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.

# Predicting atherosclerotic cardiovascular disease risk in retired civil aviation employees using the triglyceride–glucose index and machine learning models: a retrospective cross sectional study with explainable analysis

**Sunhang Cao**
Civil Aviation Administration of China
**Chunshui Huang**
836011798@qq.com

Civil Aviation Administration of China

**Research Article**

**Keywords:**

**Posted Date:** November 11th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11183490/v1

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

**Additional Declarations:** No competing interests reported.

# Predicting atherosclerotic cardiovascular disease risk in retired civil aviation employees using the triglyceride–glucose index and machine learning models: a retrospective cross-sectional study with explainable analysis

*Sunhang Cao[1], Zhengfeng Liu[2], Dongmei Zhai[3], Peng Li[1], Chunshui Huang[1]*

[1] Department of Physical Examination and Rehabilitation, Civil Aviation General Hospital, Beijing 100123, China

[2] Department of Vascular Surgery, Civil Aviation General Hospital, Beijing 100123, China

[3] China Civil Aviation Information Network Co., Ltd., Beijing 101318, China

* Correspondence: Chunshui Huang. Email: 836011798@qq.com

## Abstract

**Background:** Atherosclerotic cardiovascular disease (ASCVD) is the leading cause of death and disability worldwide. Retired civil aviation employees carry substantial accumulated cardiometabolic risk, yet ASCVD risk prediction is understudied in this group. The triglyceride–glucose (TyG) index is a reliable surrogate for insulin resistance, but its predictive value in this population and its incremental benefit over the simple combination of triglyceride (TG) and glucose (GLU) remain unverified. This study evaluated the TyG index for ASCVD risk prediction in retired civil aviation employees and compared machine-learning models.

**Methods:** A retrospective cross-sectional study was conducted using physical examination data of 1,047 retired civil aviation employees from Civil Aviation General Hospital (January 2023–June 2025). Participants were categorized into high-risk and non-high-risk ASCVD groups. Significant variables were screened by univariate logistic regression ($P<0.05$) and multivariable analysis after variance inflation factor (VIF) based collinearity exclusion. Three models were built—basic, basic + TyG, and basic + TG + GLU—and the incremental value of TyG was assessed by AUC, the DeLong test, the likelihood-ratio test, NRI, and IDI. Six machine-learning algorithms were used to build predictive models, evaluated by 5-fold cross-validation and an independent validation set, and the best model was interpreted with SHapley Additive exPlanations (SHAP).

**Results:** A total of 458 participants were in the high-risk group and 589 in the non-high-risk group. Univariate regression identified 17 significant variables (all VIF <3). In multivariable regression, the TyG model achieved an AUC of 0.8786, significantly higher than the basic model (0.8034) and the TG + GLU model (0.8636; DeLong test $P<0.001$). Among the six algorithms, binary logistic regression yielded the highest validation-set AUC (0.8495, 95% CI 0.8080–0.8898) and the most stable cross-validation performance (5-fold CV AUC = 0.8788 ± 0.0124). SHAP analysis identified the TyG index as the most important feature (mean |SHAP| = 1.1143), with higher values corresponding to greater ASCVD risk.

**Conclusions:** The TyG index is a core predictor of ASCVD risk in retired civil aviation employees and outperforms the simple combination of TG and GLU. Binary logistic regression

showed robust performance with good discrimination and interpretability, making it suitable for clinical application.

**Keywords:** Triglyceride–glucose index; Atherosclerotic cardiovascular disease; Machine learning; Logistic regression; SHAP interpretability; Retired civil aviation employees.

## Background

Atherosclerotic cardiovascular disease (ASCVD) is the leading cause of death and disability worldwide, encompassing coronary heart disease, ischemic stroke, and peripheral arterial disease. According to the Report on Cardiovascular Health and Diseases in China [1], approximately 330 million individuals in China suffer from cardiovascular disease, and ASCVD-related mortality accounts for a continuously rising proportion of total deaths among both urban and rural residents. Early identification of high-risk populations and implementation of targeted interventions are key strategies for reducing disease burden; therefore, constructing reliable ASCVD risk prediction models holds significant clinical and public health importance.

Civil aviation employees are exposed to elevated cardiometabolic risk factors owing to their occupational characteristics (chronic shift work, high psychological stress, sedentary lifestyle, and irregular diet). Although retirement removes them from the original occupational environment, long-term accumulated metabolic abnormalities continue to affect cardiovascular health. However, research on ASCVD risk prediction specifically targeting this special population of retired civil aviation employees remains scarce, and there is an urgent need to develop applicable risk assessment tools based on real-world physical examination data.

The triglyceride–glucose (TyG) index was proposed by Simental-Mendía et al. in 2008 [2] and is calculated as ln[TG (mg/L) × GLU (mg/L)/2]. It comprehensively reflects the metabolic status of triglycerides and fasting glucose and has been extensively validated as a reliable surrogate marker of insulin resistance [3–10]. In recent years, multiple studies have shown that the TyG index is closely associated with the risk of ASCVD and its subtypes (coronary heart disease, stroke, etc.), and its predictive performance may surpass that of TG or GLU alone because TyG captures the non-linear interactive effect of the two. However, the predictive value of the TyG index in retired civil aviation employees has not yet been verified, and empirical evidence comparing its incremental value with the combined TG+GLU indicator is lacking.

Traditional ASCVD risk assessment largely relies on linear models such as logistic regression, which offer strong interpretability but may be limited in handling multidimensional feature interactions and non-linear relationships. Machine learning algorithms (random forest, support vector machine, gradient boosting machine, etc.) can automatically capture complex patterns [11–14], yet their “black-box” nature constrains clinical translation. In recent years, the SHapley Additive exPlanations (SHAP) interpretability framework [15–18] has provided an effective tool to address the interpretability challenge of machine learning models, revealing both the direction and magnitude of each feature’s contribution to model decisions while preserving predictive accuracy.

Based on the above background, this study, leveraging the physical examination database of Civil Aviation General Hospital from 2023 to 2025, is the first to investigate retired employees of a civil aviation sector by: (1) verifying the predictive value of the TyG index for ASCVD risk and

comparing its incremental value with the combined TG+GLU indicator; (2) constructing and comparing the discriminatory performance of six prediction models (including traditional logistic regression and five machine learning algorithms); and (3) performing SHAP interpretability analysis on the best model to elucidate the mechanism by which key features (especially the TyG index) contribute to ASCVD risk prediction.

# Methods

## 2.1 Study design and data source

This study was a retrospective cross-sectional real-world study. Data were sourced from the annual physical examination electronic health records of retired civil aviation employees in the health examination management information system of Civil Aviation General Hospital (Beijing, China) from January 2023 to June 2025. All data were extracted after anonymization to protect participant privacy. The study protocol complied with the Declaration of Helsinki [15] and the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guideline for cross-sectional studies [14], and was supported by the Youth Project of Peking University Civil Aviation School of Clinical Medicine (No. 202430) and the Civil Aviation Medical Center of the Civil Aviation Administration of China (Project No. 2025-L-K-067).

From the 7,255 initial records extracted between January 2023 and June 2025, we excluded 6,106 records of participants who did not meet the age criteria (male <60 years, female <50 years), leaving 1,149; excluded 23 records with missing data required for ASCVD risk assessment, leaving 1,126; and excluded 79 records with missing independent variables (e.g., LDL-C, blood pressure, waist-to-hip ratio), leaving a final analytical sample of 1,047.

## 2.2 Inclusion and exclusion criteria

Inclusion criteria: (1) male age ≥60 years or female age ≥50 years; (2) prior occupational affiliation traceable to a civil aviation sector; (3) complete annual physical examination records (including fasting blood biochemistry). Exclusion criteria: (1) failure to meet age criteria; (2) known severe hepatic or renal dysfunction, malignancy, or acute critical illness; (3) missing FPG or TG required for TyG calculation, or missing core information required for ASCVD grouping.

## 2.3 Data collection

After fasting ≥8 h, all participants underwent blood collection the following morning. Biochemical assays were performed using a BECKMAN COULTER AU5800 automated biochemical analyzer (standardized enzymatic method). This study constructed an analytical framework based on multidimensional data: demographic characteristics included age, sex, marital status, education level, and prior job category; anthropometric measures included height, weight, and waist circumference (WC), from which body mass index (BMI = weight/height$^2$) was derived; vital signs included resting systolic blood pressure (SBP) and diastolic blood pressure (DBP), each taken as the mean of two measurements; biochemical indicators included fasting plasma glucose (FPG), TG, total cholesterol (TC), high-density lipoprotein cholesterol (HDL-C),

low-density lipoprotein cholesterol (LDL-C), serum creatinine, and uric acid; personal and disease history included hypertension, diabetes, coronary heart disease, and stroke.

## 2.4 Variable definition

### 2.4.1 Independent variables

Based on clinical significance and literature evidence, 28 candidate variables were included across three dimensions: (1) classic risk factors (demographics, physical examination, TyG index, blood lipids, glucose, homocysteine, etc.); (2) auxiliary indicators (liver function, renal function, other biochemical indicators, etc.); (3) exploratory indicators (routine blood parameters, etc.). Blood biochemical indicators (albumin, LDL-C, HDL-C, total cholesterol, alanine aminotransferase, aspartate aminotransferase, etc.) were dichotomized according to clinical reference ranges. TG and GLU were used as continuous variables to preserve maximal information. The TyG index was calculated as: $TyG = \ln[TG\ (mg/L) \times GLU\ (mg/L)/2]$.

### 2.4.2 Dependent variable

Based on the Chinese Guideline for Lipid Management (Primary Care Edition·2024) [20], the ACC/AHA (American College of Cardiology/American Heart Association) risk assessment framework [21], and the PREVENT (Predicting Risk of cardiovascular disease EVENTs) equation [22–24], we comprehensively evaluated the number and severity of risk factors for each participant and classified them into an ASCVD high-risk group or non-high-risk group. High-risk determination criteria included: (1) diagnosed ASCVD; (2) diabetes with target-organ damage or ≥1 major risk factor; (3) non-diabetic individuals with ≥3 major risk factors (hypertension, dyslipidemia, smoking, obesity, etc.) whose risk factor score reached the high-risk threshold.

## 2.5 Statistical analysis

Data analysis was performed using SPSS and Python 3.11 (pandas, scipy, scikit-learn). Continuous variables were assessed for normality using the Shapiro–Wilk test. Normally distributed data are presented as mean ± standard deviation and compared between groups using the independent-samples t-test; non-normally distributed data are presented as median (interquartile range) and compared using the Mann–Whitney U test. Categorical variables are presented as frequencies (percentages) and compared using the $\chi^2$ test or Fisher's exact test (when expected frequency <5). A two-sided $P<0.05$ was considered statistically significant.

Univariate logistic regression was used to screen variables one by one, with those at $P<0.05$ entering multivariate analysis. Before multivariate regression, collinearity was diagnosed using the variance inflation factor (VIF), and variables with VIF>10 were excluded. Multivariate regression used the Backward stepwise method ($\alpha=0.05$ for removal).

TyG versus TG+GLU comparison design: To ensure a fair comparison, the basic model included only univariate-significant variables with VIF<10 (excluding TG, GLU, and TyG). On this basis, we constructed: (1) the basic + TyG model; and (2) the basic + TG + GLU model. Discriminatory ability was assessed using the area under the curve (AUC) and its 95% confidence interval (Hanley–McNeil method); the DeLong test was used to compare AUC differences between

models; the likelihood-ratio (LR) test assessed incremental value relative to the basic model; and the net reclassification improvement (NRI) and integrated discrimination improvement (IDI) assessed reclassification ability.

Machine learning comparison design: The feature space comprised 11 significant basic variables selected by binary logistic regression (Backward: LR method) plus the TyG index. Six algorithms (binary logistic regression, random forest, support vector machine, k-nearest neighbors, gradient boosting machine, and naive Bayes) were used to construct prediction models. Seventy percent of the data served as the training set and 30% as the validation set to compare predictive performance.

# Results

## 3.1 Baseline characteristics

Table 1 presents the baseline characteristics of the 1,047 participants. There were 458 in the high-risk group and 589 in the non-high-risk group. The two groups differed significantly ($P<0.001$) in sex, age, TG, LDL-C, HDL-C, ALT, AST, uric acid, glucose, estimated glomerular filtration rate, lymphocyte percentage, lymphocyte count, mean corpuscular volume, mean corpuscular hemoglobin, TyG index, BMI, blood pressure, waist-to-hip ratio, and homocysteine. The high-risk group had a higher proportion of males (50.4% vs. 28.5%), was older, and had a higher TyG index (8.91 vs. 8.34), suggesting these factors are closely associated with the high-risk ASCVD state. Variables such as albumin, platelets, and total cholesterol showed no significant intergroup differences ($P>0.05$) and were not included in subsequent analyses.

## 3.2 Univariate logistic regression

Table 2 presents the univariate logistic regression results. Excluding the three comparison variables (TG, GLU, and TyG), 20 variables were statistically significant at $P<0.05$, among which waist-to-hip ratio abnormality (odds ratio [OR] = 3.110), blood pressure abnormality (OR = 2.574), male sex (OR = 2.550), and the TyG index (OR = 13.505) showed relatively large effect sizes.

## 3.3 Collinearity diagnosis

VIF diagnosis was performed on the 20 univariate-significant variables; all VIF values were <3 (maximum 2.33), indicating no severe collinearity, and no variables needed to be excluded.

## 3.4 Incremental predictive value of the TyG index

To verify the incremental predictive value of the TyG index relative to the combined TG+GLU indicator, we included the univariate-significant independent variables in binary logistic regression and set up three models for comparison: the basic model (excluding TyG and TG+GLU); the basic + TyG model; and the basic + TG + GLU model.

#### 3.4.1 Basic model

Table 3 presents the binary logistic regression results for the basic model. The basic model fit: Akaike information criterion (AIC) = 1,134.3, −2LL = 1,110.3, Nagelkerke $R^2$ = 0.357.

#### 3.4.2 Basic + triglyceride–glucose model

Table 4 presents the binary logistic regression results for the basic + TyG model. TyG index: OR = 12.226 (95% CI: 8.049–18.571), P<0.001, with a significant effect size.

#### 3.4.3 Basic + triglyceride + glucose model

Table 5 presents the binary logistic regression results for the basic + TG + GLU model. TG: OR = 3.728 (95% CI: 2.853–4.870), P<0.001; GLU: OR = 1.243 (95% CI: 1.107–1.396), P<0.001.

#### 3.4.4 Comparison of AUC among the three models

Tables 6 and 7 present the comparison of AUC and the results of statistical tests, respectively. The basic model had an AUC of 0.8034; adding the TyG index increased the AUC to 0.8786 (ΔAUC = +0.0752), and adding TG+GLU increased it to 0.8636 (ΔAUC = +0.0602). The DeLong test showed a statistically significant difference in AUC between the TyG and TG+GLU models (P<0.001), with superior predictive performance for TyG. In terms of AIC, the TyG model (929.0) was lower than the TG+GLU model (974.0), indicating that the TyG model is more parsimonious and efficient.

Figure 1 shows the comparison of ROC curves among the basic model, the basic + TyG model, and the basic + TG + GLU model. The basic + TyG model curve is clearly superior to the other two, intuitively demonstrating the incremental predictive value of the TyG index.

### 3.5 Comparison of binary logistic regression and machine learning algorithms

#### 3.5.1 Feature variables

The feature space consisted of 11 significant basic variables selected by binary logistic regression (Backward: LR method) plus the TyG index, comprising 12 feature variables in total.

#### 3.5.2 Data partitioning and algorithm settings

The training and validation sets were randomly partitioned at 70%/30% (random_state = 42): the training set comprised 732 participants (320 positive, 412 negative) and the validation set 315 (138 positive, 177 negative). Five-fold stratified cross-validation was performed within the training set for model evaluation. Table 8 lists the six algorithms and their key parameters.

#### 3.5.3 Model evaluation results

Table 9 presents the 5-fold cross-validation AUC. The 5-fold cross-validation AUC mean was highest for binary logistic regression (0.8788 ± 0.0124) with the smallest standard deviation, indicating the most stable performance across different data subsets.

Table 10 presents the performance metrics in the validation set. The validation-set AUC was highest for binary logistic regression (AUC = 0.8495, 95% CI: 0.8080–0.8898), followed by random forest (0.8401). All six algorithms achieved validation-set AUCs >0.80, indicating that the feature set including the TyG index has good discriminative ability for ASCVD risk. This model was selected as the best model for subsequent SHAP interpretability analysis [25]. Table 11 presents the confusion matrix in the validation set.

### 3.5.4 Visualization results

Figure 2 shows the AUC values of the six algorithms across the five cross-validation folds; color shading from light to dark indicates AUC from low to high, intuitively reflecting each algorithm's stability across different data subsets. Figure 3 shows the average ROC curves of the six algorithms in the 5-fold cross-validation of the training set. Figure 4 shows the ROC curves and AUC values (with 95% CIs) of the six algorithms on the independent validation set; the bolded curve is the best model, binary logistic regression (AUC = 0.8495). Figure 5 shows the built-in feature importance rankings of the two tree-based models, random forest and gradient boosting machine; both models ranked the TyG index as the most important predictive feature.

### 3.5.5 SHAP interpretability analysis

The following SHAP analysis is based on the best model, binary logistic regression (LR), using LinearExplainer to compute SHAP values. Table 12 presents the ranking of SHAP feature importance. The TyG index ranked first, with a mean |SHAP| value of 1.1143, approximately 2.3 times that of the second-ranked feature (sex). Table 13 presents the standardized coefficients of binary logistic regression.

Figure 6 is the SHAP summary plot. Each point represents a validation-set sample; the vertical axis is the feature (ranked in descending order of mean |SHAP|), the horizontal axis is the SHAP value, and color indicates feature value level (red = high, blue = low). A SHAP value >0 indicates the feature pushes the model prediction toward ASCVD-positive, and <0 toward negative. Figure 7 ranks all samples by mean absolute SHAP value, quantifying each feature's global contribution to the model prediction. Figure 8 shows the feature contribution decomposition for a high-risk individual in the validation set; red arrows indicate features that increase the predicted probability and blue arrows indicate features that decrease it, with arrow length indicating contribution magnitude. Figure 9 is the SHAP dependence plot for the TyG index, with the horizontal axis as the actual TyG value and the vertical axis as the corresponding SHAP value; this plot reveals the positive relationship between the TyG index value and predicted ASCVD risk.

## Discussion

### 4.1 Model performance comparison and clinical implications

In terms of validation-set AUC, binary logistic regression performed best (AUC = 0.8495), followed by random forest (AUC = 0.8401) and support vector machine (AUC = 0.8386). All six algorithms achieved validation-set AUCs exceeding 0.80, indicating that the feature set including the TyG index has good discriminative ability for ASCVD risk.

Notably, traditional binary logistic regression achieved the highest validation-set AUC (0.8495), surpassing all five machine learning algorithms. This result suggests that, under the current feature set (12 variables selected by logistic regression plus the TyG index), the relationship between features and outcome may be predominantly linear, and the linear combination model of logistic regression can already capture it effectively. Although machine learning algorithms can model non-linear relationships and feature interactions, they exhibited obvious overfitting tendencies on the training set (e.g., k-nearest neighbors [KNN] training-set AUC = 1.0000, random forest [RF] training-set AUC = 0.9823), resulting in validation-set generalization performance that did not exceed that of logistic regression. Furthermore, logistic regression offers strong interpretability, high computational efficiency, and low susceptibility to overfitting, giving it unique practical value in clinical applications.

### 4.2 Predictive contribution of the TyG index

Interpretability analysis showed that the TyG index was the most important feature in the binary logistic regression model prediction, with a mean |SHAP| value of 1.1143, ranking first among all 12 features. This result is consistent with the biological mechanism of the TyG index as a marker of insulin resistance: the TyG index comprehensively reflects the metabolic status of triglycerides and glucose [2,26] and can capture their non-linear interactive effect, thereby playing a core role in ASCVD risk prediction. Previous studies have also shown that the TyG index independently predicts major adverse cardiovascular and cerebrovascular events after percutaneous coronary intervention in patients with acute coronary syndrome [27]; and the TyG-BMI index, combining TyG with body mass index, also has good predictive value for the severity of coronary lesions in patients with premature coronary heart disease [28], further confirming the key role of obesity-related insulin resistance in the progression of atherosclerosis.

From the SHAP dependence plot (Figure 9), it can be further observed that the SHAP value of the TyG index increases as its feature value rises, indicating that higher TyG index values correspond to a greater model-predicted probability of ASCVD-positive, showing a consistent positive contribution trend.

Comparing the standardized coefficients of logistic regression (Table 13) with the SHAP feature importance (Table 12), the two methods yielded essentially consistent assessments of feature importance, with the TyG index ranking first in both, jointly confirming its central position in ASCVD risk prediction from different perspectives.

### 4.3 Incremental value of TyG versus TG+GLU

Multivariate regression results showed that the TyG model (AUC = 0.8786) was significantly superior to the TG+GLU model (AUC = 0.8636), with a DeLong test $P<0.001$. In terms of AIC, the TyG model (929.0) was lower than the TG+GLU model (974.0), indicating that the TyG index achieves superior predictive performance with fewer variables, offering better parsimony. Both NRI (0.158 vs. 0.126) and IDI (0.155 vs. 0.121) showed stronger reclassification ability for the TyG model. These results indicate that the TyG index, by integrating the interactive effect of TG and GLU, provides more incremental information than their simple combination, supporting the inclusion of the TyG index rather than the TG+GLU combination in ASCVD risk assessment tools.

### 4.4 Complementarity of logistic regression and machine learning

Including binary logistic regression in the machine learning algorithm comparison framework in this study has important methodological significance. As a classic statistical learning method [11,12], logistic regression has the following advantages: (1) simple model form that directly outputs OR values to interpret variable effects; (2) low susceptibility to overfitting with stable generalization performance; (3) high computational efficiency, suitable for large-scale data. The advantage of machine learning algorithms lies in automatically capturing non-linear relationships without pre-specifying the form of variable relationships; ensemble methods improve predictive accuracy by aggregating multiple weak learners.

In this study, the 5-fold cross-validation AUC of logistic regression ($0.8788 \pm 0.0124$) was higher than that of all machine learning (ML) algorithms, with the smallest CV standard deviation, indicating the most stable performance across different data subsets. This may be because the feature set itself was selected by logistic regression and is highly adapted to the logistic model; it also reflects that, with a moderate sample size (N = 1,047), simpler models may achieve a better bias–variance trade-off than complex models.

### 4.5 Limitations

This study has the following limitations: (1) the retrospective cross-sectional design cannot establish causality and can only reveal associations; (2) the sample was drawn from a single medical institution's retired civil aviation employees, limiting generalizability [22–24]; (3) some variables (e.g., smoking history, alcohol consumption, medication use) were not included, potentially omitting important confounders; (4) the moderate sample size (N = 1,047) may limit the advantage of machine learning algorithms, which might be more pronounced in larger samples. Future studies should consider multicenter prospective designs, incorporate more confounding variables, and enlarge the sample size to further validate model performance.

## Conclusions

This study is the first to verify the predictive value of the TyG index for ASCVD risk in retired employees of a civil aviation sector. The main conclusions are as follows: (1) The TyG index is a core indicator for ASCVD risk prediction, with predictive performance significantly superior to the simple combination of TG and GLU (AUC 0.8786 vs. 0.8636, DeLong test $P<0.001$), and achieves a more parsimonious and efficient model with fewer variables. (2) Among the six prediction models, binary logistic regression achieved the highest validation-set AUC (0.8495) and the best cross-validation stability, indicating that under the current feature set a linear model can already effectively capture the main predictive information. (3) SHAP interpretability analysis confirmed the TyG index as the most important feature (mean $|SHAP| = 1.1143$); higher values correspond to greater ASCVD risk, providing a clear target direction for clinical intervention. (4) Logistic regression combines good discriminative ability with interpretability and is suitable as a clinical tool for ASCVD risk assessment in retired civil aviation employees.

## List of abbreviations

ASCVD, atherosclerotic cardiovascular disease; TyG, triglyceride–glucose; TG, triglyceride; GLU, glucose; FPG, fasting plasma glucose; BMI, body mass index; WC, waist circumference; SBP, systolic blood pressure; DBP, diastolic blood pressure; TC, total cholesterol; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; VIF, variance inflation factor; AUC, area under the curve; ROC, receiver operating characteristic; CI, confidence interval; LR, likelihood ratio; NRI, net reclassification improvement; IDI, integrated discrimination improvement; SHAP, SHapley Additive exPlanations; ML, machine learning; KNN, k-nearest neighbors; RF, random forest; PREVENT, Predicting Risk of cardiovascular disease EVENTs; ACC/AHA, American College of Cardiology/American Heart Association; STROBE, Strengthening the Reporting of Observational Studies in Epidemiology; AIC, Akaike information criterion.

## Declarations

### Ethics approval and consent to participate

This study was conducted in accordance with the Declaration of Helsinki [15] and approved by the Ethics Committee of Civil Aviation General Hospital (Project No. 2025-L-K-067). Because this was a retrospective study with anonymized data, the requirement for informed consent was waived.

### Consent for publication

Not applicable.

### Availability of data and materials

The underlying dataset of this study was sourced from the physical examination information system of Civil Aviation General Hospital. Restricted by patient privacy protection regulations, the data are not publicly available, but can be obtained from the corresponding author upon reasonable request.

### Competing interests

The authors declare that they have no competing interests.

### Funding

This study was supported by the Youth Physician Project of Peking University Civil Aviation School of Clinical Medicine (No. 202430) and the Civil Aviation Medical Center of the Civil Aviation Administration of China (Project No. 2025-L-K-067).

### Authors' contributions

SC and CH conceptualized and designed the study. SC, ZL, and DZ contributed to data curation and formal analysis. SC and DZ performed the statistical modeling and machine learning analyses.

PL and CH supervised the study and acquired funding. SC drafted the manuscript. ZL, DZ, PL, and CH critically revised the manuscript. All authors read and approved the final manuscript. [AUTHOR CONTRIBUTIONS TO BE CONFIRMED BY AUTHORS]


## Acknowledgements

Not applicable.


## Authors' information


Sunhang Cao, MD, attending physician. Research interests: risk prediction of atherosclerotic cardiovascular disease and clinical application of machine learning; geriatric common diseases. Email: 116413935@qq.com.

Zhengfeng Liu, MD, attending physician. Research interests: diagnosis and treatment of vascular surgery and peripheral arterial disease; diabetic foot. Email: 116413935@qq.com.

Dongmei Zhai, MS, engineer. Research interests: safety management, data analysis, machine learning. Email: 2696709771@qq.com.

Peng Li, attending physician. Civil Aviation General Hospital, Beijing, China.

Chunshui Huang, attending physician. Civil Aviation General Hospital, Beijing, China. Email: 836011798@qq.com


## References


[1] National Center for Cardiovascular Diseases; The Writing Committee of the Report on Cardiovascular Health and Diseases in China. Report on cardiovascular health and diseases in China 2023: an updated summary. Chin Circ J. 2024;39(7):625–660. doi:10.3969/j.issn.1000-3614.2024.07.001.

[2] Simental-Mendía LE, Rodríguez-Morán M, Guerrero-Romero F. The product of fasting glucose and triglycerides as surrogate for identifying insulin resistance in apparently healthy subjects. Metab Syndr Relat Disord. 2008;6(4):299–304. doi:10.1089/met.2008.0034.

[3] Liu X, Tan Z, Huang Y, Zhao H, Liu M, Yu P, et al. Relationship between the triglyceride-glucose index and risk of cardiovascular diseases and mortality in the general population: a systematic review and meta-analysis. Cardiovasc Diabetol. 2022;21(1):124. doi:10.1186/s12933-022-01546-0.

[4] Xia X, Chen S, Tian X, Xu Q, Zhang Y, Zhang X, et al. Association of triglyceride-glucose index and its related parameters with atherosclerotic cardiovascular disease: evidence from a 15-year follow-up of Kailuan cohort. Cardiovasc Diabetol. 2024;23(1):208. doi:10.1186/s12933-024-02290-3.

[5] Che B, Zhong C, Zhang R, Pu L, Zhao T, Zhang Y, et al. Triglyceride-glucose index and triglyceride to high-density lipoprotein cholesterol ratio as potential cardiovascular disease risk factors: an analysis of UK Biobank data. Cardiovasc Diabetol. 2023;22(1):34. doi:10.1186/s12933-023-01762-2.

[6] Lopez-Jaramillo P, Gomez-Arbelaez D, Martinez-Bello D, et al. Association of the triglyceride glucose index as a measure of insulin resistance with mortality and cardiovascular disease in populations from five continents (PURE study): a prospective cohort study. Lancet Healthy Longev. 2023;4(1):e23–e33. doi:10.1016/S2666-7568(22)00247-1.

[7] Yin JL, Yang J, Song XJ, Qin X, Chang YJ, Chen X, et al. Triglyceride-glucose index and health outcomes: an umbrella review of systematic reviews with meta-analyses of observational studies. Cardiovasc Diabetol. 2024;23(1):177. doi:10.1186/s12933-024-02241-y.

[8] Liang S, Wang X, Zhang X, Liu Z, Bai Y, Chen H, et al. Triglyceride-glucose index and coronary artery disease: a systematic review and meta-analysis of risk, severity, and prognosis. Cardiovasc Diabetol. 2023;22(1):170. doi:10.1186/s12933-023-01906-4.

[9] Muhammad IF, Bao X, Nilsson PM, Zaigham S. Triglyceride-glucose (TyG) index is a predictor of arterial stiffness, incidence of diabetes, cardiovascular disease, and all-cause and cardiovascular mortality: a longitudinal two-cohort analysis. Front Cardiovasc Med. 2023;9:1035105. doi:10.3389/fcvm.2022.1035105.

[10] Nayak SS, Kuriyakose D, Polisetty LD, et al. Diagnostic and prognostic value of triglyceride glucose index: a comprehensive evaluation of meta-analysis. Cardiovasc Diabetol. 2024;23:310. doi:10.1186/s12933-024-02392-y.

[11] Mirjalili SR, Soltani S, Heidari Meybodi Z, Marques-Vidal P, Kraemer A, Sarebanhassanabadi M. An innovative model for predicting coronary heart disease using triglyceride-glucose index: a machine learning-based cohort study. Cardiovasc Diabetol. 2023;22:200. doi:10.1186/s12933-023-01939-9.

[12] Mirjalili S, Soltani S, Meybodi Z, Marques-Vidal P, Firouzabadi A, Eshraghi A, et al. Which surrogate insulin resistance indices best predict coronary artery disease? A machine learning approach. Cardiovasc Diabetol. 2024;23(1):214. doi:10.1186/s12933-024-02306-y.

[13] Yu J, Yang X, Deng Y, Krefman AE, Pool LR, Zhao L, et al. Incorporating longitudinal history of risk factors into atherosclerotic cardiovascular disease risk prediction using deep learning. Sci Rep. 2024;14(1):2554. doi:10.1038/s41598-024-51685-5.

[14] von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med. 2007;4(10):e296. doi:10.1371/journal.pmed.0040296.

[15] World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA. 2013;310(20):2191–2194. doi:10.1001/jama.2013.281053.

[16] Martin-Morales A, Yamamoto M, Inoue M, Vu T, Dawadi R, Araki M. Predicting cardiovascular disease mortality: leveraging machine learning for comprehensive assessment of health and nutrition variables. Nutrients. 2023;15(18):3937. doi:10.3390/nu15183937.

[17] Miranda E, Adiarto S, Bhatti FM, Zakiyyah AY, Aryuni M, Bernando C. Understanding arteriosclerotic heart disease patients using electronic health records: a machine learning and Shapley Additive exPlanations approach. Healthc Inform Res. 2023;29(3):228–238. doi:10.4258/hir.2023.29.3.228.

[18] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765–4774.

[19] Gazit T, Mann H, Gaber S, et al. A novel, machine-learning model for prediction of short-term ASCVD risk over 90 and 365 days. Front Digit Health. 2024;6:1485508. doi:10.3389/fdgth.2024.1485508.

[20] Chinese Joint Committee for the Revision of the Lipid Management Guidelines. Chinese guideline for lipid management (primary care edition 2024) [in Chinese]. Chin Gen Pract. 2024;27(20):2429–2436. doi:10.12114/j.issn.1007-9572.2024.0005.

[21] Arnett DK, Blumenthal RS, Albert MA, et al. 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Circulation. 2019;140(11):e596–e646. doi:10.1161/CIR.0000000000000678.

[22] Ren Q, Huang Y, Liu Q, Chu T, Li G, Wu Z. Association between triglyceride glucose-waist height ratio index and cardiovascular disease in middle-aged and older Chinese individuals: a nationwide cohort study. Cardiovasc Diabetol. 2024;23(1):247. doi:10.1186/s12933-024-02336-6.

[23] Gan YY, Zhai L, Liao Q, Huo RR. Triglyceride-glucose index as a mediator of body mass index and cardiovascular disease in middle-aged and older Chinese adults: a nationally representative longitudinal cohort study. Front Endocrinol (Lausanne). 2024;15:1431087. doi:10.3389/fendo.2024.1431087.

[24] Cui H, Liu Q, Zhao Z, et al. Interacting and joint effects of triglyceride-glucose index and blood pressure on cardiovascular diseases risk: a prospective cohort study. Diabetol Metab Syndr. 2024;16:188. doi:10.1186/s13098-024-01433-6.

[25] Khan SS, Coresh J, Pencina MJ, et al. Development and validation of the American Heart Association's PREVENT equations. Circulation. 2024;149(6):430–449. doi:10.1161/CIRCULATIONAHA.123.067626.

[26] Sun, Wu, Guo, Chen, Song, et al. Triglyceride-glucose index as a superior marker of insulin resistance for predicting long-term major adverse cardiovascular events following coronary artery bypass grafting in China. Sci Rep. 2025;15(1):6450. doi:10.1038/s41598-025-87967-9.

[27] Zhang Y, Cai X, Zhong Z, Ning F, Guo N. Predictive value of the triglyceride-glucose product index for the long-term prognosis of patients with acute coronary syndrome undergoing emergency PCI [in Chinese]. Chin J Arterioscler. 2023;31(8):687–696. doi:10.20039/j.cnki.1007-3949.2023.08.006.

[28] Wang Y, Zheng Y. Correlation between the triglyceride-glucose-body mass index and the severity of coronary lesions in patients with premature coronary heart disease [in Chinese]. Chin J Arterioscler. 2025;33(10):870–876. doi:10.20039/j.cnki.1007-3949.2025.10.006.

## Tables

**Table 1. Comparison of baseline characteristics of the study population.**

| Variable | Total (N=1047) | Non-high-risk (n=589) | High-risk (n=458) | P-value |
|---|---|---|---|---|
| Sex (male) | 399(38.1%) | 168(28.5%) | 231(50.4%)b | <0.001 |
| Age | 64.00 (59.00, 70.00) | 61.00 (56.00, 67.00) | 67.00 (61.00, 72.00)b | <0.001 |
| Albumin abnormality | 13(1.2%) | 8(1.4%) | 5(1.1%) | 0.916 |
| LDL-C abnormality | 547(52.2%) | 283(48.0%) | 264(57.6%)b | 0.003 |
| Triglyceride (continuous) | 1.31 (0.89, 1.84) | 1.06 (0.73, 1.49) | 1.66 (1.22, 2.24)b | <0.001 |
| HDL-C abnormality | 476(45.5%) | 309(52.5%) | 167(36.5%)b | <0.001 |
| Total cholesterol abnormality | 465(44.4%) | 252(42.8%) | 213(46.5%) | 0.254 |
| ALT abnormality | 86(8.2%) | 33(5.6%) | 53(11.6%)b | <0.001 |
| AST abnormality | 75(7.2%) | 30(5.1%) | 45(9.8%)b | 0.005 |

| | | | | |
|---|---|---|---|---|
| Hematocrit abnormality | 132(12.6%) | 70(11.9%) | 62(13.5%) | 0.481 |
| Uric acid abnormality | 145(13.8%) | 65(11.0%) | 80(17.5%)b | 0.004 |
| Glucose (continuous) | 5.21 (4.66, 5.91) | 5.00 (4.52, 5.61) | 5.53 (4.87, 6.46)b | <0.001 |
| Platelet abnormality | 47(4.5%) | 27(4.6%) | 20(4.4%) | 0.986 |
| Albumin/globulin abnormality | 10(1.0%) | 6(1.0%) | 4(0.9%) | 1.000 |
| Monocyte percentage abnormality | 18(1.7%) | 8(1.4%) | 10(2.2%) | 0.436 |
| eGFR abnormality | 490(46.8%) | 229(38.9%) | 261(57%)b | <0.001 |
| RDW-CV abnormality | 129(12.3%) | 79(13.4%) | 50(10.9%) | 0.261 |
| Lymphocyte percentage abnormality | 75(7.2%) | 27(4.6%) | 48(10.5%)b | <0.001 |
| Lymphocyte count abnormality | 69(6.6%) | 29(4.9%) | 40(8.7%)a | 0.019 |
| MCV abnormality | 44(4.2%) | 38(6.5%) | 6(1.3%)b | <0.001 |
| MCH abnormality | 38(3.6%) | 30(5.1%) | 8(1.7%)b | 0.007 |
| MCHC abnormality | 85(8.1%) | 55(9.3%) | 30(6.6%) | 0.127 |
| Basophil ratio abnormality | 61(5.8%) | 24(4.1%) | 37(8.1%)b | 0.009 |
| BMI | 24.60 (22.50, 26.70) | 23.80 (21.70, 25.76) | 25.44 (23.70, 27.58)b | <0.001 |
| Blood pressure abnormality | 508(48.5%) | 226(38.4%) | 282(61.6%)b | <0.001 |
| Waist-to-hip ratio abnormality | 402(38.4%) | 158(26.8%) | 244(53.3%)b | <0.001 |
| Homocysteine abnormality | 231(22.1%) | 105(17.8%) | 126(27.5%)b | <0.001 |
| TyG index | 8.60 (8.23, 8.99) | 8.34 (8.03, 8.69) | 8.91 (8.60, 9.28)b | <0.001 |

*Note: Continuous variables were compared using the independent-samples t-test or Mann–Whitney U test, and categorical variables using the $\chi^2$ test or Fisher's exact test; aP<0.05, bP<0.01 (vs. non-high-risk group).*

**Table 2. Results of univariate logistic regression.**

| Variable | Type | OR (95% CI) | P-value | Significant |
|---|---|---|---|---|
| Sex (male) | binary | 2.550 (1.974-3.294) | <0.001 | Yes |
| Age | continuous | 1.075 (1.058-1.092) | <0.001 | Yes |
| Albumin abnormality | binary | 0.802 (0.260-2.467) | 0.700 | No |
| LDL-C abnormality | binary | 1.471 (1.151-1.882) | 0.002 | Yes |
| Triglyceride (continuous) | continuous | 3.778 (3.012-4.739) | <0.001 | Yes |
| HDL-C abnormality | binary | 0.520 (0.405-0.668) | <0.001 | Yes |
| Total cholesterol abnormality | binary | 1.163 (0.909-1.486) | 0.229 | No |
| ALT abnormality | binary | 2.205 (1.402-3.469) | <0.001 | Yes |

| | | | | |
|---|---|---|---|---|
| AST abnormality | binary | 2.030 (1.257-3.278) | 0.004 | Yes |
| Hematocrit abnormality | binary | 1.161 (0.805-1.674) | 0.424 | No |
| Uric acid abnormality | binary | 1.706 (1.199-2.428) | 0.003 | Yes |
| Glucose (continuous) | continuous | 1.526 (1.366-1.705) | <0.001 | Yes |
| Platelet abnormality | binary | 0.950 (0.526-1.717) | 0.866 | No |
| Albumin/globulin abnormality | binary | 0.856 (0.240-3.052) | 0.811 | No |
| Monocyte percentage abnormality | binary | 1.621 (0.635-4.141) | 0.313 | No |
| eGFR abnormality | binary | 2.083 (1.625-2.670) | <0.001 | Yes |
| RDW-CV abnormality | binary | 0.791 (0.542-1.154) | 0.224 | No |
| Lymphocyte percentage abnormality | binary | 2.437 (1.495-3.971) | <0.001 | Yes |
| Lymphocyte count abnormality | binary | 1.848 (1.127-3.030) | 0.015 | Yes |
| MCV abnormality | binary | 0.192 (0.081-0.459) | <0.001 | Yes |
| MCH abnormality | binary | 0.331 (0.150-0.730) | 0.006 | Yes |
| MCHC abnormality | binary | 0.681 (0.428-1.081) | 0.103 | No |
| Basophil ratio abnormality | binary | 2.069 (1.219-3.511) | 0.007 | Yes |
| BMI | continuous | 1.241 (1.186-1.299) | <0.001 | Yes |
| Blood pressure abnormality | binary | 2.574 (2.002-3.308) | <0.001 | Yes |
| Waist-to-hip ratio abnormality | binary | 3.110 (2.401-4.029) | <0.001 | Yes |
| Homocysteine abnormality | binary | 1.749 (1.303-2.348) | <0.001 | Yes |
| TyG index | continuous | 13.505 (9.473-19.254) | <0.001 | Yes |

*Note: P<0.05 indicates statistical significance.*

**Table 3. Results of binary logistic regression for the basic model.**

| **Variable** | **B** | **SE** | **OR (95% CI)** | **P-value** |
|---|---|---|---|---|
| Sex (male) | 0.7647 | 0.1644 | 2.148(1.557-2.965) | <0.001 |
| Age | 0.0555 | 0.0096 | 1.057(1.037-1.077) | <0.001 |
| LDL-C abnormality | 0.3765 | 0.1477 | 1.457(1.091-1.947) | 0.011 |
| HDL-C abnormality | -0.4513 | 0.1500 | 0.637(0.475-0.854) | 0.003 |
| ALT abnormality | 0.7063 | 0.2694 | 2.026(1.195-3.436) | 0.009 |
| Lymphocyte percentage abnormality | 0.9179 | 0.3055 | 2.504(1.376-4.557) | 0.003 |
| MCV abnormality | -2.2896 | 0.5175 | 0.101(0.037-0.279) | <0.001 |
| Basophil ratio | 0.9482 | 0.3310 | 2.581(1.349-4.937) | 0.004 |

| | | | | |
|---|---|---|---|---|
| abnormality | | | | |
| BMI | 0.1179 | 0.0264 | 1.125(1.068-1.185) | <0.001 |
| Blood pressure abnormality | 0.8193 | 0.1498 | 2.269(1.692-3.043) | <0.001 |
| Waist-to-hip ratio abnormality | 1.0143 | 0.1626 | 2.757(2.005-3.792) | <0.001 |

*Note: P<0.05 indicates statistical significance.*

**Table 4. Results of binary logistic regression for the basic + triglyceride–glucose model.**

| Variable | B | OR (95% CI) | P-value |
|---|---|---|---|
| Sex (male) | 0.8848 | 2.423(1.681-3.492) | <0.001 |
| Age | 0.0621 | 1.064(1.042-1.087) | <0.001 |
| LDL-C abnormality | 0.4036 | 1.497(1.081-2.074) | 0.015 |
| HDL-C abnormality | -0.2511 | 0.778(0.556-1.088) | 0.143 |
| ALT abnormality | 0.5812 | 1.788(1.018-3.142) | 0.043 |
| Lymphocyte percentage abnormality | 0.7284 | 2.072(0.999-4.295) | 0.050 |
| MCV abnormality | -2.1481 | 0.117(0.038-0.356) | <0.001 |
| Basophil ratio abnormality | 1.0845 | 2.958(1.446-6.049) | 0.003 |
| BMI | 0.0466 | 1.048(0.987-1.112) | 0.126 |
| Blood pressure abnormality | 0.7959 | 2.216(1.593-3.084) | <0.001 |
| Waist-to-hip ratio abnormality | 0.8405 | 2.318(1.613-3.329) | <0.001 |
| TyG index | 2.5036 | 12.226(8.049-18.571) | <0.001 |

*Note: P<0.05 indicates statistical significance.*

**Table 5. Results of binary logistic regression for the basic + triglyceride + glucose model.**

| Variable | B | OR (95% CI) | P-value |
|---|---|---|---|
| Sex (male) | 0.8097 | 2.247(1.569-3.218) | <0.001 |
| Age | 0.0618 | 1.064(1.042-1.086) | <0.001 |
| LDL-C abnormality | 0.3928 | 1.481(1.078-2.034) | 0.015 |
| HDL-C abnormality | -0.3021 | 0.739(0.534-1.024) | 0.069 |
| ALT abnormality | 0.6298 | 1.877(1.069-3.296) | 0.028 |
| Lymphocyte percentage abnormality | 0.8739 | 2.396(1.202-4.778) | 0.013 |
| MCV abnormality | -2.2314 | 0.107(0.036-0.323) | <0.001 |
| Basophil ratio abnormality | 1.1311 | 3.099(1.541-6.234) | 0.002 |
| BMI | 0.0813 | 1.085(1.024-1.149) | 0.006 |
| Blood pressure abnormality | 0.8174 | 2.265(1.642-3.124) | <0.001 |
| Waist-to-hip ratio abnormality | 0.8294 | 2.292(1.611-3.260) | <0.001 |
| Triglyceride | 1.3158 | 3.728(2.853-4.870) | <0.001 |
| Glucose | 0.2178 | 1.243(1.107-1.396) | <0.001 |

*Note: P<0.05 indicates statistical significance.*

**Table 6. Comparison of the area under the curve among the three models.**

| Model | No. of variables | AUC (95% CI) | ΔAUC | AIC |
|---|---|---|---|---|
| Basic model | 11 | 0.8034(0.7771-0.8281) | — | 1 134.3 |
| Basic + TyG | 12 | 0.8786(0.8575-0.8989) | +0.0752 | 929.0 |
| Basic + TG + GLU | 13 | 0.8636(0.8414-0.8839) | +0.0602 | 974.0 |

**Table 7. Results of statistical tests.**

| Test | Statistic | Sub-item 1 | P-value / Sub-item 2 |
|---|---|---|---|
| DeLong test (TyG vs TG+GLU) | diff=0.0150 | Z=3.656 | <0.001 |
| LR test (basic→TyG) | $\chi^2$=207.3 | df=1 | <0.001 |
| LR test (basic→TG+GLU) | $\chi^2$=164.3 | df=2 | <0.001 |
| NRI(TyG) | 0.158 | Event: 0.083 | Non-event: 0.075 |
| NRI(TG+GLU) | 0.126 | Event: 0.055 | Non-event: 0.071 |
| IDI(TyG) | 0.155 | | |
| IDI(TG+GLU) | 0.121 | | |

*Note: P<0.05 indicates statistical significance.*

**Table 8. Six algorithms and their key parameters.**

| Algorithm | Principle | Key parameters | Standardized |
|---|---|---|---|
| Binary logistic regression | Traditional statistical method based on the log-linear model | L2 regularization, C=1.0, lbfgs solver | Yes |
| Random forest | Ensemble learning, Bagging | 200 trees, max depth 10 | No |
| Support vector machine | Maximum-margin classification, RBF kernel | C=1.0,gamma=scale | Yes |
| K-nearest neighbors | Distance-based lazy learning | K=7, distance weighting | Yes |
| Gradient boosting machine | Ensemble learning, Boosting | 150 trees, learning rate 0.1 | No |
| Naive Bayes | Probabilistic classifier based on Bayes' theorem | Gaussian naive Bayes | Yes |

**Table 9. Area under the curve for 5-fold cross-validation.**

| Algorithm | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Mean±SD |
|---|---|---|---|---|---|---|
| Binary logistic regression | 0.8902 | 0.8761 | 0.8666 | 0.8933 | 0.8678 | 0.8788±0.0124 |
| Random forest | 0.8790 | 0.8428 | 0.8493 | 0.8687 | 0.8321 | 0.8544±0.0191 |
| Support vector | 0.8639 | 0.8567 | 0.8523 | 0.8605 | 0.8281 | 0.8523±0.0142 |

| | | | | | | |
|---|---|---|---|---|---|---|
| machine | | | | | | |
| K-nearest neighbors | 0.8101 | 0.8255 | 0.8026 | 0.7921 | 0.8048 | 0.8070±0.0122 |
| Gradient boosting machine | 0.8579 | 0.8571 | 0.8399 | 0.8716 | 0.8394 | 0.8532±0.0136 |
| Naive Bayes | 0.8330 | 0.7948 | 0.8270 | 0.8657 | 0.8089 | 0.8259±0.0269 |

**Table 10. Performance metrics in the validation set.**

| Algorithm | Accuracy | Sensitivity | Specificity | Precision | F1-score | AUC (95% CI) |
|---|---|---|---|---|---|---|
| Binary logistic regression | 0.7778 | 0.7319 | 0.8136 | 0.7537 | 0.7426 | 0.8495(0.8080-0.8898) |
| Random forest | 0.7492 | 0.7391 | 0.7571 | 0.7034 | 0.7208 | 0.8401(0.7950-0.8837) |
| Support vector machine | 0.7556 | 0.6812 | 0.8136 | 0.7402 | 0.7094 | 0.8386(0.7936-0.8830) |
| K-nearest neighbors | 0.7238 | 0.6667 | 0.7684 | 0.6917 | 0.6790 | 0.8069(0.7548-0.8556) |
| Gradient boosting machine | 0.7429 | 0.6957 | 0.7797 | 0.7111 | 0.7033 | 0.8340(0.7895-0.8770) |
| Naive Bayes | 0.7556 | 0.7101 | 0.7910 | 0.7259 | 0.7179 | 0.8029(0.7535-0.8490) |

**Table 11. Confusion matrix in the validation set.**

| Algorithm | TP | FP | FN | TN |
|---|---|---|---|---|
| Binary logistic regression | 101 | 33 | 37 | 144 |
| Random forest | 102 | 43 | 36 | 134 |
| Support vector machine | 94 | 33 | 44 | 144 |
| K-nearest neighbors | 92 | 41 | 46 | 136 |
| Gradient boosting machine | 96 | 39 | 42 | 138 |
| Naive Bayes | 98 | 37 | 40 | 140 |

**Table 12. Ranking of SHapley Additive exPlanations (SHAP) feature importance.**

| Rank | Feature | Mean \|SHAP\| |
|---|---|---|
| 1 | TyG index | 1.1143 |
| 2 | Sex (male) | 0.4886 |
| 3 | Waist-to-hip ratio | 0.4868 |
| 4 | Age | 0.3665 |
| 5 | Blood pressure abnormality | 0.3283 |

| | | |
|---|---|---|
| 6 | LDL-C abnormality | 0.1848 |
| 7 | MCV abnormality | 0.1643 |
| 8 | Basophil ratio abnormality | 0.1519 |
| 9 | ALT abnormality | 0.1014 |
| 10 | BMI | 0.0991 |
| 11 | HDL-C abnormality | 0.0753 |
| 12 | Lymphocyte percentage abnormality | 0.0449 |

**Table 13. Standardized coefficients of binary logistic regression.**

| **Rank** | **Feature** | **Standardized coefficient** | **\|Coefficient\|** |
|---|---|---|---|
| 1 | TyG index | 1.5499 | 1.5499 |
| 2 | Sex (male) | 0.5069 | 0.5069 |
| 3 | Waist-to-hip ratio | 0.4945 | 0.4945 |
| 4 | Age | 0.4730 | 0.4730 |
| 5 | MCV abnormality | -0.3758 | 0.3758 |
| 6 | Blood pressure abnormality | 0.3280 | 0.3280 |
| 7 | Basophil ratio abnormality | 0.2703 | 0.2703 |
| 8 | ALT abnormality | 0.2623 | 0.2623 |
| 9 | LDL-C abnormality | 0.1841 | 0.1841 |
| 10 | BMI | 0.1206 | 0.1206 |
| 11 | Lymphocyte percentage abnormality | 0.0840 | 0.0840 |
| 12 | HDL-C abnormality | -0.0751 | 0.0751 |

## Figures

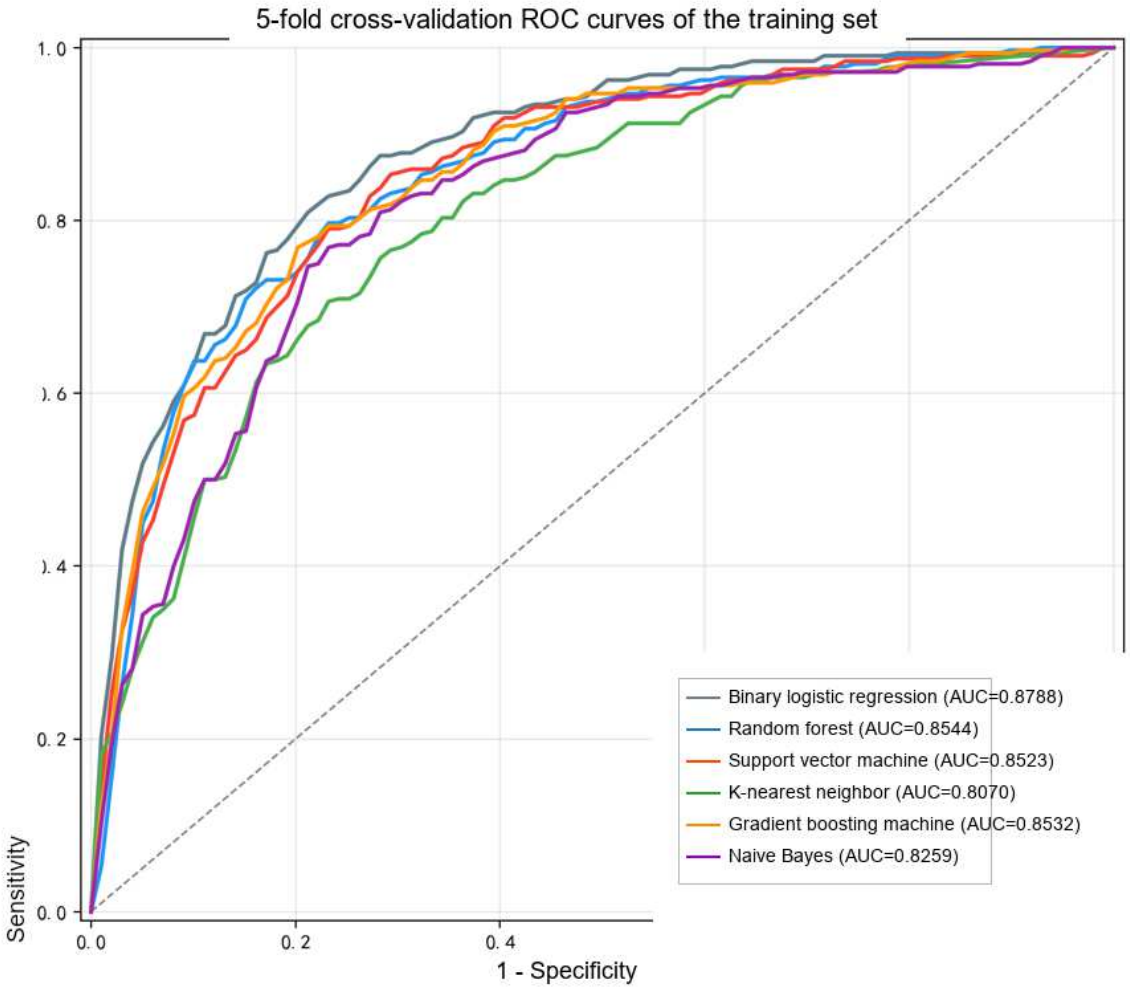



Figure 1. Comparison of receiver operating characteristic (ROC) curves among the three models.

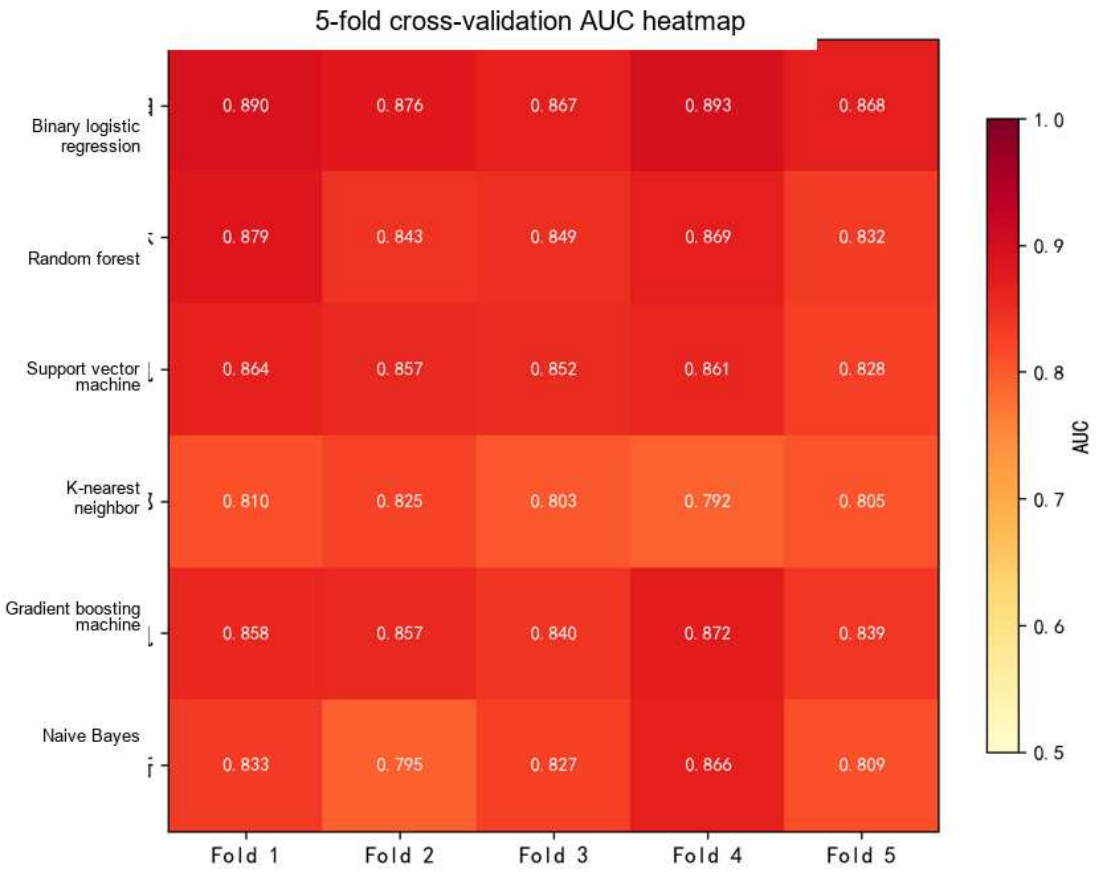



Figure 2. Heatmap of the area under the curve (AUC) for 5-fold cross-validation.

*[PLACEHOLDER – figure image not provided in source file]*

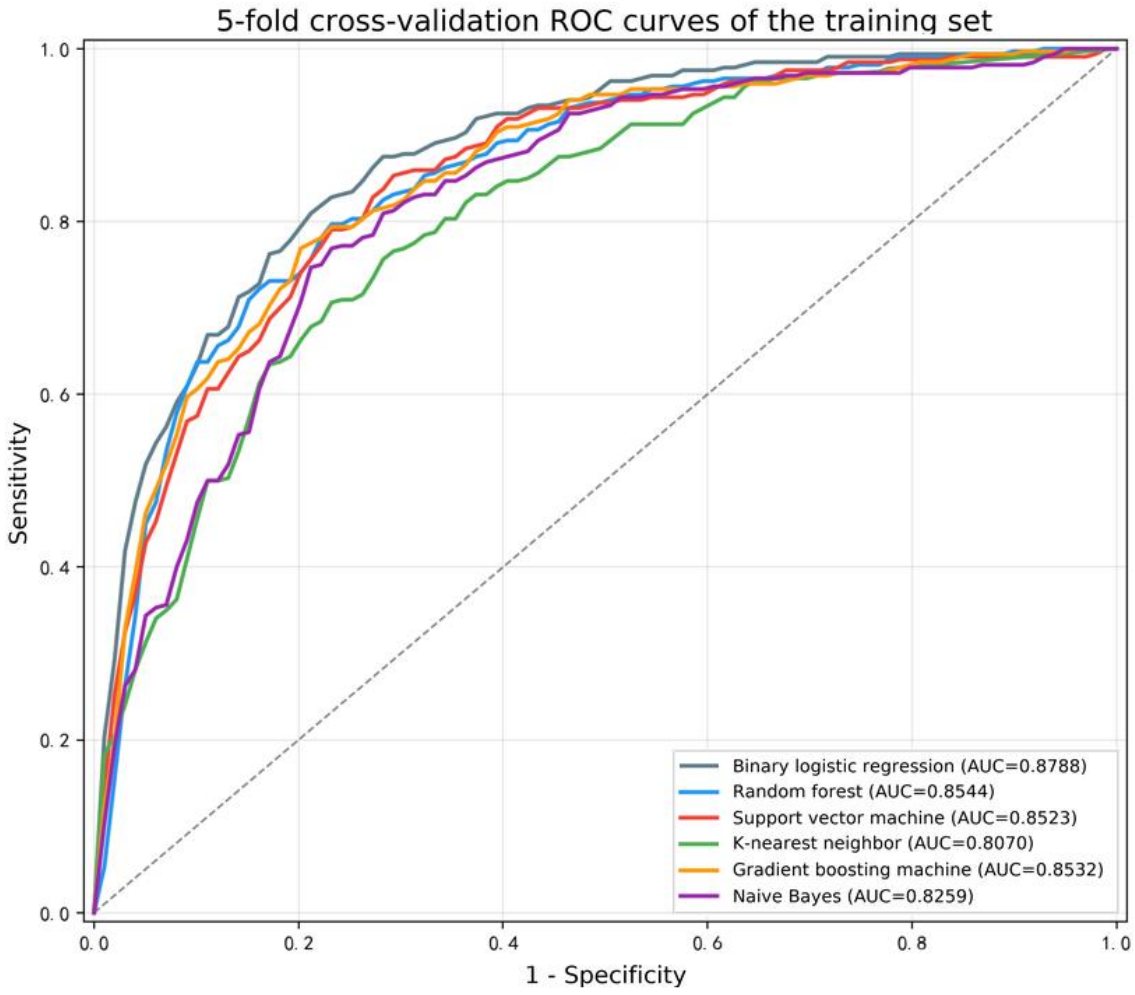
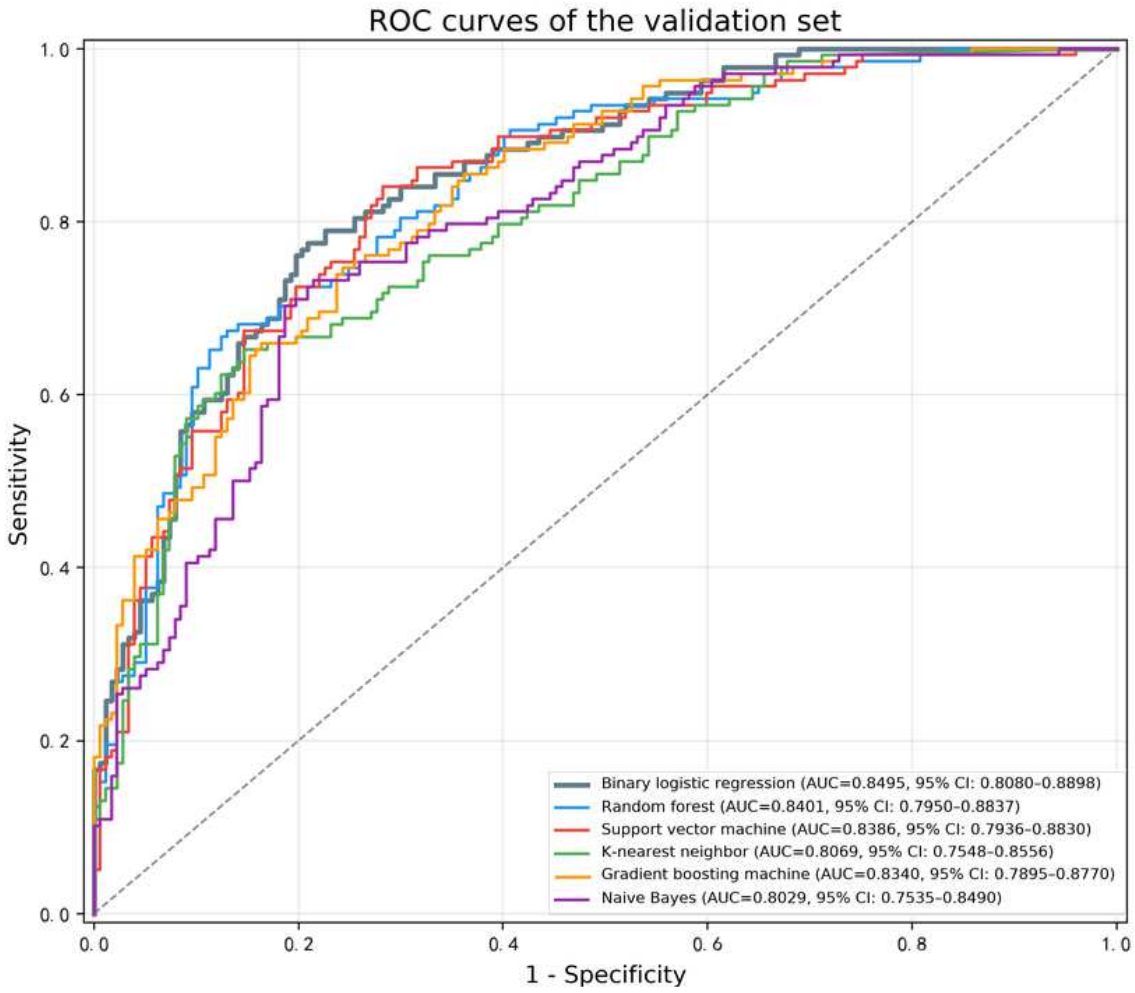



Figure 3. Receiver operating characteristic (ROC) curves for 5-fold cross-validation in the training set.
[IMAGE TO BE INSERTED BY AUTHOR]




*[PLACEHOLDER – figure image not provided in source file]*

Figure 4. Receiver operating characteristic (ROC) curves in the validation set. [IMAGE TO BE INSERTED BY AUTHOR]

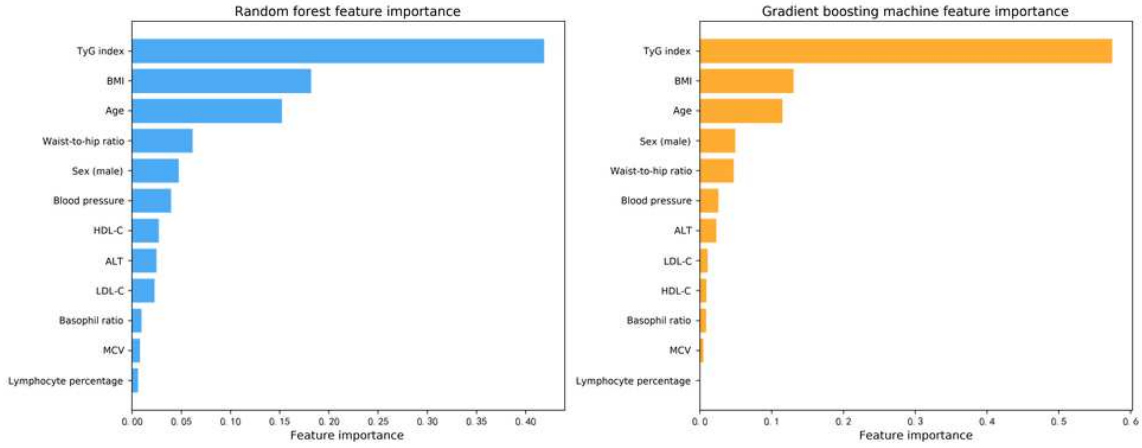



Figure 5. Feature importance of random forest and gradient boosting machine.

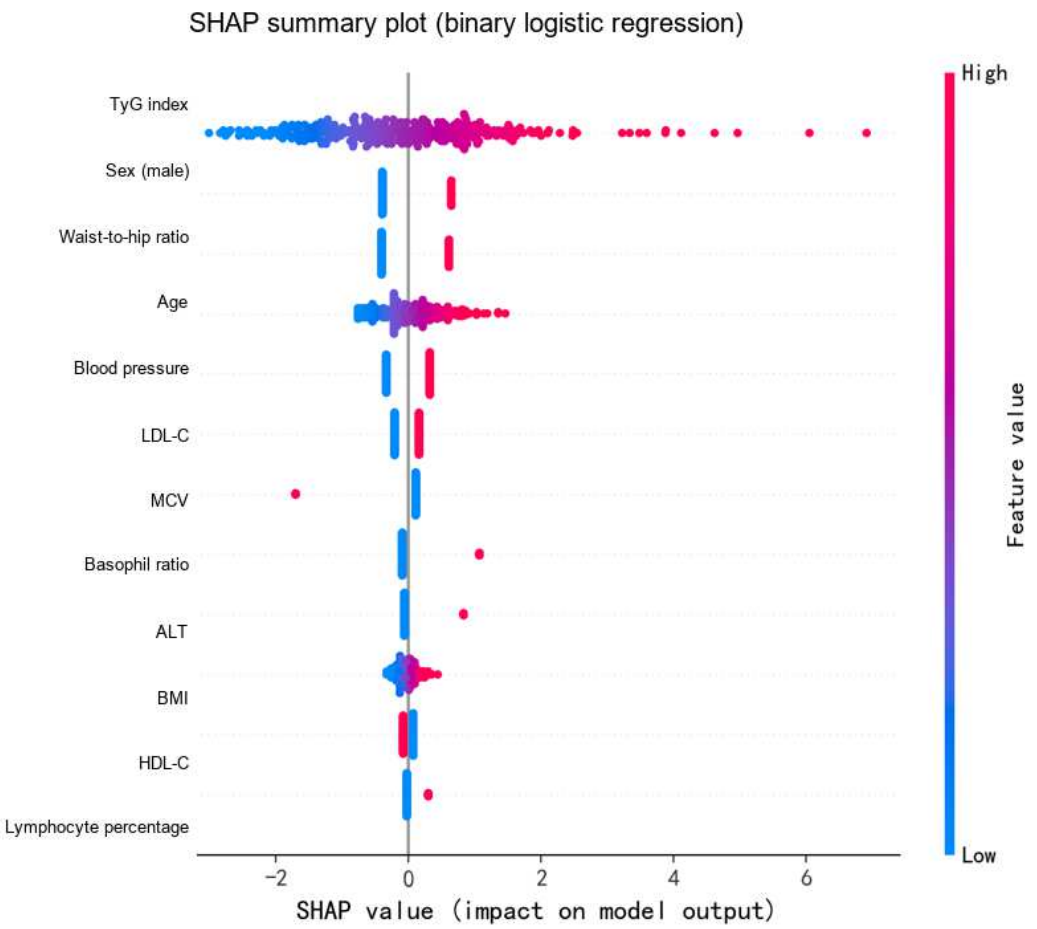



Figure 6. SHapley Additive exPlanations (SHAP) summary plot.

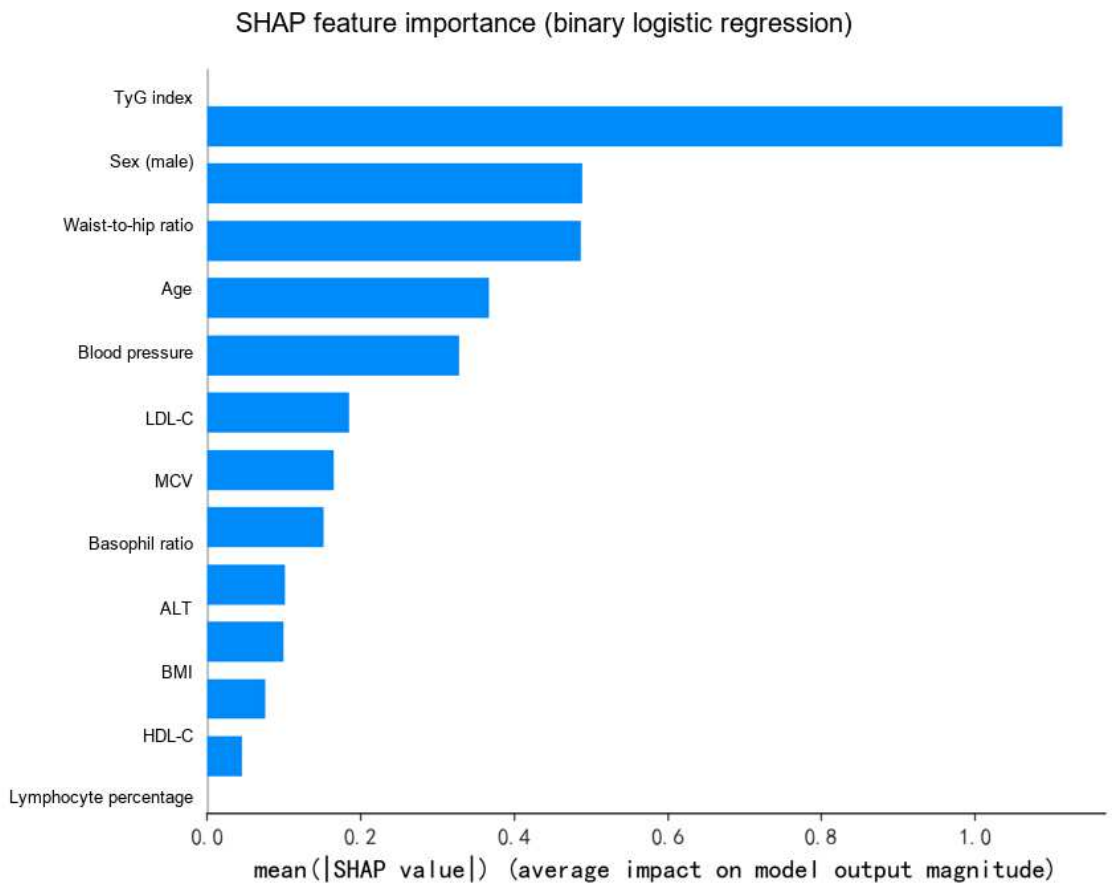



Figure 7. Bar chart of SHapley Additive exPlanations (SHAP) feature importance.

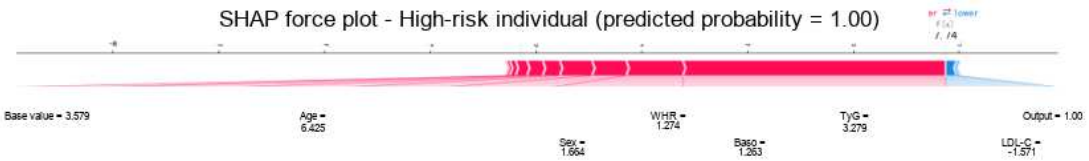



Figure 8. SHapley Additive exPlanations (SHAP) force plot (high-risk individual).

*[PLACEHOLDER – figure image not provided in source file]*

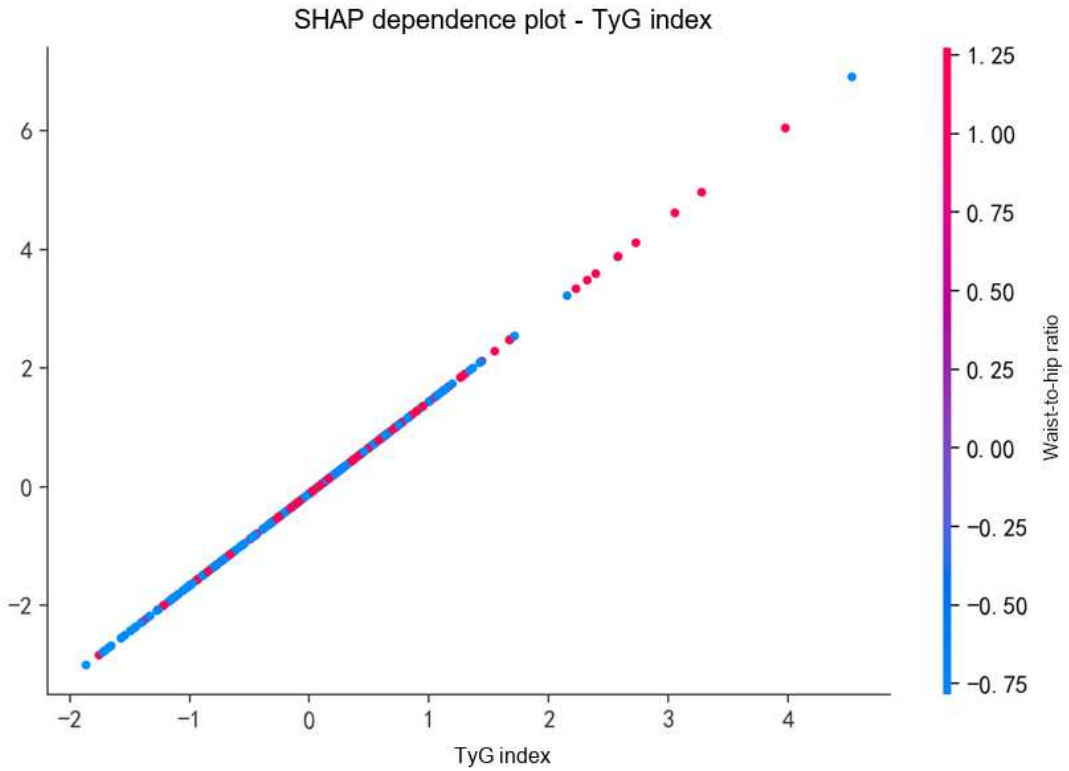



Figure 9. SHapley Additive exPlanations (SHAP) dependence plot (triglyceride–glucose index). [IMAGE TO BE INSERTED BY AUTHOR]