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.

An Interpretable Machine Learning Model for the Diagnosis of Coronary Heart Disease in Patients with Type 2 Diabetes Mellitus: A Multi-Center Cohort Study

Yibo Lin
Health，Guangdong Medical University
Xie Yanhong
Qinghai Province Haidong City Center for Disease Control and Prevention
Zheng Jianfeng
Affiliated Hospital Group of Guangdong Medical University Shenzhen Baoan Central Hospital
Yu Zhao
Affiliated Hospital Group of Guangdong Medical University Shenzhen Baoan Central Hospital
Zilong Wen
Affiliated Hospital Group of Guangdong Medical University Shenzhen Baoan Central Hospital
Junjie Zeng
Health，Guangdong Medical University
Haibing Yu
hby616688@gdmu.edu.cn

Health，Guangdong Medical University

Research Article

Keywords: T2DM, CHD, Machine learning, Gated Recurrent Unit (GRU), SHAP

Posted Date: June 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9840166/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

Objective

To develop and validate an interpretable machine learning model for predicting coronary heart disease (CHD) in patients with type 2 diabetes mellitus (T2DM), providing a clinical decision support tool for early screening and risk stratification.

Methods

This multi-center retrospective cohort study included 2,187 patients with T2DM. The training cohort (n = 1,667) was randomly partitioned into training and internal validation sets at a 7:3 ratio, while 520 patients formed the external validation cohort. Feature selection was performed using Spearman correlation analysis, Variance Inflation Factor (VIF), univariate logistic regression, Boruta, and LASSO regression. Based on 11 identified variables, 15 machine learning models were developed and compared. Model performance was comprehensively evaluated via Area Under the Curve (AUC), accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1-score, Matthews Correlation Coefficient (MCC), Brier score, calibration curves, and Decision Curve Analysis (DCA). Model interpretability was addressed using SHAP (SHapley Additive exPlanations). A sensitivity analysis was performed to evaluate the robustness of the model.

Results

Among all candidate models, the Gated Recurrent Unit (GRU) model demonstrated the superior performance. It achieved an AUC of 0.7625 and a sensitivity of 0.7466 in the test set. In the external validation set, the AUC was 0.7188, accuracy was 0.7192, sensitivity was 0.5639, specificity was 0.7726, and the Brier score was 0.1876, indicating robust discrimination, calibration, and generalizability. SHAP analysis revealed that age, sex, and LDL-C were the most significant predictors, followed by fasting blood glucose (FBG), systolic blood pressure (SBP), sleep duration, and sweet food intake. These findings underscore the critical roles of demographic characteristics, lipid metabolism, and lifestyle factors in predicting CHD risk among T2DM patients.

Conclusion

This study established an interpretable GRU-based predictive model using multi-center real-world data, which effectively identifies high-risk CHD patients within the T2DM population. With its strong predictive performance, reliable calibration, and high interpretability, the model holds potential to support non-specialists in conducting early screening and personalized interventions.

1. Background/ Introduction

Type 2 diabetes mellitus (T2DM) is a chronic metabolic disorder characterized by insufficient insulin secretion, impaired insulin action, or a combination of both. According to the International Diabetes Federation (IDF), approximately 589 million adults (aged 20–79) worldwide are living with diabetes. In 2024 alone, diabetes was responsible for 3.4 million deaths, equivalent to one death every six seconds[1].The persistent hyperglycemic state in T2DM patients promotes the formation of advanced glycation end-products (AGEs), which impair vascular endothelial function and trigger chronic inflammatory responses[2]. Concurrently, insulin resistance and associated dyslipidemia in T2DM accelerate the progression of atherosclerosis by enhancing oxidative stress and activating macrophages, ultimately leading to arterial plaque formation and vascular injury[3]. This complex interplay between metabolic and vascular dysfunction significantly elevates the risk of coronary heart disease (CHD). In China, among approximately 1.87 million patients with T2DM, about 67.5% are classified as being at very high risk for cardiovascular disease (CVD), with CHD being the most prevalent cardiovascular complication[4, 5].

T2DM patients often suffer from severe autonomic neuropathy, which leads to an increased pain threshold. Consequently, initial symptoms are often insidious, with early myocardial ischemia being mild or asymptomatic. This latency makes early detection challenging, often resulting in patients presenting only after developing severe coronary lesions[6]. Although coronary angiography (CA) remains the gold standard for CHD diagnosis, its invasive nature and associated risks limit its utility in widespread early screening[7]. Alternative modalities, such as Coronary Computed Tomography Angiography (CCTA) and Cardiac Magnetic Resonance Imaging (MRI), are commonly used. While CCTA offers high sensitivity and negative predictive value, its diagnostic accuracy remains a subject of debate[8]. Cardiac MRI provides a non-invasive assessment of coronary artery disease; however, its clinical utility is constrained by contraindications in patients with metallic implants and low accessibility[9]. Therefore, there is a critical need for novel auxiliary methods for the early diagnosis of CHD in T2DM patients. Machine learning (ML) has demonstrated a robust capacity to identify patterns and generate predictions from complex data. Compared to traditional predictive modeling, ML can capture intricate non-linear relationships between variables, thereby enabling more precise disease diagnosis[10, 11]. Nevertheless, existing early diagnostic models for this specific high-risk population often suffer from limitations such as poor interpretability and a lack of multi-center validation.

Utilizing a multi-center retrospective cohort dataset, this study aims to develop and validate an interpretable machine learning (ML) model for patients with type 2 diabetes mellitus (T2DM) complicated by coronary heart disease (CHD). By quantifying pivotal risk factors and predicting the risk of early-onset CHD, this research is designed to empower non-specialist clinicians to accurately identify occult cardiovascular risks, thereby providing robust decision support for personalized therapeutic interventions.

2. Methods

2.1 Study design and population

This multi-center retrospective study included patients with type 2 diabetes mellitus (T2DM) and T2DM complicated by coronary heart disease (CHD) collected from the inpatient departments of Endocrinology and Cardiology at the Affiliated Hospital of Guangdong Medical University and the Second Affiliated Hospital of Guangdong Medical University from July 2021 to December 2022, and from Shenzhen Baoan Central Hospital of the Guangdong Medical University Hospital Group from July 2025 to March 2026. Data from the Affiliated Hospital of Guangdong Medical University and the Second Affiliated Hospital of Guangdong Medical University were used for model training and internal validation. To rigorously evaluate model performance, we used data from Shenzhen Baoan Central Hospital of the Guangdong Medical University Hospital Group as the external validation set to ensure that the model could withstand the challenges of both geographical variations and temporal shifts in clinical practice. The training cohort initially included 4,713 subjects; after excluding 850 individuals with incomplete questionnaire information, 39 with a missing data rate > 20%, and 2,157 who did not have T2DM, 1,667 eligible participants were finally included and randomly partitioned into training and test sets at a 7:3 ratio. For the external validation, 538 subjects were initially enrolled; after excluding 15 with incomplete questionnaires and 3 with a missing data rate > 20%, 520 participants were ultimately included for external validation. (Figure 1)

2.2 Inclusion and Exclusion Criteria

Inclusion Criteria: Adults aged ≥18 years; Patients meeting the diagnostic criteria for T2DM; Availability of complete laboratory test data; Provision of written informed consent.

Exclusion Criteria: Presence of other endocrine disorders or other cardiovascular and cerebrovascular diseases; Severe comorbidities, such as malignant tumors or chronic hepatic and renal insufficiency; Severe psychiatric disorders or cognitive impairment; Inability to comply with the study protocol or lost to follow-up.

2.3 Definition and outcome

T2DM was defined according to the following criteria: fasting plasma glucose (FPG) ≥ 7.0 mmol/L, 2-hour plasma glucose (2h-PG) during a 75-g oral glucose tolerance test (OGTT) ≥ 11.1 mmol/L, or a prior diagnosis of T2DM with the current use of oral hypoglycemic agents or insulin therapy.[12].According to the guidelines of the American College of Cardiology (ACC)/American Heart Association (AHA)/European College of Cardiology (ESC), the diagnosis of CAD was made by chief physicians based on clinical histories, angina symptoms, cardiac markers, and specific examinations, including

electrocardiogram (rest and exercise), echocardiography, radionuclide imaging, computed tomography, and coronary angiography[13,14].

2.4 Data and variables

Data collection primarily encompassed three components: questionnaire surveys, physical examinations, and laboratory testing.

Questionnaire Surveys: (1) Socio-demographic characteristics, including age, sex, educational level, marital status, and occupation; (2) Family history, including family history of diabetes and coronary heart disease; (3) Lifestyle and self-care behaviors, including smoking status, alcohol consumption, dietary habits (intake of vegetables, meat, fruits, and sweets), physical activity, and sleep duration.

Physical Examinations: Measurements included height, weight, waist circumference, hip circumference, systolic blood pressure (SBP), diastolic blood pressure (DBP), and heart rate. All physical measurements were performed using standardized protocols. To ensure data quality and inter-rater reliability, all examiners underwent unified training and rigorous assessment; only those who demonstrated proficiency and passed the examination were authorized to conduct measurements.

Laboratory Testing: Biochemical assessments included fasting blood glucose (FBG), total cholesterol (TC), triglycerides (TG), high-density lipoprotein cholesterol (HDL-C), and low-density lipoprotein cholesterol (LDL-C). All biochemical indices were based on the initial measurement obtained within the first 24 hours of hospital admission.

2.5 Feature selection

Prior to model development, variables with high redundancy were filtered using Spearman correlation analysis. Multicollinearity among the candidate predictors was quantified using the Variance Inflation Factor (VIF). Subsequently, feature selection was performed using a tripartite approach: univariate logistic regression, the Boruta algorithm, and Least Absolute Shrinkage and Selection Operator (LASSO) regression. Univariate logistic regression served as a traditional statistical method, where variables with a P-value < 0.05 were considered statistically significant. The Boruta algorithm, a wrapper method based on the Random Forest architecture, was employed for feature ranking and selection. LASSO regression, a regularization technique, was utilized to shrink coefficients and eliminate non-essential variables via an L1-norm penalty[15,16]. The final feature set was determined by synthesizing the results from these three methods while simultaneously accounting for clinical relevance.

2.6 Modelling and validation

During the model development phase, the training data were randomly partitioned into a training set and an internal validation set at a 7:3 ratio. To address the inherent class imbalance within the study population, over-sampling techniques were applied to the training set to prevent the models from favoring the majority class. The training set was standardized to remove differences in scale across variables.Based on considerations of model diversity, data suitability, robustness against overfitting, and clinical utility, 15 machine learning models were constructed and compared: Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Backpropagation Neural Network (BP), Extreme Learning Machine (ELM), Radial Basis Function (RBF) network, Partial Least Squares (PLS) regression, Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), Naive Bayes (NB), K-Nearest Neighbors (KNN), Decision Tree (DT), Least Squares Support Vector Machine (LSSVM), Convolutional Neural Network (CNN), and Bidirectional Long Short-Term Memory (BiLSTM).

Model performance was comprehensively assessed using multidimensional metrics. The core evaluation focused on the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC), alongside Calibration Curves and Decision Curve Analysis (DCA). Furthermore, accuracy (ACC), sensitivity (SEN/Recall), specificity (SPE), positive predictive value (PPV), negative predictive value (NPV), F1-score, and Matthews Correlation Coefficient (MCC) were calculated to provide a holistic evaluation of model efficacy. Model interpretability was achieved via SHapley Additive exPlanations (SHAP).

To evaluate the impact of the feature selection strategy on model development, a sensitivity analysis was further performed by relaxing the inclusion criterion to variables selected by at least two of the three methods. The model was then rebuilt using the revised feature set, and its predictive performance was compared with that of the primary analysis to assess consistency.

2.7 Statistical analysis

In the initial dataset, features with a missing rate exceeding 25% were excluded from the analysis. To mitigate the potential impact of missing data on model performance, remaining missing values were addressed using Multiple Imputation by Chained Equations (MICE) via the mice package (version 3.16.0). Continuous variables were expressed as medians and interquartile ranges (IQR), with inter-group comparisons performed using the Mann-Whitney U test. Categorical variables were described as frequencies and percentages, and group differences were analyzed using the Chi-square (χ^2) test or Fisher's exact test, as appropriate. All statistical tests were two-tailed, and a P-value < 0.05 was considered statistically significant. All statistical analyses and modeling were conducted using R (version 4.5.1) and MATLAB (version R2024b).

3. Results

3.1 Characteristics of study participant

A total of 2,187 participants were included in this study. The training cohort consisted of 1,667 patients (1,163 without CHD and 504 complicated by CHD), while the external validation cohort comprised 520 patients. Within the training cohort, significant differences were observed between the CHD and non-CHD groups in age, height, weight, BMI, waist circumference, hip circumference, blood pressure, lipid profiles, blood glucose, and family history (all $P < 0.05$). Specifically, the CHD group was significantly older than the non-CHD group (median 68.00 vs. 61.00 years, $P < 0.001$), with a notably higher proportion of elderly individuals. Patients in the CHD group exhibited higher systolic blood pressure (140.00 vs. 133.00 mmHg, $P < 0.001$), significantly elevated levels of total cholesterol (TC) and LDL-C, and lower levels of HDL-C (all $P < 0.05$). Furthermore, the CHD group had a higher proportion of males (61.90% vs. 54.60%, $P < 0.05$). Occupational status, marital status, and family history of both diabetes and CHD also showed significant distribution differences between the two groups (all $P < 0.05$). The prevalence of central obesity and hypertension was significantly higher in the CHD group ($P < 0.05$) .

Marked heterogeneity in baseline characteristics was observed between the training and external validation cohorts, with significant differences in age, weight, BMI, waist circumference, hip circumference, blood pressure, lipid profiles, blood glucose, and most lifestyle indicators (all $P < 0.001$). Overall, the external validation cohort was younger than the training cohort. Regarding anthropometric indices, the validation cohort exhibited higher weight and hip circumference but significantly lower BMI and markedly higher waist circumference compared to the training cohort .(Table 1)

Feature Selection

A multi-stage feature selection strategy was implemented to identify the most robust predictors. Initially, a correlation heatmap was utilized to evaluate the associations among all candidate predictors (Fig. 2A). To ensure model stability and mitigate the impact of multicollinearity, the Variance Inflation Factor (VIF) was calculated for each variable (Fig. 2B). Although certain clinical indicators exhibited moderate correlation, no significant multicollinearity was detected among the study variables. Subsequently, univariate logistic regression was employed to assess the preliminary predictive significance of each feature for the outcome (Fig. 2C). LASSO regression, integrated with 10-fold cross-validation, was then applied for effective coefficient shrinkage. Guided by the optimal error criterion, the LASSO algorithm identified a subset of key predictors with non-zero coefficients, effectively filtering out redundant variables (Fig. 2D, 2E). Furthermore, the Boruta algorithm provided rigorous screening, where green bars indicate variables confirmed as significant predictors (Fig. 2F). Ultimately, 11 core variables were selected for the final model: sex, age, occupation, sweet food intake, family history of diabetes, sleep duration, systolic blood pressure (SBP), heart rate (HR) groups, fasting blood glucose (FBG), high-density lipoprotein cholesterol (HDL-C), and low-density lipoprotein cholesterol (LDL-C). These variables were established as the primary features for predicting CHD risk in patients with T2DM.

Table 1
Baseline Characteristics of Patients

Variables	Training Cohort		*P*	Training Cohort (n = 1667)	Validation Cohort (n = 520)	*P*
	without CHD (n = 1163)	CHD (n = 504)				
Age, M (Q_1, Q_3)	61.00 (53.00, 68.00)	68.00 (60.00, 76.00)	< 0.001	63.00 (55.00, 70.00)	57.00 (47.00, 66.00)	< 0.001
Height, M (Q_1, Q_3)	162.00 (156.00, 169.00)	165.00 (158.00, 170.00)	0.009	163.00 (157.00, 169.00)	167.00 (160.00, 172.00)	< 0.001
Weight, M (Q_1, Q_3)	61.00 (54.00, 70.00)	62.50 (55.00, 70.00)	0.042	62.00 (55.00, 70.00)	67.95 (59.60, 76.53)	< 0.001
Bmi, M (Q_1, Q_3)	23.38 (21.01, 25.63)	23.51 (21.26, 25.60)	0.631	23.39 (21.10, 25.63)	20.31 (18.26, 22.53)	< 0.001
Waist, M (Q_1, Q_3)	85.00 (79.00, 90.00)	85.00 (80.00, 91.12)	0.011	85.00 (80.00, 90.00)	92.00 (86.00, 98.00)	< 0.001
Hip, M (Q_1, Q_3)	93.00 (88.00, 99.00)	94.00 (89.00, 99.00)	0.398	93.00 (88.00, 99.00)	97.00 (92.00, 103.00)	< 0.001
Sbp, M (Q_1, Q_3)	133.00 (120.50, 150.00)	140.00 (126.00, 157.00)	< 0.001	135.00 (122.00, 152.00)	127.00 (116.00, 138.00)	< 0.001
Dbp, M (Q_1, Q_3)	77.00 (70.00, 85.00)	76.00 (69.00, 85.00)	0.447	77.00 (69.00, 85.00)	80.00 (74.00, 88.00)	< 0.001
Hr, M (Q_1, Q_3)	79.00 (74.00, 88.00)	76.00 (68.00, 82.25)	< 0.001	79.00 (72.00, 86.00)	80.00 (73.00, 89.00)	0.002
Fbg, M (Q_1, Q_3)	7.80 (5.64, 11.40)	6.93 (5.48, 8.71)	< 0.001	7.41 (5.58, 10.46)	6.70 (5.60, 8.22)	< 0.001

Variables	Training Cohort		*P*	Training Cohort (n = 1667)	Validation Cohort (n = 520)	*P*
	without CHD (n = 1163)	CHD (n = 504)				
Tg, M (Q_1, Q_3)	1.47 (1.00, 2.21)	1.48 (1.04, 2.02)	0.765	1.47 (1.01, 2.17)	1.55 (1.09, 2.52)	0.003
Tc, M (Q_1, Q_3)	5.07 (4.21, 6.00)	4.50 (3.63, 5.38)	< 0.001	4.90 (4.01, 5.85)	4.48 (3.64, 5.27)	< 0.001
Hdl, M (Q_1, Q_3)	1.14 (0.93, 1.41)	1.12 (0.93, 1.32)	0.033	1.13 (0.93, 1.38)	1.06 (0.88, 1.26)	< 0.001
Ldl, M (Q_1, Q_3)	3.05 (2.32, 3.83)	2.62 (1.93, 3.33)	< 0.001	2.94 (2.19, 3.68)	2.60 (1.89, 3.42)	< 0.001
Sex, n(%)			0.006			< 0.001
Male	635 (54.60)	312 (61.90)		947 (56.81)	357 (68.65)	
Female	528 (45.40)	192 (38.10)		720 (43.19)	163 (31.35)	
Edu, n(%)			0.129			< 0.001
No formal education	92 (7.91)	44 (8.73)		136 (8.16)	67 (12.88)	
Primary/Secondary/Vocational	936 (80.48)	418 (82.94)		1354 (81.22)	339 (65.19)	
Higher education	135 (11.61)	42 (8.33)		177 (10.62)	114 (21.92)	
Marry, n(%)			0.335			< 0.001
Married	1099 (94.50)	482 (95.63)		1581 (94.84)	459 (88.27)	
Unmarried/divorced/widowed	64 (5.50)	22 (4.37)		86 (5.16)	61 (11.73)	
Work, n(%)			< 0.001			< 0.001

Variables	Training Cohort		*P*	Training Cohort (n = 1667)	Validation Cohort (n = 520)	*P*
	without CHD (n = 1163)	CHD (n = 504)				
Agriculture/Forestry/Fishery/ Water Conservancy	368 (31.64)	120 (23.81)		488 (29.27)	191 (36.73)	
Commercial and Service Industry	166 (14.27)	30 (5.95)		196 (11.76)	265 (50.96)	
Military/Government/Public Sector	137 (11.78)	61 (12.10)		198 (11.88)	38 (7.31)	
Others	492 (42.30)	293 (58.13)		785 (47.09)	26 (5.00)	
Fh T2DM, n(%)			< 0.001			0.014
No	938 (80.65)	440 (87.30)		1378 (82.66)	405 (77.88)	
Yes	225 (19.35)	64 (12.70)		289 (17.34)	115 (22.12)	
Fh CHD, n(%)			< 0.001			0.217
No	1122 (96.47)	466 (92.46)		1588 (95.26)	502 (96.54)	
Yes	41 (3.53)	38 (7.54)		79 (4.74)	18 (3.46)	
Fh Comorbidity, n(%)			0.102			0.705
No	1153 (99.14)	495 (98.21)		1648 (98.86)	513 (98.65)	
Yes	10 (0.86)	9 (1.79)		19 (1.14)	7 (1.35)	
Smoke, n(%)			0.013			< 0.001
No	866 (74.46)	357 (70.83)		1223 (73.37)	313 (60.19)	
Former	38 (3.27)	32 (6.35)		70 (4.20)	41 (7.88)	
Yes	259 (22.27)	115 (22.82)		374 (22.44)	166 (31.92)	

Variables	Training Cohort		*P*	Training Cohort (n = 1667)	Validation Cohort (n = 520)	*P*
	without CHD (n = 1163)	CHD (n = 504)				
Drink, n(%)			0.781			< 0.001
No	944 (81.17)	412 (81.75)		1356 (81.34)	329 (63.27)	
Yes	219 (18.83)	92 (18.25)		311 (18.66)	191 (36.73)	
Meat Intake, n(%)			< 0.001			< 0.001
< 7 times/week	152 (13.07)	71 (14.09)		223 (13.38)	18 (3.46)	
7–14 times/week	486 (41.79)	255 (50.60)		741 (44.45)	358 (68.85)	
≥ 14 times/week	525 (45.14)	178 (35.32)		703 (42.17)	144 (27.69)	
Vegetable Intake, n(%)			0.001			< 0.001
< 7 times/week	35 (3.01)	29 (5.75)		64 (3.84)	6 (1.15)	
7–14 times/week	521 (44.80)	252 (50.00)		773 (46.37)	39 (7.50)	
≥ 14 times/week	607 (52.19)	223 (44.25)		830 (49.79)	475 (91.35)	
Fruit Intake, n(%)			0.042			< 0.001
< 7 times/week	945 (81.26)	431 (85.52)		1376 (82.54)	136 (26.15)	
7–14 times/week	203 (17.45)	64 (12.70)		267 (16.02)	306 (58.85)	
≥ 14 times/week	15 (1.29)	9 (1.79)		24 (1.44)	78 (15.00)	
Sweet Intake, n(%)			0.006			< 0.001
< 3 times/week.	1034 (88.91)	473 (93.85)		1507 (90.40)	294 (56.54)	

Variables	Training Cohort		*P*	Training Cohort (n = 1667)	Validation Cohort (n = 520)	*P*
	without CHD (n = 1163)	CHD (n = 504)				
3–7 times/week	87 (7.48)	23 (4.56)		110 (6.60)	203 (39.04)	
≥ 7 times/week	42 (3.61)	8 (1.59)		50 (3.00)	23 (4.42)	
Exercise, n(%)			0.299			< 0.001
No	614 (52.79)	280 (55.56)		894 (53.63)	155 (29.81)	
Yes	549 (47.21)	224 (44.44)		773 (46.37)	365 (70.19)	
Sleep, n(%)			< 0.001			< 0.001
1 < 5 hours/day or requiring sleep aids	130 (11.18)	106 (21.03)		236 (14.16)	18 (3.46)	
5–7 hours/day	614 (52.79)	217 (43.06)		831 (49.85)	176 (33.85)	
7–9 hours/day.	402 (34.57)	152 (30.16)		554 (33.23)	304 (58.46)	
≥ 9 hours/day	17 (1.46)	29 (5.75)		46 (2.76)	22 (4.23)	
Central Obesity, n(%)			0.061			< 0.001
No	719 (61.82)	287 (56.94)		1006 (60.35)	153 (29.42)	
Yes	444 (38.18)	217 (43.06)		661 (39.65)	367 (70.58)	
Sbp Group, n(%)			< 0.001			< 0.001
Normal (< 140 mmHg)	724 (62.25)	251 (49.80)		975 (58.49)	156 (30.00)	
Abnormal(≥ 140 mmHg)	439 (37.75)	253 (50.20)		692 (41.51)	364 (70.00)	
Dbp Group, n(%)			0.551			< 0.001

Variables	Training Cohort		*P*	Training Cohort (n = 1667)	Validation Cohort (n = 520)	*P*
	without CHD (n = 1163)	CHD (n = 504)				
Normal (< 90 mmHg)	986 (84.78)	433 (85.91)		1419 (85.12)	247 (47.50)	
Abnormal(≥ 90 mmHg)	177 (15.22)	71 (14.09)		248 (14.88)	273 (52.50)	
Hr Group, n(%)			< 0.001			0.504
Bradycardia (< 60 bpm)	16 (1.38)	21 (4.17)		37 (2.22)	11 (2.12)	
Normal (60–100 bpm)	1060 (91.14)	466 (92.46)		1526 (91.54)	469 (90.19)	
Tachycardia (> 100 bpm)	87 (7.48)	17 (3.37)		104 (6.24)	40 (7.69)	
Z: Mann-Whitney test, χ^2: Chi-square test, M: Median, Q_1: 1st Quartile, Q_3: 3st Quartile						
CHD, coronary heart disease;SBP, systolic blood pressure; DBP, diastolic blood pressure; HR, heart rate; FBG, fasting blood glucose; TG, triglyceride; TC, total cholesterol; HDL, high-density lipoprotein; LDL, low-density lipoprotein.						

A, Correlation heatmap of all candidate variables;B, Variance inflation factor (VIF) analysis for multicollinearity diagnostics;C, Univariate logistic regression results for all candidate variables;D, LASSO regression coefficient profile plot;E, LASSO cross-validation curve;F, Boruta variable importance ranking

Evaluation of ML models

To identify the optimal predictive model for disease screening, this study comprehensively evaluated the performance of multiple candidate models across various metrics in both the test and validation sets (Table 2). Among the 15 models evaluated, the Gated Recurrent Unit (GRU) model demonstrated superior performance. In the test set, the GRU model achieved an AUC of 0.7625 and a sensitivity of 0.7466; its F1-score, MCC, and other key indicators outperformed those of the alternative models, effectively mitigating the risk of underdiagnosis in clinical screening scenarios. In the validation set, the GRU model maintained a stable AUC of 0.7188 and consistent performance across all metrics, indicating robust generalizability. Calibration curves showed a high degree of concordance between the predicted probabilities of the GRU model and the actual outcomes, supported by a low Brier score, which confirms its excellent calibration capability. Furthermore, Decision Curve Analysis (DCA) demonstrated that the GRU model yielded a higher net clinical benefit than the other models (Fig. 3F).

Table 2
Comprehensive Performance Evaluation of Each Model

Model	ACC	SEN	SPE	PPV	NPV	F1	MCC	Brier	AUC
Train									
RF	0.7186	0.5685	0.7803	0.5155	0.8147	0.5407	0.3394	0.1822	0.7527
SVM	0.7086	0.4178	0.8282	0.5	0.7757	0.4552	0.2604	0.2043	0.7115
BP	0.6447	0.7603	0.5972	0.437	0.8583	0.555	0.3249	0.2139	0.7557
ELM	0.6607	0.6849	0.6507	0.4464	0.8339	0.5405	0.3068	0.2132	0.7204
RBF	0.6667	0.6781	0.662	0.4521	0.8333	0.5425	0.3115	0.2024	0.7522
PLS	0.6567	0.726	0.6282	0.4454	0.8479	0.5521	0.3223	0.2084	0.7495
GRU	0.6747	0.7466	0.6451	0.4638	0.8609	0.5722	0.3566	0.203	0.7625
LSTM	0.6687	0.7123	0.6507	0.4561	0.8462	0.5561	0.3313	0.2051	0.758
LR	0.6587	0.7123	0.6366	0.4464	0.8433	0.5488	0.3179	0.208	0.7492
NB	0.6008	0.774	0.5296	0.4036	0.8507	0.5305	0.2778	0.2573	0.7448
KNN	0.6208	0.6233	0.6197	0.4027	0.8	0.4892	0.2219	0.2527	0.6725
DT	0.6487	0.6233	0.6592	0.4292	0.8097	0.5084	0.2598	0.2355	0.6434
LSSVM	0.2914	1	0	0.2914	0	0.4513	0	0.2836	0.6483
CNN	0.6766	0.5616	0.7239	0.4556	0.8006	0.5031	0.2705	0.217	0.7209
BiLSTM	0.6647	0.6918	0.6535	0.4509	0.8375	0.5459	0.3156	0.2014	0.7607
Valid									
RF	0.7269	0.4436	0.8243	0.4646	0.8117	0.4538	0.2721	0.1756	0.7287
SVM	0.7288	0.218	0.9044	0.4394	0.7709	0.2915	0.1605	0.1938	0.6782
BP	0.6923	0.6015	0.7235	0.4278	0.8408	0.5	0.2955	0.203	0.7068
ELM	0.6788	0.4737	0.7494	0.3938	0.8056	0.43	0.2108	0.2068	0.6571
RBF	0.6942	0.5038	0.7597	0.4188	0.8167	0.4573	0.249	0.2025	0.6722
PLS	0.7135	0.5865	0.7571	0.4535	0.842	0.5115	0.3186	0.1918	0.7223
GRU	0.7192	0.5639	0.7726	0.4601	0.8375	0.5068	0.3165	0.1876	0.7188
LSTM	0.7077	0.5865	0.7494	0.4457	0.8406	0.5065	0.3101	0.1921	0.7165
LR	0.7096	0.5564	0.7623	0.4458	0.8333	0.495	0.2982	0.1913	0.7189
NB	0.6558	0.4887	0.7132	0.3693	0.8023	0.4207	0.1862	0.2256	0.6635

Model	ACC	SEN	SPE	PPV	NPV	F1	MCC	Brier	AUC
KNN	0.6462	0.5263	0.6873	0.3665	0.8085	0.4321	0.1934	0.2366	0.6624
DT	0.6808	0.4737	0.7519	0.3962	0.8061	0.4315	0.2137	0.24	0.598
LSSVM	0.2558	1	0	0.2558	0	0.4074	0	0.2789	0.6118
CNN	0.675	0.3985	0.77	0.3732	0.7884	0.3855	0.165	0.234	0.6398
BiLSTM	0.7192	0.5414	0.7804	0.4586	0.832	0.4966	0.3057	0.1864	0.7218

A, Test set ROC comparison;B, External validation set ROC comparison;C, Test set DCA comparison ;D, External validation set DCA comparison;E,Test set Calibration Curve;F,External validation Calibration Curve

Model interpretability

SHAP analysis revealed that age was the most influential feature for model prediction, followed by sex and LDL-C. Together, these three factors represent the primary predictors of the model, while the contributions of the remaining features were relatively marginal (Fig. 4A). The SHAP heatmap visualizes the varying contributions of each feature at the individual level, highlighting the heterogeneous impact of core features across different samples (Fig. 4C). Specifically, elderly male patients with elevated LDL-C levels were associated with a higher risk of CHD (Fig. 4B).

At the individual level, representative low-risk and high-risk samples were selected for local SHAP analysis. For the low-risk sample, the prediction was primarily driven by lower LDL-C levels and younger age, where negative SHAP values collectively reduced the individual's predicted risk (Fig. 4G). Conversely, the high-risk sample was characterized by advanced age and elevated LDL-C levels, which generated significant positive SHAP values, leading the model to classify the individual as high-risk (Fig. 4).The sensitivity analysis indicated that, following relaxation of the feature selection threshold, the model’s discriminatory performance, calibration, and clinical utility remained largely stable, supporting the robustness of the primary results(Supplement1).

A, Global feature importance ranking;B, SHAP summary beeswarm plot;C,SHAP heatmap of feature contributions;D, Dependence plot for age;E,Dependence plot for sex;F,Dependence plot for LDL;G,Local SHAP explanation for low-risk sample;H,Local SHAP explanation for high-risk sample

4. Discussion

This study focused on the early screening of CHD in patients with T2DM. By leveraging a multi-center retrospective cohort, we developed and validated 15 machine learning models and utilized the SHAP method to elucidate the key risk factors. Our analysis identified 11 highly discriminative patient parameters essential for supervised machine learning. Following rigorous validation across training, testing, and external datasets, the Gated Recurrent Unit (GRU) model emerged as the superior tool,

demonstrating robust screening efficacy and generalizability. Key predictors identified by the model included age, sex, LDL-C, frequency of sweet food intake, and family history of diabetes. These findings suggest that elderly males with elevated LDL-C levels, a family history of diabetes, and a high intake of sweets are at a significantly higher risk for CHD.

The GRU model demonstrated superior performance by achieving the highest AUC in the test set and maintaining consistent discriminative power, robust sensitivity, and acceptable calibration within the external validation cohort. In contrast, certain traditional algorithms, such as LSSVM, exhibited pronounced classification bias, characterized by high sensitivity but exceptionally low specificity. These findings justify the selection of the GRU as the optimal model for this study, as it effectively balances predictive accuracy with clinical reliability across heterogeneous datasets.

While previous studies have explored CHD prediction, few have focused specifically on the T2DM population, and many lacked objective CHD validation or failed to account for population heterogeneity[17–19]. This study conducted standardized recruitment of patients from three hospitals and performed objective classification based on clinical expertise and authoritative guidelines. Notably, comparisons between the training and external validation cohorts revealed significant differences in key variables such as age, body habitus, blood pressure, lipid profiles, and lifestyle habits, indicating marked population heterogeneity. Notably, the AUC of the GRU model exhibited a marginal decline from 0.7625 in the test set to 0.7188 in the external validation set. This shift is primarily attributable to the use of a geographically and temporally independent validation cohort, which represents a population distinctly different from the training sites. Despite these inherent spatio-temporal discrepancies, the GRU model maintained consistent and reliable predictive performance. This stability underscores the model's robust clinical generalizability and its exceptional resilience to data drift, confirming its potential for reliable long-term implementation across diverse clinical settings.Compared to the model developed by Nabrdalik et al. for T2DM patients using the Silesia Diabetes-Heart Project data (which achieved an AUC of 0.72 with RUSBoost), our model demonstrates an improvement [20]. Furthermore, while Hyunji Sang achieved an AUC of 0.830 with a Random Forest model in the validation set, the AUC dropped significantly to 0.722 in the external validation set, suggesting potential overfitting and weak generalization in that study[21]. Additionally, the interpretability and clinical integration potential of our model benefit from its foundation in real-world, routinely collected data, ensuring high generalizability.

Based on the feature selection results, age emerged as the most significant predictor, consistent with previous literature. Aging is typically associated with a decline in pancreatic β-cell function, exacerbated insulin resistance, and an increased cumulative risk of atherosclerosis [22, 23]. Furthermore, males generally exhibit a higher risk for CHD, a disparity potentially linked to sex hormone profiles[24]. Classic metabolic and cardiovascular risk factors, including LDL-C, FBG, and SBP, also demonstrated high importance in the model. Notably, lifestyle variables such as sleep duration, sweet food intake, and occupational status were also selected. These findings suggest that CHD risk is intricately linked to circadian disruption, poor dietary patterns, and work-related stress, which may further elevate

cardiovascular risk by modulating systemic inflammation, metabolic homeostasis, and autonomic function[25–27].

SHAP analysis further bolstered the interpretability of our model. Global explanations revealed that age, sex, and LDL-C were the core variables driving model predictions, indicating that the screening primarily relies on demographic characteristics and lipid metabolism, consistent with prior research[28, 29]. Local interpretation results revealed significant heterogeneity in risk-driving factors across different samples. The key feature combinations relied upon by the model were not identical for patients with T2DM alone versus those with T2DM complicated by CHD. Notably, the direction of LDL in SHAP was not entirely consistent with traditional clinical perception in some individuals; specifically, LDL exhibited a negative contribution in certain samples. This does not imply that LDL has a protective effect; rather, it likely reflects treatment confounding within the retrospective real-world data—patients with CHD are more likely to receive lipid-lowering therapy, resulting in lower measured LDL levels despite their high baseline cardiovascular risk. The GRU model is capable of not only identifying global risk patterns but also providing differentiated explanations at the individual level. This is of significant importance for implementing precision screening and stratified interventions in clinical practice.

This study has several limitations. First, as a retrospective study, although multi-center data and external validation were utilized, inherent selection and information biases cannot be entirely eliminated. Second, several critical variables—such as medication history, inflammatory biomarkers, and imaging findings—were not incorporated into the model. This omission may affect predictive precision, particularly given that LDL levels are often confounded by lipid-lowering therapies. Third, this research relied primarily on structured clinical data. While the machine learning models demonstrated strong performance, these findings warrant further validation in prospective, large-scale, multi-center cohorts. Future research should aim to integrate laboratory biomarkers, imaging informatics, and longitudinal medication records to further enhance the model's stability and clinical utility.

Conclusion

Leveraging a multi-center real-world retrospective cohort, this study developed and validated an interpretable machine learning model for predicting coronary heart disease (CHD) in patients with type 2 diabetes mellitus (T2DM). Compared to other candidate models, the GRU model demonstrated superior performance in terms of discrimination, calibration, and net clinical benefit across both the test and external validation sets, making it a well-suited tool for early screening. SHAP analysis further elucidated the significant contributions of key predictors—such as age, sex, and LDL—thereby enhancing the model’s interpretability. This model has the potential to assist non-specialist clinicians in the early identification of high-risk patients, providing robust decision support for personalized interventions and precision prevention strategies.

Abbreviations

ACC
Accuracy / American College of Cardiology*
AGEs
Advanced glycation end-products
AHA
American Heart Association
AUC
Area Under the Curve
BiLSTM
Bidirectional Long Short-Term Memory
BP
Backpropagation (Neural Network)
CA
Coronary angiography
CAD
Coronary artery disease
CCTA
Coronary Computed Tomography Angiography
CHD
Coronary heart disease
CNN
Convolutional Neural Network
CVD
Cardiovascular disease
DBP
Diastolic blood pressure
DCA
Decision Curve Analysis
DT
Decision Tree
ELM
Extreme Learning Machine
ESC
European Society of Cardiology
F1
F1-score
FBG
Fasting blood glucose
GRU
Gated Recurrent Unit

HDL-C
High-density lipoprotein cholesterol
HR
Heart rate
IDF
International Diabetes Federation
IQR
Interquartile range
KNN
K-Nearest Neighbors
LASSO
Least Absolute Shrinkage and Selection Operator
LDL-C
Low-density lipoprotein cholesterol
LR
Logistic Regression
LSSVM
Least Squares Support Vector Machine
LSTM
Long Short-Term Memory
MCC
Matthews Correlation Coefficient
MICE
Multiple Imputation by Chained Equations
ML
Machine learning
MRI
Magnetic Resonance Imaging
NB
Naive Bayes
NPV
Negative predictive value
OGTT
Oral glucose tolerance test
PLS
Partial Least Squares
PPV
Positive predictive value
RBF
Radial Basis Function

RF
Random Forest
ROC
Receiver Operating Characteristic
SBP
Systolic blood pressure
SEN
Sensitivity
SHAP
SHapley Additive exPlanations
SPE
Specificity
SVM
Support Vector Machine
T2DM
Type 2 diabetes mellitus
TC
Total cholesterol
TG
Triglyceride
VIF
Variance Inflation Factor

Declarations

Acknowledgements

The authors acknowledge invaluable contributions of the participants and staff

Author Contributions

Yibo Lin was responsible for the study conception, formal analysis, visualization, and writing of the first draft of the manuscript. Yanhong Xie and Jianfeng Zheng were responsible for data collation and investigation. Zilong Wen and Junjie Tsang were responsible for data preparation, methodology, and model validation. Zhao Yu was responsible for methodology, project management, paper review and revision, and funding acquisition. Haibing Yu was responsible for methodology, project management, supervision, paper review and revision, and funding acquisition

Funding

the State Key Laboratory of Pathogenesis, Prevention, The State Key Laboratory of Pathogenesis, Prevention, Treatment of Central Asian High Incidence Diseases Fund (SKL-HIDCA-2024-GD7B)

Clinical & Basic Science Technology Innovation Project of Guangdong Medical University (GDMULCJC2025167,GDMULCJC2024142)

Guangdong Province Basic and Applied Basic Research Fund Provincial-Municipal Joint Fund Project (2024A1515140126, 2022A1515012407)

Dongguan Science and Technology of Social Development Program (20221800905642)

Innovation and Entrepreneurship Training Program of Guangdong Medical University in 2024 (202510571030)

2024 Guangdong Medical University Undergraduate Innovation and Entrepreneurship Education Base Project (JJDXM2024064, DXM2024069F)

2025 Innovation and Entrepreneurship Training Program for college students of Guangdong Medical University

Data availability

The dataset used in this study is subject to ethical and project restrictions and cannot be publicly shared.If there is a reasonable request and it has been approved by ethics, an application for access can be made to the corresponding author.

Ethics approval and consent to participate:

This study was reviewed and approved by the Medical Research Ethics Review Committee of Shenzhen Baoan District Central Hospital, which is affiliated to Shenzhen Baoan District Central Hospital. The research was conducted in accordance with the Measures for the Ethical Review of Life Sciences and Medical Research Involving Human Subjects (National Health Science and Education Document No. 4 of 2023), the Declaration of Helsinki, and the International Ethical Guidelines for Biomedical Research Involving Human Subjects. Since this was a multi-center retrospective study using anonymized and de-identified routine clinical data, the requirement for individual informed consent was waived by the ethics committee.

Consent for publication: Not applicable.

Competing interests: The authors declare no competing interests.

References

1. IDF Diabetes Atlas | Global Diabetes Data. & Statistics[EB/OL]. [2026-04-08]. https://diabetesatlas.org/
2. Type 2 diabetes. as an inflammatory disease | Nature Reviews Immunology[EB/OL]. [2026-04-08]. https://www.nature.com/articles/nri2925

3. FARUP J, JUST J, DE PAOLI F. Human skeletal muscle CD90 + fibro-adipogenic progenitors are associated with muscle degeneration in type 2 diabetic patients[J/OL]. Cell Metabol. 2021;33(11):2201–e221411. 10.1016/j.cmet.2021.10.001.
4. WANG W, QIAO J. ZHANG L, et,al. Prevalence of very high cardiovascular disease risk in patients with type 2 diabetes mellitus: A population-based cross-sectional screening study[J/OL]. [2026] [2026-04-08]. https://dom-pubs.onlinelibrary.wiley.com/doi/10.1111/dom.15763. DOI:10.1111/dom.15763.
5. HONG T, YAN Z. The Prevalence of Cardiovascular Disease in Adults with Type 2 Diabetes in China: Results from the Cross-Sectional CAPTURE Study[J/OL]. Diabetes Therapy. 2022;13(5):969. 10.1007/s13300-022-01243-x.
6. G K. Diabetic cardiovascular autonomic neuropathy: clinical implications.[J/OL]. Expert Rev Cardiovasc Ther. 2012;10(6):747–65. 10.1586/erc.12.53.
7. OEING C U, MATHESON M B, OSTOVANEH M R. Coronary artery disease grading by cardiac CT for predicting outcome in patients with stable angina[J/OL]. J Cardiovasc Comput Tomogr. 2023;17(5):310–7. 10.1016/j.jcct.2023.07.004.
8. OCHS MM, SIEPEN F A dem FRITZT. Limits of the possible: diagnostic image quality in coronary angiography with third-generation dual-source CT[J/OL]. Clin Res Cardiology: Official J German Cardiac Soc. 2017;106(7):485–92. 10.1007/s00392-017-1077-2.
9. RIZK J. 4D flow MRI applications in congenital heart disease[J/OL]. Eur Radiol. 2021;31(2):1160–74. 10.1007/s00330-020-07210-z.
10. HELM JM, SWIERGOSZ A M, HAEBERLE HS. Machine Learning and Artificial Intelligence: Definitions, Applications, and Future Directions[J/OL]. Curr Rev Musculoskelet Med. 2020;13(1):69. 10.1007/s12178-020-09600-8.
11. PATEL KOKORIE. Machine learning in predicting heart failure survival: a review of current models and future prospects[J/OL]. Heart Fail Rev. 2025;30(2):431–42. 10.1007/s10741-024-10474-y.
12. SCHLEICHER E, GERDES C. Definition, Classification and Diagnosis of Diabetes Mellitus[J/OL]. Exp Clin Endocrinol Diabetes. 2022;130(S 1):S1–8. 10.1055/a-1624-2897.
13. FISHER MA, BORGNAKKE W S, TAYLOR G W. Periodontal disease as a risk marker in coronary heart disease and chronic kidney disease[J/OL]. Curr Opin Nephrol Hypertens. 2010;19(6):519–26. 10.1097/MNH.0b013e32833eda38.
14. VALLÉE A, CINAUD A. Coronary heart disease diagnosis by artificial neural networks including aortic pulse wave velocity index and clinical parameters[J/OL]. J Hypertens. 2019;37(8):1682–8. 10.1097/HJH.0000000000002075.
15. KAMKAR I, GUPTA S K PHUNGD. Stabilizing l1-norm prediction models by supervised feature grouping[J/OL]. J Biomed Inform. 2016;59:149–68. 10.1016/j.jbi.2015.11.012.
16. DEGENHARDT F, SEIFERT S. Evaluation of variable selection methods for random forests and omics data sets[J/OL]. Brief Bioinform. 2019;20(2):492–503. 10.1093/bib/bbx124.

17. ALIZADEHSANI R, ABDAR M. Machine learning-based coronary artery disease diagnosis: A comprehensive review[J/OL]. Comput Biol Med. 2019;111:103346. 10.1016/j.compbiomed.2019.103346.
18. MIRJALILI S R SOLTANIS, HEIDARI MEYBODI Z. An innovative model for predicting coronary heart disease using triglyceride-glucose index: a machine learning-based cohort study[J/OL]. Cardiovasc Diabetol. 2023;22(1):200. 10.1186/s12933-023-01939-9.
19. SONG X, TONG Y, LUO Y. Predicting 7-day unplanned readmission in elderly patients with coronary heart disease using machine learning[J/OL]. Front Cardiovasc Med. 2023;10:1190038. 10.3389/fcvm.2023.1190038.
20. NABRDALIK K, KWIENDACZ H. Machine Learning Predicts Cardiovascular Events in Patients With Diabetes: The Silesia Diabetes-Heart Project[J/OL]. Curr Probl Cardiol. 2023;48(7):101694. 10.1016/j.cpcardiol.2023.101694.
21. SANG H, LEE H. Prediction model for cardiovascular disease in patients with diabetes using machine learning derived and validated in two independent Korean cohorts[J/OL]. Sci Rep. 2024;14(1):14966. 10.1038/s41598-024-63798-y.
22. LAKATTA E G, LEVY D. Arterial and cardiac aging: major shareholders in cardiovascular disease enterprises: Part II: the aging heart in health: links to heart disease[J/OL]. Circulation, 2003, 107(2): 346–54. 10.1161/01.cir.0000048893.62841.f7
23. CHANG A M, HALTER JB. Aging and insulin secretion[J/OL]. Am J Physiol Endocrinol Metab. 2003;284(1):E7–12. 10.1152/ajpendo.00366.2002.
24. VITALE C, MENDELSOHN M E, ROSANO G M. C. Gender differences in the cardiovascular effect of sex hormones[J/OL]. Nat Rev Cardiol. 2009;6(8):532–42. 10.1038/nrcardio.2009.105.
25. JOHNSON R K, APPEL L J BRANDSM. Dietary sugars intake and cardiovascular health: a scientific statement from the American Heart Association[J/OL]. Circulation. 2009;120(11):1011–20. 10.1161/CIRCULATIONAHA.109.192627.
26. JIKE M, ITANI O. Long sleep duration and health outcomes: A systematic review, meta-analysis and meta-regression[J/OL]. Sleep Med Rev. 2018;39:25–36. 10.1016/j.smrv.2017.06.011.
27. KELSALL H L, FERNANDO P H S, GWINI SM. Cardiovascular Disease and Type 2 Diabetes Risk Across Occupational Groups and Industry in a Statewide Study of an Australian Working Population[J/OL]. J Occup Environ Med. 2018;60(3):286–94. 10.1097/JOM.0000000000001228.
28. JI Y, SHANG H. Machine learning-based models to predict type 2 diabetes combined with coronary heart disease and feature analysis-based on interpretable SHAP[J/OL]. Acta Diabetol. 2025;62(10):1631–46. 10.1007/s00592-025-02496-1.
29. CHEN Y, WANG C, LIU X. Machine learning-based coronary heart disease diagnosis model for type 2 diabetes patients[J/OL]. Front Endocrinol. 2025;16:1550793. 10.3389/fendo.2025.1550793.

Figures

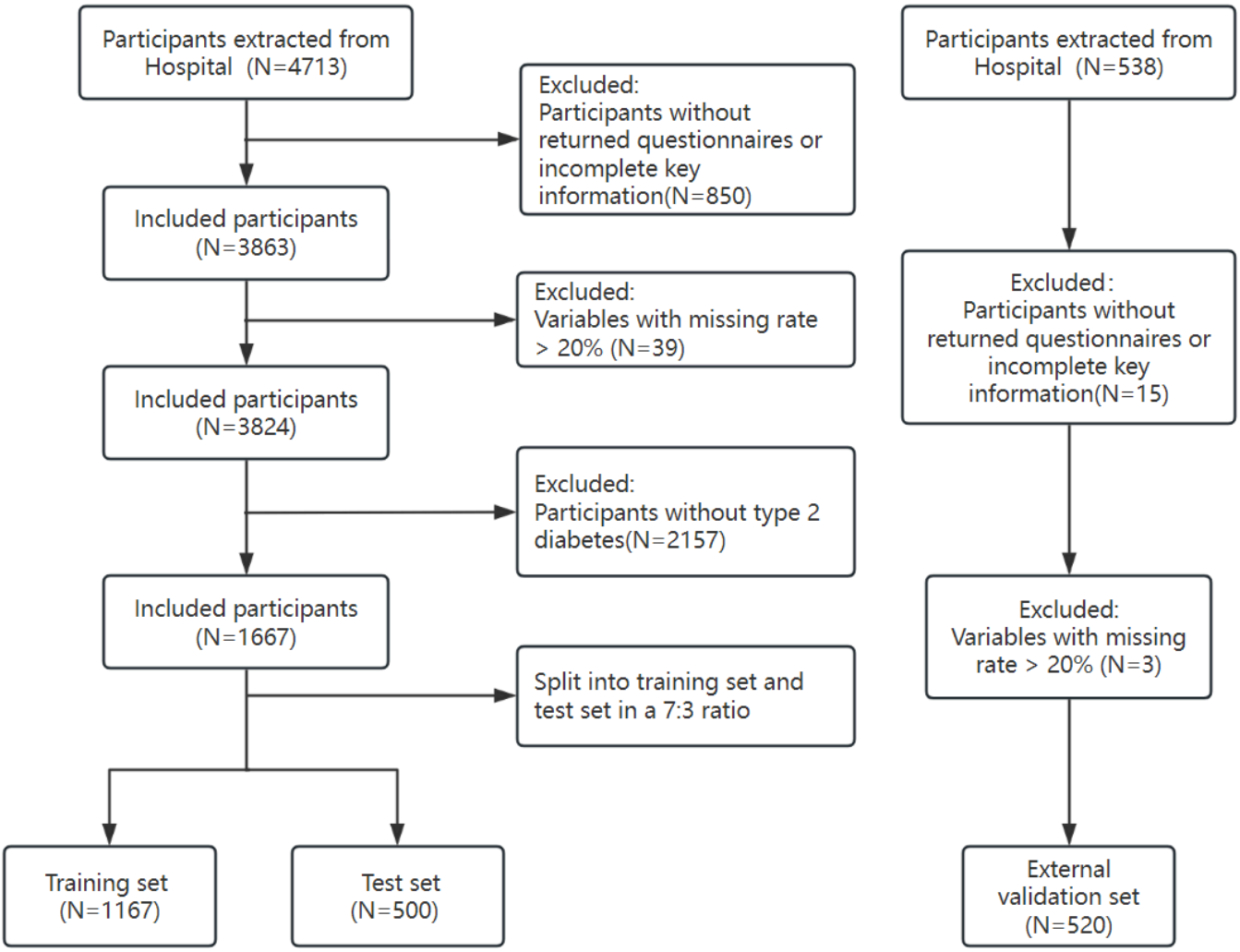


Figure 1

Flowchart of the research population selection process

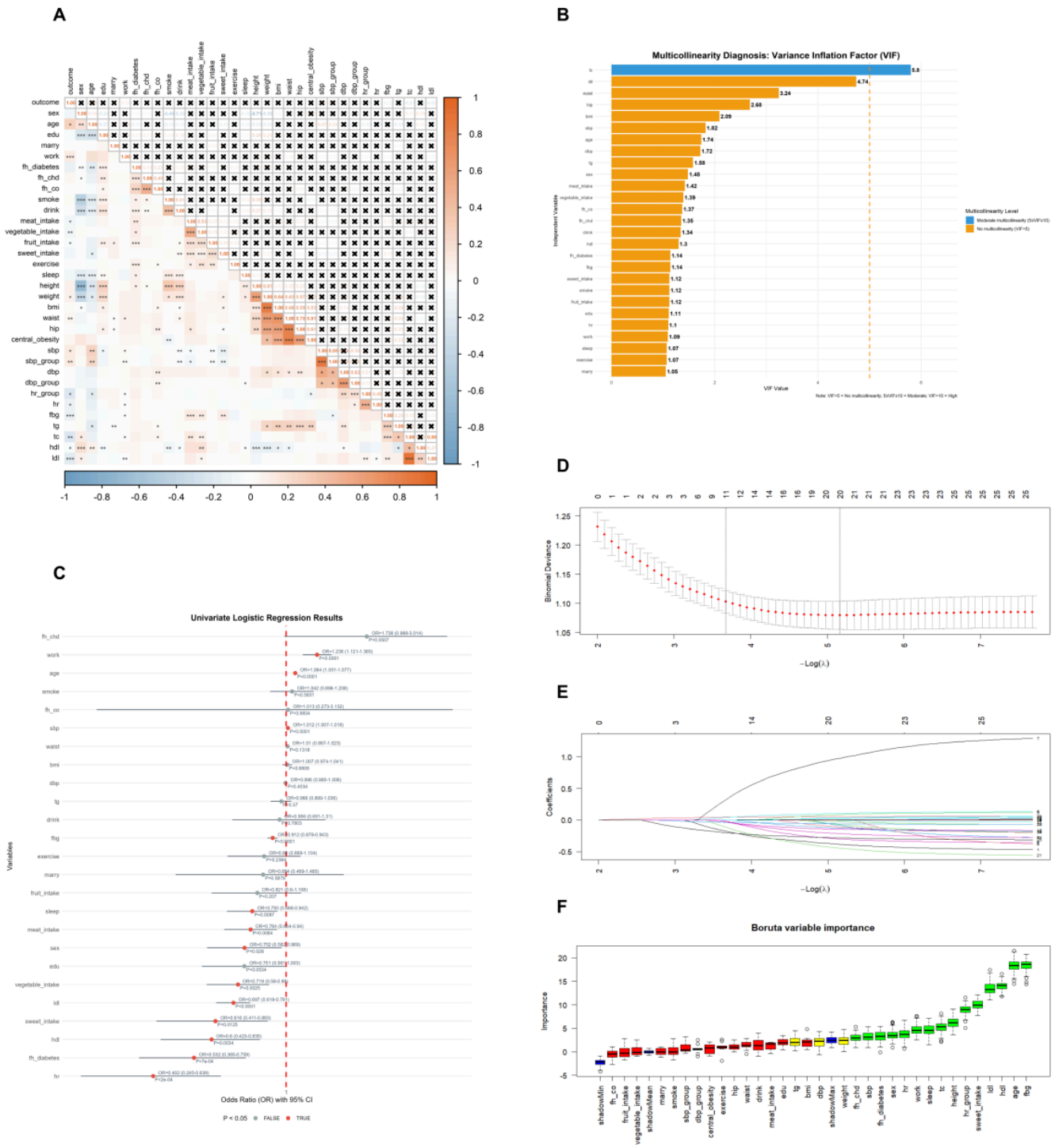


Figure 2

Variable preprocessing and feature selection

A, Correlation heatmap of all candidate variables;B, Variance inflation factor (VIF) analysis for multicollinearity diagnostics;C, Univariate logistic regression results for all candidate variables;D, LASSO regression coefficient profile plot;E, LASSO cross-validation curve;F, Boruta variable importance ranking

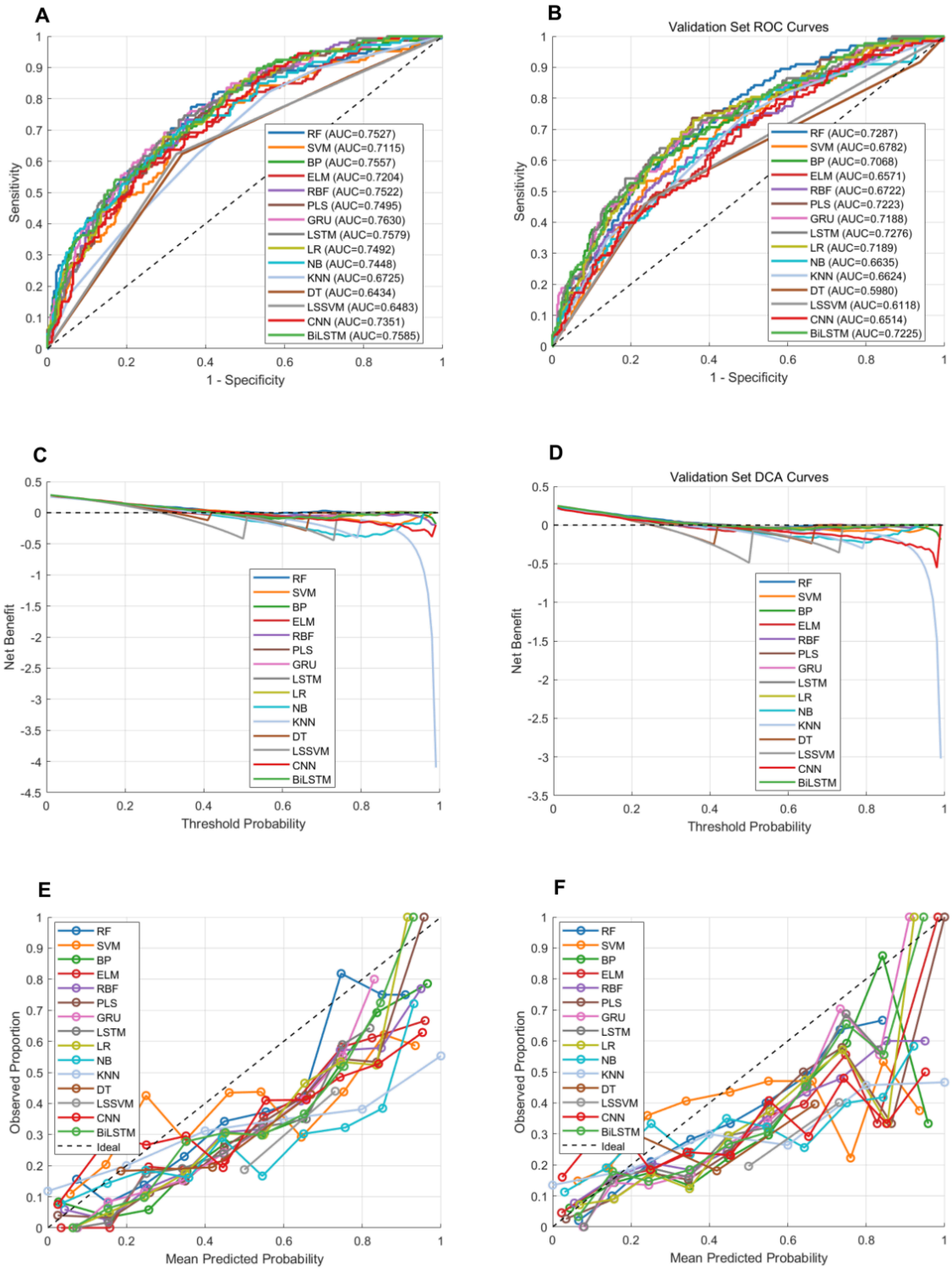


Figure 3

ROC curves, calibration curves, and decision curve analysis (DCA) of all models

A, Test set ROC comparison;B, External validation set ROC comparison;C, Test set DCA comparison ;D, External validation set DCA comparison;E,Test set Calibration Curve;F,External validation Calibration Curve

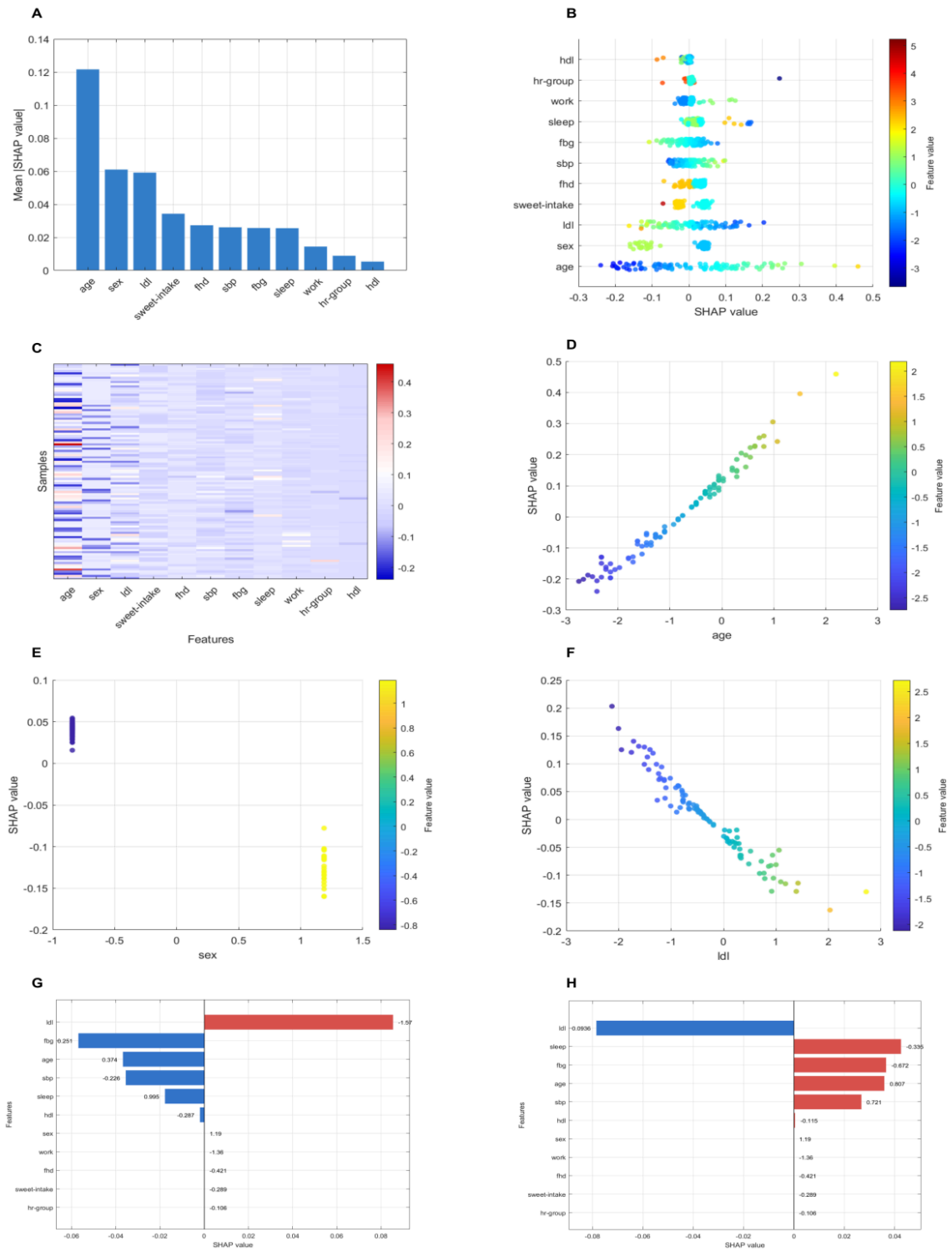


Figure 4

SHAP analysis for model interpretation

A, Global feature importance ranking;B, SHAP summary beeswarm plot;C,SHAP heatmap of feature contributions;D, Dependence plot for age;E,Dependence plot for sex;F,Dependence plot for LDL;G,Local SHAP explanation for low-risk sample;H,Local SHAP explanation for high-risk sample

Supplementary Files

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

- Supplementary1.docx