Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Machine Learning-Based Model for Cardiovascular Disease Risk Stratification in Middle-Aged and Older Adults with Diabetes: A Cross-Sectional Study based on CHARLS

Changli Wang
School of Health Science and Engineering, University of Shanghai for Science and Technology (USST)
Lvkan Weng
kevinwlk210126.com
Shanghai Chest Hospital

Article

Keywords: Cardiovascular Disease(CVD), Diabetes Mellitus(DM), Machine Learning, Explainable AI (XAI), Clinical Decision Support

Posted Date: June 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9792321/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

cardiovascular disease(CVD) is the leading cause of mortality among patients with diabetes mellitus(DM).The primary aim of this study was to develop a machine learning-based risk assessment model for CVD in the diabetic population.The model is intended to facilitate early risk stratification and support clinical decision-making.

Methods

We performed a retrospective research based on data from the China Health and Retirement Longitudinal Study(CHARLS) and comprised middle-aged and older adults with DM.We applied LASSO regression and the Boruta algorithm for feature selection from the initial 29 clinical variables. Six machine learning algorithms—Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), Light Gradient Boosting Machine (LightGBM), and eXtreme Gradient Boosting (XGBoost) were trained and comparatively evaluated.Furthermore,we employed SHapley Additive exPlanations (SHAP) to provide model interpretability.

Results

The feature selection process identified 8 robust clinical variables for model training.Among these algorithms,the RF model showed robust classification capabilities..On the test set,the RF model achieved an area under the receiver operating characteristic curve(AUC) of 0.752,with an accuracy of 66.12%, sensitivity of 72.53%, and specificity of61.64%.

Conclusion

This study identified crucial risk factors for CVD in patients with diabetes and successfully translated these findings into an explainable, machine learning-driven visual risk assessment tool. This user-friendly system provides a practical approach for personalized cardiovascular risk assessment in clinical settings.

Clinical trial registration

Clinical trial number: not applicable.

1 Introduction

According to estimates from the International Diabetes Federation (IDF) Diabetes Atlas, 11th Edition, the global number of adults aged 20 to 79 living with diabetes is projected to reach approximately 589 million in 2024 [1].Diabetes mellitus(DM) is a common condition closely associated with aging, comorbidities affect more than 80% of patients with diabetes and the number increases with the increasing age and duration of diabetes [4–5]. Among these comorbidities, cardiovascular disease (CVD) was the most common cause of mortality [3].DM is a significant risk factor for complications such as CVD [2], and The intersection of DM and CVD is a growing clinical concern. Therefore, developing a CVD risk stratification model specifically tailored for middle-aged and elderly diabetic patients is of critical importance for promoting precision interventions and improving clinical management.

Although many validated cardiovascular disease (CVD) prediction models currently exist [6,7], there is still a severe lack of tools calibrated specifically for this high-risk group of older adults living with diabetes.In addition, traditional cardiovascular risk models are primarily built upon classical statistical frameworks, including the Cox proportional hazards model and

linear regression models.These approaches are limited by rigid linear and proportional hazards assumptions. Consequently, they exhibit significant limitations in parsing high-dimensional and non-linear clinical datasets.Additionally, standard statistical models are typically derived from population-level averages, which often fail to capture the profound pathophysiological heterogeneity of individual patients [8].Machine learning (ML) provides a robust methodological alternative to overcome these limitations. By leveraging its inherent ability to process high-dimensional datasets and identify complex, non-linear interaction patterns [9], ML holds significant promise for developing highly accurate and personalized predictive instruments.

With the rapid advancement of big data and artificial intelligence technologies, machine learning (ML) models based on electronic health records have demonstrated significant potential in cardiovascular disease (CVD) prediction [10,11].For example, Zhu et al. [12] developed a CVD risk prediction model for chronic kidney disease patients using routine clinical data, whereas Huang et al. [13] built a similar model for elderly individuals with chronic diseases based on the China Health and Retirement Longitudinal Study(CHARLS) cohort.Despite the abundance of existing cardiovascular prediction models, dedicated tools specifically calibrated for middle-aged and older diabetic patients remain scarce. To fill this research gap, Hanna et al. [14] employed unsupervised machine learning (ML) algorithms to develop a CVD prediction model for this population. However,this study's small sample size (n = 238) necessitates further evaluation of its predictive performance and generalization capabilities.Crucially, the translation of such algorithms into actionable point-of-care clinical tools remains largely underexplored.

To overcome these critical limitations, we drew upon data from the CHARLS—a nationally representative, high-quality longitudinal cohort.To systematically evaluate and develop predictive models for cardiovascular disease (CVD), this study employed six distinct machine learning (ML) algorithms: logistic regression (LR), k-nearest neighbors (KNN), support vector machine (SVM), random forest (RF), light gradient boosting machine (LightGBM), and extreme gradient boosting (XGBoost).Because the complex, non-linear architectures of advanced ML algorithms often create a "black-box" dilemma that hinders clinical trust [15], we integrated explainable artificial intelligence (XAI) frameworks to elucidate model decision-making.Finally, to bridge the gap between model development and real-world clinical application, we deployed the best-performing model as an interactive web application built with Streamlit.This visual clinical decision support system generates personalized risk profiles from routine clinical inputs, ultimately empowering accessible and precision-driven cardiovascular care for middle-aged and older patients with diabetes.

2 Methods

2.1 Data Source and Study Design

This study conducted a cross-sectional analysis based on the 2020 wave data from the China Health and Retirement Longitudinal Study (CHARLS) (https://charls.pku.edu.cn/).CHARLS is an ongoing, large-scale, interdisciplinary longitudinal survey project jointly conducted by Peking University and Wuhan University.The survey aims to establish a nationally representative, high-quality micro-database covering individuals aged 45 and older and their households in China [16].The project completed its national baseline survey in 2011, covering 450 villages and communities across 28 provinces. Subsequently, multiple waves of follow-up surveys were conducted in 2013, 2015, 2018, 2020, and 2021–2023.The CHARLS protocol was approved by the Peking University Institutional Review Board (IRB00001052-11015), with written informed consent obtained from all participants before data collection.

2.2 Study Population and Variable Definitions

2.2.1 Ascertainment of the Primary Outcome

Study participants were diagnosed with DM if they met at least one of the following clinical criteria a self-reported physician diagnosis of diabetes.The primary outcome variable of this study was CVD status. Following established

methodologies from previous research [17], the CVD status of the participants was determined using standardized items from the CHARLS questionnaire.Specifically, study participants were classified as having CVD if they self-reported a physician diagnosis of any of the following conditions: (1) myocardial infarction, coronary heart disease, angina, congestive heart failure, or other heart diseases; or (2) stroke.

2.2.2 Data Inclusion and Exclusion Criteria

This study utilized survey data from the fifth wave of the CHARLS cohort, which initially included a total of 19,395 study participants.As illustrated in the participant selection flowchart in Fig. 1, participants were sequentially excluded based on the following criteria: (1) individuals without a diabetes diagnosis across all five survey waves (n = 16,540);(2) age < 45 or > 85 years (n = 52); and (3) extreme self-reported daily sleep duration (< 3 or > 12 hours) (n = 178).After applying these screening criteria, a total of 2,625 eligible participants were ultimately included to serve as the final analytic cohort for this study.

2.2.3 Study Variables and Feature Preselection

The CHARLS database contains comprehensive phenotypic data covering multiple dimensions, including demographic characteristics, family structure, health status and physical functioning, as well as cognitive and psychological health status.To guarantee data quality and the robustness of the predictive models, we performed an initial feature selection on the original pool of 171 variables.First, variables with a missing data rate exceeding 20% were excluded. Subsequently, informed by well-established clinical literature, a set of known and potential risk factors for CVD was systematically selected.Ultimately, a total of 29 variables were selected and categorized into three core dimensions (variable information is summarized in Table 1).

Demographic characteristics: age, marital status, urban/rural residence, education level, and household size.

Health status and physical functioning: self-rated health (SRH), activities of daily living (ADL), self-reported history of chronic diseases (hypertension, cancer, lung disease, asthma, dyslipidemia, liver disease, kidney disease, stomach disease, arthritis, and memory impairment), and recent healthcare utilization (history of hospitalization and outpatient visits within the past year).

Lifestyle factors: physical activity intensity, smoking status, and alcohol consumption.

2.3 Data Preprocessing

To handle the residual missing values after preliminary data cleaning, multivariate iterative imputation was employed in this study (maximum iterations = 5) [18].Following the imputation process, the dataset was randomly split into a training set (70%) and a testing set (30%).To accelerate model convergence and eliminate scale disparities, continuous variables were subjected to Z-score standardization, while categorical and ordinal features were scaled using Min-Max normalization.

2.4 Feature Selection

Feature selection is necessary to minimize computational overhead, reduce data dimension, and improve model robustness [19]. To prevent data leakage, all selection procedures were restricted solely to the training set. We employed a dual-method approach combining Least Absolute Shrinkage and Selection Operator (LASSO) regression with the Boruta algorithm [20].

For LASSO regularization, convergence was reached at 54 iterations. The minimum penalty parameter (λ) was established at 0.000841, driving non-informative coefficients to zero and retaining 17 candidate variables (Figs. 2a and 2b).

In parallel, we evaluated features using Boruta, a wrapper method built upon a Random Forest (RF) base estimator. To handle potential data imbalance, the RF classifier was configured with balanced class weights and a maximum tree depth

of 5. The Boruta algorithm was executed with a maximum of 100 iterations at a significance threshold of α = 0.1. To ensure selection stability, this process was repeated 20 times using different random seeds, isolating a final subset of 9 features.

Finally, by calculating the intersection of the feature sets from both LASSO and Boruta, we identified an optimal, non-redundant set of eight predictive variables for model construction: age, self-rated health (SRH), activities of daily living (ADL), hypertension, dyslipidemia, renal disease, hospitalizations within the past year, and chest pain.

2.5 Machine Learning Algorithms and Model Development

To develop the CVD predictive model, we selected six different machine learning algorithms: Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM). We included LR as the traditional epidemiological benchmark for binary health outcomes, which established a baseline for performance comparison [21]. KNN and SVM were used for their respective strengths in instance-based spatial classification [22] and the construction of optimal high-dimensional decision boundaries [23].Furthermore, we incorporated three advanced ensemble learning frameworks to capture complex, non-linear clinical interactions. RF aggregates multiple decision trees via bootstrap aggregating (bagging) to improve predictive stability and reduce overfitting [24]. XGBoost and LightGBM are cutting-edge gradient boosting techniques; XGBoost is well known for its execution speed and predictive accuracy in clinical modeling [25], while LightGBM uses histogram-based algorithms specifically optimized for computational efficiency in high-dimensional datasets [26].

Model development and evaluation were implemented using the scikit-learn library in Python 3.12. After feature selection, we optimized the hyperparameters for each of the six algorithms using a grid search strategy combined with five-fold cross-validation on the training set. This robust internal validation approach ensures optimal parameter tuning while strictly preventing model overfitting.

To translate continuous predicted probabilities into binary clinical decisions, we determined the optimal classification threshold for each model using the Youden Index (J=sensitivity+specificity − 1), a standard metric in medical diagnostics for balancing true positive and true negative rates [27]. Finally, we evaluated the overall predictive performance on the independent test set using a comprehensive set of metrics: the area under the receiver operating characteristic curve (AUC-ROC), accuracy, sensitivity, specificity, precision, and the F1 score. The algorithm that showed the best overall performance was then selected as the final model for XAI interpretation and visualization.

2.6 Model Interpretability and Clinical Deployment

To clarify the decision-making process of the optimal machine learning model, we used SHapley Additive exPlanations (SHAP). Rooted in cooperative game theory, the SHAP framework calculates the marginal contribution of each feature to the final prediction, effectively resolving the inherent opacity of complex algorithms [28]. This approach evenly distributes the predictive output among the input features, providing both global and local interpretability for clinical stakeholders.

After identifying the optimal model, we deployed it as an interactive web application using the Streamlit framework(https://cvdpredfordmmae.streamlit.app). This user-friendly tool allows clinicians to input patient-specific parameters and instantly obtain a personalized CVD risk prediction, along with a visual breakdown of the key factors driving the risk score.

3 Results

3.1 Baseline Characteristics of the Study Population

The final analytical cohort comprised 2,625 patients with diabetes, among whom 1,099 (41.9%) had a confirmed diagnosis of CVD. Compared to individuals without CVD, patients with CVD were generally older (mean age: 66.5 years) and exhibited a significantly higher burden of comorbidities. Specifically, the CVD group demonstrated a higher prevalence of

hypertension (78.1%) and dyslipidemia (70.9%), alongside a greater proportion reporting poor self-rated health (SRH; 48.3%). As summarized in Table 1, statistically significant differences ($P < 0.05$) were observed between the CVD and non-CVD groups across multiple domains, including demographic characteristics, clinical health status, and lifestyle factors.

Furthermore, to validate the integrity of the data partitioning for machine learning, baseline characteristics were compared between the training and testing datasets (detailed in Supplementary Table S1). No statistically significant differences were detected across any variables between the two sets (all $P > 0.05$). This balanced distribution confirms that the random split successfully preserved the representative proportions of both the target outcome and predictive features, effectively mitigating potential selection bias.

Table 1
Baseline Characteristics of Cardiovascular and Non-Cardiovascular Patients

	Total	**Non CVD**	**CVD**	**p-value**
	(n = 2625)	***(n = 1526)***	***(n = 1099)***	
Age	64.5 (8.52)	63.1 (8.58)	66.5 (8.04)	< 0.001
Gender(%):				0.002
female	1422 (54.2)	787 (51.6)	635 (57.8)	
male	1203 (45.8)	739 (48.4)	464 (42.2)	
Marry(%):				< 0.001
No	402 (15.3)	186 (12.2)	216 (19.7)	
Yes	2223 (84.7)	1340 (87.8)	883 (80.3)	
Education(%):				0.436
Middle school and below	2215 (84.4)	1280 (83.9)	935 (85.1)	
High school and above	410 (15.6)	246 (16.1)	164 (14.9)	
Family size(IQR)	2.00 [2.00;3.00]	2.00 [2.00;4.00]	2.00 [2.00;3.00]	< 0.001
Rural(%):				0.289
No	1259 (48.0)	718 (47.1)	541 (49.2)	
Yes	1366 (52.0)	808 (52.9)	558 (50.8)	
Retire(%):				< 0.001
No	1966 (74.9)	1187 (77.8)	779 (70.9)	
Yes	659 (25.1)	339 (22.2)	320 (29.1)	
Self Reported Health Status(%):				< 0.001
very poor	280 (10.7)	110 (7.21)	170 (15.5)	
poor	679 (25.9)	318 (20.8)	361 (32.8)	
fair	1380 (52.6)	885 (58.0)	495 (45.0)	
good	156 (5.94)	116 (7.60)	40 (3.64)	
very good	130 (4.95)	97 (6.36)	33 (3.00)	
High Intensity Exercise(%):				< 0.001
No	1862 (70.9)	1008 (66.1)	854 (77.7)	
Yes	763 (29.1)	518 (33.9)	245 (22.3)	
Moderate Intensity Exercise(%):				< 0.001
No	1237 (47.1)	628 (41.2)	609 (55.4)	
Yes	1388 (52.9)	898 (58.8)	490 (44.6)	
Light Intensity Exercise(%):				0.016
No	526 (20.0)	281 (18.4)	245 (22.3)	

	Total	Non CVD	CVD	p-value
	(n = 2625)	*(n = 1526)*	*(n = 1099)*	
Yes	2099 (80.0)	1245 (81.6)	854 (77.7)	
Smoking(%):				< 0.001
No	2075 (79.0)	1169 (76.6)	906 (82.4)	
Yes	550 (21.0)	357 (23.4)	193 (17.6)	
Drinking(%):				< 0.001
No	1788 (68.1)	963 (63.1)	825 (75.1)	
Yes	837 (31.9)	563 (36.9)	274 (24.9)	
Hypertension(%):				< 0.001
No	931 (35.5)	690 (45.2)	241 (21.9)	
Yes	1694 (64.5)	836 (54.8)	858 (78.1)	
Dyslipidemia(%):				< 0.001
No	1078 (41.1)	758 (49.7)	320 (29.1)	
Yes	1547 (58.9)	768 (50.3)	779 (70.9)	
Cancer(%):				0.227
No	2521 (96.0)	1472 (96.5)	1049 (95.5)	
Yes	104 (3.96)	54 (3.54)	50 (4.55)	
Lung disease(%):				< 0.001
No	2165 (82.5)	1336 (87.5)	829 (75.4)	
Yes	460 (17.5)	190 (12.5)	270 (24.6)	
Arthritis(%):				< 0.001
No	1466 (55.8)	930 (60.9)	536 (48.8)	
Yes	1159 (44.2)	596 (39.1)	563 (51.2)	
Liver diseases(%):				< 0.001
No	2341 (89.2)	1418 (92.9)	923 (84.0)	
Yes	284 (10.8)	108 (7.08)	176 (16.0)	
Kidney disease(%):				< 0.001
No	2164 (82.4)	1345 (88.1)	819 (74.5)	
Yes	461 (17.6)	181 (11.9)	280 (25.5)	
Gastric disease(%):				< 0.001
No	1718 (65.4)	1084 (71.0)	634 (57.7)	
Yes	907 (34.6)	442 (29.0)	465 (42.3)	
Asthma(%):				< 0.001

	Total	Non CVD	CVD	p-value
	(n = 2625)	*(n = 1526)*	*(n = 1099)*	
No	2416 (92.0)	1444 (94.6)	972 (88.4)	
Yes	209 (7.96)	82 (5.37)	127 (11.6)	
Memory disorders(%):				< 0.001
No	2397 (91.3)	1451 (95.1)	946 (86.1)	
Yes	228 (8.69)	75 (4.91)	153 (13.9)	
Hospitalization within one year(%):				< 0.001
No	1823 (69.4)	1161 (76.1)	662 (60.2)	
Yes	802 (30.6)	365 (23.9)	437 (39.8)	
Outpatient visits in the past month(%):				< 0.001
No	1899 (72.3)	1160 (76.0)	739 (67.2)	
Yes	726 (27.7)	366 (24.0)	360 (32.8)	
ADL Score(IQR)	0.00 [0.00;1.00]	0.00 [0.00;0.00]	0.00 [0.00;1.00]	< 0.001
Cesd10(IQR)	9.00 [5.00;13.0]	9.00 [4.00;11.0]	10.0 [6.00;15.0]	< 0.001
Head pain(%):				< 0.001
No	2016 (76.8)	1242 (81.4)	774 (70.4)	
Yes	609 (23.2)	284 (18.6)	325 (29.6)	
Chest pain(%):				< 0.001
No	2336 (89.0)	1430 (93.7)	906 (82.4)	
Yes	289 (11.0)	96 (6.29)	193 (17.6)	
Sleep time	6.10 (1.72)	6.23 (1.65)	5.92 (1.79)	< 0.001

3.2 Model Performance and Evaluation

We successfully evaluated six diverse machine learning algorithms (LR, KNN, SVM, RF, LightGBM, and XGBoost) to classify prevalent CVD status in middle-aged and older adults with diabetes. The Youden index was applied to determine the optimal diagnostic threshold for each algorithm. Comprehensive performance metrics, including receiver operating characteristic (ROC) curves, calibration plots, and decision curve analysis (DCA), are illustrated in Fig. 3 .

As detailed in Table 2, all models exhibited acceptable discriminative capacity on the testing cohort (AUC range: 0.733–0.752). The RF algorithm emerged as the most robust overall performer, achieving the highest AUC of 0.752 (95% CI: 0.716–0.786) on the testing set. Notably, the RF model maintained a comparable AUC of 0.773 (95% CI: 0.751–0.795) during training, indicating strong generalization capabilities without evidence of overfitting. While the XGBoost model yielded a marginally higher overall accuracy (68.4%) and F1 score (0.646), the RF model demonstrated significantly superior sensitivity (72.53%). In the context of early cardiovascular screening, maximizing sensitivity to minimize false negatives is of paramount clinical importance; thus, RF was deemed clinically superior.

Further analysis revealed distinct limitations in the alternative algorithms. For instance, although the KNN model achieved the highest specificity (74.35%) on the test set, it suffered from unacceptably low sensitivity (59.88%), leading to

unbalanced performance and the lowest F1 score. Conversely, the LR model showed strong sensitivity (73.77%), yet its AUC and specificity were inferior to those of the RF model.Consequently, the RF algorithm was selected as the optimal model for downstream interpretability analysis and practical deployment.

Table 2
Comparison of Performance Among Different Models

Data	threshold	Model	AUC	Accuracy(%)	Sensitivity(%)	Specificity(%)	Precision(%)	F1 Score
Training	0.451	KNN	0.747	69.24	60.00	**75.99**	**64.58**	0.622
	0.405	LR	0.743	68.21	72.26	65.25	60.28	0.657
	0.422	SVM	0.743	68.48	70.45	67.04	60.94	0.654
	0.428	LightGBM	0.767	70.11	69.42	70.62	63.29	0.662
	0.492	Random Forest	**0.773**	69.90	**74.32**	66.67	61.94	**0.676**
	0.423	XGBoost	0.764	**70.17**	70.71	69.77	63.06	0.667
Test	0.451	KNN	0.733	68.40	59.88	**74.35**	**61.98**	0.609
	0.405	LR	0.748	67.01	**73.77**	62.28	57.73	**0.648**
	0.422	SVM	0.749	68.02	70.68	66.16	59.33	0.645
	0.428	LightGBM	0.749	**69.16**	68.52	69.61	61.16	0.646
	0.492	Random Forest	**0.752**	66.12	72.53	61.64	56.90	0.638
	0.423	XGBoost	0.751	68.4	70.06	67.24	59.89	0.646

3.3 Interpretable Analysis of the Model

SHAP plots provide an intuitive analysis of machine learning by illustrating the direction and strength of feature influence on prediction outcomes. Therefore, we conducted SHAP analysis on the optimal model to generate both SHAP feature importance plots and swarm plots (see Fig. 4). In both figures, the vertical axis displays feature names sorted by importance. Figure 4a shows the average SHAP value on its horizontal axis, where higher values indicate greater influence on the outcome. Figure 4b plots SHAP values horizontally, with each data point representing a sample. Colors indicate feature value levels, enabling analysis of the relationship between features and prediction outcomes through their distribution patterns.

3.4 Deployment of the Visual Clinical Decision Support System

Translating these predictive insights into practical utility, we developed a user-friendly, web-based visual decision support system (Fig. 4). The interface is structurally divided into a patient data input panel on the left and a dynamic prediction dashboard on the right. Clinicians can seamlessly input continuous clinical variables via interactive sliders and select categorical variables from dropdown menus.

Upon execution, the system computes the individualized probability of CVD. To ensure clinical transparency, this probability is accompanied by a localized SHAP force plot (Fig. 4d). This customized visualization elegantly dissects the patient's specific risk profile: the length of the arrows denotes the magnitude of influence for each variable, with red arrows driving the prediction toward a higher risk of CVD (positive impact) and blue arrows mitigating that risk (negative impact).

4 Discussion

In this research, we built and validated a machine learning-based model for assessing cardiovascular disease (CVD) risk among middle-aged and older adults with diabetes. This group of patients faces a notably higher risk of developing CVD compared to the general population. Among all the algorithms we tested, the random forest (RF) model stood out as the best performer, showing consistent predictive performance. We then integrated this model into an interpretable web-based visualization tool, which we designed to support clinical decision-making.

For missing values in the dataset, we used multivariate iterative imputation. After finishing the imputation step, we carried out feature selection using a combined strategy of LASSO regression and the Boruta algorithm. We then took the intersection of the variables selected by these two methods to get our final feature set. This dual-step selection approach substantially improved both the generalization ability and parsimony of the final model.

We used the key features we identified to build the final model, and ranked their relative predictive importance using SHAP plots (Fig. 4). As shown in Fig. 4-a, the feature importance, ordered from highest to lowest, was: hypertension, dyslipidemia, SRH, age, ADL score, renal disease, hospitalization in the preceding year, and chest pain.As shown in Fig. 4b, the presence of hypertension, dyslipidemia, renal disease, hospitalization in the preceding year, and chest pain markedly elevated the predicted probability of assessed probability of having CVD.In contrast, better self-rated health and younger age reduced this risk. The links between these variables and CVD have been well documented in earlier research.

Epidemiological research has confirmed that diabetes is not just a metabolic disorder, but also makes people more likely to develop related conditions such as hypertension and arterial stiffness [29]. What's more, the perspective put forward by Fuchs and colleagues [30]—that CVD risk is primarily driven by a rightward shift in blood pressure distribution—aligns perfectly with our findings, which showed hypertension had the highest SHAP value.

Similarly, research by Kalra et al. [31] found that diabetic dyslipidemia, which typically involves lower HDL cholesterol and higher triglyceride levels, is closely linked to atherosclerotic cardiovascular disease, which in turn drives up CVD risk. The strong predictive power of self-rated health (SRH) in our model matches the findings of van der Linde et al. [32], who identified SRH as a reliable predictor of both fatal and non-fatal CVD events over a 10-year period in middle-aged populations.

Age, one of the most critical demographic factors, had a clear impact on the model's predictions. Previous research has shown that CVD prevalence rises exponentially with age across different age groups [33], which matches what we observed in our analysis.

Our results also showed that more severe impairment in activities of daily living (ADL) had a stronger effect on CVD risk. Impaired ADL may indirectly increase CVD incidence through multiple pathways. These include lower physical activity levels, faster structural aging of the heart and blood vessels, less social engagement, and higher risk of depression and anxiety.Diabetes-related chronic kidney disease also substantially increases CVD risk [35]. Diabetic patients have nearly twice the risk of developing chronic kidney disease (CKD) compared to non-diabetic individuals. This is because kidney disease contributes to cardiovascular pathology through a combination of traditional risk factors—hypertension, diabetes, dyslipidemia—and CKD-specific mechanisms, such as inflammatory responses and oxidative stress [35,36]

Chest pain, a common symptom of CVD, is also a known risk factor for future cardiovascular morbidity and mortality [36]. Based on our findings, hospitalization in the past year may be linked to multiple diabetes-related health issues in patients. We think this acts as a surrogate marker for the overall severity and instability of the disease.

After comparing the performance of all six models, we selected RF as our optimal model. On the independent test set, the RF model achieved excellent AUC-ROC and sensitivity, which reflects its strengths in both discrimination and sensitivity. What's more, decision curve analysis (DCA) of the test set showed that the RF model provided a higher net clinical benefit for most intervention thresholds, except for a small number of low-preference intervals.

In terms of calibration, the RF model reached a Brier score of 0.202 (95% CI 0.195–0.211) on the test set, which indicates good calibration between predicted probabilities and observed outcomes. Taken together, these results show that RF has strong generalization capabilities and superior performance, which is why we chose it as our final model.

Early work on CVD prediction in diabetic patients mostly relied on traditional predictive models. Zethelius did an observational study using the Swedish National Diabetes Register. Through Cox regression analysis, he identified 12 significant CVD risk factors, including prior CVD, age at onset, diabetes duration, lipid ratios, and HbA1c, and built a 5-year CVD risk prediction model [38]. Kengne et al. developed a similar model for type 2 diabetes patients, using Cox proportional hazards models based on the ADVANCE study cohort [39].

While these earlier models have been shown to be effective, the inherent linear assumptions of traditional statistical models make it hard for them to capture complex patterns in high-dimensional data. Compared with these previous studies, our work brings several unique advantages.

First, we used the CHARLS dataset, which has long follow-up periods, broad geographic coverage, rich phenotypic data, and rigorous quality control. This high-quality data makes our findings more reliable. Second, we used machine learning models, which are better than traditional statistical models at handling high-dimensional data and non-linear relationships.

Third, we combined Boruta and LASSO regression for feature selection, which let us identify a small, parsimonious set of key predictors. This helped us boost both model efficiency and generalization. Finally, we went beyond simple black-box predictions. We used SHAP analysis to interpret the model's predictions, and built an interactive visual predictive system that can deliver personalized risk assessments to patients.

Even with these strengths, our study has some limitations that should be noted. First, as a cross-sectional study that only used 2020 data, we cannot establish causal relationships between the variables we identified and prevalent CVD, nor can we evaluate the model's long-term prognostic performance.Second, while the CHARLS database covers various regions across China, it focuses exclusively on East Asian populations. This may limit how well our model can generalize to other ethnic groups.

Additionally, while missing data prevented the inclusion of clinical biomarkers (e.g., laboratory blood test results) and BMI, this limitation paradoxically highlights the model's unique clinical utility. By relying exclusively on easily accessible, non-invasive features (such as self-reported history and ADL scores), our model serves as a zero-cost, rapid screening tool highly suitable for resource-limited primary care settings or community-based eldercare.The self-rated health variable also has inherent subjective bias. Even though it can predict outcomes, some patients may misjudge their own health status, which could lead to inaccurate results.

In this study, we addressed a critical gap in diabetes and cardiovascular disease research for middle-aged and older populations, using machine learning. We built a machine learning model validated on our cohort, and used XAI methods to quantify the contribution of each feature. This let us clarify how each feature affects CVD risk prediction. This approach offers new insights into the academic question of how DM is linked to cardiovascular disease.

What's more, we developed a visualization system that can quickly provide patients with CVD risk assessments and interpretable results, based on their clinical features. This tool has the potential to support rapid clinical triage and personalized screening strategies, which could improve therapeutic efficiency by identifying high-risk patients for further laboratory testing.Additionally, by providing clear risk assessments, this work may help diabetic patients better understand their own risk, which could in turn increase their motivation for early intervention.

Ultimately, this line of work could help reduce mortality from diabetes and its cardiovascular complications, ease the strain on healthcare resources and economic burdens, and improve population health standards in the long term.

5 Conclusion

This study employed six machine learning algorithms on the CHARLS dataset to assess cardiovascular disease risk in middle-aged and elderly diabetic patients. The RF model emerged as the optimal model, demonstrating superior performance.To enhance interpretability and clinical utility,, we applied SHAP analysis for visualization and deployed an online predictive system. This tool assists healthcare professionals in assessing CVD risk efficiently, facilitating timely intervention.

Declarations

Funding

This work was supported by the Shanghai Jiao Tong University Talent Cultivation Program (Grant No. 2023QN040) for the project "Research on Chronic Disease Health Promotion Strategies Based on Population Medicine Theory".

Clinical trial registration

Clinical trial number: not applicable.

Ethics statement

The studies involving humans were approved by Peking University Institutional Review Board (IRB00001052-11015). The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

Author contributions Changli Wang: Methodology, Data curation, Writing – original draft. Lvkan Weng : Supervision, Writing – review & editing. All authors have read and approved the final manuscript.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here:https://charls.charlsdata.com/pages/data/111/zh-cn.html

References

1. International Diabetes Federation. IDF Diabetes Atlas, 11th edn. Brussels, Belgium: 2025. Available at: https://diabetesatlas.org
2. Xue L, Wang H, He Y, et al. Incidence and risk factors of diabetes mellitus in the Chinese population: a dynamic cohort study. *BMJ Open*. 2022;12(11):e060730. doi:10.1136/bmjopen-2021-060730
3. Magnan EM, Bolt DM, Greenlee RT, Fink J, Smith MA. Stratifying Patients with Diabetes into Clinically Relevant Groups by Combination of Chronic Conditions to Identify Gaps in Quality of Care. *Health Serv Res*. 2018;53(1):450-468. doi:10.1111/1475-6773.12607

4. Raghavan S, Vassy JL, Ho YL, et al. Diabetes Mellitus-Related All-Cause and Cardiovascular Mortality in a National Cohort of Adults. *J Am Heart Assoc*. 2019;8(4):e011295. doi:10.1161/JAHA.118.011295
5. Stratton IM, Adler AI, Neil HA, et al. Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): prospective observational study. *BMJ*. 2000;321(7258):405-412. doi:10.1136/bmj.321.7258.405
6. Chipayo-Gonzales D, Ramakrishna H, Nuñez-Gil IJ. Score2: A New Updated Algorithm to Predict Cardiovascular Disease Risk in Europe. *J Cardiothorac Vasc Anesth*. 2022;36(1):18-21. doi:10.1053/j.jvca.2021.09.033
7. McGorrian C, Yusuf S, Islam S, et al. Estimating modifiable coronary heart disease risk in multiple regions of the world: the INTERHEART Modifiable Risk Score. *Eur Heart J*. 2011;32(5):581-589. doi:10.1093/eurheartj/ehq448
8. Wan S, Wan F, Dai X jian. Machine learning approaches for cardiovascular disease prediction: A review. *Arch Cardiovasc Dis*. 2025;118(10):554-562. doi:10.1016/j.acvd.2025.04.055
9. Chen Z, Li T, Guo S, Zeng D, Wang K. Machine learning-based in-hospital mortality risk prediction tool for intensive care unit patients with heart failure. *Front Cardiovasc Med*. 2023;10:1119699. doi:10.3389/fcvm.2023.1119699
10. Heseltine-Carp W, Courtman M, Browning D, et al. Machine learning to predict stroke risk from routine hospital data: A systematic review. *Int J Med Inf*. 2025;196:105811. doi:10.1016/j.ijmedinf.2025.105811
11. Ahsan MM, Siddique Z. Machine learning-based heart disease diagnosis: A systematic literature review. *Artif Intell Med*. 2022;128:102289. doi:10.1016/j.artmed.2022.102289
12. Zhu H, Qiao S, Zhao D, et al. Machine learning model for cardiovascular disease prediction in patients with chronic kidney disease. *Front Endocrinol*. 2024;15:1390729. doi:10.3389/fendo.2024.1390729
13. Huang Q, Jiang Z, Shi B, et al. Characterisation of cardiovascular disease (CVD) incidence and machine learning risk prediction in middle-aged and elderly populations: data from the China health and retirement longitudinal study (CHARLS). *BMC Public Health*. 2025;25(1):518. doi:10.1186/s12889-025-21609-7
14. Kwiendacz H, Wijata AM, Nalepa J, et al. Machine learning profiles of cardiovascular risk in patients with diabetes mellitus: the Silesia Diabetes-Heart Project. *Cardiovasc Diabetol*. 2023;22(1):218. doi:10.1186/s12933-023-01938-w
15. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206-215. doi:10.1038/s42256-019-0048-x
16. Chen X, Wang Y, Strauss J, Zhao Y. China Health and Retirement Longitudinal Study (CHARLS). In: *Encyclopedia of Gerontology and Population Aging*. Springer, Cham; 2021:948-956. doi:10.1007/978-3-030-22009-9_333
17. Dai X, He S, Wu Y, Zhuang J, Xu G. Association between sarcopenia and cardiovascular disease according to menopausal status: findings from the China Health and Retirement Longitudinal Study (CHARLS). *BMC Public Health*. 2025;25(1):730. doi:10.1186/s12889-025-21933-y
18. IterativeImputer — scikit-learn 1.7.2 documentation. Accessed November 20, 2025. https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html
19. Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: A new perspective. *Neurocomputing*. 2018;300:70-79. doi:10.1016/j.neucom.2017.11.077
20. Sun T, Liu J, Yuan H, Li X, Yan H. Construction of a risk prediction model for lung infection after chemotherapy in lung cancer patients based on the machine learning algorithm. *Front Oncol*. 2024;14:1403392. doi:10.3389/fonc.2024.1403392
21. Boakye NF, O'Toole CC, Jalali A, Hannigan A. Comparing logistic regression and machine learning for obesity risk prediction: A systematic review and meta-analysis. *Int J Med Inf*. 2025;199:105887. doi:10.1016/j.ijmedinf.2025.105887
22. Afolabi S, Ajadi N, Jimoh A, Adenekan I. Predicting diabetes using supervised machine learning algorithms on E-health records. *Inform Health*. 2025;2(1):9-16. doi:10.1016/j.infoh.2024.12.002
23. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20(3):273-297. doi:10.1007/BF00994018

24. Salman HA, Kalakech A, Steiti A. Random Forest Algorithm Overview. *Babylon J Mach Learn*. 2024;2024:69-79. doi:10.58496/BJML/2024/007
25. Song X, Shi J, Zhu C, Xian F, Dong Z, Li J. XGBoost machine learning algorithm for predicting unplanned readmission in elderly patients with coronary heart disease. *Geriatr Nur (Lond)*. 2025;66:103609. doi:10.1016/j.gerinurse.2025.103609
26. Ge X, Zhou Y, Li Q, Tan Y, Luo Y, Hong H. Machine learning for food flavor prediction and regulation: models, data integration, and future perspectives. *J Adv Res*. Published online October 2025:S209012322500815X. doi:10.1016/j.jare.2025.10.018
27. Smits N. A note on Youden's J and its cost ratio. *BMC Med Res Methodol*. 2010;10:89. doi:10.1186/1471-2288-10-89
28. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56-67. doi:10.1038/s42256-019-0138-9
29. Jia G, Sowers JR. Hypertension in Diabetes: An Update of Basic Mechanisms and Clinical Disease. *Hypertens Dallas Tex 1979*. 2021;78(5):1197-1205. doi:10.1161/HYPERTENSIONAHA.121.17981
30. Fuchs FD, Whelton PK. HIGH BLOOD PRESSURE AND CARDIOVASCULAR DISEASE. *Hypertens Dallas Tex 1979*. 2020;75(2):285-292. doi:10.1161/HYPERTENSIONAHA.119.14240
31. Kalra S, Raizada N. Dyslipidemia in diabetes. *Indian Heart J*. 2024;76(Suppl 1):S80-S82. doi:10.1016/j.ihj.2023.11.002
32. van der Linde RM, Mavaddat N, Luben R, et al. Self-Rated Health and Cardiovascular Disease Incidence: Results from a Longitudinal Population-Based Cohort in Norfolk, UK. *PLoS ONE*. 2013;8(6):e65290. doi:10.1371/journal.pone.0065290
33. Zhang J, Tong H, Jiang L, Zhang Y, Hu J. Trends and disparities in China's cardiovascular disease burden from 1990 to 2019. *Nutr Metab Cardiovasc Dis*. 2023;33(12):2344-2354. doi:10.1016/j.numecd.2023.07.039
34. Han W, Mao K, Li Y, et al. Trajectories of activities of daily living/ instrumental activities of daily living and the risk of cardiovascular diseases. *Nutr Metab Cardiovasc Dis*. 2025;35(6):103969. doi:10.1016/j.numecd.2025.103969
35. Koye DN, Magliano DJ, Nelson RG, Pavkov ME. The Global Epidemiology of Diabetes and Kidney Disease. *Adv Chronic Kidney Dis*. 2018;25(2):121-132. doi:10.1053/j.ackd.2017.10.011
36. Marx-Schütt K. Cardiovascular disease in chronic kidney disease.
37. Barnett LA, Prior JA, Kadam UT, Jordan KP. Chest pain and shortness of breath in cardiovascular disease: a prospective cohort study in UK primary care. *BMJ Open*. 2017;7(5):e015857. doi:10.1136/bmjopen-2017-015857
38. Zethelius B, Eliasson B, Eeg-Olofsson K, Svensson AM, Gudbjörnsdottir S, Cederholm J. A new model for 5-year risk of cardiovascular disease in type 2 diabetes, from the Swedish National Diabetes Register (NDR). *Diabetes Res Clin Pract*. 2011;93(2):276-284. doi:10.1016/j.diabres.2011.05.037
39. Contemporary model for cardiovascular risk prediction in people with type 2 diabetes | European Journal of Preventive Cardiology | Oxford Academic. Accessed November 20, 2025. https://academic.oup.com/eurjpc/article-abstract/18/3/393/5931558

Figures

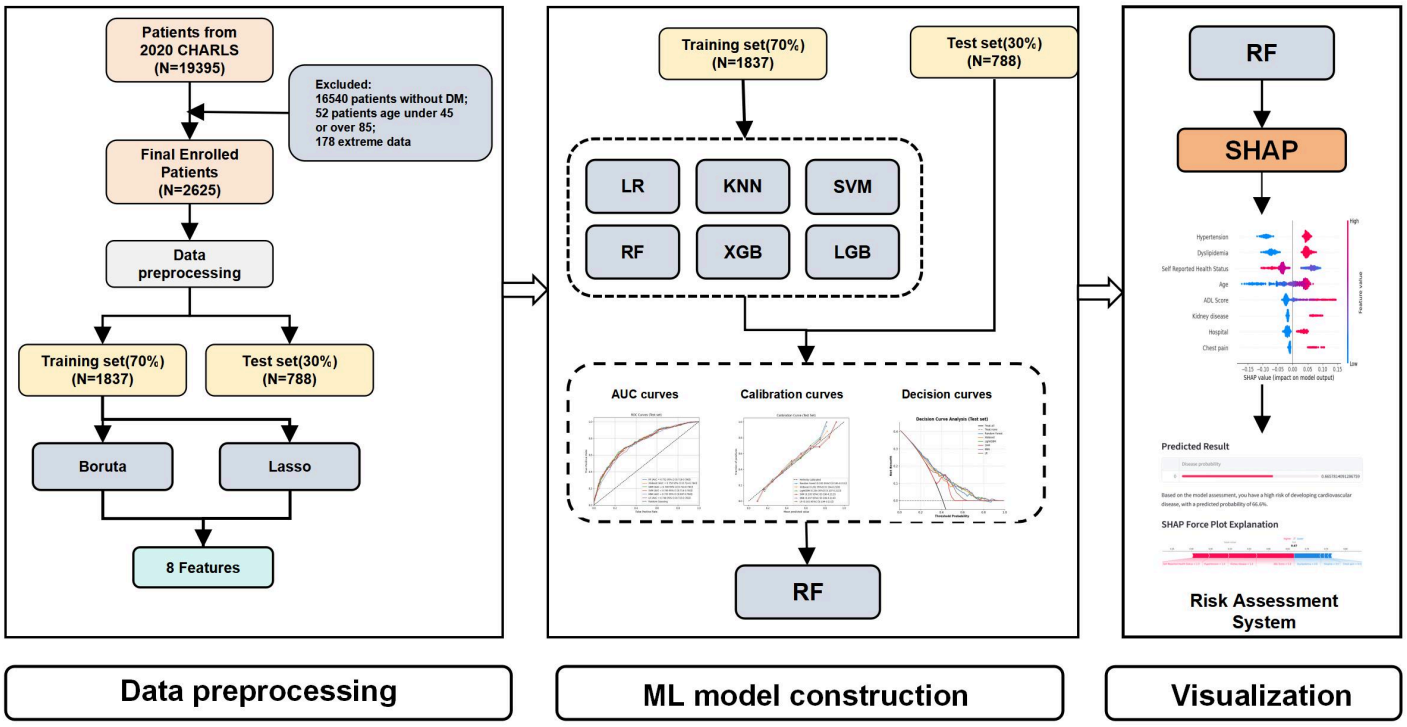


Figure 1

The overall flowchart of the study.

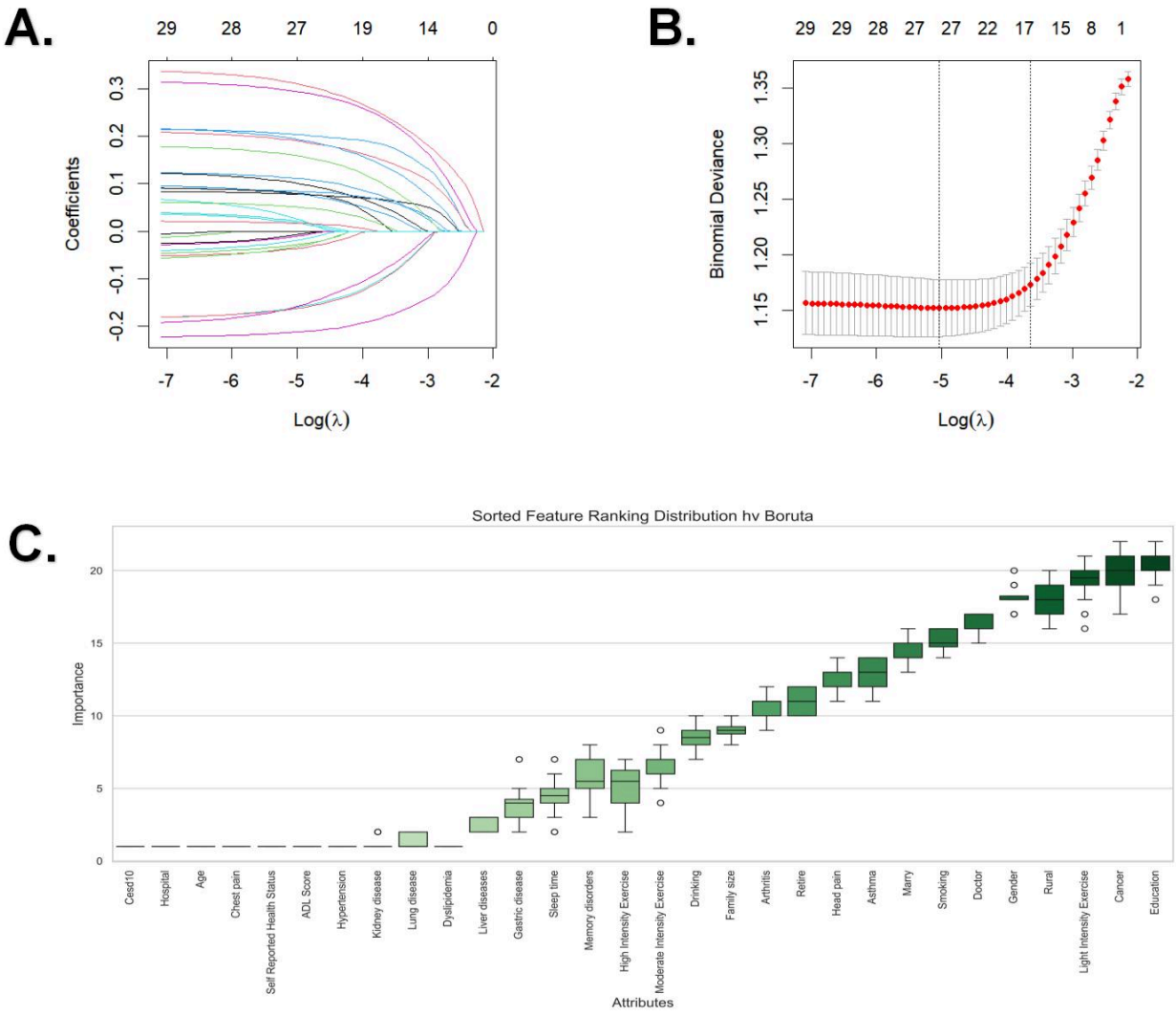


Figure 2

Feature Selection by the LASSO Regression Model and Boruta Algorithm(A) Trajectory plot of variable screening by the LASSO regression model. The bottom x-axis is logλ, the y-axis is binomial deviance, and the top x-axis is the number of variables.(B) Selection of the optimal parameter (λ) in the LASSO regression model, with logλ as the x-axis and regression coefficients as the y-axis. The left dashed line represents the optimal λ value, while the right dashed line denotes the λ value within one standard error of the optimal value.□C□Boruta

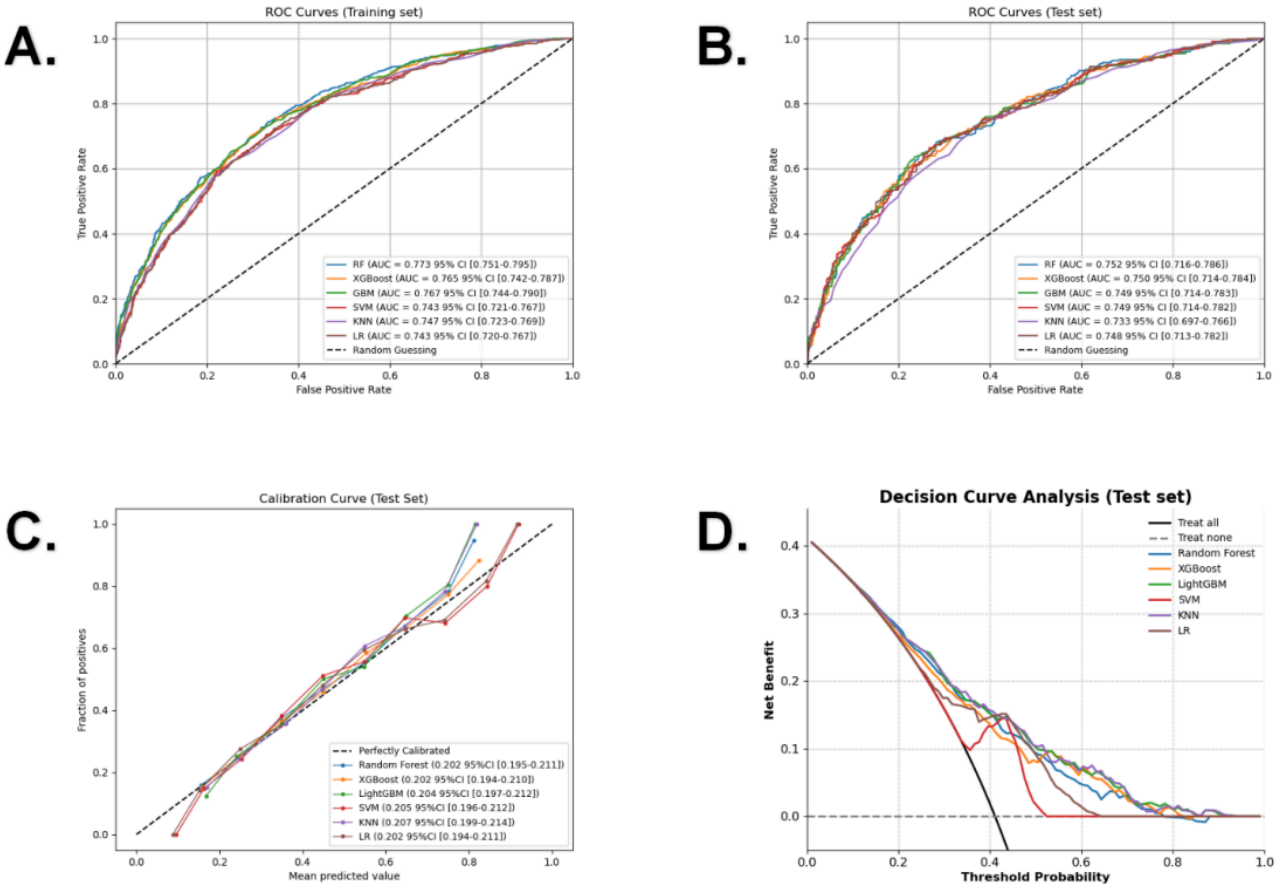


Figure 3

Performance and Comparison of Machine Learning Predictive Models(A) ROC Curve of the Training Set; (B) ROC Curve of the Test Set; (C) Calibration Curve of the Test Set; (D) Decision Curve Analysis of the Test Set

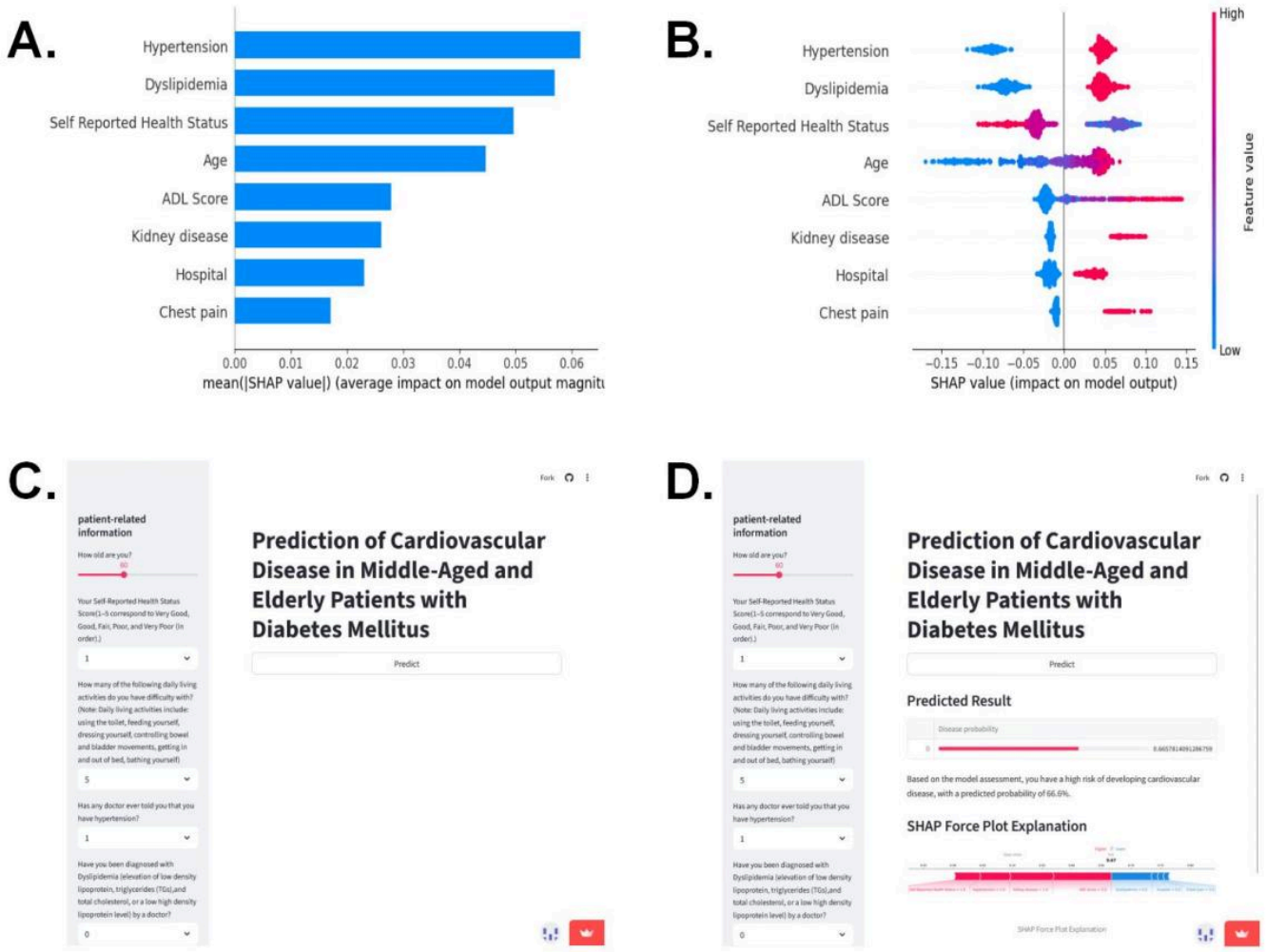


Figure 4

SHAP Analysis and Visualization System for the Model (A)The SHAP dependence plot of the RF model.With the x-axis representing variable names sorted by SHAP values and the y-axis denoting SHAP values. (B)SHAP Summary Plot: The x-axis is the SHAP value, and the y-axis is the variable name. Each dot represents a sample, with red dots indicating higher feature values and blue dots indicating lower feature values. (C)Main Interface of the Visualization System: On the left is the information input panel, and on the right is the result output area.□D□Presentation of Predictive Results: Visualization of individual results via SHAP force plot analysis.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.docx