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.

Identifying Distinct Cardiometabolic Groups of Coronary Artery Calcification in Patients with Coronary Artery Disease: A Machine Learning Approach

Lexuan Wang
Beijing Friendship Hospital
Shaodong He
Beijing Friendship Hospital
Wenbo Yue
The second clinical college of Capital Medical University
Xinran Li
The second clinical college of Capital Medical University
Lichen Su
The second clinical college of Capital Medical University
Ruifeng Liu
fengziliu06@163.com

Beijing Friendship Hospital

Research Article

Keywords: Coronary Artery Disease, Coronary Artery Calcification, Cardiometabolic Phenotypes, Machine Learning, Inflammation, Risk Stratification

Posted Date: June 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9941903/v1

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

Additional Declarations:
No competing interests reported.

Figures are available in the Supplementary Files section.

Identifying Distinct Cardiometabolic Groups of Coronary Artery Calcification in Patients with Coronary Artery Disease: A Machine Learning Approach

Lexuan Wang[1,11], Shaodong He[1,11], Wenbo Yue[2], Xinran Li[2], Lichen Su[2], Ruifeng Liu[1]

1. Department of Cardiology, Beijing Friendship Hospital Affiliated to Capital Medical University, Beijing, China.

2. The second clinical college of Capital Medical University, Beijing, China.

11. These authors contributed equally: Lexuan Wang, Shaodong He.

Correspondence:

Ruifeng Liu, MD, Department of Cardiology, Beijing Friendship Hospital, Capital Medical University, 95 Yong An Road, Xicheng District, Beijing, 100050, P.R. China.

Telephone: 86-10-63138299, Fax: 86-10-63138299,

Email: fengziliu06@163.com

Running Title: Phenotyping Coronary Calcification with Machine Learning

Abstract

Background: Patients with coronary artery disease (CAD) exhibit significant clinical heterogeneity. We aimed to identify distinct clinical groups in CAD patients based on their comprehensive profiles and to explore the unique risk factors for severe coronary artery calcification (CAC) within each group.

Methods: We retrospectively enrolled a cohort of 550 patients with CAD. Unsupervised K-means clustering was used to identify patient subgroups from 76 standardized baseline variables. The identified groups were characterized based on demographic, lifestyle, and biochemical data. The association between these groups and severe CAC (Agatston score > 400) was assessed using multivariable logistic regression. To develop a clinically actionable tool for classifying patients into these distinct groups, a multiclass prediction model was also constructed. Feature selection was performed using a consensus approach combining LASSO (Least Absolute Shrinkage and Selection Operator) regression, SHAP (SHapley Additive exPlanations) analysis, and Random Forest. Model development and internal validation were conducted using nested k-fold cross-validation on this 550-patient cohort.

Results: Unsupervised clustering identified three distinct groups: Group 1 (n=117), a "Metabolism-Driven" subgroup; Group 2 (n=194), an "Age- and Inflammation-Related" subgroup; and Group 3 (n=239), a "Lifestyle- and Inflammation-Driven" subgroup. Group 3 was independently associated with the highest risk of severe CAC (Adjusted OR: 1.79, 95% CI: 1.12-2.85, P=0.015) in the multivariable logistic regression model. The multinomial logistic regression model, developed for multiclass classification of patient groups, demonstrated robust

performance in internal validation, achieving a Micro-AUC (Micro-averaged Area Under the Curve) of 0.983 (95% CI: 0.978-0.989) and excellent calibration (Brier Score: 0.126, 95% CI: 0.117-0.136). Other models, including support vector machine (SVM), random forest and XGBoost, also performed well but were outperformed by logistic regression in calibration. A clinical risk score system was developed based on the multinomial logistic regression model's predicted probabilities, which showed superior interpretability and calibration. This system effectively stratified patients into risk tiers and, when applied with an optimal cutoff for identifying Group 3, yielded a sensitivity of 95.4% and specificity of 89.1%.

Conclusion: This study identified three distinct pathophysiological groups within a CAD cohort, revealing that a "Lifestyle- and Inflammation-Driven" group carries the highest risk for severe calcification, suggesting inflammation induced by lifestyle factors as a critical, non-lipid-centric driver. These findings pave the way for more personalized risk stratification and targeted therapeutic strategies based on a patient's specific pathophysiological phenotype.

Keywords: Coronary Artery Disease; Coronary Artery Calcification; Cardiometabolic Phenotypes; Machine Learning; Inflammation; Risk Stratification

Clinical Perspective

(1) The population with coronary artery calcification is diverse and can be divided into three distinct groups.

(2) Inflammation and poor lifestyle habits are the two factors that cause severe coronary artery calcification with the highest probability.

(3) Machine learning is highly effective in clustering and grouping patients.

1. Introduction

Coronary artery disease (CAD) remains the leading cause of morbidity and mortality worldwide [1,2]. However, CAD is not a monolithic entity; patients exhibit substantial heterogeneity in clinical presentation, risk factor profiles, and underlying pathophysiological mechanisms [3]. This diversity suggests that a "one-size-fits-all" approach to risk stratification is suboptimal and underscores the need for a more personalized understanding of the disease. Coronary artery calcification (CAC), a hallmark of advanced atherosclerosis, is a powerful predictor of future cardiovascular events [4,5]. While factors like dyslipidemia and hypertension are known drivers of CAC [6,7], the relative contribution of different pathological pathways—such as metabolic derangement versus systemic inflammation—in distinct patient subgroups remains poorly understood. It is plausible that different "types" of patients develop severe atherosclerosis through different mechanisms.

In recent years, data-driven approaches using unsupervised machine learning have emerged as powerful tools for objectively classifying patients into more homogeneous groups based on their complete clinical data [8,9]. This methodology allows for the discovery of novel, data-defined patient subgroups, which may

correspond to distinct pathophysiological pathways that are not apparent through traditional risk factor analysis [10]. Therefore, this study aimed to utilize an unsupervised K-means clustering approach to identify distinct clinical groups among patients with CAD. Our primary hypothesis was that we could uncover subgroups driven by different core mechanisms (e.g., metabolic vs. inflammatory). We then sought to characterize these groups and determine their respective associations with severe CAC. Finally, we aimed to develop an interpretable clinical risk score to effectively identify these distinct patient types in a clinical setting.

2. Methods

2.1. Study Population and Data Collection

This retrospective cohort study enrolled a total of 550 consecutive patients diagnosed with CAD who were admitted to the Beijing Friendship Hospital, Capital Medical University, between January 2015 and December 2023. The diagnosis of CAD was confirmed by coronary angiography showing ≥50% stenosis in at least one major coronary artery. Inclusion criteria were age ≥18 years and a confirmed diagnosis of CAD. Exclusion criteria included acute infection, severe hepatic or renal dysfunction (defined as alanine aminotransferase >3 times the upper limit of normal or an estimated glomerular filtration rate <30 mL/min/1.73 m2), malignancy, autoimmune disease, or incomplete clinical data. The specific research methods are shown in the flowchart (Figure 1).

The study protocol was approved by the Ethics Committee of Beijing Friendship Hospital, Capital Medical University (Approval No. 2021-P2-280-02), and adhered to the principles of the Declaration of Helsinki. Due to the retrospective nature of the analysis of de-identified data, the requirement for individual informed consent was waived.

Figure 1. Flow Diagram for Identifying Different Cardiometabolic Groups Using Machine Learning

2.2. Variable Collection and Definitions

A total of 76 baseline variables were systematically extracted from the hospital's electronic medical record system for each patient upon admission. These variables were chosen based on comprehensive clinical practice, existing literature, and their known association with cardiometabolic risk factors and coronary artery disease, aiming to capture a broad spectrum of biological pathways and clinical indicators relevant to CAD and CAC. These included:

Demographics and Lifestyle: Age (years), sex, smoking history (current or former), and drinking history (regular alcohol consumption).

Comorbidities: A documented history of hypertension, type 2 diabetes mellitus, and dyslipidemia.

Physical Examination: Body mass index (BMI), heart rate, systolic blood pressure (SBP), and diastolic blood pressure (DBP).

Laboratory Parameters: A comprehensive panel was collected and defined as follows:

Renal and Liver Function: Creatinine, urea, uric acid, albumin, globulin, total protein, total bilirubin, direct bilirubin, indirect bilirubin, alanine aminotransferase (ALT), and aspartate aminotransferase (AST).

Lipid Profile: Total cholesterol (TC), triglycerides (TG), high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), apolipoprotein A (ApoA), apolipoprotein B (ApoB), and lipoprotein(a) [Lp(a)]. A derived indicator was calculated: the atherogenic index of plasma (AIP) as log10(TG/HDL-C), which reflects the balance between atherogenic and anti-atherogenic lipoproteins [11].

Inflammatory and Hematological Markers: White blood cell count and its differentials, red blood cell count, hematocrit (HCT), hemoglobin, high-sensitivity C-reactive protein (hs-CRP). A derived indicator was calculated: neutrophil percentage to albumin ratio (NPAR), which reflects a combined indicator of both inflammatory and nutritional conditions as a warning indicator.[12].

Coagulation Function: D-dimer, fibrinogen, prothrombin time (PT), activated partial thromboplastin time (APTT), international normalized ratio (INR), and fibrinogen degradation product (FDP).

Myocardial Injury Markers and Others: Creatine kinase (CK), creatine kinase-MB (CK-MB), troponin T, and N-terminal pro-B-type natriuretic peptide (NT-proBNP).

Thyroid Function: Thyroid-stimulating hormone (TSH), total triiodothyronine (T3), total thyroxine (T4), free T3 (FT3), and free T4 (FT4).

2.3. Missing Data Handling

Missing data for categorical variables were imputed using the mode, and for continuous variables using the median. For unsupervised clustering, a complete-case analysis was performed to ensure robust group identification. For predictive modeling, missing data imputation was performed within each training fold of the nested cross-validation to prevent data leakage and ensure that imputation was based only on information available in the training set.

2.4. Coronary Artery Calcification Assessment

All patients underwent non-contrast cardiac computed tomography (CT) for CAC scoring using a 64-slice CT scanner (GE Healthcare). The Agatston score was calculated using dedicated software (SmartScore 4.0, GE Healthcare). Severe CAC was defined as a total Agatston score > 400[13,14].

2.5. Unsupervised Clustering and Phenotype Identification

To identify latent subgroups, we employed the K-means clustering algorithm. The analysis was performed in Python (version 3.9.12) using the ‘scikit-learn` library. First, all 76 continuous variables were standardized using Z-score transformation. Next, the optimal number of clusters (k) was determined by assessing the

convergence of two independent methods: the Elbow Method (Figure 2A), which identifies the point of diminishing returns in the sum of squared errors (SSE), and the Silhouette Score (Figure 2B), which measures cluster cohesion and separation. Both methods consistently supported k=3 as the most appropriate number of clusters. Finally, Principal Component Analysis (PCA) was used to visualize the separation of the three final groups (Figure 2C). While K-means was run with multiple initializations (n_init=10) to find robust local optima, formal cluster stability assessment (e.g., through bootstrap resampling) was not performed in this study, which is acknowledged as a limitation.

2.6. Feature Selection for Predictive Modeling

To build an interpretable model for classifying patients into the identified groups, a rigorous, consensus-based feature selection process was conducted. This process was performed strictly within each training fold of the nested cross-validation to prevent data leakage. We used three methods:

1. LASSO (Least Absolute Shrinkage and Selection Operator) Regression: A penalized multinomial logistic regression with ten-fold cross-validation was used to shrink the coefficients of non-important features to zero.

2. SHAP (SHapley Additive exPlanations): The SHAP framework is applied to quantify the magnitude and direction of each feature's contribution to the model prediction.

3. Random Forest: A Random Forest model ranked features based on their

mean decrease in Gini impurity.

The final 23 subsets of predictive factors were selected by analyzing the intersection of the results identified by the three methods of LASSO, SHAP and Random Forest using Venn diagrams (Figure 4D), ensuring robustness and consistency.

2.7. Development and Validation of a Clinical Risk Score

Using the 23 selected features, we developed and compared four machine learning models for multiclass classification: multinomial Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost. Model development and internal validation were performed using a nested k-fold cross-validation strategy on the 550-patient cohort. Specifically, an outer 5-fold cross-validation was used to evaluate overall model performance. Within each outer training fold, an inner 5-fold cross-validation with GridSearchCV was employed to optimize hyperparameters for SVM, Random Forest, and XGBoost, preventing selection bias. Model performance was evaluated on discrimination, measured by the micro-averaged Area Under the Receiver Operating Characteristic Curve (Micro-AUC) and macro-averaged AUC (Macro-AUC), and calibration, assessed with calibration curves and Brier Score. The 95% confidence intervals for AUC values were estimated through bootstrap resampling (1000 iterations) within each validation fold. The multinomial logistic regression model was ultimately chosen for its superior calibration and interpretability. A points-based risk score was derived using the formula: Risk Score = Σ ($\beta i \times xi$),

where βi is the coefficient for predictor i.

2.8. Statistical Analysis

Continuous variables were reported as median and interquartile range (IQR) and compared using the Kruskal-Wallis test. Categorical variables were reported as frequency and percentage (%) and compared using the Chi-squared test or Fisher's exact test. To assess the independent association between groups and severe CAC, multivariable logistic regression models were constructed with severe CAC (Agatston score > 400) as the dependent variable and Group 1 as the reference. Model 1 was unadjusted; Model 2 was adjusted for age and sex; Model 3 was fully adjusted for age, sex, BMI, smoking, drinking, hypertension, diabetes, and dyslipidemia. Statistical significance was set at a two-sided P-value < 0.05. All analyses were performed using Python 3.9.12, with key libraries including pandas 1.5.3, numpy 1.23.5, scikit-learn 1.2.2, xgboost 1.7.5, SHAP 0.41.0, matplotlib 3.7.1, and seaborn 0.12.2 (all obtained from the Python Package Index (PyPI)).

3. Results

3.1. Identification of Three Distinct Clinical Groups

K-means clustering analysis partitioned the 550 CAD patients into three distinct and clinically coherent groups (Figure 1). By identifying variables with significant

differences among the three groups, we summarized the characteristics of each group based on the pathophysiological features behind these variables:

Group 1 (n=117): "Metabolism-Driven" Type. This group consisted predominantly of younger males (median age 60.00 (54.00 - 67.00), 71.8% male) and was defined by a severe metabolic derangement. They exhibited the highest levels of AIP (0.34 (0.19 - 0.48)), TG (2.10 (1.59 - 2.93)), TC (4.67 (4.05 - 5.36)), and ApoB (0.92 (0.83 - 1.16)), identifying dyslipidemia as the core pathological feature. This group also had a distinct hematological profile with high HCT (42.60 (40.00 - 45.20)) (Table 1, 2).

Group 2 (n=194): "Age- and Inflammation-Related" Type. This group was composed mainly of the oldest patients (median age 69.00 (63.00 - 73.00)) and was overwhelmingly female (72.2%). While this group had the lowest rates of high-risk behaviors, their biochemical profile suggested age-related physiological changes and chronic low-grade inflammation. They had the highest AST/ALT ratio (1.34 (1.10 - 1.59)) and the lowest HCT levels (37.50 (35.52 - 39.90)) (Table 1, 2).

Group 3 (n=239): "Lifestyle- and Inflammation-Driven" Type. This largest subgroup was almost exclusively male (97.5%) and was defined by extreme lifestyle risk factor exposure, with the highest rates of smoking (82.8%) and drinking (69.9%). Interestingly, this group had the most favorable traditional lipid profile (lowest TC (3.68 (3.09 - 4.28)), TG (1.11 (0.83 - 1.55)), ApoB (0.65 (0.55 - 0.79))). However, their cardiovascular risk appeared to be driven by a profound systemic inflammatory response, as evidenced by the significantly highest levels of leukocytes (6.13 (5.44 -

7.34)) and indirect bilirubin (11.27 (8.73 - 14.61)) (Table 1, 2).

Figure 2. Patient Subgroup Identification through K-Means Clustering Analysis.

The study population (n=550) was stratified into three distinct subgroups, which exhibited significant heterogeneity across multiple clinical and biochemical parameters ($P < 0.05$ for all listed). Specifically, statistically significant differences were observed among the groups in baseline demographics including age, sex, and lifestyle factors such as smoking and drinking history. Key cardiovascular risk factors also varied, including the prevalence of hypertension, systolic blood pressure, heart rate, and BMI. Furthermore, the subgroups demonstrated profound differences in metabolic and biochemical profiles. These included significant variations in liver function markers (total protein, globulin, ALT, AST/ALT ratio, and all forms of bilirubin), renal function indicators (creatinine), and a comprehensive panel of lipid measurements (total cholesterol, triglycerides, HDL-C, LDL-C, apolipoprotein A, apolipoprotein B, and lipoprotein(a)). Electrolyte levels, including sodium, chloride, calcium, and potassium, also differed significantly across the groups. Notably, parameters such as diastolic blood pressure, prevalence of diabetes and hyperlipidemia, aspartate aminotransferase, and blood urea nitrogen did not show statistically significant differences.

Table 1. Baseline Clinical Characteristics of Patients in Different Subgroups

Further analysis revealed significant inter-group differences in coronary artery calcification severity and related biomarkers. The prevalence of severe calcification (Agatston score > 400) and the median Agatston scores themselves varied significantly among the three subgroups (P=0.033 and P=0.003, respectively), although the number of diseased vessels did not differ. The groups were distinct in terms of the atherogenic index of plasma (AIP) and the neutrophil percentage to albumin ratio (NPAR) (P<0.001 for both). Significant heterogeneity was also evident in inflammatory markers, with all major white blood cell components (total count, monocytes, lymphocytes, eosinophils, basophils, neutrophils), hematocrit, procalcitonin, and C-reactive protein showing significant variation (P<0.001 for most). The coagulation profile, including INR, APTT, antithrombin III, prothrombin time, PTA, FDP, fibrinogen, and D-dimer, also differed significantly across the groups. Moreover, thyroid function tests (TSH, T4, T3, FT3, FT4) and markers of myocardial injury, specifically troponin T and CK-MB, were significantly different, whereas troponin I and creatine kinase levels were not.

Table 2. Angiographic, Inflammatory, Coagulation, and Myocardial Injury Markers of Patients in Different Subgroups

3.2. Association of Groups with Severe Coronary Calcification

This forest plot illustrates the odds ratios (ORs) for the association between the identified patient subgroups and the presence of severe coronary artery calcification (Figure 3). The analysis was conducted using multivariable logistic regression with severe CAC as the dependent variable. Three sequential models were applied: a crude model (Model 1), a partially adjusted model (Model 2), and a fully adjusted model (Model 3). Across all three models, Group 3 consistently demonstrated a significantly higher risk of severe calcification when compared to Group 1 (reference group), with the odds ratio remaining robust after full adjustment (OR 1.79, 95% CI 1.12-2.85, P=0.015). The overall association for the group variable remained statistically significant in all models, indicating that patient subgroup is an independent predictor of severe calcification. No significant interactions were found with hypertension or diabetes in the final model.

Figure 3. Association of Patient Subgroups with Severe Coronary Artery Calcification.

To identify the most robust predictors for classifying patient subgroups, we implemented a multi-pronged feature selection strategy, applying three distinct and parallel analytical methods (Figure 4). First, LASSO regression was used to systematically penalize and eliminate less important variables. Second, a Random

Forest model ranked features based on their Gini impurity reduction, providing a measure of variable importance. Third, we utilized SHAP to quantify the impact of each feature on the model's predictions. Finally, the outputs from these three independent methods were synthesized. A Venn diagram was constructed to visualize the intersection of the feature sets identified by each algorithm. This process revealed a core set of 23 features that were consistently identified as important. This consensus approach ensures that the variables selected for the final model are not only predictive but also robust across different analytical frameworks, forming a solid basis for our subsequent risk score development.

Figure 4. Comprehensive Feature Selection Using LASSO, SHAP, and Random Forest Algorithms.

3.3. Predictive Model Performance and Calibration

To translate the predictive features into a clinically actionable tool, we developed a risk scoring system based on the multinomial logistic regression model for multiclass classification of patient groups, which was selected over other machine learning algorithms (i.e., SVM, Random Forest, and XGBoost, detailed in the Supplementary Material) for its superior interpretability and calibration. Table 4 summarizes the performance of the four models. Based on the formula "Risk Score = Σ ($\beta i \times xi$), where i represents the 23 characteristic variables screened out by the Venn

diagram" derived from this multinomial logistic regression model, the comprehensive risk score of each patient was calculated (Table 3). We then established both a two-tier and a three-tier risk stratification system. The primary two-tier system, designed to identify individuals corresponding to the high-risk "Group 3" group, was developed using ROC curve analysis based on the predicted probability of belonging to Group 3. By maximizing the Youden's Index, an optimal cut-off value of 30.46 was determined, which yielded a high sensitivity of 95.4% and a specificity of 89.1%. Consequently, patients with a score ≥ 30.46 are classified as "High Risk". For a more granular risk assessment, a three-tier system was also proposed based on the 25th and 75th percentiles of the total score distribution across the entire cohort. This defined "Low Risk" as scores < -137.65, "Medium Risk" for scores between -137.65 and 154.74, and "High Risk" for scores ≥ 154.74. These stratification systems provide clinicians with a practical and interpretable framework to efficiently categorize patients, enabling tailored monitoring and intervention strategies.

Table 3. Risk Stratification System Based on the Logistic Regression Model

Table 4. Performance Metrics of Machine Learning Models (Internal Validation)

To select the most suitable classification algorithm, we evaluated the performance of four different models: Multinomial Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost (Figure 5). The program files

of the four models are detailed in the attachment. In internal validation, all four models demonstrated robust discriminative ability. The Multinomial Logistic Regression model achieved a Micro-AUC of 0.983 (95% CI: 0.978-0.989) and Macro-AUC of 0.982 (95% CI: 0.975-0.989), with a Brier Score of 0.126 (95% CI: 0.117-0.136). Random Forest and XGBoost showed slightly higher Micro-AUCs (0.969 and 0.972 respectively) but exhibited poorer calibration. As shown by the micro-averaged Receiver Operating Characteristic (ROC) curves, and detailed in Table 4, all models showed good discriminative power. However, since the goal was to create an interpretable clinical risk score, model calibration—the agreement between predicted probabilities and actual outcomes—was equally critical. The calibration curves for each of the three groups revealed that the multinomial Logistic Regression model consistently showed the best calibration, with its predictions closely following the ideal diagonal line (Figure 5B, C, D). In contrast, the more complex tree-based models (Random Forest and XGBoost) exhibited poorer calibration. Given its strong discriminative power and superior calibration, the multinomial Logistic Regression model was chosen for the development of the final clinical risk scoring system.

Figure 5. Comparative Performance and Calibration of Machine Learning Models.

4. Discussion

Our study represents a significant step forward in understanding the profound heterogeneity of coronary artery disease (CAD). By applying unsupervised machine learning to a diverse CAD cohort, we successfully moved beyond traditional, generalized risk factor analysis to identify and robustly characterize three distinct and clinically coherent patient phenotypes, each distinguished by unique pathophysiological drivers. These data-driven groups are: a "Metabolism-Driven" type (Group 1), an "Age- and Inflammation-Related" type (Group 2), and a "Lifestyle- and Inflammation-Driven" type (Group 3). The discovery and detailed characterization of these novel phenotypes are central to our findings, providing a new framework for dissecting CAD complexity. Our most significant and clinically provocative finding is that the "Lifestyle- and Inflammation-Driven" group, not the classic dyslipidemic group, was associated with the highest risk of severe coronary artery calcification.

This result critically challenges the conventional lipid-centric view of atherosclerosis, powerfully highlighting the independent and dominant role of systemic inflammation, largely induced by lifestyle factors, as a primary driver of advanced coronary disease in a substantial subset of patients [15,16]. The paradox of Group 3—having the most favorable lipid panel yet the worst calcific burden—is central to our findings. It strongly suggests their risk is mediated through non-lipid pathways. The extreme exposure to smoking and alcohol in this group likely induces a state of chronic, low-grade systemic inflammation. This is further supported by their elevated leukocyte count, a well-established marker of inflammation linked to

atherosclerosis progression and adverse cardiovascular events [17]. Furthermore, their elevated indirect bilirubin, while traditionally viewed as an antioxidant, can also reflect increased heme catabolism and cellular turnover associated with systemic stress and inflammation [18]. Our work provides compelling, data-driven evidence that for many male patients, high-risk behaviors are the principal trigger for a pro-inflammatory state that aggressively accelerates vascular calcification. This aligns with large-scale clinical trials such as CANTOS, which demonstrated that targeting inflammation via the interleukin-1β pathway can reduce cardiovascular events, confirming inflammation as a causal factor in atherosclerosis [19].

The identification and characterization of our other two distinct groups (Group 1 and Group 2) further validate the robustness of our clustering approach and offer important context to the multifaceted nature of CAD. Group 1 ("Metabolism-Driven") represents the classical patient profile described in textbooks, where atherogenesis is clearly linked to severe dyslipidemia, as reflected by the highest levels of AIP and ApoB. The AIP, in particular, is a strong indicator of small, dense LDL particles, which are highly atherogenic [20]. The elevated HCT in this group may also contribute to risk through increased blood viscosity and endothelial shear stress [21]. This group confirms the validity of our clustering approach by successfully isolating a well-established pathophysiological pathway. Group 2 ("Age-Related") offers a more nuanced picture, portraying an elderly, largely female group whose risk may be tied to "inflammaging"—a chronic, low-grade inflammation associated with the aging process—and altered organ function rather than overt metabolic syndrome or

high-risk behaviors [22]. Their low HCT could indicate early stages of anemia of chronic disease, while the elevated AST/ALT ratio might suggest subtle, age-related changes in hepatic function. This group underscores the importance of considering age-specific physiological changes as a distinct contributor to cardiovascular risk, particularly in women[23].

To translate these newly identified phenotypes into actionable clinical tools, we employed a robust, multi-faceted feature selection process. Rather than relying on a single metric of importance, we integrated three independent methods, each based on distinct mathematical principles: LASSO regression, SHAP, and Random Forest. The final feature set was determined by the intersection of the results from these parallel analyses. Specifically, LASSO regression was utilized for its outstanding ability to handle multicollinearity and perform dimensionality reduction on high-dimensional clinical data. SHAP was applied to quantify the magnitude and direction of each feature's contribution to the model prediction. Random Forest provided a non-parametric, data-driven perspective, capturing complex nonlinear relationships and interactions between feature variables and outcomes, thereby complementing the potential shortcomings of linear models. Through this rigorous, triple-validation approach, we successfully condensed an initial pool of 76 variables down to a core set of 23 features for subsequent model construction.

At the model level, the outstanding predictive performance of multinomial logistic regression, XGBoost, SVM, and random forest models indicates that the underlying mechanisms distinguishing these different patient subgroups are likely

complex and nonlinear, involving the interaction of a large number of features. The multinomial logistic regression model we ultimately chose, due to its unique interpretability, aligns well with clinical decision-making thinking. It provides a clear formula and a risk score system. Although it sacrifices a little accuracy compared to more complex models, it gains significant transparency, which is crucial for clinical adoption. Based on this multinomial logistic regression model's predicted probabilities for each group, we calculated a comprehensive risk score for each patient and established a multi-tier phenotypical grouping system. This system allows for the classification of patients into Group 1, Group 2, or Group 3, and can additionally be used to identify patients at high risk for Group 3 (binary classification). This approach achieves "personalized" and "precise" risk level prediction for each patient. The significance of this innovative process lies in its ability to bridge the gap from "model construction" to "clinical application," transforming a predictive machine into an auxiliary tool whose decision-making logic can be reviewed, understood, and trusted by physicians, which is paramount for its future clinical translation.

Ultimately, our research provides both a theoretical basis and practical tools for advancing the concept of 'precision medicine' in CAD by facilitating 'different treatments for the same disease' based on distinct patient phenotypes. For the identified high-risk Group 3 (Lifestyle- and Inflammation-Driven type), the focus of intervention will be on aggressive lifestyle improvement (quitting smoking and drinking), and future anti-inflammatory treatment may be considered, rather than

merely focusing on lipid-lowering drugs, which may be less critical for this specific phenotype [24]. We envision that in the future, when issuing coronary computed tomography angiography (CTA) reports, we will not only be able to provide coronary calcification scores but also attach the specific patient subgroup, offering cardiologists real-time and personalized calcification risk reports. Our model can also be applied in hospital information systems to achieve automatic phenotyping of large-scale populations, providing a data foundation for future precise clinical research targeting these specific subgroups of people.

5.Limitations

This study has several limitations that must be acknowledged. First, its retrospective, single-center design may limit the generalizability of our findings. Although we have adopted the nested k-fold cross-validation method to simulate the external verification situation as closely as possible, external validation in larger, multi-center, and ethnically diverse cohorts is crucial. Second, while K-means was run with multiple initializations, a formal assessment of cluster stability (e.g., through bootstrap resampling) was not performed. This limits our ability to definitively quantify the robustness of the identified groups. Third, while we controlled for numerous confounders, the possibility of residual confounding from unmeasured variables (e.g., diet, physical activity, psychosocial stress) cannot be entirely excluded. Fourth, CAC was assessed at a single time point; longitudinal data are needed to determine if our groups predict not only the burden but also the rate of atherosclerotic

progression. Finally, our study did not assess hard clinical outcomes. Future research must prospectively link these groups and the derived risk score to the incidence of myocardial infarction, stroke, and cardiovascular mortality to fully establish their prognostic value.

6.Conclusion

In conclusion, our study reframes the heterogeneity of CAD by identifying three distinct pathophysiological phenotypes. We demonstrate that a lifestyle-induced inflammatory state, independent of traditional dyslipidemia, represents a major pathway to severe coronary calcification. The clinical risk score developed from these findings offers a practical tool to identify these patient subtypes, paving the way for more targeted and personalized therapeutic strategies in the management of coronary artery disease.

7.Declarations

List of Abbreviations

Abbreviation	Full Term
AIP	Atherogenic Index of Plasma
ALT	Alanine Aminotransferase
ApoA	Apolipoprotein A
ApoB	Apolipoprotein B
APTT	Activated Partial Thromboplastin Time

AST Aspartate Aminotransferase

AUC Area Under the Curve

BMI Body Mass Index

CAC Coronary Artery Calcification

CAD Coronary Artery Disease

CI Confidence Interval

CK Creatine Kinase

CK-MB Creatine Kinase-MB Isoenzyme

CRP C-Reactive Protein

CT Computed Tomography

DBP Diastolic Blood Pressure

FDP Fibrinogen Degradation Product

FT3 Free Triiodothyronine

FT4 Free Thyroxine

HCT Hematocrit

HDL-C High-Density Lipoprotein Cholesterol

INR International Normalized Ratio

IQR Interquartile Range

LASSO Least Absolute Shrinkage and Selection Operator

LDL-C Low-Density Lipoprotein Cholesterol

Lp(a) Lipoprotein(a)

NPAR Non-HDL-C to Apolipoprotein B Ratio

NT-proBNPN-Terminal pro-B-type Natriuretic Peptide

OR Odds Ratio

PCA Principal Component Analysis

PT Prothrombin Time

ROC Receiver Operating Characteristic

SBP Systolic Blood Pressure

SHAP SHapley Additive exPlanations

SSE Sum of Squared Errors

SVM Support Vector Machine

T3 Total Triiodothyronine

T4 Total Thyroxine

TC Total Cholesterol

TG Triglycerides

TSH Thyroid-Stimulating Hormone

Ethical Approval Statement

This study was conducted in strict accordance with the ethical guidelines and principles set forth by our institution's ethical review bodies. All protocols and procedures were reviewed and approved by the Ethics Committee of Beijing Friendship Hospital (Approval No. 2018-P2-030-01). The study adhered to the principles established under the Declaration of Helsinki and relevant institutional guidelines for the care and use of human subjects.

Consent to Participate

In the case of human studies, informed consent was obtained from all participants prior to their inclusion in the research. Participants were briefed on the study's aims, methods, possible benefits, and potential risks. A clear explanation of their participation was provided, and they were informed of their right to withdraw from the study at any time without facing any penalty.

Consent for Publication

Participants gave their explicit consent to have their anonymized data included in the study's findings and subsequent publication. They were assured that their identities would remain confidential, and any identifying information would not be disclosed in any reports, presentations, or publications resulting from this research.

Data Availability Statement

Data available on request. The data underlying this article will be shared on reasonable request to the corresponding author.

Competing Interest

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

Funding

This study was supported by the National Natural Science Foundation of China (Grant No. 81600276).

Authors' Contributions

Conceptualization: Lexuan Wang

Methodology: Shaodong He

Software: Wenbo Yue

Validation: Shaodong He

Formal Analysis: Lexuan Wang

Investigation: Shaodong He

Data Curation: Lexuan Wang

Writing – Original Draft: Lexuan Wang

Writing – Review & Editing: Shaodong He

Supervision: Lichen Su

Project Administration: Ruifeng Liu

Funding Acquisition: Xinran Li

Acknowledgments

The authors would like to express their gratitude to the entire cardiovascular team at Beijing Friendship Hospital, including all physicians and nurses, for their

tireless clinical work and dedication to patient care, which made this study possible.

References

[1] GBD 2021 Diseases and Injuries Collaborators. Global incidence, prevalence, years lived with disability (YLDs), disability-adjusted life-years (DALYs), and healthy life expectancy (HALE) for 371 diseases and injuries in 204 countries and territories and 811 subnational locations, 1990-2021: a systematic analysis for the Global Burden of Disease Study 2021. Lancet. 2024 May 18;403(10440):2133-2161.

[2] GBD 2021 Causes of Death Collaborators. Global burden of 288 causes of death and life expectancy decomposition in 204 countries and territories and 811 subnational locations, 1990-2021: a systematic analysis for the Global Burden of Disease Study 2021. Lancet. 2024 May 18;403(10440):2100-2132.

[3] Frąk W, Wojtasińska A, Lisińska W, Młynarska E, Franczyk B, Rysz J. Pathophysiology of Cardiovascular Diseases: New Insights into Molecular Mechanisms of Atherosclerosis, Arterial Hypertension, and Coronary Artery Disease. Biomedicines. 2022 Aug 10;10(8):1938.

[4] Onnis C, Virmani R, Kawai K, Nardi V, Lerman A, Cademartiri F, et al. Coronary Artery Calcification: Current Concepts and Clinical Implications. Circulation. 2024 Jan 16;149(3):251-266.

[5] Budoff MJ, Young R, Burke G, Jeffrey Carr J, Detrano RC, Folsom AR, et al. Ten-year association of coronary artery calcium with atherosclerotic cardiovascular disease (ASCVD) events: the multi-ethnic study of atherosclerosis (MESA). Eur Heart J. 2018 Jul 1;39(25):2401-2408.

[6] Mihăilă RG. Pragmatic Analysis of Dyslipidemia Involvement in Coronary Artery Disease: A Narrative Review. Curr Cardiol Rev. 2020;16(1):36-47.

[7] Mallikethi-Reddy S, Rubenfire M, Jackson LA, Brook RD. Coronary artery calcium in hypertension: a review.

J Am Soc Hypertens. 2015 Dec;9(12):993-1000.

[8] Komuro J, Kusumoto D, Hashimoto H, Yuasa S. Machine learning in cardiology: Clinical application and basic research. J Cardiol. 2023 Aug;82(2):128-133.

[9] Pantanowitz L, Pearce T, Abukhiran I, Hanna M, Wheeler S, Soong TR, et al. Nongenerative Artificial Intelligence in Medicine: Advancements and Applications in Supervised and Unsupervised Machine Learning. Mod Pathol. 2025 Mar;38(3):100680.

[10] Bresso E, Lacomblez C, Duarte K, Monzo L, Baudry G, Tromp J, et al. Unsupervised machine learning for cardiovascular disease: A framework for future studies. Eur J Heart Fail. 2025 Nov 6.

[11] Saini S, Rai S, Gulati T, Agrawal V, Goel T, Agarwal A, et al. Association of Atherogenic Index of Plasma with Coronary Artery Disease: A Systematic Review and Meta-Analysis. Curr Probl Cardiol. 2023 Feb;48(2):101460.

[12] Intravaia RCM, Tognola C, Maloberti A, Brucato F, Campione E, Giannattasio C, et al. The role of inflammation in acute coronary syndrome: A systematic review on biochemical inflammatory assessment and anti-inflammatory therapies. Nutr Metab Cardiovasc Dis. 2026 Feb;36(2):104409.

[13] Suzuki Y, Matsumoto N, Yoda S, Amano Y, Okumura Y. Coronary artery calcium score: Current status of clinical application and how to handle the results. J Cardiol. 2022 May;79(5):567-571.

[14] Arad Y, Goodman KJ, Roth M, Newstein D, Guerci AD. Coronary calcification, coronary disease risk factors, C-reactive protein, and atherosclerotic cardiovascular disease events: the St. Francis Heart Study. J Am Coll Cardiol. 2005 Jul 5;46(1):158-65.

[15] Ference BA, Ginsberg HN, Graham I, Ray KK, Packard CJ, Bruckert E, et al. Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement from the European Atherosclerosis Society Consensus Panel. Eur Heart J. 2017 Aug

21;38(32):2459-2472.

[16] Attiq A, Afzal S, Ahmad W, Kandeel M. Hegemony of inflammation in atherosclerosis and coronary artery disease. Eur J Pharmacol. 2024 Mar 5;966:176338.

[17] Willeit P, Thompson SG, Agewall S, Bergström G, Bickel H, Catapano AL, et al. Inflammatory markers and extent and progression of early atherosclerosis: Meta-analysis of individual-participant-data from 20 prospective studies of the PROG-IMT collaboration. Eur J Prev Cardiol. 2016 Jan;23(2):194-205.

[18] Wu B, Wu Y, Tang W. Heme Catabolic Pathway in Inflammation and Immune Disorders. Front Pharmacol. 2019 Jul 24;10:825.

[19] Everett BM, MacFadyen JG, Thuren T, Libby P, Glynn RJ, Ridker PM. Inhibition of Interleukin-1β and Reduction in Atherothrombotic Cardiovascular Events in the CANTOS Trial. J Am Coll Cardiol. 2020 Oct 6;76(14):1660-1670.

[20] Vekic J, Zeljkovic A, Cicero AFG, Janez A, Stoian AP, Sonmez A, et al. Atherosclerosis Development and Progression: The Role of Atherogenic Small, Dense LDL. Medicina (Kaunas). 2022 Feb 16;58(2):299.

[21] Malek AM, Alper SL, Izumo S. Hemodynamic shear stress and its role in atherosclerosis. JAMA. 1999 Dec 1;282(21):2035-42.

[22] Henein MY, Vancheri S, Longo G, Vancheri F. The Role of Inflammation in Cardiovascular Disease. Int J Mol Sci. 2022 Oct 26;23(21):12906.

[23] Tyrrell DJ, Goldstein DR. Ageing and atherosclerosis: vascular intrinsic and extrinsic factors and potential role of IL-6. Nat Rev Cardiol. 2021 Jan;18(1):58-68.

[24] Deroissart J, Porsch F, Koller T, Binder CJ. Anti-inflammatory and Immunomodulatory Therapies in Atherosclerosis. Handb Exp Pharmacol. 2022;270:359-404.

Figures and Tables

Figure 1: Flow Diagram for Identifying Different Cardiometabolic Groups Using Machine Learning

Legend: This flow diagram briefly summarizes the entire process from the selection of the research subjects, clustering, the screening of key variables, all the way to the construction of the prediction model, as well as the evaluation and comparison of the prediction efficacy.

Figure 2: Patient Subgroup Identification through K-Means Clustering Analysis.

Legend: (A) The Elbow Method for Optimal k. This plot shows the Sum of Squared Errors (SSE) against the number of clusters (k). The "elbow point" indicates the optimal number of clusters where adding more clusters no longer significantly improves the clustering solution. (B) Silhouette Scores for Various Numbers of Clusters. This plot displays the average silhouette score for different numbers of clusters (k), providing another metric to evaluate clustering quality and identify the optimal k. (C) K-Means Clustering Results (k=3) via Principal Component Analysis (PCA). The scatter plot visualizes the three identified patient phenotypes (Group 1: Metabolism-Driven, Group 2: Age-Related, Group 3: Lifestyle- and Inflammation-Driven) in a 2D space, reduced by PCA to Principal Component 1 and Principal Component 2. Each point represents a patient, colored according to their assigned cluster. (D) Top 20 Most Important Features for Clustering. This bar chart

illustrates the top 20 features contributing most significantly to the differentiation of the identified patient phenotypes, ranked by their importance score (e.g., ANOVA F-score or similar metric).

Figure 3: Association of Patient Subgroups with Severe Coronary Artery Calcification.

Lgend: This forest plot presents the odds ratios (OR) and 95% confidence intervals (CI) for the association between patient subgroups (phenotypes) and the presence of severe coronary artery calcification. Three models are displayed: Model 1 (Crude): Unadjusted analysis. Model 2 (Partially Adjusted): Adjusted for age and sex. Model 3 (Fully Adjusted): Adjusted for age, sex, BMI, smoking, drinking, hypertension, diabetes, and dyslipidemia. The results show the adjusted ORs for Group 2 vs Group 1, Group 3 vs Group 1, and Group 3 vs Group 2, as well as an overall group effect, highlighting the independent risk conferred by each phenotype.

Figure 4: Comprehensive Feature Selection Using LASSO, SHAP, and Random Forest Algorithms.

Legend: (A) LASSO Analysis. This panel shows the results of Least Absolute Shrinkage and Selection Operator (LASSO) regression for feature selection, typically presented as cross-validation error curves and coefficient shrinkage paths for distinguishing different groups (e.g., Group 1 vs 3, Group 2 vs 3, Group 1 vs 2). This identifies key features influencing the classification between specific phenotypes. (B)

SHAP Summary Plot for Groups 1, 2, and 3. These plots display SHapley Additive exPlanations (SHAP) values for the most important features, illustrating their impact on the model's output for each identified phenotype. Each point represents a feature's SHAP value for a patient, with color indicating feature magnitude (e.g., high red, low blue). (C) Top 20 Most Important Features in Random Forest Model. This bar chart ranks the top 20 features by their importance in a Random Forest model, providing an alternative perspective on feature relevance for distinguishing the phenotypes. (D) Comparison of Features Selected by SHAP, LASSO, and Random Forest. This Venn diagram visualizes the overlap and unique features identified as important by three different feature selection methods: SHAP, LASSO, and Random Forest, demonstrating the robustness of critical markers across various analytical approaches.

Figure 5: Comparative Performance and Calibration of Machine Learning Models.

Legend: (A) Micro-Averaged Receiver Operating Characteristic (ROC) Curves. This plot displays the micro-averaged ROC curves for the multi-class classification models (Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost) on out-of-fold (OOF) predictions. The area under the curve (Micro-AUC) quantifies the overall discriminative ability of each model across all patient phenotypes. (B) Calibration Curves for Group 1. This plot shows the calibration curves for the models, specifically for predicting membership in Group 1 (Metabolism-Driven). A perfectly calibrated model's predictions would fall along the diagonal dashed line. (C) Calibration Curves for Group 2. This plot presents the

calibration curves for the models, for predicting membership in Group 2 (Age-Related). (D) Calibration Curves for Group 3. This plot illustrates the calibration curves for the models, for predicting membership in Group 3 (Lifestyle- and Inflammation-Driven).

Table 1: Baseline Clinical Characteristics of Patients in Different Subgroups

Item	Total (n = 550)	Group 1 (n = 117)	Group 2 (n = 194)	Group 3 (n = 239)	P value
Age (years)	65.00 (59.00 - 70.00)	60.00 (54.00 - 67.00)	69.00 (63.00 - 73.00)	64.00 (58.00 - 68.00)	<0.001
Sex					<0.001
Female, n (%)	179 (32.5)	33 (28.2)	140 (72.2)	6 (2.5)	
Male, n (%)	371 (67.5)	84 (71.8)	54 (27.8)	233 (97.5)	
Hypertension					0.014
No, n (%)	147 (26.7)	19 (16.2)	59 (30.4)	69 (28.9)	
Yes, n (%)	403 (73.3)	98 (83.8)	135 (69.6)	170 (71.1)	
Diastolic Blood Pressure (mmHg)	130.00 (121.00 - 141.00)	133.00 (123.00 - 145.00)	130.00 (121.00 - 138.75)	130.00 (121.00 - 139.50)	0.098
Systolic Blood Pressure (mmHg)	75.00 (68.00 - 83.00)	79.00 (73.00 - 87.00)	70.50 (65.00 - 79.00)	76.00 (69.00 -	<0.001

				83.50)	
Diabetes Mellitus					0.228
No, n (%)	341 (62.0)	67 (57.3)	129 (66.5)	145 (60.7)	
Yes, n (%)	209 (38.0)	50 (42.7)	65 (33.5)	94 (39.3)	
Hyperlipidemia					0.929
No, n (%)	241 (43.8)	50 (42.7)	87 (44.8)	104 (43.5)	
Yes, n (%)	309 (56.2)	67 (57.3)	107 (55.2)	135 (56.5)	
Smoking History					<0.001
No, n (%)	255 (46.4)	54 (46.2)	160 (82.5)	41 (17.2)	
Yes, n (%)	295 (53.6)	63 (53.8)	34 (17.5)	198 (82.8)	
Drinking History					<0.001
No, n (%)	314 (57.1)	74 (63.2)	168 (86.6)	72 (30.1)	
Yes, n (%)	236 (42.9)	43 (36.8)	26 (13.4)	167 (69.9)	
BMI (kg/m²)	25.61 (23.44 - 27.76)	27.10 (24.77 - 29.52)	24.83 (22.90 - 27.06)	25.35 (23.37 - 27.35)	<0.001
Heart Rate (beats/min)	70.00 (64.00 - 78.00)	72.00 (67.00 - 81.00)	70.00 (63.00 - 77.00)	69.00 (63.00 - 77.00)	0.013
Total Protein (g/L)	65.70 (62.73 - 68.90)	69.10 (66.50 - 71.50)	65.00 (61.73 - 67.40)	64.60 (62.30 - 67.90)	<0.001

Globulin (g/L)	26.50 (23.90 - 28.90)	28.00 (26.50 - 30.60)	27.05 (24.20 - 29.38)	25.00 (22.65 - 27.60)	<0.001
Alanine Aminotransferase (U/L)	23.00 (16.00 - 33.00)	31.00 (24.00 - 44.00)	17.00 (13.00 - 24.75)	25.00 (19.00 - 33.00)	<0.001
Aspartate Aminotransferase (U/L)	19.60 (16.32 - 23.40)	20.30 (17.20 - 24.40)	19.00 (15.75 - 23.45)	19.40 (16.70 - 22.80)	0.112
AST/ALT Ratio	1.11 (0.88 - 1.40)	0.92 (0.70 - 1.18)	1.34 (1.10 - 1.59)	1.04 (0.88 - 1.23)	<0.001
Total Bilirubin (μmol/L)	13.16 (10.71 - 16.65)	12.61 (10.18 - 15.75)	11.40 (9.71 - 13.76)	15.27 (12.34 - 18.81)	<0.001
Direct Bilirubin (μmol/L)	3.71 (2.44 - 4.79)	3.59 (2.10 - 4.49)	3.56 (2.26 - 4.42)	3.78 (2.67 - 5.10)	0.005
Indirect Bilirubin (μmol/L)	9.66 (7.28 - 12.49)	9.46 (7.02 - 11.97)	7.97 (6.51 - 10.43)	11.27 (8.73 - 14.61)	<0.001
Creatinine (μmol/L)	71.25 (61.70 - 82.62)	71.40 (63.20 - 83.50)	62.00 (55.78 - 72.78)	76.00 (69.80 - 85.55)	<0.001
Blood Urea	5.23 (4.48 -	5.01 (4.30 -	5.24 (4.29 -	5.34 (4.62 -	0.356

Nitrogen (mmol/L)	6.25)	6.46)	6.13)	6.27)	
Total Cholesterol (mmol/L)	4.06 (3.41 - 4.64)	4.67 (4.05 - 5.36)	4.14 (3.56 - 4.74)	3.68 (3.09 - 4.28)	<0.001
Triglycerides (mmol/L)	1.31 (0.93 - 1.78)	2.10 (1.59 - 2.93)	1.17 (0.91 - 1.63)	1.11 (0.83 - 1.55)	<0.001
HDL-C (mmol/L)	1.06 (0.89 - 1.20)	0.96 (0.82 - 1.17)	1.17 (1.01 - 1.31)	1.00 (0.86 - 1.17)	<0.001
LDL-C (mmol/L)	2.43 (1.91 - 2.79)	2.85 (2.43 - 3.36)	2.43 (2.00 - 2.76)	2.09 (1.73 - 2.51)	<0.001
Apolipoprotein A (g/L)	1.25 (1.10 - 1.37)	1.19 (1.05 - 1.30)	1.27 (1.19 - 1.43)	1.20 (1.06 - 1.34)	<0.001
Apolipoprotein B (g/L)	0.75 (0.60 - 0.88)	0.92 (0.83 - 1.16)	0.75 (0.60 - 0.85)	0.65 (0.55 - 0.79)	<0.001
Lipoprotein(a) (mg/L)	230.62 (108.25 - 230.62)	230.62 (140.00 - 230.62)	230.62 (141.00 - 230.62)	230.62 (82.00 - 230.62)	0.018
Albumin (g/L)	13.40 (12.10 - 14.70)	14.30 (13.10 - 15.70)	12.80 (11.60 - 13.90)	13.50 (12.30 - 14.70)	<0.001
Chloride (mmol/L)	106.00 (104.00 -	105.00 (103.00 -	107.00 (106.00 -	106.00 (104.00 -	<0.001

	108.00)	107.00)	108.00)	107.00)	
Sodium (mmol/L)	140.80 (139.60 - 142.00)	140.30 (138.90 - 141.30)	141.30 (140.00 - 142.38)	140.80 (139.65 - 141.95)	<0.001
Calcium (mmol/L)	2.28 (2.22 - 2.35)	2.34 (2.28 - 2.40)	2.26 (2.19 - 2.32)	2.27 (2.21 - 2.34)	<0.001
Potassium (mmol/L)	4.00 (3.84 - 4.20)	4.09 (3.89 - 4.25)	4.03 (3.85 - 4.21)	3.96 (3.80 - 4.16)	0.002

BMI: Body Mass Index; HDL-C: High-Density Lipoprotein Cholesterol; LDL-C: Low-Density Lipoprotein Cholesterol.

Table 2: Angiographic, Inflammatory, Coagulation, and Myocardial Injury Markers of Patients in Different Subgroups

Item	Total (n = 550)	Group 1 (n = 117)	Group 2 (n = 194)	Group 3 (n = 239)	P value
Severe Calcification					0.033
No, n (%)	289 (52.5)	73 (62.4)	102 (52.6)	114 (47.7)	
Yes, n (%)	261 (47.5)	44 (37.6)	92 (47.4)	125 (52.3)	
Agatston Score	344.06 (128.19 - 879.91)	254.35 (68.94 - 689.62)	343.86 (134.84 - 939.97)	442.62 (171.43 - 929.32)	0.003

Number of Diseased Vessels					0.749
Single-vessel disease, n (%)	135 (24.5)	27 (23.1)	53 (27.3)	55 (23.0)	
Two-vessel disease, n (%)	142 (25.8)	29 (24.8)	52 (26.8)	61 (25.5)	
Three-vessel disease, n (%)	273 (49.6)	61 (52.1)	89 (45.9)	123 (51.5)	
AIP	0.09 (-0.09 - 0.28)	0.34 (0.19 - 0.48)	0.02 (-0.13 - 0.14)	0.05 (-0.11 - 0.20)	<0.001
NPAR	1.57 (1.39 - 1.74)	1.46 (1.30 - 1.57)	1.66 (1.48 - 1.84)	1.59 (1.39 - 1.74)	<0.001
White Blood Cell Count (10^9/L)	6.20 (5.36 - 7.37)	7.18 (6.20 - 8.63)	5.74 (4.90 - 6.58)	6.13 (5.44 - 7.34)	<0.001
Monocyte Count (10^9/L)	1.49 (1.34 - 1.66)	1.45 (1.31 - 1.56)	1.42 (1.26 - 1.61)	1.59 (1.44 - 1.75)	<0.001
Lymphocyte Count (10^9/L)	0.39 (0.31 - 0.49)	0.45 (0.37 - 0.55)	0.35 (0.28 - 0.43)	0.40 (0.32 - 0.50)	<0.001

Eosinophil Count (10^9/L)	1.77 (1.40 - 2.17)	2.23 (1.82 - 2.61)	1.60 (1.29 - 1.93)	1.74 (1.40 - 2.08)	<0.001
Basophil Count (10^9/L)	0.13 (0.08 - 0.20)	0.17 (0.11 - 0.25)	0.11 (0.07 - 0.15)	0.14 (0.09 - 0.22)	<0.001
Neutrophil Count (10^9/L)	0.02(0.01-0.03)	0.03(0.02-0.04)	0.02(0.01-0.03)	0.02(0.01-0.03)	<0.001
Hematocrit (%)	3.85 (3.06 - 4.76)	4.41 (3.40 - 5.33)	3.63 (2.79 - 4.28)	3.83 (3.06 - 4.79)	<0.001
Procalcitonin (ng/mL)	41.00 (37.82 - 43.70)	42.60 (40.00 - 45.20)	37.50 (35.52 - 39.90)	42.40 (40.20 - 44.70)	<0.001
C-Reactive Protein (mg/L)	0.20 (0.17 - 0.23)	0.23 (0.20 - 0.26)	0.19 (0.17 - 0.23)	0.19 (0.16 - 0.21)	<0.001
INR	1.66 (0.69 - 3.67)	3.13 (1.35 - 4.95)	1.81 (0.72 - 3.67)	1.18 (0.50 - 3.10)	<0.001
APTT (s)	1.02 (0.97 - 1.07)	1.01 (0.95 - 1.05)	1.03 (0.99 - 1.09)	1.02 (0.98 - 1.07)	0.006
Antithrombin III (%)	27.50 (25.60 - 30.18)	28.50 (25.80 - 31.60)	27.70 (25.50 - 30.98)	27.10 (25.60 - 29.30)	0.010
Prothrombin	93.05	95.80 (89.00	93.44 (85.30 -	91.60 (83.95	0.002

Time (s)	(85.03 - 101.00)	- 104.00)	100.92)	- 99.20)	
PTA (%)	11.30 (10.80 - 11.70)	11.00 (10.50 - 11.50)	11.30 (10.83 - 11.80)	11.36 (10.80 - 11.70)	<0.001
FDP (μg/mL)	97.73 (91.83 - 104.30)	99.60 (94.80 - 106.80)	97.73 (90.75 - 103.30)	96.80 (91.80 - 103.15)	0.011
Fibrinogen (g/L)	2.50 (1.00 - 2.50)	1.58 (0.91 - 2.50)	2.50 (1.19 - 2.50)	2.50 (0.97 - 2.50)	0.014
D-Dimer (ng/mL)	2.90 (2.52 - 3.33)	3.25 (2.71 - 3.87)	2.93 (2.63 - 3.33)	2.74 (2.39 - 3.12)	<0.001
TSH (μIU/mL)	0.22 (0.14 - 0.44)	0.20 (0.10 - 0.43)	0.28 (0.15 - 0.47)	0.20 (0.19 - 0.40)	0.045
Total Thyroxine (T4) (nmol/L)	1.98 (1.25 - 2.71)	2.07 (1.50 - 2.85)	2.35 (1.36 - 2.96)	1.72 (1.12 - 2.64)	<0.001
Total Triiodothyronine (T3) (nmol/L)	96.85 (83.09 - 108.38)	99.17 (89.81 - 114.36)	96.89 (83.51 - 111.31)	90.95 (80.27 - 103.62)	<0.001

Free Triiodothyronine (FT3) (pmol/L)	74.50 (1.64 - 95.08)	57.67 (1.54 - 93.77)	60.68 (1.62 - 91.05)	81.47 (48.09 - 99.35)	<0.001
Free Thyroxine (FT4) (pmol/L)	3.33 (2.88 - 4.21)	3.77 (3.20 - 4.84)	3.35 (2.81 - 4.13)	3.15 (2.84 - 3.79)	<0.001
Troponin T (ng/mL)	1.04 (0.85 - 9.95)	4.89 (0.88 - 10.83)	1.22 (0.86 - 10.37)	0.97 (0.83 - 4.89)	<0.001
Troponin I (ng/mL)	0.01 (0.01 - 0.01)	0.01 (0.01 - 0.01)	0.01 (0.01 - 0.01)	0.01 (0.01 - 0.01)	0.269
Creatine Kinase (U/L)	0.00 (0.00 - 0.01)	0.00 (0.00 - 0.01)	0.00 (0.00 - 0.01)	0.00 (0.00 - 0.01)	0.990
CK-MB (U/L)	0.80 (0.60 - 1.41)	0.70 (0.50 - 1.20)	0.80 (0.60 - 1.60)	0.90 (0.60 - 1.41)	0.009
	79.50 (58.00 - 109.00)	73.00 (57.00 - 105.00)	73.00 (54.25 - 106.00)	86.00 (65.00 - 115.50)	0.004

AIP: Atherogenic Index of Plasma; NPAR: Neutrophil Percentage to Albumin Ratio; INR: International Normalized Ratio; APTT: Activated Partial Thromboplastin Time; PTA: Prothrombin Activity; FDP: Fibrinogen Degradation Product; TSH:

Thyroid-Stimulating Hormone; CK-MB: Creatine Kinase-MB Isoenzyme.

Variable	Score	Group 1 (n = 117)	Group 2 (n = 194)	Group 3 (n = 239)	P value
Albumin (g/L)	4	14.30 (13.10 - 15.70)	12.80 (11.60 - 13.90)	13.50 (12.30 - 14.70)	<0.001
AIP	10	0.34 (0.19 - 0.48)	0.02 (-0.13 - 0.14)	0.05 (-0.11 - 0.20)	<0.001
AST/ALT Ratio	-24	0.92 (0.70 - 1.18)	1.34 (1.10 - 1.59)	1.04 (0.88 - 1.23)	<0.001
Age (years)	11	60.00 (54.00 - 67.00)	69.00 (63.00 - 73.00)	64.00 (58.00 - 68.00)	<0.001
Apolipoprotein B (g/L)	-36	0.92 (0.83 - 1.16)	0.75 (0.60 - 0.85)	0.65 (0.55 - 0.79)	<0.001
Basophil Count (10^9/L)	-28	0.17 (0.11 - 0.25)	0.11 (0.07 - 0.15)	0.14 (0.09 - 0.22)	<0.001
Calcium	-5	2.34 (2.28 -	2.26 (2.19 -	2.27 (2.21 -	<0.001

Variable	Score	Group 1 (n = 117)	Group 2 (n = 194)	Group 3 (n = 239)	P value
(mmol/L)		2.40)	2.32)	2.34)	
Drinking History	20	43 (36.8)	26 (13.4)	167 (69.9)	<0.001
Eosinophil Count (10^9/L)	-1	0.13 (0.08 - 0.20)	0.11 (0.07 - 0.15)	0.14 (0.09 - 0.22)	<0.001
Fibrinogen (g/L)	-14	2.50 (1.00 - 2.50)	2.50 (1.19 - 2.50)	2.50 (0.97 - 2.50)	0.014
Globulin (g/L)	-8	28.00 (26.50 - 30.60)	27.05 (24.20 - 29.38)	25.00 (22.65 - 27.60)	<0.001
Hematocrit (%)	22	42.60 (40.00 - 45.20)	37.50 (35.52 - 39.90)	42.40 (40.20 - 44.70)	<0.001
Indirect Bilirubin (μmol/L)	48	9.46 (7.02 - 11.97)	7.97 (6.51 - 10.43)	11.27 (8.73 - 14.61)	<0.001
White Blood Cell Count	27	7.18 (6.20 - 8.63)	5.74 (4.90 - 6.58)	6.13 (5.44 - 7.34)	<0.001

Variable	Score	Group 1 (n = 117)	Group 2 (n = 194)	Group 3 (n = 239)	P value
(10^9/L)					
Lymphocyte Count (10^9/L)	2	1.77 (1.40 - 2.17)	1.60 (1.29 - 1.93)	1.74 (1.40 - 2.08)	<0.001
NPAR	1	1.46 (1.30 - 1.57)	1.66 (1.48 - 1.84)	1.59 (1.39 - 1.74)	<0.001
Procalcitonin (ng/mL)	-17	0.23 (0.20 - 0.26)	0.19 (0.17 - 0.23)	0.19 (0.16 - 0.21)	<0.001
Systolic Blood Pressure (mmHg)	14	79.00 (73.00 - 87.00)	70.50 (65.00 - 79.00)	76.00 (69.00 - 83.50)	<0.001
Sex	50	84 (71.8)	54 (27.8)	233 (97.5)	<0.001
Smoking History	50	63 (53.8)	34 (17.5)	198 (82.8)	<0.001
Total Cholesterol (mmol/L)	-39	4.67 (4.05 - 5.36)	4.14 (3.56 - 4.74)	3.68 (3.09 - 4.28)	<0.001
Triglycerides	-57	2.10 (1.59 -	1.17 (0.91 -	1.11 (0.83 -	<0.001

Variable	Score	Group 1 (n = 117)	Group 2 (n = 194)	Group 3 (n = 239)	P value
(mmol/L)		2.93)	1.63)	1.55)	
Total Protein (g/L)	10	69.10 (66.50 - 71.50)	65.00 (61.73 - 67.40)	64.60 (62.30 - 67.90)	<0.001

Table 3: Risk Stratification System Based on the Logistic Regression Model

The risk score was derived from a logistic regression model. The two-tier system cut-off is defined by maximizing the Youden's Index from the ROC curve analysis for predicting membership in the high-risk Group 3 (Optimal Score = 30.46; Sensitivity=0.954, Specificity=0.891). The three-tier system is based on score quartiles from the total patient population: Low Risk (<25th percentile, score < -137.65), Medium Risk (25th-75th percentile, score -137.65 to 154.74), and High Risk (>75th percentile, score ≥ 154.74).

Table 4: Performance Metrics of Machine Learning Models (Internal Validation)

Model	Micro-AUC (Mean ± SD)	Macro-AUC (Mean ± SD)	AUC for Group	Accuracy (Mean ± SD)	F1-score (Macro) (Mean ±	Brier Score (Multinomial) (Mean ±

			2 (95% CI)		SD)	SD)
Logistic Regression	0.9831 (±0.0054)	0.9819 (±0.0072)	0.9875 (0.9807 - 0.9931)	0.9182 (±0.0152)	0.9114 (±0.0125)	0.1264 (±0.0093)
SVM	0.9820 (±0.0036)	0.9820 (±0.0029)	0.9800 (0.9708 - 0.9873)	0.8818 (±0.0182)	0.8704 (±0.0192)	0.1553 (±0.0218)
Random Forest	0.9686 (±0.0092)	0.9695 (±0.0072)	0.9686 (0.9576 - 0.9788)	0.8636 (±0.0237)	0.8506 (±0.0253)	0.2383 (±0.0165)
XGBoost	0.9686 (±0.0082)	0.9676 (±0.0081)	0.9682 (0.9567 - 0.9783)	0.8527 (±0.0416)	0.8397 (±0.0520)	0.2103 (±0.0385)

)

Average performance metrics (Micro-AUC, Macro-AUC, AUC for Group 2 (95% CI), Accuracy, F1-score (Macro), Brier Score (Multinomial)) and their standard deviations for all four machine learning models, derived from stratified 5-fold cross-validation on held-out validation folds.

Supplementary Files

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

- Figures.pdf
- LogisticRegressionfinaltrained3Groups.joblib
- SVMfinaltrained3Groups.joblib
- XGBoostfinaltrained3Groups.joblib
- RandomForestfinaltrained3Groups.joblib