

**Supplementary table 1** The results of feature screening using the Boruta algorithm and LASSO regression.

| <b>Boruta</b>       |         |           |        |        |          |           | <b>LASSO</b>        |             |
|---------------------|---------|-----------|--------|--------|----------|-----------|---------------------|-------------|
| Feature             | meanImp | medianImp | minImp | maxImp | normHits | Decision  | Feature             | coefficient |
| Age                 | 20.55   | 20.502    | 17.761 | 22.728 | 1        | Confirmed | Age                 | 0.0148      |
| Sex                 | 4.671   | 4.616     | 2.719  | 8.025  | 0.98     | Confirmed | Sex                 | 0.5967      |
| Education level     | 5.492   | 5.429     | 2.263  | 6.974  | 0.98     | Confirmed | Education level     | -0.0079     |
| Smoking status      | 2.769   | 2.783     | 0.302  | 5.395  | 0.606    | Tentative | Smoking status      | -0.078      |
| Alcohol consumption | 0.352   | 0.426     | -1.517 | 1.319  | 0        | Rejected  | Alcohol consumption | 0.1427      |
| Residence           | 1.869   | 1.848     | -1.061 | 4.499  | 0.354    | Tentative | Residence           | -0.0393     |
| Physical activity   | 4.117   | 4.191     | 1.014  | 6.644  | 0.899    | Confirmed |                     |             |
| Hypertension        | 19.547  | 19.457    | 17.724 | 21.494 | 1        | Confirmed | Hypertension        | 0.3313      |
| Diabetes            | 7.523   | 7.439     | 5.258  | 9.078  | 1        | Confirmed | Diabetes            | 0.1166      |
| SBP                 | 27.22   | 27.046    | 24.738 | 30.098 | 1        | Confirmed |                     |             |
| DBP                 | 23.768  | 23.77     | 21.708 | 26.386 | 1        | Confirmed | DBP                 | 0.0259      |
| FPG                 | 16.562  | 16.483    | 13.945 | 19.476 | 1        | Confirmed |                     |             |
| AIP                 | 26.908  | 26.918    | 24.616 | 29.156 | 1        | Confirmed | AIP                 | 0.713       |
| Non-HDL-C           | 9.989   | 9.795     | 7.862  | 13.146 | 1        | Confirmed | Non-HDL-C           | -0.0571     |
| AC                  | 22.843  | 22.896    | 20.817 | 24.906 | 1        | Confirmed |                     |             |
| CRI-I               | 22.786  | 23.047    | 20.774 | 25.529 | 1        | Confirmed |                     |             |
| CRI-II              | 19.093  | 19.055    | 17.399 | 21.015 | 1        | Confirmed | CRI-II              | 0.3522      |
| LCI                 | 21.061  | 20.943    | 18.501 | 23.519 | 1        | Confirmed | LCI                 | -0.0047     |
| TG/HDLC             | 27.711  | 27.777    | 25.098 | 29.912 | 1        | Confirmed | TG/HDLC             | -0.0184     |
| HbA1c               | 21.658  | 21.623    | 18.833 | 24.495 | 1        | Confirmed | HbA1c               | 0.262       |
| UA                  | 13.914  | 13.857    | 10.704 | 16.857 | 1        | Confirmed | UA                  | 0.0029      |
| CALPCPRO            | 2.158   | 2.184     | -0.037 | 4.647  | 0.414    | Tentative |                     |             |
| CALPCFAT            | 5.096   | 5.143     | 2.175  | 7.211  | 0.98     | Confirmed | CALPCFAT            | 0.0014      |
| CALPCCAR            | 5.512   | 5.465     | 3.755  | 8.112  | 0.98     | Confirmed | CALPCCAR            | -0.001      |

**Abbreviations:** LASSO, least absolute shrinkage and selection operator; SBP, systolic blood pressure; DBP, diastolic blood pressure; FPG, fasting plasma glucose; AIP, atherogenic index of plasma; Non-HDL-C, non-high-density lipoprotein cholesterol; AC, atherogenic coefficient; CRI-I, castelli risk index I; CRI-II, castelli risk index II; LCI, lipid cumulative index; TG/HDLC, the ratio of triglycerides to high-density lipoprotein cholesterol; HbA1c, glycated hemoglobin; UA, uric acid; CALPCPRO, energy comes from protein; CALPCFAT, energy comes from fat; CALPCCAR, energy comes from carbohydrates.

**Supplementary table 2** The results of the Log\_Loss, Brier score, and Hosmer-Lemeshow test for the training and the testing set.

| Model    | Log_Loss<br>(training) | Log_Loss<br>(testing) | Brier<br>(training) | Brier<br>(testing) | Hosmer–Lemeshow<br>(training) | Hosmer–Lemeshow<br>(testing) |
|----------|------------------------|-----------------------|---------------------|--------------------|-------------------------------|------------------------------|
| LR       | 0.5726                 | 0.5787                | 0.1146              | 0.1954             | 0.105                         | 0.396                        |
| SVM      | 0.5352                 | 0.5840                | 0.1772              | 0.1982             | $P < 0.01$                    | 0.116                        |
| LightGBM | 0.4018                 | 0.5824                | 0.1227              | 0.1985             | $P < 0.01$                    | 0.088                        |
| XGBoost  | 0.3202                 | 0.5902                | 0.0895              | 0.2004             | $P < 0.01$                    | $P < 0.01$                   |
| RF       | 0.2608                 | 0.5865                | 0.0614              | 0.2006             | $P < 0.01$                    | 0.132                        |
| NB       | 1.1630                 | 1.2413                | 0.2643              | 0.2653             | $P < 0.01$                    | $P < 0.01$                   |

**Abbreviations:** LR: logistic regression; SVM: support vector machine; LightGBM, Light Gradient Boosting Machine; XGBoost: extreme gradient boosting; RF: random forest; NB: naive Bayes; Log\_Loss, logarithmic loss.

**Supplementary table 3** Delong test results of AUC differences between various machine learning models in the test set.

| Model1   | Model2   | AUC1   | AUC2   | Z statistic | P value | Significant |
|----------|----------|--------|--------|-------------|---------|-------------|
| LR       | LightGBM | 0.7683 | 0.7593 | 1.6356      | 0.1019  | No          |
| LR       | XGBoost  | 0.7683 | 0.7581 | 1.8015      | 0.0716  | No          |
| LR       | SVM      | 0.7683 | 0.7571 | 2.2379      | 0.0252  | Yes         |
| LR       | RF       | 0.7683 | 0.7524 | 2.8177      | 0.0048  | Yes         |
| LR       | NB       | 0.7683 | 0.7365 | 5.2840      | 0.0000  | Yes         |
| LightGBM | XGBoost  | 0.7593 | 0.7581 | 0.3172      | 0.7511  | No          |
| LightGBM | SVM      | 0.7593 | 0.7571 | 0.4182      | 0.6758  | No          |
| LightGBM | RF       | 0.7593 | 0.7524 | 1.4767      | 0.1398  | No          |
| LightGBM | NB       | 0.7593 | 0.7365 | 3.1428      | 0.0017  | Yes         |
| XGBoost  | SVM      | 0.7581 | 0.7571 | 0.2014      | 0.8404  | No          |
| XGBoost  | RF       | 0.7581 | 0.7524 | 1.2580      | 0.2084  | No          |
| XGBoost  | NB       | 0.7581 | 0.7365 | 2.9161      | 0.0035  | Yes         |
| SVM      | RF       | 0.7571 | 0.7524 | 0.8811      | 0.3783  | No          |
| SVM      | NB       | 0.7571 | 0.7365 | 2.9315      | 0.0034  | Yes         |
| RF       | NB       | 0.7524 | 0.7365 | 2.4721      | 0.0134  | Yes         |

**Abbreviations:** LR: logistic regression; SVM: support vector machine; LightGBM, Light Gradient Boosting Machine; XGBoost: extreme gradient boosting; RF: random forest; NB: naive Bayes.

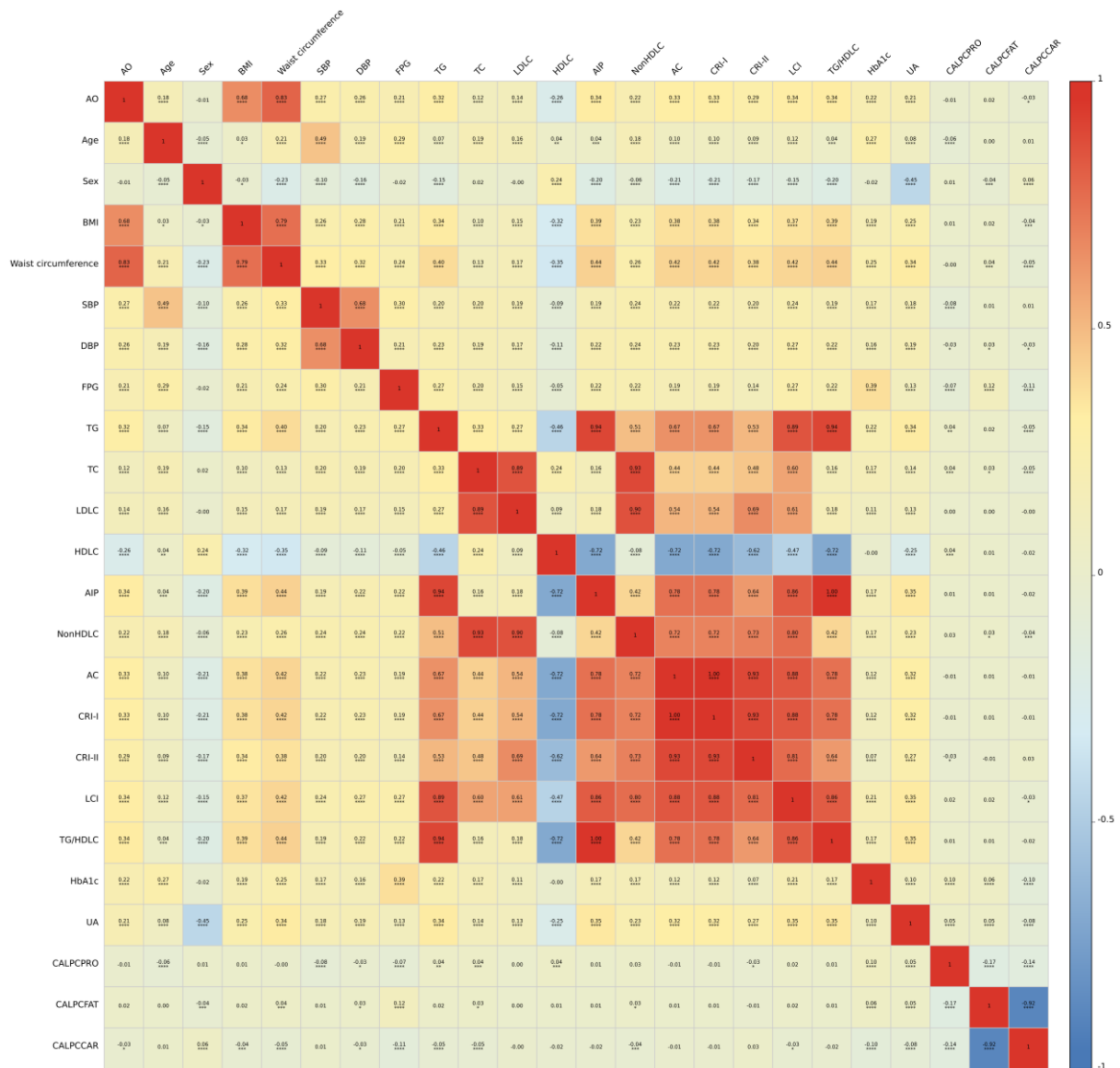


**Figure S1.** Correlation heatmap illustrating pairwise Pearson correlations among candidate continuous variables used in the feature selection process. Pairwise Pearson correlation coefficients among candidate continuous variables are displayed. Color intensity represents the magnitude and direction of correlations (red: positive correlation; blue: negative correlation). This analysis was conducted to systematically evaluate relationships among predictors and assess potential multicollinearity prior to feature selection. As expected, several lipid-derived indices sharing common lipid components showed strong correlations.

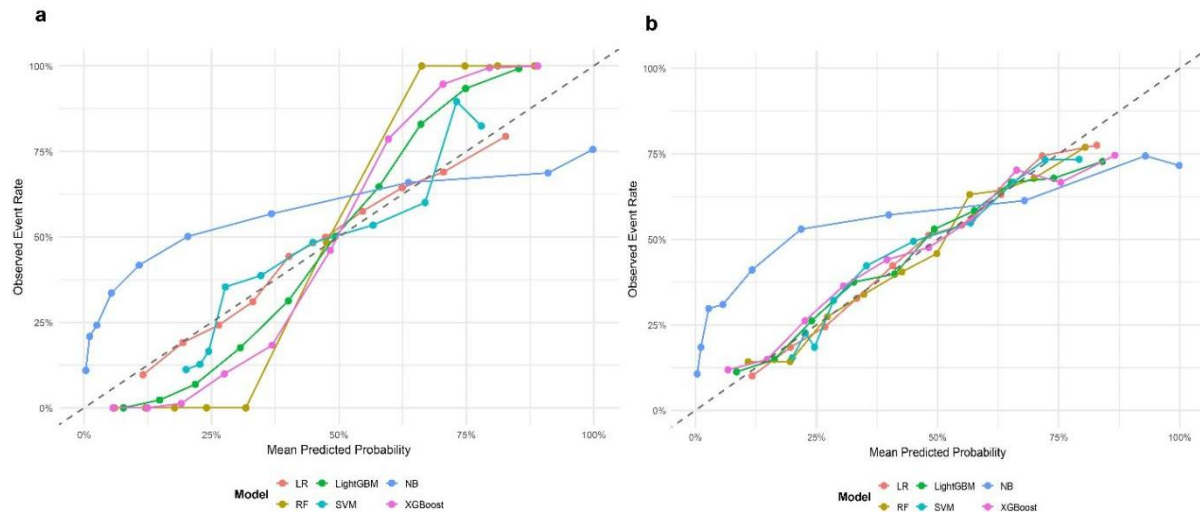


**Figure S2.** Calibration curves assessing the agreement between predicted probabilities and observed outcomes for all machine learning models in the (a) training set and (b) test set. The dashed diagonal line represents perfect calibration.

**Abbreviations:** LR, logistic regression; SVM, support vector machine; RF, random forest; LightGBM, Light Gradient Boosting Machine; NB, Naive Bayes; XGBoost, eXtreme Gradient Boosting.

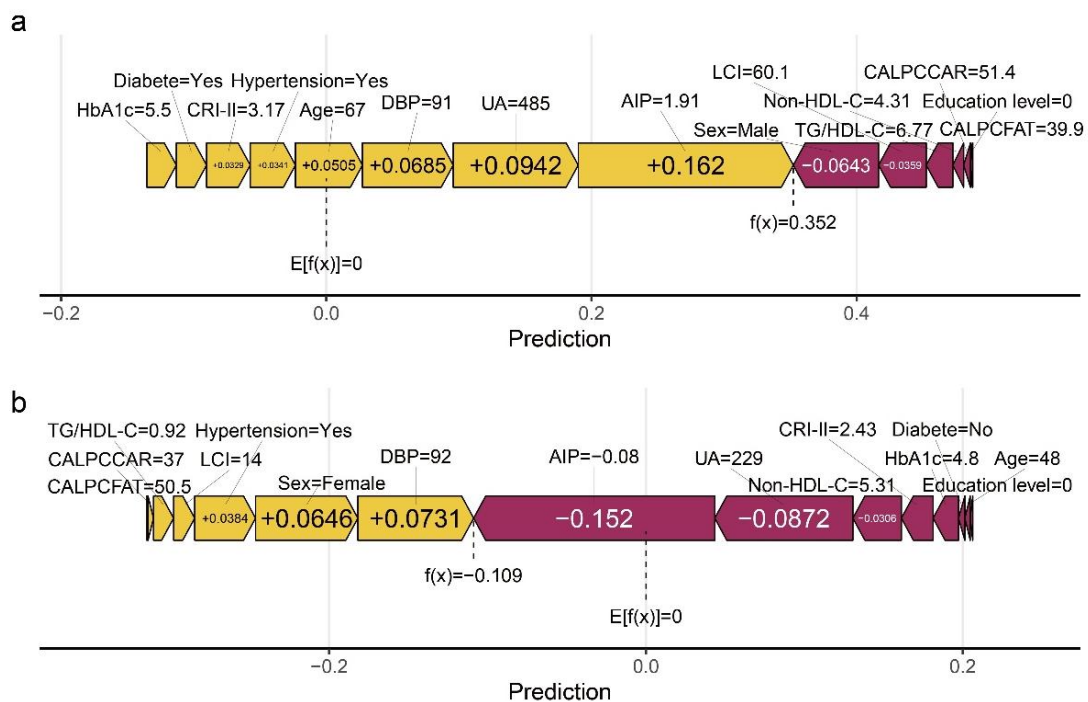


**Figure S3.** The SHAP analysis provides individual explanations of the model. (a) and (b) Force plots display the contribution of each feature's SHAP value to the prediction for two different individuals.

**Abbreviations:** AIP, atherogenic index of plasma; CALPCCAR, energy comes from carbohydrates; CALPCFAT, energy comes from fat; CRI-II, castelli risk index II; DBP, diastolic blood pressure; HbA1c, glycated hemoglobin; LCI, lipid cumulative index; Non-HDL-C, non-high-density lipoprotein cholesterol; TG/HDL-C, triglycerides to high-density lipoprotein cholesterol ratio; UA, uric acid.

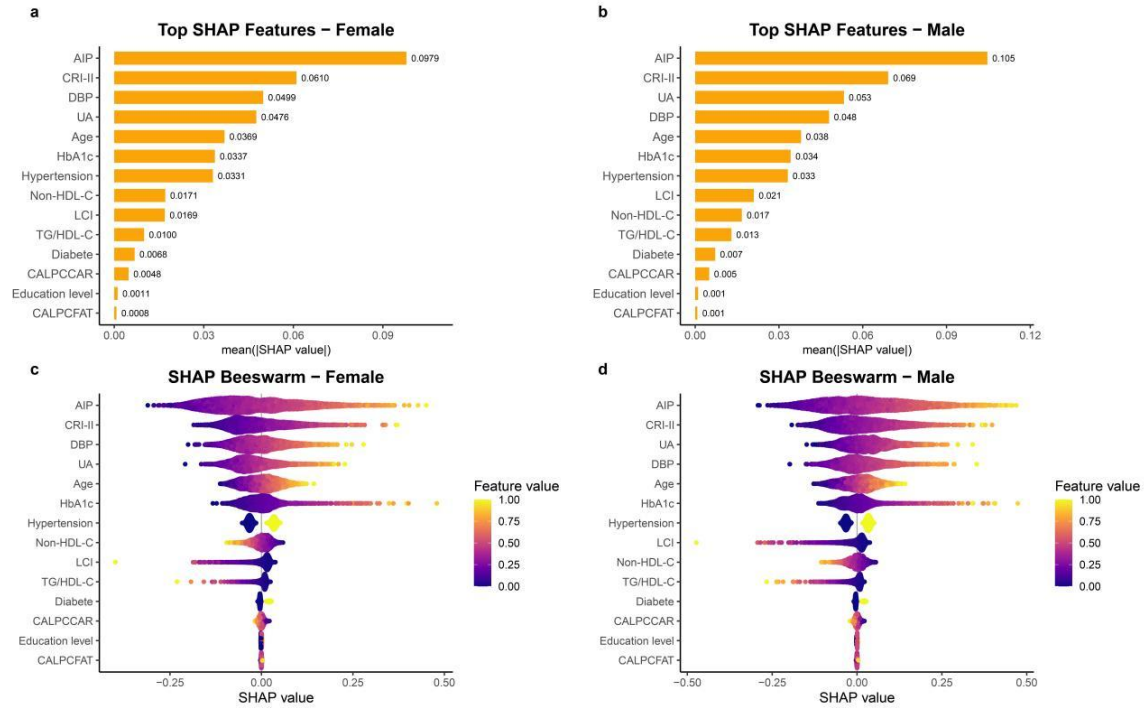


**Figure S4.** The model was globally explained using SHAP analysis stratified by sex. (a) and (b) The summary bar charts depict the overall importance of each feature, ranked by the mean absolute SHAP value, for the female and male subgroups, respectively. (c) and (d) The beeswarm plots illustrate the distribution and directional impact of individual feature values on the predicted probability in the female and male subgroups, respectively.

**Abbreviations:** AIP, atherogenic index of plasma; CALPCCAR, energy comes from carbohydrates; CALPCFAT, energy comes from fat; CRI-II, castelli risk index II; DBP, diastolic blood pressure; HbA1c, glycated hemoglobin; LCI, lipid cumulative index; Non-HDL-C, non-high-density lipoprotein cholesterol; TG/HDL-C, triglycerides to high-density lipoprotein cholesterol ratio; UA, uric acid.

# Supplement material: iThenticate check report

---

|            |      |                |                     |
|------------|------|----------------|---------------------|
| WORD COUNT | 4472 | TIME SUBMITTED | 11-MAR-2026 05:25PM |
|            |      | PAPER ID       | 120728141           |

15%

SIMILARITY INDEX

PRIMARY SOURCES

|   |                                                                                                                                                                                                                                                                                                        |                 |
|---|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------|
| 1 | <a href="https://cardiab.biomedcentral.com">cardiab.biomedcentral.com</a><br>Internet                                                                                                                                                                                                                  | 89 words — 2%   |
| 2 | <a href="https://www.mdpi.com">www.mdpi.com</a><br>Internet                                                                                                                                                                                                                                            | 86 words — 2%   |
| 3 | <a href="https://assets-eu.researchsquare.com">assets-eu.researchsquare.com</a><br>Internet                                                                                                                                                                                                            | 51 words — 1%   |
| 4 | <a href="https://www.frontiersin.org">www.frontiersin.org</a><br>Internet                                                                                                                                                                                                                              | 35 words — 1%   |
| 5 | <a href="https://ouci.dntb.gov.ua">ouci.dntb.gov.ua</a><br>Internet                                                                                                                                                                                                                                    | 25 words — < 1% |
| 6 | <a href="https://www.researchsquare.com">www.researchsquare.com</a><br>Internet                                                                                                                                                                                                                        | 25 words — < 1% |
| 7 | Saif Ghabisha, Noaman Almashraki, Qasem Alyhari, Ahmed Ateik et al. "Development and Validation of a Preoperative Risk Score for Difficult Laparoscopic Cholecystectomy: A Retrospective Study in Resource-Limited Settings", Springer Science and Business Media LLC, 2026<br>Crossref Posted Content | 24 words — < 1% |
| 8 | <a href="https://www.preprints.org">www.preprints.org</a><br>Internet                                                                                                                                                                                                                                  | 24 words — < 1% |

9

Internet

21 words — &lt; 1%

- 
- 10 Yang Qu, Jie Lian, Tianli Liu, Ying Xu et al. "Binary and Ternary Classification Prediction for Breast Cancer and Breast Sclerosing Adenosis With Interpretable Artificial Intelligence From Clinical and Imaging Features: A Retrospective, Diagnostic Accuracy Cohort Study", Cancer Innovation, 2026

Crossref

- 
- 11 doctorpenguin.com

Internet

20 words — &lt; 1%

- 
- 12 www.panafrican-med-journal.com

Internet

20 words — &lt; 1%

- 
- 13 elibrary.trf.or.th

Internet

19 words — &lt; 1%

- 
- 14 www.researchgate.net

Internet

19 words — &lt; 1%

- 
- 15 cmjournal.biomedcentral.com

Internet

18 words — &lt; 1%

- 
- 16 public-pages-files-2025.frontiersin.org

Internet

17 words — &lt; 1%

- 
- 17 B. G. Nagaraja, S. Kannadhasan. "Information and Communication Systems", CRC Press, 2026

Publications

16 words — &lt; 1%

- 
- 18 Lian Li, Liuyun Wu, Yin Wang, Hulin Wang, Xingyue Zheng, Lizhu Han, Qinan Yin, Xingwei Wu, Yuan Bian. "Development of a venous thromboembolism risk prediction model for patients with primary membranous nephropathy based on machine learning", Frontiers in Pharmacology, 2025

19 J. Anitha Menon, Benita Veronica, Ben Milbourn, Mithra S. Gnana Sanga. "Sustainable Well-being - Advancing Health through Psychology, Medicine, and Intelligent Technologies", CRC Press, 2026 13 words — < 1%

Publications

20 Muzaffer Bayhatun, Sadettin Selçuk Baysal. "The Association of Atherogenic Indices with Coronary Slow Flow: Evidence from a Large Cohort Study", Diagnostics, 2026 12 words — < 1%

Crossref

21 Zhenyuan Yu, Huan Tang, Wenjia Ye, Zixin Gu, Yu Fu, Rong Yao, Ying Guan, Yonghong Shen. "Development and External Validation of a Nurse-Friendly Machine Learning Model for Early Identification of Intradialytic Hypotension in ICU Patients Receiving Renal Replacement Therapy", Springer Science and Business Media LLC, 2025 12 words — < 1%

Crossref Posted Content

22 Chunqiang Gu, Dongmei Tang, Fang Yuan. "Comparative Evaluation of Obesity-Related Indices for Predicting Incident Hypertension: Evidence from Chinese and UK Longitudinal Cohorts With Machine Learning Interpretation", Springer Science and Business Media LLC, 2026 11 words — < 1%

Crossref Posted Content

23 Sanhu Wang, Xiqiang Gao, Fuxing Li. "Extended Oral Glucose Tolerance Test Unmasks Intermediate Metabolic Phenotypes in Polycystic Ovary Syndrome", Diabetes, Metabolic Syndrome and Obesity, 2026 11 words — < 1%

Crossref

24 [link.springer.com](https://link.springer.com) 11 words — < 1%

Internet

25 [www.ncbi.nlm.nih.gov](https://www.ncbi.nlm.nih.gov) 11 words — < 1%

Internet

11 words — < 1%

26 Lan Lan, Yuenan Ni, Yubei Zhou, Fengming Luo. "Associations of body roundness index and body mass index with obstructive sleep apnea: evidence from two cohorts study", *Frontiers in Neurology*, 2025

Crossref

27 Wenlong Ding, Tao Li, Caoyang Fang, Xin-Xin Zhang, Enyang Wang. "Association of non-high-density lipoprotein cholesterol to high-density lipoprotein cholesterol ratio with coronary heart disease: Establishment and validation of a clinical nomogram model", *Medicine*, 2025

Crossref

28 Yuancheng He, Xiaojuan Zhang, Yuexin Mei, Deng Qianyun, Xiuqing Zhang, Yuehua Chen, Jie Li, Zhou Meng, Yuehong Wei. "Construction of a machine learning-based prediction model for unfavorable discharge outcomes in patients with ischemic stroke", *Heliyon*, 2024

Crossref

29 [www.biochemjournal.com](http://www.biochemjournal.com)

Internet

30 [www.dovepress.com](http://www.dovepress.com)

Internet

31 [www.nature.com](http://www.nature.com)

Internet

32 Junxiang Cheng, Qia Liu, Zhouli Qiu, Mengyi Pan, Jie Song, Yu Zhao, Niuniu Dong, Suping Li, Shifan Han. "Dietary management improves sleep quality in patients with metabolic syndrome: the mediating roles of metabolic, inflammatory, and oxidative stress changes", *Frontiers in Nutrition*, 2025

Crossref

- 
- 33 Kenan Toprak, Mustafa Kaplangöray, Mesut Karataş, Ayten Dursun et al. "Atherogenic Combined Index: Validation of a Novel Predictive Lipid Biomarker for the Presence and Severity of Coronary Artery Disease", Archives of Medical Research, 2024  
Crossref 8 words — < 1%
- 
- 34 Md. Shahidul Islam. "Organic Acids in Diabetes - Applications and Mechanisms", CRC Press, 2025  
Publications 8 words — < 1%
- 
- 35 Tianfeng Zhang, Chenghua Wang, Zhenghui Wang, Long Zhao, Hao Xia, Yiming Wang, Chunyang Yin. "Associations between various lipid indices and coronary collateral circulation in patients with acute ST-segment elevation myocardial infarction: a cross-sectional study", International Journal of Cardiology Cardiovascular Risk and Prevention, 2026  
Crossref 8 words — < 1%
- 
- 36 [bmcmedinformdecismak.biomedcentral.com](https://bmcmedinformdecismak.biomedcentral.com)  
Internet 8 words — < 1%
- 
- 37 [bmcpublichealth.biomedcentral.com](https://bmcpublichealth.biomedcentral.com)  
Internet 8 words — < 1%
- 
- 38 [researchonline.jcu.edu.au](https://researchonline.jcu.edu.au)  
Internet 8 words — < 1%
- 
- 39 [shrishikshayatancollege.org](https://shrishikshayatancollege.org)  
Internet 8 words — < 1%
- 
- 40 [www.pjoes.com](https://www.pjoes.com)  
Internet 8 words — < 1%
- 
- 41 [www.science.gov](https://www.science.gov)  
Internet 8 words — < 1%

42 Eugene Lin, Yu-Ting Yan, Mu-Hong Chen, Albert C. Yang, Po-Hsiu Kuo, Shih-Jen Tsai. "Gene clusters linked to insulin resistance identified in a genome-wide study of the Taiwan Biobank population", Nature Communications, 2025

7 words — < 1%

Crossref

43 L. Wilhelmsen. "Factors associated with reaching 90years of age: a study of men born in 1913 in Gothenburg, Sweden : Factors associated with reaching age 90 years", Journal of Internal Medicine, 04/2011

7 words — < 1%

Crossref

44 Mudassir Hussain Tahir, Mahmoud A.A. Ibrahim, Shaban R.M. Sayed, Denis Magero, Anthony Pembere. "Dielectric constant prediction of polymers for organic solar cells and generation of library of new organic compounds", Journal of Solid State Chemistry, 2025

7 words — < 1%

Crossref

45 Yazeed Alshuweishi, Salihah H. Khobrani, Muath Alsaidan, Tahani M. Alharthi, Mohannad G. Abdelgader, Abdulaziz M. Almuqrin. "Lipid-Derived Cardiometabolic Indices in Normouricemic and Hyperuricemic Adults: A Retrospective Cross-Sectional Association Study", Healthcare, 2025

7 words — < 1%

Crossref

46 Binqi Li, Xin Zhou, Yang Liu, Yue Zhang, Yiming Mu. "Remnant Cholesterol is More Strongly Associated with Arterial Stiffness than Traditional Lipids and Lipid Ratios in the General Chinese Population", Journal of Atherosclerosis and Thrombosis, 2023

6 words — < 1%

Crossref

47 Chae Hong Lim, Sang-Won Um, Hong Kwan Kim, Yong Soo Choi, Hong Ryul Pyo, Myung-Ju Ahn, Joon Young Choi. "18F-Fluorodeoxyglucose Positron Emission

6 words — < 1%

48

Wei-Meng Cai, Dong Rui, Li-Bo Zhao, Ying-Hui Gao, Zhe Zhao, Xin Xue, Tian-Jiao Li, Ting-Yu Nie, Yao Ma, Lin Liu. "SHAP Analysis of Machine Learning-Based Carotid Atherosclerosis Prediction of Older Adults with OSA, a Multicenter Study", Nature and Science of Sleep, 2025

6 words — < 1%

Crossref

41  
1 An Explainable Machine Learning Method for Identifying Key Atherogenic Lipid Biomarkers in  
2 Abdominal Obesity Among the Southern Chinese Population

3 **Abstract**

4 Background Abdominal obesity (AO) significantly contributes to cardiometabolic diseases and poses  
5 an increasing public health challenge. Composite lipid-derived indices have been proposed as simple  
6 indicators of atherogenic dyslipidemia, but their relative importance for identifying AO at  
7 population-level remains incompletely understood. This study evaluated the link between various  
8 atherogenic lipid indices and AO, identifying key lipid predictors through interpretable machine  
9 learning methods.

45  
10 Methods: This cross-sectional study analyzed 5,612 adults from a population-based survey in  
11 Guangdong Province, China. Atherogenic lipid indices encompass the atherogenic index of plasma  
12 (AIP), non-HDL cholesterol (non-HDL-C), atherogenic coefficient (AC), Castelli risk indices I and  
13 II (CRI-I and CRI-II), lipid composite index (LCI), and the triglyceride-to-HDL cholesterol ratio  
14 (TG/HDL-C). Feature selection was conducted using Boruta and LASSO methods prior to machine  
15 learning modeling. Model performance was evaluated using the area under the receiver operating  
16 characteristic curve (AUC), accuracy, and F1 score. SHapley Additive exPlanations (SHAP) analysis  
17 was applied to quantify feature importance.

18 Results Individuals with AO were generally older and more likely to exhibit lower education levels,  
19 reduced physical activity, cardiometabolic comorbidities, and unfavorable metabolic profiles. Feature  
20 selection identified 15 key predictors. Logistic regression demonstrated the most stable performance  
21 (training AUC: 0.767; testing AUC: 0.768; accuracy: 0.712; F1 score: 0.667) with good calibration  
22 and clinical utility. SHAP analysis consistently identified AIP, sex, CRI-II, uric acid, and diastolic  
23 blood pressure as the most influential predictors of AO.

24 **Conclusion** Composite lipid indices, particularly AIP and CRI-II, are strongly associated with of AO  
25 and may serve as practical indicators for identifying individuals at elevated metabolic risk. Because  
26 these indices are derived from routinely measured lipid parameters, they may support scalable  
27 approaches for metabolic risk screening and monitoring in both clinical and community health  
28 settings.

29 **Keywords:** Abdominal obesity; Atherogenic Index of Plasma; Dyslipidemias; Machine learning;  
30 Risk Assessment

31 **Introduction**

Abdominal obesity (AO), marked by excessive fat in the abdominal area, significantly increases the risk of cardiometabolic conditions such as diabetes and cardiovascular disease [1–4]. In China, the prevalence of AO has increased substantially in recent decades, reaching approximately 46.9% in 2015 [5] and affecting an estimated 229.2 million adults aged over 40 years by 2020 [6]. This growing burden highlights the need for improved strategies to identify individuals at increased metabolic risk.

Guangdong Province, one of the most economically developed regions in southern China, has experienced rapid urbanization and lifestyle transitions that may influence metabolic health. Previous studies have reported strong associations between AO and metabolic abnormalities in this region, particularly dyslipidemia [7, 8]. Identifying lipid-related biomarkers associated with AO may therefore help improve risk identification and targeted prevention strategies.

Atherogenic indices combine data from both atherogenic and anti-atherogenic lipoproteins, potentially offering a more thorough evaluation of lipid metabolism compared to individual lipid parameters. The atherogenic index of plasma (AIP), non-high-density lipoprotein cholesterol (non-HDL-C) and coronary risk indices (CRI-I and CRI-II) are linked to various metabolic disorders, such as visceral adiposity, insulin resistance, non-alcoholic fatty liver disease (NAFLD), and cardiovascular outcomes. However, their relative importance in relation to AO remains incompletely understood.

We hypothesized that atherogenic lipid indices may be associated with AO and may serve as informative biomarkers for identifying individuals at higher metabolic risk. Using a population-based cross-sectional dataset from Guangdong Province, this study applied interpretable machine learning (ML) with SHapley Additive exPlanations (SHAP) [13] to identify key lipid biomarkers associated with AO, examine potential sex-specific differences, and quantify the relative contribution of different predictors.

## Methods

### Population and design

This study utilized cross-sectional data from the 2015 and 2022 waves of the China Health and Nutrition Survey (CHNS). Data from the two waves were pooled for analysis. The CHNS, a nationally representative survey initiated in 1989, is an ongoing population-based study conducted in collaboration between the Chinese Center for Disease Control and Prevention and international

research partners. Administered at five- to ten-year intervals, the survey employs a multistage stratified cluster sampling design to ensure representation across diverse regions of China. Comprehensive details on the survey's background, objectives, design, and methodology have been previously published [14]. For this analysis, data were extracted exclusively from the Guangdong Province component of the CHNS.

A stepwise exclusion process was applied. First, participants with missing or implausible values in the core variables for outcome definition—waist circumference, lipid profile indicators, and dietary data—were excluded. The remaining dataset was then checked for completeness in all other analyzed variables (e.g., smoking status, physical activity, systolic blood pressure, and other covariates). At this stage, 24 participants with missing smoking status and 2 with extreme BMI values considered implausible ( $> 50 \text{ kg/m}^2$ ) were additionally excluded. This resulted in a complete-case dataset with no missing values. To ensure data integrity for ML modeling, a complete-case dataset was used, and no statistical imputation was performed. Ultimately, a total of 5,612 participants were retained for the final analysis. The participant selection process is illustrated in Figure 1.

## Data collection and measurement

Trained personnel collected anthropometric measurements and venous blood samples. The study evaluated sociodemographic factors such as age, gender (male/female), residence type (urban/rural), and education level, categorized into primary school or below, junior middle school, and senior high school or above. Participants were asked to report any physician-diagnosed chronic conditions, such as hypertension, diabetes, and dyslipidemia. Behavioral factors included smoking status (current smoker vs. non-smoker) and alcohol consumption (current drinker vs. non-drinker). Physical activity levels were classified as mild, moderate, or vigorous based primarily on occupational physical activity intensity. Dietary intake was evaluated using a 3-day 24-hour dietary recall method, which included one weekend day. The energy contribution percentages of carbohydrates, protein, and fat were determined using the "Chinese Food Composition Table" [15], representing each macronutrient's share of total energy intake. Because the available dietary data were limited to macronutrient energy contribution ratios, detailed food-based dietary quality indicators (e.g., dietary fiber intake, whole-grain proportion, sugar-sweetened beverages, or ultra-processed foods) were not available in the present dataset. Therefore, the dietary variables in this study primarily reflected the macronutrient composition of energy intake rather than detailed dietary quality characteristics.

Physical examinations were performed following an overnight fast. Body weight was measured without shoes using calibrated scales and recorded to the nearest 0.1 kg. Height was recorded to the nearest 0.1 cm using a vertical stadiometer. Waist circumference was recorded to the nearest 0.1 cm using a non-stretchable tape measure at the umbilical level, with participants standing upright and breathing normally. Blood pressure was measured three times after at least 5 minutes of rest, at one-minute intervals, using calibrated devices, and the average of the three readings was used in analysis. BMI was calculated as weight in kilograms divided by the square of height in meters ( $\text{kg/m}^2$ ).

Venous blood samples (8 mL) were collected to assess fasting plasma glucose, triglycerides, total cholesterol, LDL-C, HDL-C, glycated hemoglobin, and uric acid levels. All specimens were processed uniformly and analyzed in a centralized laboratory using standardized enzymatic methods.

## Assessment of abdominal obesity

104 AO was identified based on the metabolic syndrome criteria outlined in the <sup>23</sup>Chinese Guideline for  
105 the Prevention and Treatment of Type 2 Diabetes (CDS 2020) [16], using waist circumference  
106 thresholds of  $\geq 90$  cm for men and  $\geq 85$  cm for women.

## 107 Atherogenic indices

108 The following formulas were used to calculate atherogenic indices:

109 <sup>1</sup> Atherogenic index of plasma (AIP) =  $\log_{10} (TG / HDL-C)$  [12];

110 (1) Non-HDL-C = TC - HDL-C [17];

111 (2) Atherogenic coefficient (AC) = Non-HDL-C / HDL-C [18];

112 (3) Castelli's risk index-I (CRI-I) = TC / HDL-C [10];

113 (4) Castelli's risk index-II (CRI-II) = LDL-C / HDL-C [10];

114 (5) Lipoprotein combined index (LCI) =  $(TC \times TG \times LDL-C) / HDL-C$  [11];

115 <sup>42</sup> TG / HDL-C ratio = TG / HDL-C [19].

## 116 Selection of variables

117 Pearson correlation coefficients were employed to initially assess the correlation among  
118 candidate continuous variables, and a correlation heatmap was used to visualize potential  
119 relationships and detect multicollinearity among predictors.

120 Variable selection was first performed using the Boruta feature selection algorithm, which iteratively  
121 assesses feature importance via random forests by comparing each original feature's significance to  
122 that of randomly permuted "shadow features" serving as noise references. A feature is retained only if  
123 its permutation importance, expressed as a Z-score, consistently exceeds the maximum importance of  
124 all shadow features across iterations, ensuring statistically robust and reliable selection [20]. The  
125 Boruta method effectively removes irrelevant variables, improves stability in high-dimensional data,  
126 and identifies both strong and weak predictors.

127 LASSO regression was also utilized for feature selection. This approach uses an L1 penalty on  
128 regression coefficients to shrink non-informative coefficients towards zero, enhancing model  
129 interpretability and minimizing overfitting and redundancy among correlated predictors. The optimal  
130 regularization parameter  $\lambda$  was determined through ten-fold cross-validation by identifying the value  
131 associated with the minimum cross-validation error according to the minimum criterion  
132 rule. Variables with non-zero coefficients were retained. To ensure robustness and stability of the

selected predictors, only variables identified by both the Boruta and LASSO regression were included in the subsequent analytical models. This conservative strategy was used to enhance robustness and reduce the risk of including spurious predictors.

## Model development and validation

In this study, categorical variables were encoded using one-hot encoding. To prevent data leakage, model development, feature selection, and hyperparameter tuning were strictly confined to the training set, with performance evaluation conducted on a separate testing set. Following a 7:3 ratio of the dataset into training and testing subsets, the Boruta algorithm and LASSO regression were applied solely to the training subset to identify the most informative predictors. The finalized feature set derived from this process was then used to transform the hold-out testing set, ensuring that no information from the testing data influenced the feature selection process.

Six machine learning algorithms were employed to create models: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Light Gradient Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost), and Naive Bayes (NB). These algorithms encompass linear, kernel-based, ensemble, and probabilistic learning methods. Hyperparameter tuning and model optimization were conducted using ten-fold cross-validation solely on the training set. The dataset was divided into 10 mutually exclusive subsets, with 9 folds used for training and 1 for validation in each iteration; this process was repeated across all 10 folds to ensure comprehensive evaluation, minimize overfitting, and improve model generalizability. A grid search was performed over predefined hyperparameter spaces to find the optimal configuration, which was subsequently used to retrain the model on the complete training set.

## Model performance evaluation

The ML models' performance was assessed using various metrics to thoroughly evaluate both discrimination and calibration. Model discrimination was assessed using the area under the receiver operating characteristic curve (AUC), along with accuracy and F1 score. The study also reported sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).

Model calibration was evaluated using calibration curves to illustrate the alignment between predicted probabilities and observed outcomes, supplemented by Brier scores and the Hosmer-Lemeshow test. To compare AUC values across models, the DeLong testing was applied to evaluate statistical differences.

Decision curve analysis (DCA) assessed the models' net benefit over various decision thresholds. Decision curve analysis (DCA) assesses the net benefit of model-guided decisions over various threshold probabilities, facilitating comparison with the default strategies of 'treat all' or 'treat none.' This method was utilized to evaluate if its application in screening offers a significant clinical net benefit over traditional methods.

## Model interpretation

SHapley Additive exPlanations (SHAP) [13] were used to improve interpretability of the top-performing model, chosen for its superior performance on various evaluation metrics such as AUC, accuracy, and F1 score in the testing dataset. This technique quantifies the contribution of each feature by evaluating its individual impact on model output, enabling a comprehensive assessment of feature importance at both the global (dataset-level) and local (instance-level) levels.

## Statistical analysis

Descriptive analyses of participant characteristics utilized survey weights to address the CHNS's complex sampling design, while ML models were constructed with unweighted data. Continuous variables are reported as means  $\pm$  standard deviation for normal distributions and as medians with interquartile ranges for skewed distributions, while categorical variables are shown as frequencies and percentages. Independent samples t-tests or Mann–Whitney U tests were used for continuous variables, while chi-square tests or Fisher's exact tests were applied to categorical variables. All analyses were conducted in R (version 4.5.1) using established statistical packages. Statistical significance was defined as a two-tailed P-value less than 0.05.

## Results

### Baseline characteristics of participants

Table 1 details the baseline characteristics of the study population, divided into non-AO (n = 3,096) and AO (n = 2,516) groups. The overall mean age was  $52.26 \pm 14.71$  years. The AO group was older (55.22 vs. 49.85 years), had a greater percentage of individuals with primary education or less (44.9% vs. 36.7%), and engaged in less moderate-to-vigorous physical activity compared to the non-AO group. The AO group showed notably higher rates of cardiovascular and metabolic disorders, such as dyslipidemia, hypertension, and type 2 diabetes. Additionally, they displayed elevated levels of BMI, waist circumference, blood pressure, FPG, TG, TC, UA, and HbA1c, along with reduced HDL-C. All calculated atherogenic lipid indices were also significantly higher in the AO group. These findings collectively indicate that individuals with AO are more likely to present with an adverse cardiometabolic profile.

### Feature selection

Figure 2a presents the outcomes of clinical feature selection via the Boruta algorithm, categorizing features into three groups according to their importance compared to randomized 'shadow' features (blue area). Box plots depict the z-score distributions of each variable across iterative model evaluations. Features within the green area are validated as statistically significant predictors, with importance scores surpassing the shadowMax value. Variables in the yellow area show marginal significance, with scores near the shadowMax threshold, and are retained as provisional predictors pending further validation. Features within the red area are deemed irrelevant as their importance falls below the threshold set by shadow feature importance. Boruta identified 21 confirmed and 3 tentative features. Only the confirmed features were retained for subsequent intersection with LASSO results.

Subsequently, a 10-fold cross-validation and LASSO regression analysis with 1,000 iterations were performed to identify the most relevant variables. The coefficient path (Figure 2b) illustrates how coefficients shrink toward zero as the regularization penalty increases. Figure 2c illustrates the cross-validated deviance for each  $\lambda$  value, aiding in the selection of the  $\lambda$  that minimizes prediction error. In the end, 18 features were kept for additional analysis. The intersection of the top 21 features identified by Boruta and the 18 features selected by LASSO yielded 15 features for subsequent ML modeling (Figure 2d). Detailed variable screening results are available in Supplementary Table 1. Otherwise, the correlation matrix of lipid-related variables is shown in Figure S1.

### Model evaluation and comparison

A total of six ML models were developed and evaluated. Table 2 illustrates that all models exhibited moderate and similar discriminative ability, with AUC values between 0.718 and 0.767 in the training set and 0.720 to 0.768 in the testing set, reflecting consistent discrimination across datasets. LR achieved the highest AUC in the training set (0.767, 95% CI: 0.751–0.783) (Table 2 and Figures 3a and 3c) and the testing set (0.768, 95% CI: 0.746–0.791) (Table 2 and Figures 3b and 3d), followed closely by LightGBM (testing AUC = 0.759, 95% CI: 0.736–0.782) (Table 2 and Figures 3a–3d).

In the testing set, LR also demonstrated the best overall classification performance, achieving the highest accuracy (0.712) and F1 score (0.667), together with well-balanced sensitivity (0.643) and specificity (0.768) (Table 2). SVM showed a similar balance between sensitivity and specificity, whereas tree-based models (RF, LightGBM, XGBoost) exhibited slightly lower performance. The NB model was characterized by lower sensitivity (0.481) but higher specificity (0.820).

Model calibration and goodness-of-fit further differentiated performance across algorithms. LR demonstrated superior calibration performance, achieving a Brier score of 0.1146 on the training set and 0.1954 on the testing set (Table S2). In contrast, tree-based models, despite lower training-set log loss values (0.26–0.40), exhibited substantially higher log loss in the testing set (0.58–0.59), indicating overfitting (Table S2). Calibration curves supported these findings, showing close agreement between predicted and observed risks for LR across both datasets, whereas tree-based models displayed marked deviations in the testing set, particularly at moderate-to-high predicted risk levels (Figure S2).

Consistently, the Hosmer-Lemeshow testing confirmed good calibration for LR in both the training ( $P = 0.105$ ) and testing sets ( $P = 0.396$ ) (Table S2), while significant lack of fit was observed for several alternative models, especially XGBoost and RF, in the testing data ( $P < 0.01$ ) (Table S2). DeLong testing showed no significant differences in AUC values between LR and other models ( $P > 0.05$ ) (Table S3), suggesting that discrimination alone was inadequate for assessing model performance.

DCA further evaluated the potential clinical usefulness of the models in identifying individuals with AO. The LR model demonstrated a superior net benefit across various clinically relevant threshold probabilities compared to treat-all and treat-none strategies in both training and testing sets, and generally surpassed other ML models (Figure 4). In contrast, several tree-based models demonstrated lower and less stable net benefit in the testing set, which was consistent with their poorer calibration performance and reduced stability across datasets.

These findings indicate that the LR model offers a stable and interpretable framework for identifying lipid-related biomarkers linked to AO in the current dataset. Consequently, LR was chosen as the best model for further interpretation with SHAP analysis.

## Model explanation

The SHAP method was utilized to elucidate the final model's output by assessing each variable's contribution to the model's output for individual participants. This model-agnostic interpretability approach offers complementary explanations: global interpretation at the feature level and local interpretations at the instance level. Global explanations describe the overall behavior of the model across the entire dataset. The SHAP summary bar chart (Figure 5a) ranks features by average absolute SHAP values, highlighting AIP, sex, CRI-II, UA, and DBP as the top five predictors in terms of importance. Additionally, the beeswarm plot (Figure 5b) visually illustrates the direction and magnitude of each feature's impact on the predicted outcome. The model identified higher AIP levels, female sex, elevated CRI-II, and increased UA and DBP levels as factors associated with a greater predicted probability of AO.

Local explanations improve comprehension of the model's decision-making by calculating and illustrating each feature's contribution to the predicted outcome for a specific sample. The feature contributions to the predicted probability of AO for the 3rd and 8th participants are depicted in the waterfall plots (Figure 5c-5d) and force plots (Figure S3a-3b). The SHAP values and corresponding feature values reveal the direction and magnitude of each feature's impact on the prediction outcome. In Figure 5c and Figure S3a, the participant's elevated AIP (1.91), high UA level (485  $\mu\text{mol/L}$ ), and DBP of 91 mmHg resulted in notable positive contributions (+0.162, +0.0942, and +0.0685, respectively), whereas male and an LCI value of 60.1 were associated with negative contributions (-0.0643 and -0.0359, respectively). In Figure 5d and Figure S3b, the identification was positively influenced by the participant's DBP level of 92 mmHg, female gender, and history of hypertension, with SHAP values of +0.0731, +0.0646, and +0.0384, respectively. An AIP value of -0.08 and a UA level of 229  $\mu\text{mol/L}$  had significant negative impacts, with SHAP values of -0.152 and -0.0872, respectively. When SHAP analysis was conducted within sex subgroups, AIP remained the top-ranked predictor in both males and females, with CRI-II closely following (Figure S4). UA and DBP alternated as the third and fourth most influential predictors across the subgroups, while age persisted as a significant risk factor in both sexes. This pattern is largely consistent with the global feature importance ranking observed in the overall population.

## Discussion

This study applied ML approaches to explore the relationship between abdominal obesity (AO) and atherogenic dyslipidemia while accounting for demographic, lifestyle, and dietary factors. Individuals with AO were characterized by a less favorable cardiometabolic profile, consistent with well-established links between visceral adiposity and disturbances in blood pressure, glucose metabolism, and lipid regulation. Among the evaluated algorithms, LR model exhibited comparatively stable performance with discrimination and calibration. Model interpretation using SHAP indicated that AIP, sex, CRI-II, UA, and DBP were among the most influential predictors in identifying AO. These findings indicate that composite lipid indices may more comprehensively capture lipid-related metabolic disturbances linked to visceral adiposity than individual lipid parameters.

The research underscores the significance of composite lipid indices in identifying metabolic changes associated with AO. Among the evaluated indicators, AIP and CRI-II ranked among the most influential predictors in the model. AIP integrates triglyceride-rich lipoproteins and HDL-C-related clearance capacity [12], and previous studies have reported significant correlations with waist circumference and visceral adiposity indices [21-23], body fat composition [24], and epicardial adipose tissue [25]. CRI-II, which indicates the balance between LDL-C and HDL-C, complements AIP by measuring the atherogenic cholesterol burden linked to visceral obesity. Elevated CRI-II has been associated with increased visceral fat area [19], insulin resistance, and atherosclerosis risk [10]. Together, these indices capture complementary aspects of adiposity-related dyslipidemia—triglyceride metabolism and atherogenic cholesterol balance—and may therefore provide biologically plausible and population-relevant indicators of AO risk, especially in settings when direct assessment of visceral fat is not feasible.

These associations are biologically plausible due to known mechanisms connecting visceral adiposity with insulin resistance, chronic low-grade inflammation, and altered adipokine signaling [2, 34]. Insulin resistance may decrease lipoprotein lipase activity and HDL-C levels, while inflammatory cytokines like TNF- $\alpha$  and IL-6 can hinder HDL biogenesis and enhance hepatic VLDL production [9, 27-29]. These processes contribute to elevated triglyceride-rich lipoproteins and reduced HDL-mediated lipid clearance, metabolic alterations that are reflected by increases in AIP and CRI-II. Such convergent pathways may partly explain the consistent importance of these composite indices in identifying individuals with AO.

Notably, sex also emerged as an influential contributor, consistent with known sexual dimorphism in fat distribution. Men typically accumulate greater amounts of visceral adipose tissue, partly influenced by androgen-related fat partitioning, whereas premenopausal women are relatively protected by estrogen mediated metabolic regulation [30, 31]. Nevertheless, once visceral obesity develops, many of the associated lipid disturbances appear broadly similar across sexes. In this context, sex may function primarily as an upstream biological modifier influencing adipose distribution patterns, while composite lipid indices capture downstream metabolic consequences.

Dietary composition may also contribute to the observed metabolic patterns. Increasing evidence suggests that the metabolic effects of carbohydrate intake may depend not only on total quantity but also on carbohydrate quality, including dietary fiber content and glycemic characteristics [32]. In the present study, dietary intake was characterized only by the energy contribution ratio of macronutrients, without detailed information on carbohydrate subtypes or broader dietary quality indicators. Within this limitation, the inverse association observed between carbohydrate energy ratio and abdominal obesity may reflect broader dietary patterns rather than the isolated effect of carbohydrate quantity. Research indicates that diets high in whole grains and legumes, which are fiber-rich and have a low glycemic load, are linked to a reduced risk of metabolic syndrome and visceral adiposity. In Southern China, traditional dietary patterns often include mixed cereals, legumes, and relatively minimally processed staple foods, which may be broadly consistent with the observed pattern. However, because detailed measures of carbohydrate quality were unavailable in the

present dataset, these interpretations should be considered exploratory, and future studies incorporating more refined dietary quality indicators are needed to clarify these relationships.

Serum uric acid (UA) emerged as a key predictor in the model, aligning with evidence that associates hyperuricemia with metabolic dysregulation related to adiposity [37]. Mendelian randomization studies support bidirectional relationships between adiposity, lipid metabolism, and serum urate levels [38], and visceral fat accumulation itself has been implicated in increased gout risk [39]. In the context of Guangdong dietary habits, purine-rich foods such as poultry, seafood, and slow-cooked soups may contribute to elevated UA production. Elevated UA has been associated with insulin resistance, enhanced lipogenesis, and increased oxidative stress [40], suggesting that UA may function both as a metabolic biomarker and as a potential mediator within broader cardiometabolic pathways.

The relatively favorable performance of regularized LR compared with more complex ML models is aligns with findings from previous clinical prediction research. Meta-analyses have shown that ML models do not consistently outperform conventional regression approaches when applied to structured clinical or epidemiological datasets, particularly when sample sizes are moderate and predictor–outcome relationships are relatively linear [41, 42]. Similarly, studies in Asian populations have reported comparable performance between LR and ML in cardiometabolic risk prediction [43]. In the present analysis, tree-based models showed signs of overfitting, suggesting that they may have captured dataset-specific patterns rather than generalizable signals [44]. In contrast, LR demonstrated a favorable balance between model complexity, stability, and interpretability, with superior calibration and consistent performance across datasets. This observation aligns with the approximately linear associations reported between lipid indices such as AIP and CRI-II and measures of visceral adiposity [45, 46], highlighting that increased algorithmic complexity does not necessarily translate into improved predictive utility.

These findings also align with recent developments in cardiometabolic risk stratification that emphasize integrated lipid-related biomarkers. Emerging indicators such as triglyceride-glucose index and related composite measures have been proposed as accessible markers of insulin resistance and atherogenic burden [47]. Within this landscape, AIP and CRI-II may offer complementary insights into lipid remodeling associated with visceral adiposity. Moreover, SHAP-based interpretation enables transparent evaluation of feature contributions, supporting a more interpretable modeling framework for population-based risk assessment [48]. Together, these results underscore the potential value of integrating interpretable machine learning methods with accessible lipid biomarkers to

improve early identification of individuals at increased risk of AO and related cardiometabolic disorders.

### <sup>3</sup>Strengths and limitations

This study has several notable strengths. First, an interpretable ML framework was applied by sequentially combining Boruta and LASSO algorithms for feature selection, which helped reduce redundancy among predictors and improve model stability. Second, the analysis was based on a large population-based dataset derived from stratified multistage sampling, providing adequate statistical power and supporting population-level inference within the study region. Third, SHAP was used to improve model interpretability by quantifying the relative contribution of individual predictors, allowing clearer identification of lipid-related variables associated with AO.

<sup>6</sup>Several limitations should also be acknowledged. First, the single-province, cross-sectional design limits generalizability to other populations and precludes <sup>20</sup>sal inference regarding the observed associations between lipid indices and AO; therefore, the findings should be considered hypothesis-generating. Prospective cohort studies and intervention trials are needed to determine whether these lipid-related indicators can predict the development or regression of AO over time. Despite employing advanced feature selection and explainable modeling techniques, residual confounding from unmeasured genetic, behavioral, or environmental factors remains a possibility. Ultimately, AO was determined through anthropometric measurements instead of direct imaging techniques like MRI or CT for quantifying visceral fat. Future studies incorporating imaging-based visceral fat quantification may help further clarify the biological mechanisms underlying these associations.

### Conclusions

<sup>33</sup>In conclusion, this study suggests that composite atherogenic lipid indices, particularly the AIP and CRI-II, are significantly associated with AO in a population from southern China. By integrating multiple lipid-related indicators with explainable machine learning approaches, the analysis highlights the relative contribution of different lipid-derived markers in characterizing metabolic profiles linked to AO. From a population health perspective, these findings support the potential utility of routinely available lipid parameters, such as triglycerides and HDL-C, for identifying individuals with increased cardiometabolic risk associated with AO in large-scale health screening settings. These indices, easily derived from standard laboratory measurements, may serve as practical indicators for monitoring metabolic risk in community populations. Future research should further evaluate the generalizability of these findings through external validation in diverse populations and explore whether incorporating composite lipid indices into risk assessment frameworks may improve early detection and prevention strategies for abdominal obesity-related metabolic disorders.