scientific reports

Check for updates

OPEN

Application of interpretable machine learning algorithms to predict macroangiopathy risk in Chinese patients with type 2 diabetes mellitus

Ningjie Zhang[1], Yan Wang[2], Hui Zhang[3,4,5], Huilong Fang[6], Xinyi Li[6], Zhifen Li[6], Zhenghang Huan[6], Zugui Zhang[7], Yongjun Wang[1], Wei Li[2]✉ & Zheng Gong[8,9]✉

Macrovascular complications are leading causes of morbidity and mortality in patients with type 2 diabetes mellitus (T2DM), yet early diagnosis of cardiovascular disease (CVD) in this population remains clinically challenging. This study aims to develop a machine learning model that can accurately predict diabetic macroangiopathy in Chinese patients. A retrospective cross-sectional analytical study was conducted on 1566 hospitalized patients with T2DM. Feature selection was performed using recursive feature elimination (RFE) within the mlr3 framework. Model performance was benchmarked using 29 machine learning (ML) models, with the ranger model selected for its superior performance. Hyperparameters were optimized through grid search and 5-fold cross-validation. Model interpretability was enhanced using SHAP values and PDPs. An external validation set of 106 patients was used to test the model. Key predictive variables identified included the duration of T2DM, age, fibrinogen, and serum urea nitrogen. The predictive model for macroangiopathy was established and showed good discrimination performance with an accuracy of 0.716 and an AUC of 0.777 in the training set. Validation on the external dataset confirmed its robustness with an AUC of 0.745. This study establish an approach based on machine learning algorithm in features selection and the development of prediction tools for diabetic macroangiopathy.

Keywords T2DM, Macroangiopathy, Machine learning methods, Prediction model, Risk factor

With improving living standards and evolving lifestyle, diabetes has emerged as a rapidly growing global health crisis among adult populations. According to report from the International Diabetes Federation (IDF), approximately 783 million adults worldwide - representing 1 in 8 individuals - are anticipated to be affected by diabetes by 2045[1]. Notably, Type 2 diabetes mellitus (T2DM) constitutes over 90% of diabetes cases, with its pathogenesis involving complex interactions among demographic, environmental, socioeconomic and genetic factors[1]. Recent epidemiological studies reveal particularly concerning trends in China. A nationwide investigation published in JAMA demonstrated a significant surge in diabetes prevalence, rising from 10.9 to 12.4% between 2013 and 2018[2]. This alarming progression underscores the urgent need for effective public health strategies. Current evidence suggests that implementing early diagnostic protocols combined with personalized interventions can significantly enhance disease management outcomes.

[1]Department of Blood Transfusion, The Second Xiangya Hospital, Central South University, Changsha, China. [2]Department of Rheumatology, The First Affiliated Hospital of Zhengzhou University, NO. 1, Jianshe East Road, Zhengzhou 450052, Henan, China. [3]'The 14th Five-Year Plan' Application Characteristic Discipline of Hunan Province (Clinical Medicine), Changsha, China. [4]Hunan Provincial Key Laboratory of the Traditional Chinese Medicine Agricultural Biogenomics, Changsha, China. [5]Department of Basic Medical Sciences, Changsha Medical University, Changsha, China. [6]School of Basic Medical Sciences, Xiangnan University, Chenzhou 423000, China. [7]Institute for Research on Equity and Community Health, Christiana Care Health System, Newark, USA. [8]Sino-Cellbiomed Institutes of Medical Cell & Pharmaceutical Proteins Qingdao University, Qingdao, Shandong, China. [9]Department of Basic Medicine, Xiangnan University, 889 Chenzhou Avenue, Chenzhou, Hunan, China. ✉email: libuwei2011@163.com; xblong2000@gmail.com

Macroangiopathy constitutes a leading contributor to mortality in the T2DM populations, responsible for 70–80% of diabetes-related deaths[3]. Macrovascular complications in T2DM include coronary artery disease, peripheral arterial disease, and cerebrovascular disease, which not only shorten the lifespan but also severely impair their quality of life. Early prediction and intervention are crucial for controlling these complications and reducing mortality rates[4,5]. However, the slow progression and subtle presentation of the vascular complications often lead to missed opportunities for early intervention. Although carotid Intima-media thickness(cIMT) measurement via ultrasonography serves as a validated noninvasive method for evaluating macroangiopathy, it requires advanced training, specialized equipment, and high-cost instrumentation. This necessitates the development of simple, cost-effective screening and treat approaches.

Recent advancements in artificial intelligence (AI), particularly machine learning (ML) methodologies, have revolutionized predictive analytics in healthcare. These technologies enable the construction of sophisticated models capable of deciphering complex datasets and detecting nonlinear patterns that elude conventional statistical approaches[6]. Despite advancements in the complication risk prediction models, the impact of epidemiological risk factors and biomarkers exhibits variability across diverse populations and races, highlighting the need for population-specific prediction models for diabetic macroangiopathy control. The epidemiological landscape of Central China presents unique challenges in this domain. Despite the high disease burden of T2DM macrovascular complications, risk prediction models remains unexplored. Current models predominantly employ traditional multivariable regression techniques, display only modest predictive performance. Emerging ML algorithms like random forest (RF) offer potential to risk predictions by exploring large dataset and uncover novel risk predictors. Among the various ML frameworks, mlr3 has gained prominence due to its robust features and flexibility. Mlr3 is an R package that provides a comprehensive tools for ML, including data preprocessing, model training, hyperparameter tuning, and performance evaluation. Its modular design allows seamlessly integrate of different components to meet the specific research needs, and it supports a wide range of ML algorithms, offering advanced functionalities for handling imbalanced datasets, a common challenge in medical research[7].

Existing prediction models for diabetic macroangiopathy often rely on traditional statistical methods, which have limited predictive power. Our study leverages machine learning and interpretable AI methods to develop a robust, non-invasive risk prediction model. Unlike previous studies, this is one of the first to use mlr3 and external validation in a Chinese T2DM population. This study aimed to establish a rapid diagnostic model base on questionnaire and laboratory parameters to screen high-risk patients with diabetic macroangiopathy.

Methods

Study design and population

A cross-sectional cohort of 2328 participants, aged over 18 and diagnosed with T2DM, was recruited from the First Affiliated Hospital of Zhengzhou University between 2018 and 2020. T2DM was defined by hemoglobin A1c (HbA1c) level ≥ 6.5% or fasting blood glucose (FBG) level ≥ 7.0 mmol/L[8]. Participants were divided into the macroangiopathy (MA) group and the non-macroangiopathy (NM) group. Exclusion criteria for the MA group included:①cerebrovascular disease: cerebral infarction, transient ischemic attack, cerebral hemorrhage, subarachnoid hemorrhage, and other cerebrovascular accidents diagnosed by medical history, physical examination, cranial CT or MRI; ②Coronary atherosclerotic heart disease: diagnosis of coronary artery disease confirmed by ambulatory electrocardiogram, previous history of angina pectoris, coronary artery disease or myocardial infarction; ③peripheral vascular disease: atherosclerosis, plaques formation, or stenosis of the carotid arteries and/or both lower extremities; diabetic foot or presence of clinical signs of lower extremity ischemia (e.g., intermittent claudication, pain, gangrene, etc.). Exclusion criteria for the MA group included: acute infectious disease, uncontrolled hypertension, severe dyslipidemia, autoimmune disease, malignant tumor, evident microangiopathy included diabetic retinopathy or nephropathy, or any condition requiring immediate treatment. Patients with missing laboratory results and questionnaire data were excluded. Ultimately, 1566 T2DM patients were included. The study adhered to the Declaration of Helsinki and received ethics approval from the First Affiliated Hospital of Zhengzhou University Ethics Committee.

Data collection

To collect comprehensive information about participants, a series of questionnaires was employed to document their medical history, family disease predispositions, cardiovascular disease (CVD) history, and lifestyle factors. Smoking habits were classified as never smoking, current smoker, or previous smoker. Individuals who had smoked at least 100 cigarettes in their lifetime and presently smoking were defined as current smokers. Alcohol intake was categorized as never drinking, current drinker or previous drinker. Alcohol drinking was defined as the consumption of alcohol ≥ 18 g in the past month. Anthropometric measurements, including weight, height, and blood pressure, were recorded; and the body mass index (BMI) [weight/height squared] was calculated. Diastolic blood pressure (DBP) and systolic blood pressure (SBP) were determined utilizing an automatic blood pressure meter after 10 min of seating, with the average of three consecutive measurements recorded.

Data pre-processing

We removed variables with more than 80% missing values. The remaining missing data were imputed using the "imputehist" and "regr.kknn" function in the mlr3 piplines. Skewed distributed variables underwent log2 transformation to mitigate the impact of skewness, ensuring a more normalized distribution. All numerical-type variables were normalized by Z score transformation using the "scale" function. To avoid sample imbalance, various methods (undersampling, oversampling, and SMOTE) were tested, with undersampling (positive to negative sample ratio 1:1) showing the highest F-beta Score (0.5873).

Feature selection

Figure 1 shows the variables selection and model construction flowchart. Feature and model benchmark selection were performed using mlr3. We applied Recursive feature elimination (RFE) using seven ML methods (XGBoost-RFE, SVM-RFE, gbm-RFE, Catboost-RFE, Ranger-RFE, LightGBM-RFE, and rpart-RFE) under five-fold cross-validation. The feature importance rankings were based on AUC performance, and the top-ranked variables were selected from the intersection of the three highest-performing RFE models (gbm, SVM, Ranger). Correlation analysis and variance inflation factor (VIF) coefficients were used to assess multicollinearity, excluding features with VIF > 10. The final selected features—duration of T2DM, age, fibrinogen, and BUN—were consistent across methods and passed collinearity checks.

Model benchmarking and selection

We benchmarked 29 ML classifiers in the mlr3 framework, including Random Forest (Ranger), XGBoost, CatBoost, LightGBM, SVM, and others. Models were evaluated using 5-fold cross-validation. The key performance metrics included: AUC (area under the ROC curve)-primary indicator of model discrimination, Accuracy (ACC)-correct classification rate, Cross-entropy loss (CE)-model calibration, Sensitivity & specificity-classification quality. The ranger ML model demonstrated the best performance with the highest median AUC value.

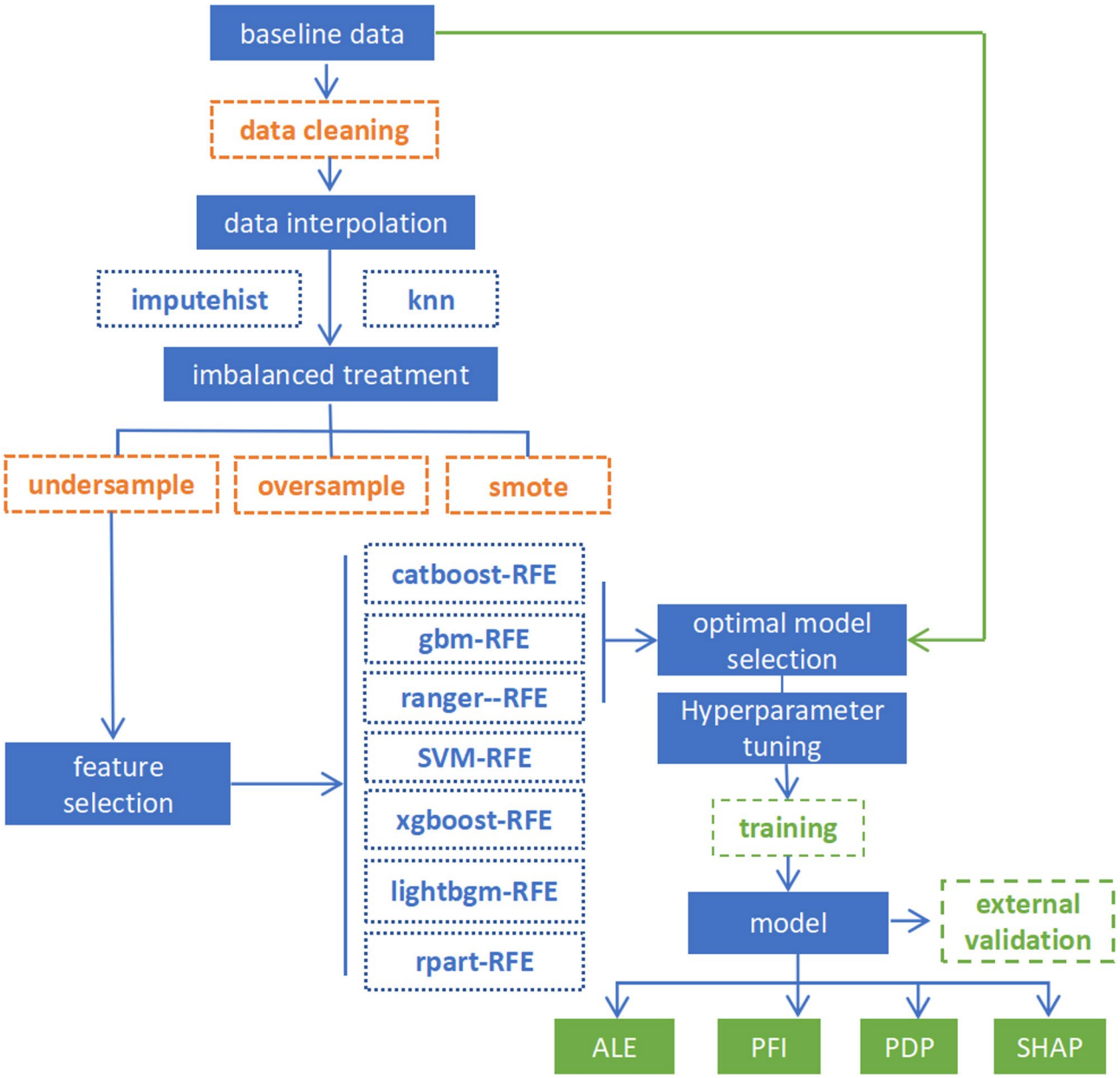


Fig. 1. Workflow of the ML model development and validation process.

Model development, hyperparameter tuning, and external validation

All patient samples formed the training set, and an external validation set ($n = 106$) was used to test the model. We used a grid search and 5-fold cross-validation strategy over 1,000 iterations to optimize the ranger classification model hyperparameters, including: "num.trees" (1–2,000), "sample.fraction" (0.1-1), "mtry.ratio" (0–1), "min.node.size" (1-100), and "num.random.splits" (1-100). Tuning was done using the AutoTuner class in mlr3tuning, with AUC as the optimization target. The best parameter set was: num.trees = 1112, sample.fraction = 0.1, mtry.ratio = 0.556, min.node.size = 34, num.random.splits = 1. The external validation set, consisting of 106 samples collected independently from the training set was utilized to evaluate the model's performance. CE is used as a performance metric for classification models, particularly in medical applications where imbalanced datasets are common. A lower CE value indicates a better-fitted model with more accurate probability predictions. In our study, CE was used alongside AUC to assess model performance, ensuring both discrimination ability and calibration were optimized.

Model interpretation techniques

To enhance interpretability, we used: The Shapley Additive Explanation (SHAP) values for global and local feature impact and Partial Dependence Plots (PDP) to visualize interaction effects. The SHAP values, calculated using the iml R package and visualized with the SHAP Python library, elucidated feature contribution to predictions. The PDPs generated using the PDPbox Python library, visualized interaction effects between key features. Accumulated local effects (ALE) plots to understand nonlinear dependencies and permutation feature importance (PFI) scores to assess feature robustness, which further interpret the model's behavior and the importance of each feature.

Statistical analysis

The Statistical information that consisted of normally distributed continues variables is represented as means along with standard deviation (SD), while non-normally distributed continuous variables were represented by medians and interquartile ranges (IQR). The categorical and qualitative variables were reported as numbers with percentages (%). Statistical differences were assessed via one-way ANOVA for normally distributed continuous variables. For non-normally distributed variables, we used the Mann-Whitney U test. For categorical variables, we used the chi-square test. All p-values were two-tailed, with significance set at $p < 0.05$. All statistical analyses were performed in R (ver 3.3.1).

Results

The study populations characteristics

The present study population was composed of 1566 participants, with a mean age of 48.9 years, including 1063 males (67.9%) and 503 females (32.1%). Participant baseline characteristics are described in Table 1. There are statistically significant differences in several key variables between the macroangiopathy (MA) group and the no macroangiopathy (NM) group. The MA group was older (mean age 52.71 vs. 42.31 years, $p < 0.001$), with a lower BMI (25.60 vs. 26.85 kg/m^2, $p < 0.001$), and higher systolic blood pressure (134.49 vs. 131.76 mmHg, $p = 0.002$). The duration of T2DM was significantly longer in the MA group (median 6.58 vs. 1.00 years, $p < 0.001$). Comorbid conditions such as hypertension, cardiac dysfunction, and cerebrovascular disease all higher in the MA group compared with the non-MA group ($p < 0.001$ for all). There was no significant difference in diastolic blood pressure, smoking status, or alcohol intake.

Pronounced differences in laboratory indicators between the MA and NM groups included lower triglycerides (1.59 vs. 1.76 mmol/L, $p = 0.037$) and higher HDL-C (1.10 vs. 1.05 mmol/L, $p = 0.005$) in the MA group. Blood glucose indicators showed higher OGTT 120 values in the MA group (14.60 vs. 13.88 mmol/L, $p < 0.001$). Renal function indicators showed higher creatinine (67.34 vs. 64.33 μmol/L, $p = 0.010$) and lower GFR (101.38 vs. 110.99 ml/min, $p < 0.001$) in the MA group. Liver function indicators demonstrated lower ALT (18.00 vs. 21.00 U/L, $p < 0.001$) and lower AST (17.00 vs. 19.00 U/L, $p = 0.022$) in the MA group. Additionally, the MA group had higher fibrinogen (3.17 vs. 2.88 mg/L, $p = 0.029$) and FT3 levels (4.89 vs. 4.85 pmol/L, $p = 0.017$). These data indicate that there are significant demographic and clinical differences between the T2DM patients with and without macroangiopathy.

Data imbalance

We observed that our dataset had an imbalance problem with far less in without macroangiopathy samples than macroangiopathy samples. It would severely affect the prediction performance of ML model. To address this issue, we compare the different approaches to handle imbalanced data sets, including: undersample, oversample and SMOTE. AutoTuner class were created from the learner using the mlr3, mlr3pipelines and mlr3tuning R package to tune the graph based on 3-fold cross-validation using F-beta Score as a performance metric (random forest learner + imbalance correction method). As seen in Figure S1 the undersample method exhibits the best effect with higher F-beta Score (F-beta = 0.5873). Therefore, the undersample strategy was further adopted to balance the dataset, and the data imbalance is resolved by undersampling so that the number of samples in both the MA and NM groups was equalized to $n = 584$.

The risk factors and features selection by ML method

We used a recursive feature elimination (RFE) method for feature selection (with a five-fold cross-validation scheme). Seven distinct machine learning methods, namely XGBoost-RFE, SVM-RFE, bgm-RFE, Catboost-RFE, Ranger-RFE, Lightbgm-RFE and rpart-RFE, were applied to identify potential key characters from a plethoral of macroangiopathy features. Following the initial screening of variables, univariate ROC curve and AUC value were utilized to evaluate the predictive abilities of different variable combinations. The univariable ROC analysis

Characteristics	Overall ($n=1566$)	MA group ($n=982$)	NM group ($n=584$)	p-value
Gender, n(%), male	1063 (67.9)	666 (67.8)	397 (68)	0.948
Age, mean (SD), years	48.90 (12.04)	52.71 (9.85)	42.31 (12.68)	< 0.001*
BMI, mean (SD), kg/m^2	26.06 (4.19)	25.60 (3.54)	26.85 (5.00)	< 0.001*
SBP, mean (SD), mmHg	133.47 (16.89)	134.49 (16.91)	131.76 (16.74)	0.002*
DBP, mean (SD), mmHg	83.71 (22.26)	83.85 (26.60)	83.49 (11.75)	0.759
Duration of T2DM, n (%), years	4.0 (0, 31.00)	6.58 (0, 31.00)	1.00 (0, 30.00)	< 0.001*
Comorbidities				
Hypertension, n (%)	675 (43.1)	474 (48.3)	201 (34.4)	< 0.001*
Cardiac function, n (%)	145 (9.31)	118 (12.0)	27 (4.6)	< 0.001*
Cerebrovascular disease, n (%)	117 (7.5)	105 (10.7%)	12 (2.1%)	< 0.001*
Smoking status, n (%)	410 (26.2)	265 (27.0%)	145 (24.8%)	0.348
Alcohol status, n (%)	322 (20.6)	213 (21.7%)	109 (18.7%)	0.152
Laboratory index				
TC, mean (SD), mmol/L	4.97 (16.8)	5.18 (21.21)	4.62 (1.18)	0.525
TG, median [IQR], mmol/L	1.65 (0.18, 24.06)	1.59 (0.18, 23.27)	1.76 (0.22, 24.06)	0.037*
HDL-C, mean (SD), mmol/L	1.08 (0.33)	1.10 (0.33)	1.05 (0.32)	0.005*
LDL-C, mean (SD), mmol/L	2.70 (0.97)	2.67 (1.00)	2.74 (0.90)	0.171
FBG, median [IQR], mmol/L	7.71 (3.61, 31.40)	7.70 (3.79, 31.40)	7.73 (3.61, 29.55)	0.767
HbA1c, mean (SD)	9.08 (2.28)	9.03 (2.08)	9.16 (2.57)	0.297
OGTT 0, median [IQR], mmol/L	7.00 (1.91, 88.50)	7.20 (1.91, 88.50)	6.80 (4.10, 77.00)	0.107
OGTT 30, mean (SD), mmol/L	9.39 (3.37)	9.51 (3.79)	9.20 (2.51)	0.075
OGTT 60, mean (SD), mmol/L	12.47 (4.53)	12.55 (3.13)	12.35 (6.20)	0.395
OGTT 120, mean (SD), mmol/L	14.34 (3.69)	14.60 (3.47)	13.88 (3.98)	< 0.001*
25(OH)D, median [IQR], pmol/L	16.91 (3.00, 321.00)	17.03 (3.00, 49.83)	16.87 (3.00, 321.00)	0.353
Fibrinogen, mean (SD), mg/L	3.06 (2.45)	3.17 (3.06)	2.88 (0.63)	0.029*
FT3, median [IQR], pmol/L	4.75 (2.21, 16.01)	4.89 (2.21, 15.68)	4.85 (2.91, 16.01)	0.017*
FT4, mean (SD), pmol/L	11.46 (2.37)	11.41 (2.49)	11.54 (2.16)	0.333
TSH, median [IQR], mIU/L	1.92 (0.01, 1037.00)	1.90 (0.01, 1037.00)	1.94 (0.04, 18.57)	0.451
PTH, mean (SD), pg/mL	36.38 (16.44)	35.95 (16.70)	37.15 (15.94)	0.194
BUN, mean (SD), mmol/L	6.35 (18.89)	7.03 (23.82)	5.21 (1.51)	0.065
Cr, mean (SD), μmol/L	66.21 (22.33)	67.34 (24.79)	64.33 (17.30)	0.010*
UA, mean (SD), μmol/L	302.86 (95.67)	294.84 (85.45)	316.28 (109.44)	< 0.001*
GFR, mean (SD), ml/min	104.94 (18.08)	101.38 (17.66)	110.99 (17.30)	< 0.001*
ALT, median [IQR], U/L	19.00 (1.00, 657.00)	18.00 (1.00, 657.00)	21.00 (3.00, 185.00)	< 0.001*
AST, median [IQR], U/L	18.00 (3.00, 584.00)	17.00 (3.00, 584.00)	19.00 (4.00, 163.00)	0.022*
GGT, median [IQR], U/L	23.00 (4.00, 542.00)	23.00 (5.00, 542.00)	25.00 (4.00, 328.00)	0.548
ALP, mean (SD), U/L	75.22 (24.00)	75.27 (23.19)	75.12 (25.34)	0.902
TP, mean (SD), g/L	68.08 (6.44)	67.84 (6.50)	68.47 (6.32)	0.063*
ALB, mean (SD)B, g/L	43.31 (4.39)	42.99 (4.38)	43.86 (4.35)	< 0.001*
TBIL, median [IQR], μmol/L	9.60 (1.53, 799.00)	9.40 (1.53, 799.00)	25.00 (4.00, 328.00)	0.899
DBIL, mean (SD), μmol/L	4.25 (1.93)	4.19 (1.87)	4.34 (2.04)	0.133

Table 1. Baseline demographic characteristics of the study population. * means significantly different between the MA group and NM group. *BMI* body mass index, *SBP* systolic blood pressure, *DBP* diastolic blood pressure, *TC* total cholesterol, *TG* triglyceride, *HDL-C* high-density lipoprotein cholesterol, *LDL-C* low-density lipoprotein cholesterol, *FBG* fasting blood glucose, *HbA1c* glycosylated hemoglobin c, *OGTT* oral glucose tolerance test, *CRP* C-reactive protein, *TT3* total triiodothyronine, *FT4* free thyroxine, *TSH* thyroid stimulating hormone, *PTH* parathyroid hormone, *BUN* blood urea nitrogen, *Cr* creatinine, *UA* uric Acid, *GFR* glomeruar filtration rate, *ALT* alanine aminotransferase, *AST* aspartate aminotransferase, *GGT* gamma-glutamyl transpeptidase, *ALP* alkaline phosphatase, *TP* total protein, *ALB* albumin, *TBIL* total bilirubin, *DBIL* direct bilirubin.

indicated the number of features at which the AUC values peaked. The crucial evaluation parameters, including of ce and the best AUC value, were summarize in Table 2. The best variables for discriminating macroangiopathy were determined by intersecting the results of gbm-RFE, Ranger-RFE and SVM-RFE, each showcasing the top 3 AUC values (Fig. 2A-C). A Venn diagram (Fig. 2D) illustrated the overlapping key features, identifying four crucial variables: duration of T2DM, age, fibrinogen and serum urea nitrogen. Feature importance value after

Model	CE	Best variable number	AUC
classif.catboost	0.300	50	0.764
classif.lightgbm	0.347	12	0.757
classif.xgboost	0.369	3	0.743
classif.rpart	0.312	14	0.730
classif.gbm	0.303	8	0.781
classif.ranger	0.328	23	0.769
classif.svm	0.294	13	0.791

Table 2. Results of the seven ML models for feature selection. CE (cross-entropy loss) measures classification model performance, with lower values indicating better fit. AUC (Area Under the Curve) represents the model's ability to distinguish between macroangiopathy and non-macroangiopathy cases.

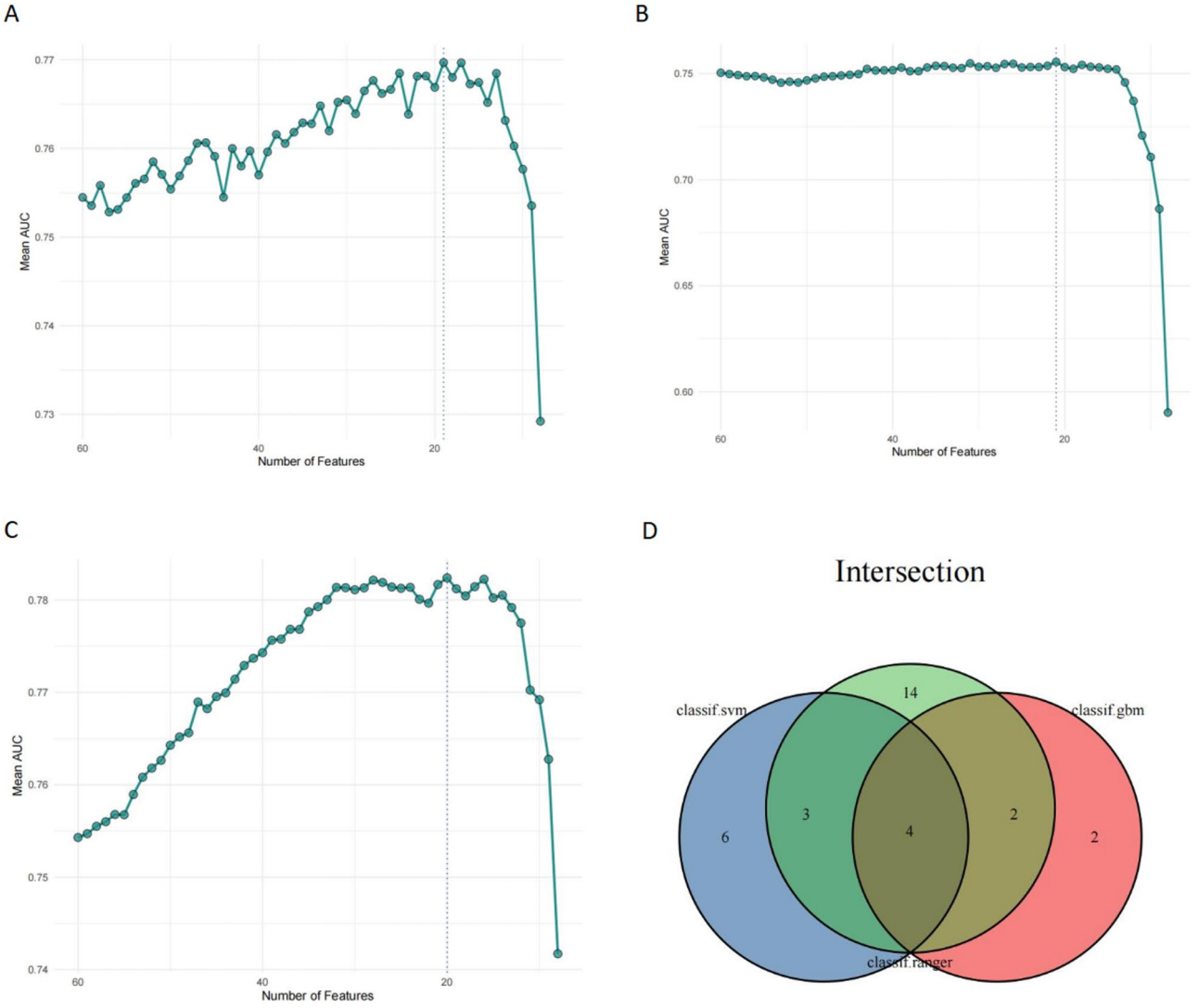


Fig. 2. Screening key variables using different RFE machine learning methods. (**A**) The number of features screened by gbm-RFE. (**B**) The number of features screened by ranger-RFE. (**C**) The number of features screened by SVM-RFE. (**D**) Venn diagram showing the overlap among the variables selected by the gbm-RFE, ranger-RFE and SVM-RFE. The intersection highlights four key variables consistently identified across all three models: duration of T2DM, age, fibrinogen, and serum urea nitrogen. These common variables were used to construct the final predictive model.

performing feature selection were presented in Fig. 3. Correlation analysis were conducted both between the all features and among the final selected variables (Figure S2 and Fig. 3). Furthermore, the variance inflation factor (VIF) coefficients were used to diagnose multicollinearity, revealing that none of the final selected variables exhibited VIF indicators of multicollinearity (all VIF < 10) (Table S1).

Screening the optimal benchmark ML model

A comparison of 29 different ML classification models was performed, encompassing models such as abess, Adaboost, catboost, glmnet, et al. Internal cross-validation using five-fold cross-validation was conducted. As described in Table S2, the ACC, AUC, CE, sensitivity and specificity of the 29 models were 0.500 ~ 0.765, 0.292 ~ 0.527, 0.400 ~ 0.659, and 0.600 ~ 0.847, respectively. While most models showed comparable performance, the ranger ML model exhibited the best discrimination performance with the maximum median AUC value (Fig. 4A). Consequently, the Ranger Classification ML model was selected for further analysis.

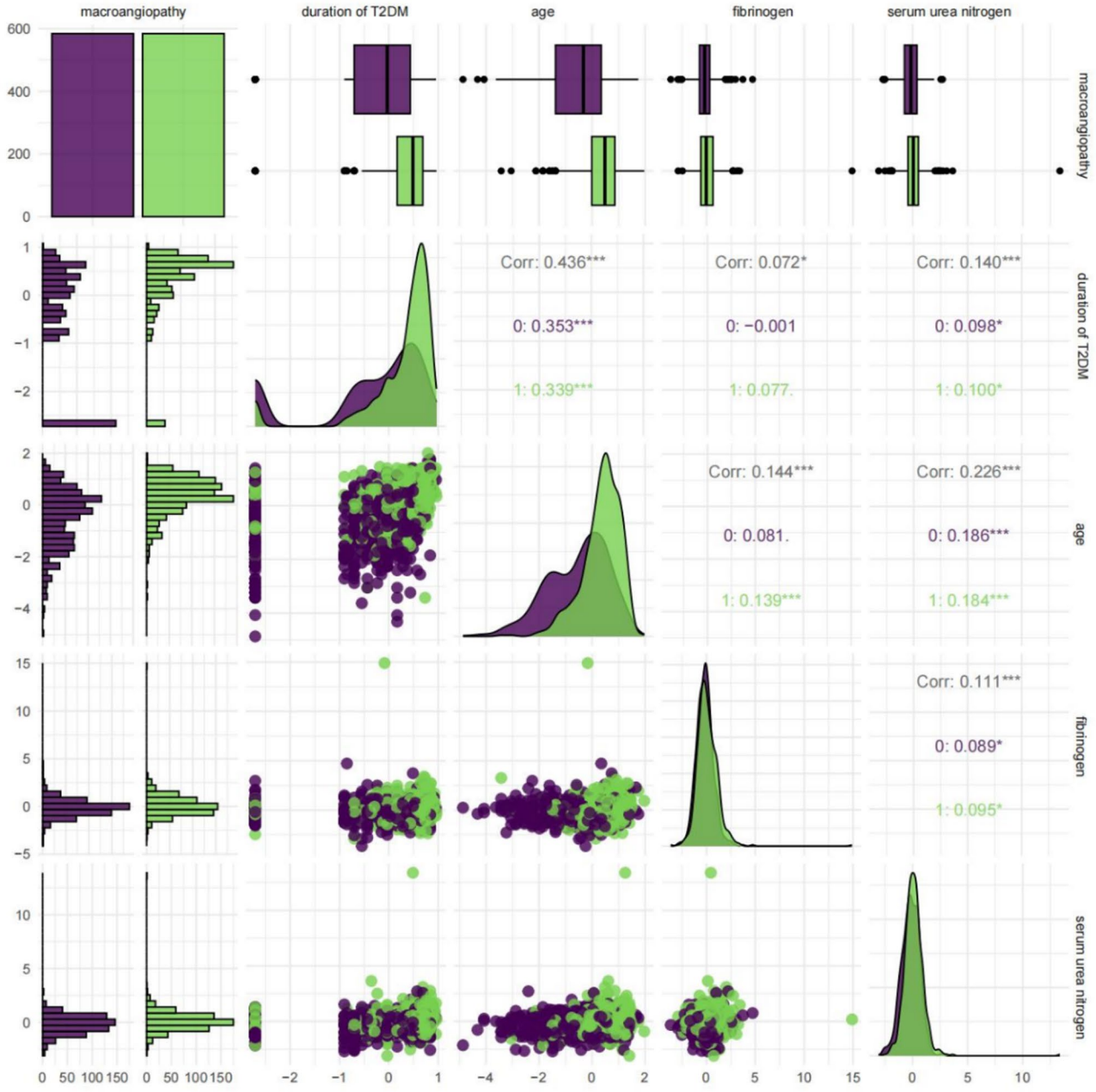


Fig. 3. Illustrates the pairwise correlation analysis of key variables identified for macroangiopathy prediction. The color purple and the values '0' indicating the absence of macroangiopathy; the color green and the values '1' indicating presence of macroangiopathy. * means p-value < 0.05, ** means p-value < 0.01, *** means p-value < 0.001.

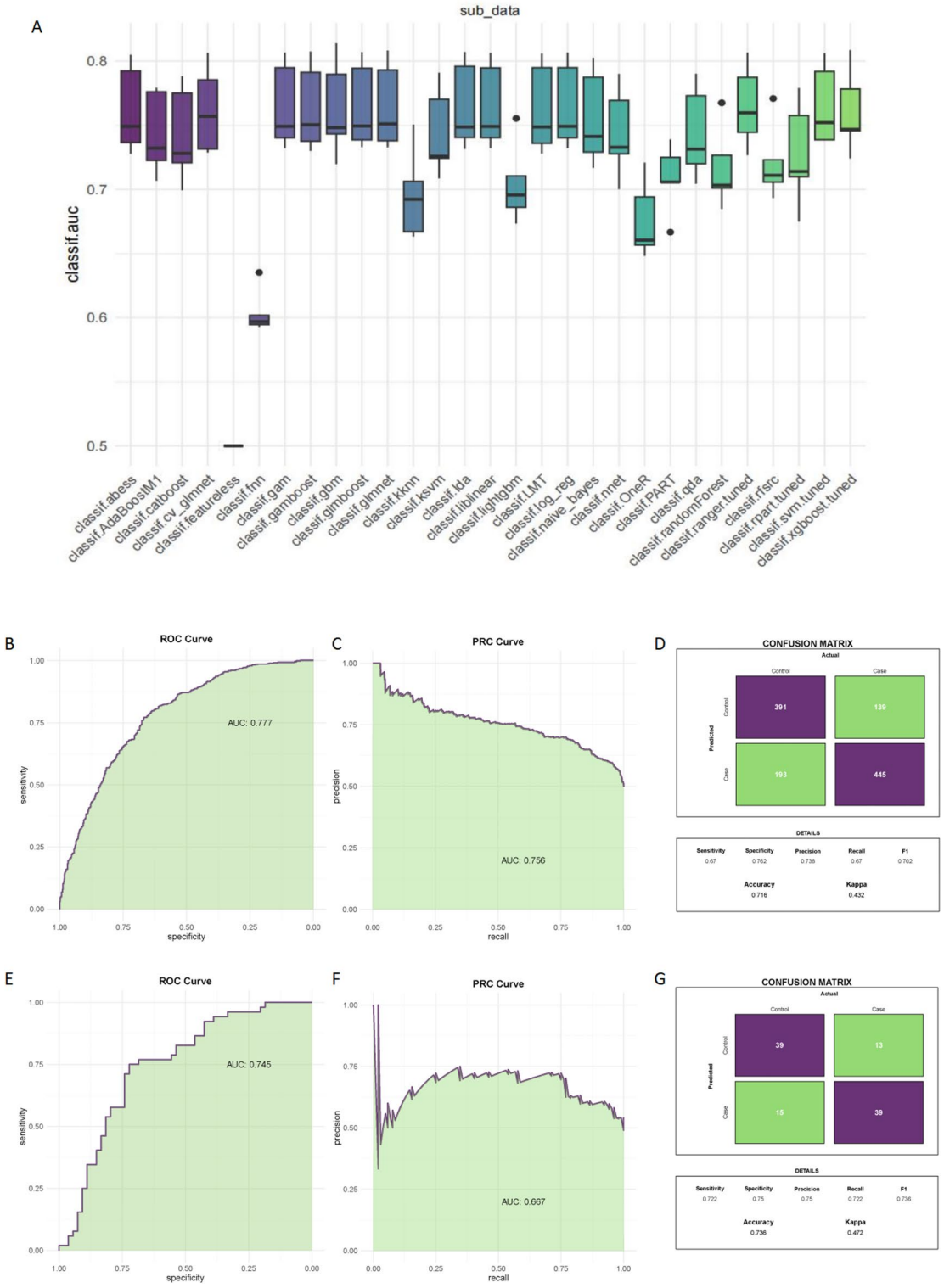


Fig. 4. Performance evaluation of the optimal machine learning model for predicting macroangiopathy in T2DM patients. (**A**) Screening of the optimal benchmark ML model using the final four selected features. Box plots depict the distribution of the Area Under the Curve (AUC) scores across 29 machine learning models, with the optimal model identified based on median performance and variability. (**B**) ROC curve for the training set of the optimal model after hyperparameter tuning. (**C**) PRC for the training set of the optimal model after hyperparameter tuning. (**D**) Confusion matrix for the training set of the optimal model. The matrix presents counts of true positives, true negatives, false positives, and false negatives, along with detailed performance metrics including sensitivity, specificity, precision, recall, F1 score, accuracy, and kappa coefficient. (**E**) ROC curve for the external validation set of the optimal model. (**F**) PRC for the external validation set of the optimal model. (**G**) Confusion matrix for the external validation set of the optimal model.

Model development and hyperparameter tuning

All patients were put into the training set and used to building the model and tune hyper-parameters to optimize the model. The target imbalance was handled by undersampling. The best hyperparameters were determined using 5-fold internal cross-validation and grid search. For the ranger classification model, we tuned the parameters including: "num.trees"(1–2,000), "sample.fraction"(0.1-1), "mtry.ratio"(0–1), "min.node.size"(1-100), and "num.random.splits"(1-100) in the framework mlr3. We used 1,000 sampling iterations to test all possible combinations of hyperparameter values, used cross-validation in the 5 folds used for model development, and selected the hyperparameter combination with the best tested AUC value. The tuning results showed that the optimal hyperparameters for the ranger classification model were "num.trees" = 1112, "sample.fraction"=0.1, "mtry.ratio"= 0.5555556, "min.node.size"=34, "num.random.splits"=1. With this set of parameters, the ranger model's accuracy was 0.716 and its AUC was 0.777 in the training dataset (Fig. 4B-D).

External validation and interpretation of personalized predictions

The final model in the external validation dataset demonstrated an accuracy of 0.736 and an AUC of 0.745 (Fig. 4E-G), which was slightly lower than the AUC of the training set. To improve interpretability of our final model, we utilized the iml R package to generate accumulated local effects (ALE) plots and permutation feature importance (PFI) scores. The results revealed that the duration of T2DM and age had significant influences on the occurrence and outcome of vascular disease complications, with BUN showing the next greatest effect and fibrinogen indicating a smaller effect (Fig. 5A-B). Additionally, SHAP values were calculated to illustrate the contribution of each feature to the final prediction, effectively clarifying and interpreting the model's predictions. The SHAP force plot for an individual sample (Fig. 5C) demonstrated the impact of specific feature values on the prediction, while the SHAP summary plot for the entire cohort (Fig. 5D) confirmed the importance of these features. These findings suggest that our final model possesses strong predictive ability and offers valuable insights into the factors influencing macroangiopathy in T2DM patients.

Synergistic effects of the key factors on macroangiopathy risk

PDPs revealed the interaction dynamics of model performance. The Fig. 6 displayed the synergistic effects of duration of T2DM, age, fibrinogen and BUN levels on macroangiopathy risk. Figures 6 (A-C) illustrated the synergy between T2DM duration and other factors. In Fig. 6A, macroangiopathy risk increased with longer T2DM duration and higher age, excluding extreme values. The T2DM duration primarily influenced predictions when its log and scale value was between − 1 and 1 (original: 0.05 to 31 year). Figure 6B showed a similar trend, comparing the interaction between T2DM duration and fibrinogen, but the interaction effect of fibrinogen was noticeably weaker than that of age. Figure 6C focused on T2DM duration and BUN levels, further indicating a greater impact of T2DM duration on macroangiopathy risk. These findings suggested that controlling T2DM duration is critical for macroangiopathy management.

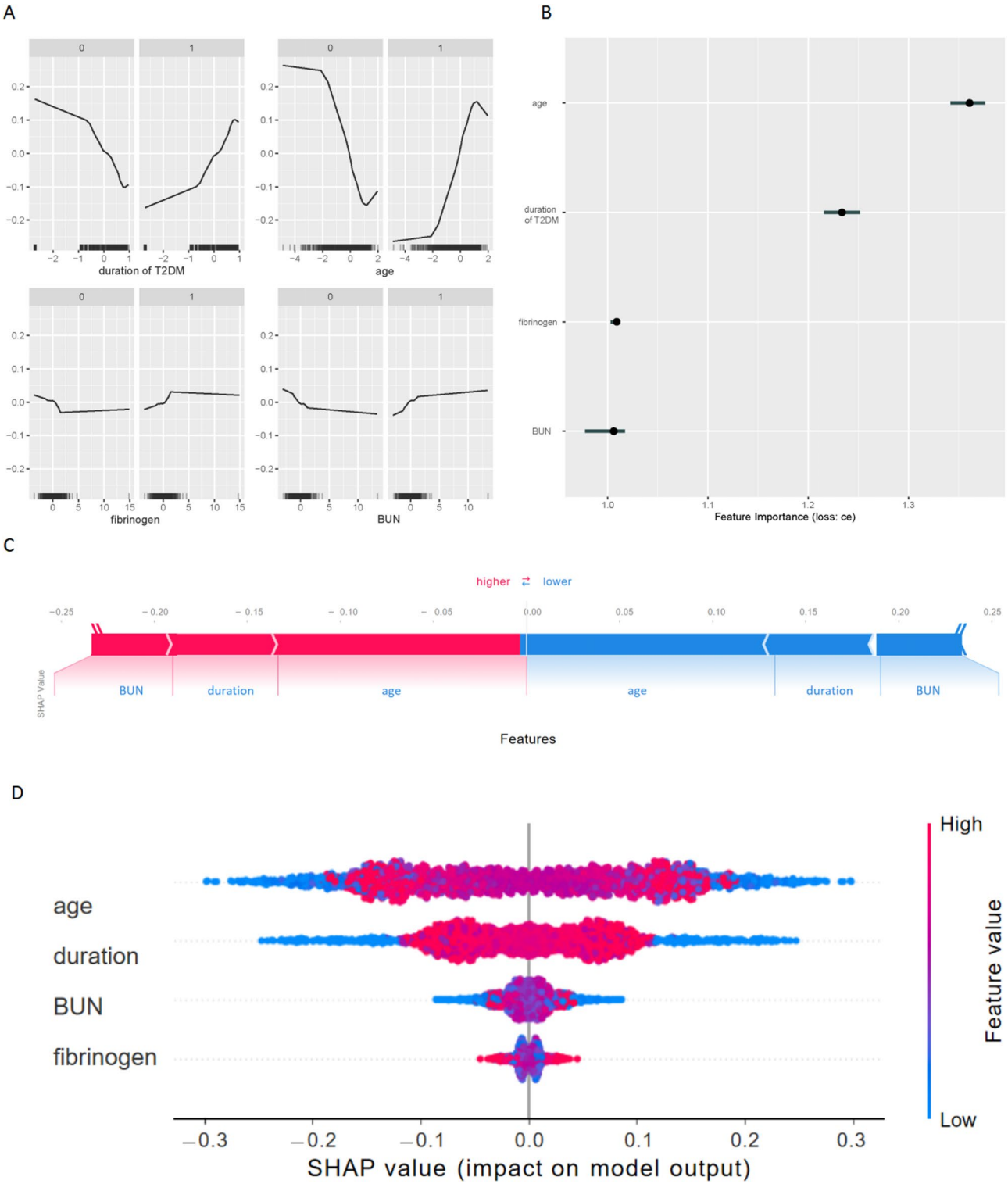
Figure 6 (D-F) illustrates the synergistic effects of age, fibrinogen, and BUN on macroangiopathy risk. As shown in Fig. 6D, macroangiopathy risk is higher with increasing age and fibrinogen levels. The nearly vertical color bars in Figs. 6(D-F) suggest that the predicted macroangiopathy risk is more sensitive to changes in age than to fibrinogen or BUN levels.

Discussion

Increasingly prevalent diabetic complications, particularly cardio-renal complications, contribute significantly to morbidity and mortality[2]. Macrovascular complications, notably in T2DM patients, account for a substantial portion of deaths, constituting 62.7% in this study. As a result of the insidiousness of onset of symptoms, macroangiopathy is often not diagnosed in the early stages. Besides, the diagnosis requires a multi-systematic comprehensive evaluation to define the site, nature and extent. Arterial angiography is the gold standard for the diagnosis of macroangiopathy, but it is limited by its invasive, expensive, and difficult to repeat, limiting its use in early screening, clinical evaluation, and epidemiologic studies. Therefore, there is an urgent need for a rapid, user-friendly screening method to prevent and control macrovascular complications of type 2 diabetes mellitus.

Machine learning, renowned for its fexibility and predictive capabilities, has been employed as a valuable tool for predicting complications among diabetes patients. In this study, we have successfully construed a diagnostic model for predicting vascular complications in patients with T2DM through demographic and laboratory data. In the mlr3 framework, we adopted seven RFE machine learning (ML) methods for feature selection and benchmarked 29 ML models. We identified the most significant predictors were the duration of disease (DD), age, fibrinogen, and serum urea nitrogen. Our final model demonstrated high predictive capability, and we used different interpretative methods including Accumulated Local Effects (ALE), Permutation Feature Importance (PFI), Partial Dependence Plot (PDP), and SHAP values to overcome the black-box nature of ML models. This study endeavored to create a machine learning diagnostic model through features collected from questionnaires and laboratory parameters. Chinese adults with T2DM were included to identify critical variables associated with macroangiopathy. The gbm-RFE, Ranger-RFE and SVM-RFE methods were utilized for feature selection, and a prediction model was constructed based on the intersection of variables selected by these methods. The resulting prediction models demonstrated robust performance, and the optimal ranger model with the highest AUC (AUC = 0.77184) in the training dataset and 0.701 in the external verification dataset. These results suggested the model's reliability across diverse datasets, showcasing high accuracy and generalization ability.

A recursive feature elimination (RFE) method, employing a five-fold cross-validation scheme, was utilized for feature selection. The best variables for macroangiopathy discrimination obtained by taking the intersection with a Venn diagram of three result of gbm-RFE, Ranger-RFE and SVM-RFE with the top 3 AUC values. The top four most important clinical predictive features for diabetic macroangiopathy were identified as duration of disease (DD), age, fibrinogen and serum urea nitrogen. Duration of disease and age aligned



with known risk factors for diabetic macroangiopathy, consistent with previous studies[9]. Furthermore, long-term exposure to hyperglycemia will lead to chronic inflammation, endothelial dysfunction, and subsequent atherosclerosis. Studies have consistently shown that longer diabetes duration correlates with an increased risk of both microvascular and macrovascular complications[10]. Chronic hyperglycemia leads to glycation of proteins and lipids, resulting in the formation of advanced glycation end products (AGEs) that contribute to vascular damage[11]. And the glycation process increase of vascular oxidative stress, promotes inflammation, and influence the vascular function and structure. This continuous vascular insult is also responsible for exacerbating the progression of atherosclerosis and other vascular complications. Thus, it is obvious that the duration of diabetes directly affects the progression and severity of vascular complications, stressing the importance of early and sustained glycemic control in the management of T2DM.

◀**Fig. 5**. Interpretation and influence of key features on the optimal model's predictions. (**A**) Accumulated Local Effects (ALE) plots for the final four selected features. These plots demonstrate the influence of each feature on the optimal model's predictions. The y-axis represents the effect size, while the x-axis shows the normalized value of each feature. (**B**) Permutation Feature Importance (PFI) plot showing the importance of the final four selected features in the optimal model. The x-axis represents the feature importance measured by the change in cross-entropy loss (ce), while the y-axis lists the features. (**C**) SHAP force plot for a single randomly selected sample, illustrating the contribution of each feature to the model's prediction of macroangiopathy for that specific patient. The plot shows how higher feature values (red) and lower feature values (blue) impact the prediction, with the horizontal axis reflecting the SHAP value. (**D**) SHAP summary plot of the proposed model on the entire cohort. Each dot represents a single patient, with colors indicating the feature value (blue for lower values, red for higher values). The horizontal axis represents the SHAP value, indicating the direction and magnitude of the feature's effect on the prediction. Positive SHAP values suggest a protective effect, while negative values indicate an increased risk of severe complications. "0" means without macroangiopathy, "1" means with macroangiopathy.

Age is another important factor in predicting the vascular complications in T2DM. The age is directly related to increased prevalence of comorbid, such as hypertension and dyslipidemia, which have a simultaneous, synergic effect in increasing cardiovascular risk. The aging process itself is linked to vascular stiffening, increased arterial calcification, and a decline in endothelial function[12]. These age-related changes exacerbate the effects of hyperglycemia and other metabolic disturbances in diabetic patients. As the arterial walls thicken and lose elasticity, the risk of plaque formation and rupture increases, leading to higher incidences of cardiovascular events. Our results recommend tailored interventions that consider risks associated with advancing age in the management of T2DM. Effective management strategies for older patients with T2DM should include comprehensive cardiovascular risk assessments and aggressive management of comorbid conditions to mitigate these risks.

Fibrinogen emerged as a potential risk factor in this study, reflecting its role as a major coagulation protein[13] and an inflammatory marker with implications for cardiovascular disease[14,15]. Several studies have demonstrated the biological impact of fibrinogen levels on cardiovascular disease risk[16–18]. As blood clotting is associated with coronary heart disease, elevated fibrinogen levels may increase the risk of coronary heart disease. In T2DM patients, high levels of fibrinogen reflect an increased state of heightened inflammation and procoagulable status, which contribute to atherosclerosis development. Fibrinogen further play its role in the pathogenesis of macrovascular complications by causing platelet aggregation and stabilization of thrombi[19]. The interaction between fibrinogen and endothelial cells triggers inflammatory pathways that accelerate the formation of atherosclerotic plaques. This points suggest the potential benefits of anti-inflammatory and anti-thrombotic therapies in reducing cardiovascular risk in diabetic patients. Targeted therapies that reduce fibrinogen levels could potentially lower the incidence of thrombotic events and improve cardiovascular outcomes.

The serum urea nitrogen, a protein metabolite was generated by the liver and excreted by the kidneys, commonly used to assess renal function[20,21]. Renal dysfunction is common among patients with T2DM and is strongly associated with cardiovascular morbidity and mortality[22]. Impaired kidney function may lead to fluid retention, hypertension, and abnormal lipid metabolism, all these factors raise the cardiovascular risk. Elevated serum urea nitrogen levels indicate decreased renal clearance, reflecting the extent of kidney damage and its systemic effects on the vascular system[7]. Our study found BUN could be an important risk factor for diabetic macroangiopathy. Previous studies have shown that BUN levels were higher in patients with CVD and that BUN can be a valuable predictor of CVD prognosis[20,23,24]. Another studies have also shown that people with higher serum urea nitrogen level come with higher cardiovascular disease risk and all-cause mortality[25]. Actually, serum urea nitrogen as a causal variable has been controversial. The correlation between BUN and macroangiopathy were supported by the large-scale population evidence in this study, but the causality between serum BUN and development of macroangiopathy remains unclear. Further exploration is warranted in future clinical and epidemiologic studies. The causal relationship between serum BUN and macroangiopathy is controversial, but our study provides novel large-scale epidemiological evidence supportive of an association between the serum BUN and the development of macroangiopathy. Further research is also required in the light of future clinical and epidemiological studies. The association between renal dysfunction and cardiovascular disease is complex and mediated by multiple pathways such as increased sympathetic nervous system activity, oxidative stress, and the accumulation of uremic toxins that directly damage the vasculature. The causal relationship between serum BUN and macroangiopathy is controversial, but our study provides novel large-scale epidemiological evidence supportive of an association between the serum BUN and the development of macroangiopathy. Further research is also required in the light of future clinical and epidemiological studies. Early detection and treatment of renal impairment may prevent the development of end-stage renal disease and reduce the associated cardiovascular risks.

The mutual relationships between the identified predictors—duration of disease, age, fibrinogen, and serum urea nitrogen—point to the complex pathophysiology of vascular complications in T2DM. Chronic hyperglycemia mediates the formation of mediates(AGEs), which, in turn, contribute to oxidative stress and inflammation, exacerbating vascular damage[26]. Age-related changes in the vascular system further aggravated these derangements, creating a vicious cycle that accelerates the deterioration of atherosclerotic progression. The upregulated levels of fibrinogen are indicative of an underlying state of inflammation that not only promotes thrombosis but also aggravates the endothelial injury. High levels of serum urea nitrogen indicate renal dysfunction, which increase cardio-related risks by causing fluid overload, hypertension, and dyslipidemia. The

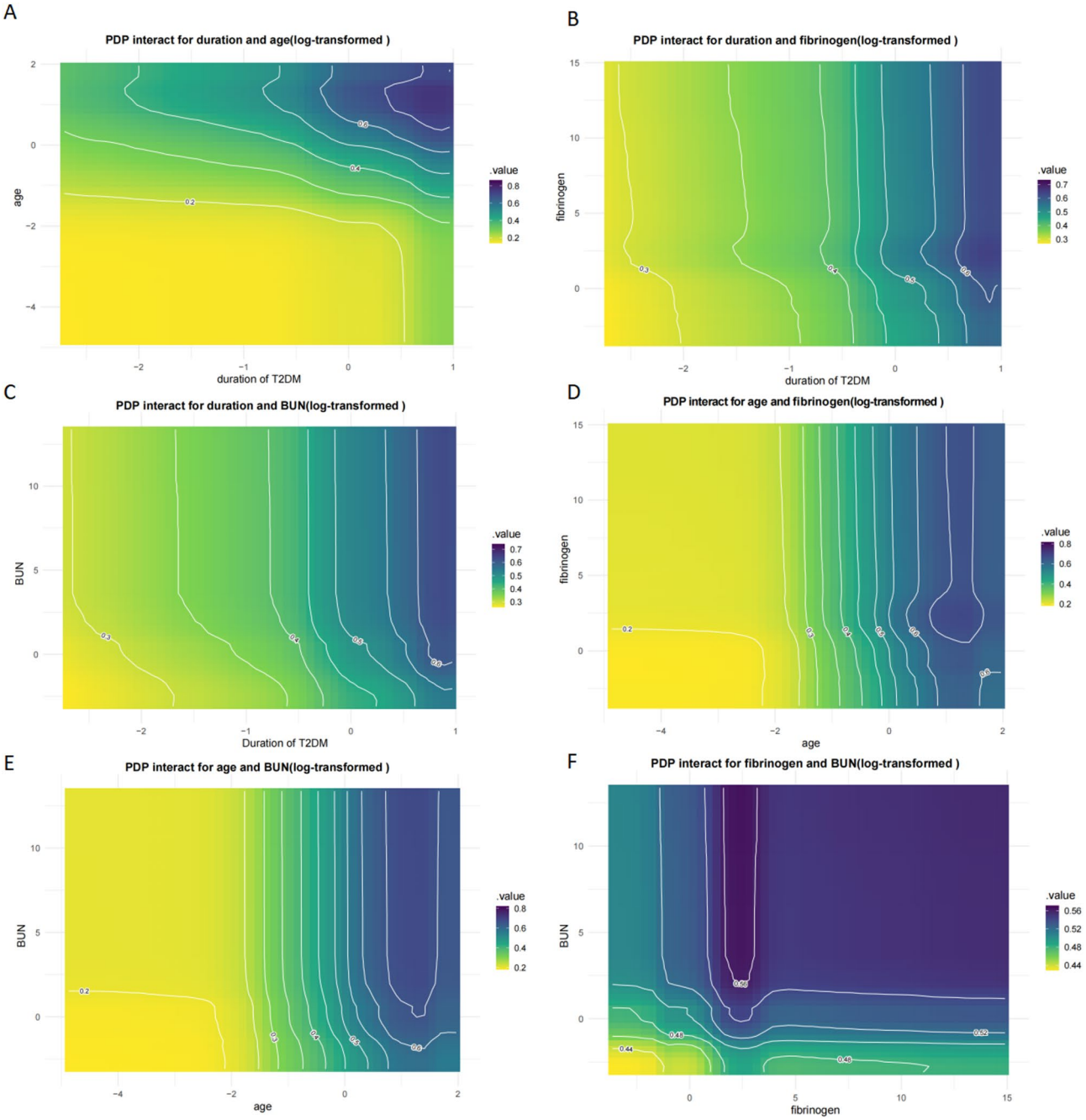


Fig. 6. Partial dependence plots (PDP) illustrating the synergistic effects between various risk factors and macroangiopathy. (**A**) Interaction between age and duration of T2DM on the risk of macroangiopathy. (**B**) Interaction between fibrinogen levels and duration of T2DM. (**C**) Interaction between BUN and duration of T2DM. (**D**) Interaction between fibrinogen levels and age. (**E**) Interaction between BUN levels and age. (**F**) Interaction between BUN and fibrinogen levels. The color bar on the right indicates the predicted values of macroangiopathy, with the x-axis and y-axis representing log and Z score transformation values of each risk factor. The contour plots visually demonstrate the combined influence of these pivotal factors on the predicted values of macroangiopathy, highlighting areas of higher and lower risk.

synergistic effects is strong between these features, demonstrate the vascular complications in T2DM that are multifactorial and necessitate critical analysis for risk assessment and management.

Our findings are consistent with most studies that emphasize good glycemic control, together with the management of comorbid conditions such as hypertension and dyslipidemia, would reduce the risk of vascular complications in T2DM[27]. Uniquely, we included these risk factors available at diagnosis in our predictive model, which providing a more valid tools for risk assessment than conventional statistical methods. Machine learning methods that allow for complex interactions between characteristics can be contributed to improve the accuracy

and practicability of prediction model. The development of a predictive model that can be conveniently applied at the primary healthcare level, without the need for specialized equipment or extensive training, has significant advancement the management of T2DM. Early identification of patients with high risk will facilitate prompt initiation of interventions, thereby minimizing serious cardiovascular events and thus greatly improving the outcome of patients. This approach aligns with the current trend towards personalized medicine, and improved provides tailored interventions based on the individual risk profile to improve outcomes.

One of the strengths of our study is the large sample size and comprehensive demographic and laboratory parameters parameter collection, making our findings generalizable. In addition, the use of the mlr3 framework in combination with several ML methods for feature selection and benchmarking provides robustness and the guarantee of high-quality prediction. However, several limitations were worthy of consideration when interpreting our findings. First, the observational nature of this study limits causal inference, further longitudinal research is warranted. The participants in this study form a relatively limited region (only central China). For this reason, the current research might be unrepresentative of the overall T2DM population and might also have a certain degree of bias. External validation sample size constraints in future study are need. Moreover, although the analysis was adjusted for many potential confounding factors, we cannot exclude the role of residual confounders, which were caused by covariate measurement error and other factors (i.e., medication) that were not assessed in the present cohort. Future studies should focus on longitudinal studies to validate the predictive model and explore causal relationships between the identified risk factors and macrovascular complications. In addition, genetic and molecular data for improving predictive values would be quite interesting. And further research on the impact of interventions based on the predictive model on clinical outcomes, will yield valuable insights into their practical application in healthcare.

Conclusions

This study underscores the potential of an approach based on machine learning algorithm in features selection and the development of prediction tools for diabetic macroangiopathy. Key predictive variables identified included the duration of T2DM, age, fibrinogen, and serum urea nitrogen. The predictive model showed good discrimination performance. It was based on a large population with high risk of diabetic macroangiopathy in middle region of China using multiple machine learning algorithms, which would provide reference for the work of diabetic macroangiopathy prediction and prevention in China. This model provides a practical screening tool for diabetic macroangiopathy, allowing early intervention in high-risk patients. Future research should focus on clinical implementation and further validation across diverse populations.

Data availability

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Received: 24 September 2024; Accepted: 5 May 2025
Published online: 12 May 2025

References

1. Ogurtsova, K. et al. IDF diabetes atlas: global estimates of undiagnosed diabetes in adults for 2021. *Diabetes Res. Clin. Pract.* **183**, 109118 (2022).
2. Wang, L. et al. Prevalence and treatment of diabetes in China, 2013–2018. *JAMA* **326** (24), 2498–2506 (2021).
3. Wu, C. Z. et al. Epidemiologic relationship between periodontitis and type 2 diabetes mellitus. *BMC Oral Health*. **20** (1), 204 (2020).
4. Bahardoust, M. et al. Medication time of Metformin and sulfonylureas and incidence of cardiovascular diseases and mortality in type 2 diabetes: a pooled cohort analysis. *Sci. Rep.* **15** (1), 8401 (2025).
5. Bahardoust, M. et al. Effect of Metformin (vs. Placebo or sulfonylurea) on all-cause and cardiovascular mortality and incident cardiovascular events in patients with diabetes: an umbrella review of systematic reviews with meta-analysis. *J. Diabetes Metab. Disord.* **23** (1), 27–38 (2023).
6. He, D. & Cui, L. Assessing the causal role of selenium in amyotrophic lateral sclerosis: A Mendelian randomization study. *Front. Genet.* **12**, 724903 (2021).
7. Karabaeva, R. Z. et al. Epigenetics of hypertension as a risk factor for the development of coronary artery disease in type 2 diabetes mellitus. *Front. Endocrinol. (Lausanne)*. **15**, 1365738 (2024).
8. American Diabetes Association Professional Practice Committee. 2. Diagnosis and classification of diabetes: standards of care in Diabetes-2024. *Diabetes Care*. **47** (Suppl 1), S20–S42 (2024).
9. Takamatsu, K. Renal status in elderly patients with type 2 diabetes. *Clin. Exp. Nephrol.* **24** (1), 53–62 (2020).
10. He, Y. et al. The Inflamm-Aging model identifies key risk factors in atherosclerosis. *Front. Genet.* **13**, 865827 (2022).
11. Yang, L. et al. Study of cardiovascular disease prediction model based on random forest in Eastern China. *Sci. Rep.* **10** (1), 5245 (2020).
12. Sattar, N. et al. Cardiovascular and kidney risks in individuals with type 2 diabetes: contemporary Understanding with greater emphasis on excess adiposity. *Diabetes Care*. **47** (4), 531–543 (2024).
13. Wolberg, A. S. Fibrinogen and fibrin: synthesis, structure, and function in health and disease. *J. Thromb. Haemost.* **21** (11), 3005–3015 (2023).
14. De Luca, G. et al. High fibrinogen level is an independent predictor of presence and extent of coronary artery disease among Italian population. *J. Thromb. Thrombolysis*. **31** (4), 458–463 (2011).
15. Bielak, L. F. et al. Association of fibrinogen with quantity of coronary artery calcification measured by electron beam computed tomography. *Arterioscler. Thromb. Vasc Biol.* **20** (9), 2167–2171 (2000).
16. Green, D. et al. Elevated fibrinogen levels and subsequent subclinical atherosclerosis: the CARDIA study. *Atherosclerosis* **202** (2), 623–631 (2009).
17. Cho, H. M. et al. Association between fibrinogen and carotid atherosclerosis according to smoking status in a Korean male population. *Yonsei Med. J.* **56** (4), 921–927 (2015).
18. Wang, J. et al. Novel biomarkers for cardiovascular risk prediction. *J. Geriatr. Cardiol.* **14** (2), 135–150 (2017).

19. Vieira, I. H. et al. Diabetes and stroke: impact of novel therapies for the treatment of type 2 diabetes mellitus. *Biomedicines* **12** (5), 1102 (2024).
20. Matsue, Y. et al. Blood Urea nitrogen-to-creatinine ratio in the general population and in patients with acute heart failure. *Heart* **103** (6), 407–413 (2017).
21. You, S. et al. Prognostic significance of blood Urea nitrogen in acute ischemic stroke. *Circ. J.* **82** (2), 572–578 (2018).
22. Strati, M. et al. Early onset type 2 diabetes mellitus: an update. *Endocrine* **85** (3), 965–978 (2024).
23. Bhatia, K. et al. Predictors of early neurological deterioration in patients with acute ischaemic stroke with special reference to blood Urea nitrogen (BUN)/creatinine ratio & urine specific gravity. *Indian J. Med. Res.* **141** (3), 299–307 (2015).
24. Lin, H. J. et al. Elevated blood Urea nitrogen-to-creatinine ratio increased the risk of hospitalization and all-cause death in patients with chronic heart failure. *Clin. Res. Cardiol.* **98** (8), 487–492 (2009).
25. Hong, C. et al. Association of blood Urea nitrogen with cardiovascular diseases and All-Cause mortality in USA adults: results from NHANES 1999–2006. *Nutrients* **15** (2), 461 (2023).
26. Gami, A. et al. New perspectives in management of cardiovascular risk among people with diabetes. *J. Am. Heart Assoc.* **13** (12), e034053 (2024).
27. Dardano, A. et al. The current landscape for diabetes treatment: preventing diabetes-associated CV risk. *Atherosclerosis* **394**, 117560 (2024).

Author contributions
WL, YW and NZ designed of the study and data collection. YW, HZ, HF, XL, ZL and ZH recruited the subjects and supervised the study. ZG, ZZ, NZ and WL analyzed the data. NZ wrote the article. ZG and LW contributed to the revising of the manuscript. All authors contributed to the article and approved the submitted version.

Funding
This study was supported the National Natural Science Foundation of China (number 82102281),, the Natural Scientific Foundation of Hunan Province (Grant Number 2021JJ40867, number 2021JJ40893), the Scientific Research Fund of Hunan Province Health Commission (202101062143), Innovation and Entrepreneurship Training Program for College Students of Hunan Province(S202310545023S).

Declarations

Competing interests
The authors declare no competing interests.

Ethics approval and consent to participate
This study was approved by the Institutional Review Board of the First Affiliated Hospital of Zhengzhou University. Written informed consent to participate was obtained from all participants.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-01161-5.

Correspondence and requests for materials should be addressed to W.L. or Z.G.