02.008.

20. Shen T, Liu D, Lin Z,et al. A Machine Learning Model to Predict Cardiovascular Events during Exercise Evaluation in Patients with Coronary Heart Disease[J]. Journal of clinical medicine, 2022, 11(20).https://doi.org/10.3390/jcm11206061
21. Wu Y, Xin B, Wan Q, et al. Risk factors and prediction models for cardiovascular complications of hypertension in older adults with machine learning: A cross-sectional study[J]. Heliyon. 2024;10(6): e27941. https://doi.org/10.1016/j.heliyon.2024.e27941.
22. Das A, Dhillon P. Application of machine learning in measurement of ageing and geriatric diseases: a systematic review[J]. BMC Geriatr. 2023;23(1):841. https://doi.org/10.1186/s12877-023-04477-x.
23. Dybiec J, Baran W, Dąbek B,et al. Advances in Treatment of Dyslipidemia[J]. International journal of molecular sciences, 2023, 24(17). https://doi.org/10.3390/ijms241713288.
24. Sepúlveda C, Palomo I, Fuentes E. Antiplatelet activity of drugs used in hypertension, dyslipidemia and diabetes: Additional benefit in cardiovascular diseases prevention[J]. Vascul Pharmacol. 2017;91:10–7. https://doi.org/10.1016/j.vph.2017.03.004.
25. ESC/EAS guidelines for the management of dyslipidaemias. Lipid modification to reduce cardiovascular risk[J]. Atherosclerosis. 2019;290:140–205. https://doi.org/10.1016/j.atherosclerosis.2019.08.014.
26. Liu W, Yang C, Lei F, et al. Major lipids and lipoprotein levels and risk of blood pressure elevation: a Mendelian Randomisation study[J]. EBioMedicine. 2024;100: 104964. https://doi.org/10.1016/j.ebiom.2023.104964.
27. Auala T, Zavale BG, Mbakwem A,et al. Acute Rheumatic Fever and Rheumatic Heart Disease: Highlighting the Role of Group A Streptococcus in the Global Burden of Cardiovascular Disease[J]. Pathogens (Basel, Switzerland), 2022, 11(5). https://doi.org/10.3390/pathogens11050496.
28. Franczyk B, Gluba-Brzózka A, Rysz-Górzyńska M,et al. The Role of Inflammation and Oxidative Stress in Rheumatic Heart Disease[J]. International journal of molecular sciences, 2022, 23(24). https://doi.org/10.3390/ijms232415812.
29. Melissaropoulos K, Bogdanos D, Dimitroulas T, et al. Primary Sjögren's Syndrome and Cardiovascular Disease[J]. Curr Vasc Pharmacol. 2020;18(5):447–54. https://doi.org/10.2174/1570161118666200129125320.
30. Pepe MS, Feng Z, Huang Y, et al. Integrating the predictiveness of a marker with its performance as a classifier[J]. Am J Epidemiol. 2008;167(3):362–8. https://doi.org/10.1093/aje/kwm305.
31. Wessler BS, Lai YhL, Kramer W, et al. Clinical Prediction Models for Cardiovascular Disease: Tufts Predictive Analytics and Comparative Effectiveness Clinical Prediction Model Database[J]. Circ Cardiovasc Qual Outcomes. 2015;8(4):368–75. https://doi.org/10.1161/circoutcomes.115.001693.
32. Vickers AJ. Decision analysis for the evaluation of diagnostic tests, prediction models and molecular markers[J]. Am Stat. 2008;62(4):314–20. https://doi.org/10.1198/000313008x370302.

Publisher's Note