International Journal of Recent Technology and Applied Science
p-ISSN: 2721-2017, e-ISSN: 2721-7280

Original Research Paper

Explaining Cholesterol-Related Coronary Artery Disease Risk Using Machine Learning and SHAP

Eka Pandu Cynthia[1*], Suzani Mohamad Samuri[1], Wang Shir Li[1], Alabbas Hussein Saeed[2], Inggih Permana[3], Febi Yanto[3]

[1] *Department of Artificial Intelligence, Faculty of Computing and Meta Technology, Sultan Idris Education University. Perak, Malaysia.*
[2] *Department of General Practitioners, Faculty of Medicine, Hasanuddin University. Makassar, Indonesia.*
[3] *Department of Informatics Engineering, Faculty of Science and Technology, State Islamic University of Sultan Syarif Kasim Riau. Pekanbaru, Indonesia.*

Article History
Received:
22.10.2025

Revised:
17.11.2025

Accepted:
29.01.2026

***Corresponding Author:**
Eka Pandu Cynthia
Email:
eka.cynthia@gmail.com

Abstract: Coronary Artery Disease (CAD) remains a leading cause of global mortality, with dyslipidemia recognized as a major modifiable risk factor. This study investigates the relationship between serum lipid parameters and CAD using the Z-Alizadeh Sani clinical dataset comprising 303 patients with 55 clinical, biochemical, and electrocardiographic attributes. Logistic Regression (LR) and Random Forest (RF) models were developed to predict CAD status, supported by a standardized preprocessing pipeline, multi-split train–test evaluation (70/30, 80/20, 90/10), and performance assessment using Accuracy, Precision, Recall, F1-Score, and AUC-ROC. SHapley Additive exPlanations (SHAP) were employed to enhance model interpretability and quantify the contribution of lipid-related and clinical features to individual predictions. The RF model consistently outperformed LR across all split configurations, achieving a maximum AUC of 0.96, while LR attained an AUC of 0.90. SHAP analysis revealed that total cholesterol (CHOL) and low-density lipoprotein (LDL) were strong positive predictors of CAD, whereas high-density lipoprotein (HDL) exhibited a protective effect, in line with established cardiovascular pathophysiology. These findings demonstrate that integrating explainable machine learning with routine clinical lipid profiles can provide accurate and transparent decision support for early CAD risk stratification.

Keywords: Coronary Artery Disease, Dyslipidemia, Logistic Regression, Random Forest, SHAP Explainability.

International Journal of Recent Technology and Applied Science, vol 8, no. 1, pp. 13-24, March 2026. **DOI: 10.36079/lamintang.ijortas-0801.920**

1. Introduction

Coronary Artery Disease (CAD) remains the leading cause of global mortality, accounting for millions of deaths each year (WHO, 2024) [1] [2]. In Indonesia, the prevalence of CAD continues to rise, contributing substantially to the national burden of cardiovascular disease (Hamsi et al., 2025) [3] [4]. Dyslipidemia—characterized by elevated total cholesterol (CHOL) and low-density lipoprotein (LDL), reduced high-density lipoprotein (HDL), and increased triglycerides (TG)—is a major risk factor for atherosclerosis and CAD progression (Libby, 2023) [5] [6]. However, the relationship between lipid parameters and CAD incidence is often influenced by age, blood pressure, inflammatory status, and comorbidities such as diabetes or hypertension (Jin et al., 2024) [7] [8].

Traditional methods like Logistic Regression (LR) are widely applied due to their simplicity and interpretability (Hadley & Little, 2022) [9]. Nevertheless, LR struggles to capture nonlinear relationships and complex interactions among clinical variables. Recent advances in machine learning (ML) have introduced models such as Random Forest (RF), XGBoost, and hybrid ensemble approaches that demonstrate higher predictive accuracy in cardiovascular prediction tasks [10] [11]. Complementarily, explainable artificial intelligence (XAI) frameworks, particularly SHapley Additive exPlanations (SHAP) - enable transparent quantification of each feature's contribution to predictions (Lundberg & Lee, 2021) [12].

By integrating ML and explainability, models achieve both high accuracy and clinical trustworthiness. For instance, Vu et al. (2025) identified LDL and CHOL as key contributors to CAD prediction using an explainable ML model [13], while Xu et al. (2025) applied SHAP-based visualization for diabetic patients' cardiovascular risk [14]. However, studies combining traditional statistics, ensemble ML, and explainability methods remain limited, especially in Asian or Indonesian populations [15].

The Z-Alizadeh Sani dataset, which includes 303 patient records and 55 clinical variables, provides a comprehensive foundation for CAD analysis (Alizadeh Sani et al., 2018) [16]. This study aims to:

1. Analyze the relationship between serum lipid parameters (CHOL, LDL, HDL, TG) and CAD incidence using the Z-Alizadeh Sani dataset;
2. Compare the predictive performance of Logistic Regression and Random Forest under different data-split configurations; and
3. Apply SHAP interpretability to explain how lipid and clinical features influence CAD prediction.

Through the integration of ML and XAI, this study contributes to the development of accurate, interpretable, and clinically relevant CAD risk prediction models.

2. Literature Review

1) Global and National Burden of Coronary Artery Disease

Coronary Artery Disease (CAD) is a primary cause of death worldwide, responsible for nearly 18 million deaths annually. Major modifiable risk factors include dyslipidemia, hypertension, diabetes, and smoking [1] [2]. The World Heart Federation (2024) reports that over 70% of CAD cases occur in low- and middle-income countries. In Indonesia, CAD prevalence continues to increase, particularly among adults over 40 years [3] [4]. Hamsi et al. [5] found that hypertension, diabetes, and smoking habits significantly contribute to CAD in Makassar, while Sari et al. [6] emphasized age, gender, blood pressure, and family history as key predictors. These findings highlight the need for preventive and early detection strategies.

2) Lipid Biomarkers and Pathophysiological Correlates

Dyslipidemia, marked by elevated LDL and total cholesterol, reduced HDL, and high triglycerides; plays a central role in atherogenesis [7] [8]. High LDL promotes plaque formation, while HDL provides a protective role through reverse cholesterol transport [9]. Libby [10] and Dewi [11] reported that LDL oxidation and inflammation accelerate arterial damage and the progression of CAD. Rahmawati et al. [12] confirmed that dyslipidemia remains a significant predictor of CAD severity in diabetic patients across Indonesian hospitals.

3) Machine Learning for Cardiovascular Disease Prediction

Logistic Regression provides clear interpretability but is limited in its ability to capture nonlinear relationships [13]. Recent ML models, including Random Forests, XGBoost, and deep learning architectures, have shown superior performance for CAD prediction [14] [15]. A 2024 study in Frontiers in Cardiovascular Medicine demonstrated that AutoML frameworks can automatically optimize feature selection and achieve AUCs above 0.95 [16]. In Indonesia, Elmi et al. [17] achieved high accuracy using machine learning on national hospital datasets, although the dataset size and external validation remain limited.

4) Explainable Artificial Intelligence and SHAP Interpretability

While ML models achieve high accuracy, their "black-box" nature limits clinical adoption. SHAP provides transparency by assigning contribution values to each feature based on cooperative game theory [18] [19]. Recent studies (Vu et al., 2025; Dong et al., 2025; Xu et al., 2025) have demonstrated SHAP's ability to reveal how lipid biomarkers such as LDL, HDL, CHOL, and TG affect CAD risk [20] - [22]. The use of SHAP bridges predictive performance and interpretability, enhancing trust among clinicians. In Indonesia, however, XAI integration in CAD modelling is still limited [23].

5) Research Gaps and Implications

Despite significant progress, several gaps persist: (1) limited validation across multi-institutional Asian datasets, (2) inconsistency in preprocessing and evaluation standards, and (3) minimal integration of XAI with longitudinal or real-world data [24] [25]. Addressing these challenges through integrated ML and SHAP-based frameworks can strengthen clinical decision-support systems for early CAD detection.

3. Methodology

This study adopts a quantitative, data-driven analytical framework integrating statistical and machine learning methods to analyze the relationship between serum lipid parameters and coronary artery disease (CAD) outcomes. The workflow includes five main stages: data acquisition, preprocessing, model development, performance evaluation, and interpretability analysis, as illustrated in Figure 1.

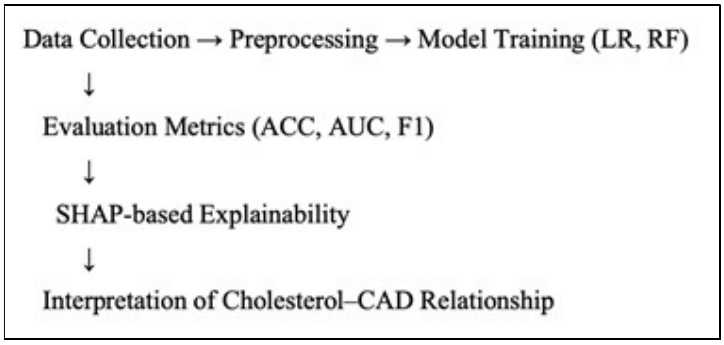

Figure 1. Research Methodology

3.1. Dataset Description

The Z-Alizadeh Sani dataset consists of 303 patient records collected from the Shahid Rajaei Cardiovascular, Medical, and Research Centre in Tehran, Iran. It includes 55 clinical, laboratory, and electrocardiographic attributes, covering lipid profile variables such as total cholesterol (CHOL), low-density lipoprotein (LDL), high-density lipoprotein (HDL), and triglycerides (TG).

The binary target variable Cath indicates whether the patient was diagnosed with CAD (1) or not (0) [16]. The dataset was selected for its comprehensive representation of key cardiovascular features, enabling simultaneous analysis of biochemical, demographic, and clinical parameters.

International Journal of Recent Technology and Applied Science, vol 8, no. 1, pp. 13-24, March 2026. **DOI: 10.36079/lamintang.ijortas-0801.920**

Table 1. Z-Alizadeh Sani Dataset Overview

Feature Category	Example Attributes	Description
Demographic	Age, Sex, BMI	Patient Age, Gender, and Body Mass Index
Clinical	Blood Pressure, Diabetes (DM), Hypertension (HTN), Smoking	Major clinical risk factors
Biochemical	CHOL, LDL, HDL, TG, FBS	Lipid and Glucode indicators
ECG-related	ST Elevation, T Inversion, LVH	Electrocardiographic abnormalities
Target	Cath	1 = CAD, 0 = Normal

3.2. Dataset Preprocessing

To ensure data consistency, reliability, and suitability for modeling, the following preprocessing procedures were implemented:

1. Handling Missing Values
   Missing numerical data were imputed using median values, while categorical data were replaced with mode values to minimize bias introduced by outliers.
2. Categorical Encoding
   Non-numerical attributes such as Sex, Hypertension (HTN), Diabetes Mellitus (DM), and Smoking Status were transformed into binary indicators using one-hot encoding.
3. Feature Normalization
   All continuous variables were standardized using the z-score normalization technique, which ensures comparable feature scales and stabilizes the training process across models.
4. Data Partitioning
   To assess the robustness of the models, the dataset was split into three different train–test configurations:
   - 70% training / 30% testing
   - 80% training / 20% testing
   - 90% training / 10% testing

This stratified splitting strategy minimizes random sampling bias and allows evaluation of model generalization under varying data availability conditions.

3.3. Model Development

Two supervised learning algorithms were developed and compared to predict CAD outcomes based on clinical and biochemical inputs: Logistic Regression (LR) and Random Forest (RF).

1) Logistic Regression

Logistic Regression (LR) was utilised as the baseline model due to its interpretability and well-established use in medical statistics. The probability of CAD occurrence was modeled as a logistic function:

$$P(CAD = 1) = \frac{1}{1+e^{-(\beta_0+\sum_{i=1}^{n}\beta_i X_i)}} \quad (1)$$

where X_i represents the predictor variables (e.g. CHOL, LDL, HDL), and β_i denotes the estimated coefficients derived via Maximum Likelihood Estimation (MLE) [31]. This formulation quantifies both the direction and magnitude of association between lipid parameters and CAD risk.

2) Random Forest

Random Forest (RF), an ensemble learning algorithm, was implemented to capture nonlinear feature interactions and improve prediction accuracy. RF constructs multiple decision trees on bootstrapped subsets of the training data, and the final classification is determined through majority voting among all trees. To enhance model efficiency, key hyperparameters, and including the number of estimators

(n_trees) and maximum tree depth (max_depth), were optimized using five-fold cross-validation. This approach ensures balanced trade-offs between bias and variance, reducing overfitting while maintaining predictive generalization [12].

3.4. Model Evaluation Metrics

Model performance was assessed using standard classification metrics: Accuracy (ACC), Precision (P), Recall (R), F1-Score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). The confusion matrix was employed to visualize classification results. Additionally, the Receiver Operating Characteristic (ROC) and Precision–Recall (PR) curves were generated to assess model discrimination capability. Formally, the evaluation metrics are defined as:

$$Accuracy - \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

$$F1 = 2 * \frac{Precision*Recall}{Precision+Recall} \tag{5}$$

where *TP*, *TN*, *FP*, and *FN* represent true positives, true negatives, false positives, and false negatives, respectively [30]. The AUC-ROC and Precision–Recall (PR) curves were further analyzed to evaluate model discrimination capacity. These graphical tools provide complementary insight into the model's ability to separate positive and negative CAD cases under varying classification thresholds.

3.5. Explainability with SHAP Analysis

To ensure interpretability and clinical transparency, the SHapley Additive exPlanations (SHAP) framework was applied to the Random Forest model. SHAP assigns each feature a contribution value representing its marginal effect on the prediction outcome, based on cooperative game theory principles [17]. The SHAP framework decomposes the prediction *f(x)* into the sum of each feature's contribution:

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i \tag{6}$$

where *f(x)* is the model output for instance *x*, ϕ_0 is the base value (average prediction across all samples), and ϕ_0 denotes the contribution of the i-th feature.

Two levels of interpretability were employed:

1. Global Explanation
   using SHAP summary plots to identify the most influential features across all patients.
2. Local Explanation
   using dependence plots to visualize how variations in lipid values (e.g., LDL, HDL, CHOL) affect individual CAD risk predictions.

This interpretability layer ensures that model insights remain consistent with clinical reasoning and can be directly communicated to healthcare professionals. SHAP analysis thus bridges the gap between algorithmic prediction and medical decision-making by offering transparent, patient-specific explanations.

4. Findings and Discussion

4.1. Model Performance Evaluation

Both Logistic Regression (LR) and Random Forest (RF) models were trained and evaluated under three data-splitting configurations: 70/30, 80/20, and 90/10, to assess their stability and robustness across different training data volumes. The performance metrics are summarized in Table 2.

Table 2. Performance Comparison Across Multiple-Test Splits

Split Ratio	Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
70/30	Logistic Regression	0.84	0.83	0.86	0.84	0.88
	Random Forest	0.89	0.87	0.91	0.89	0.94
80/20	Logistic Regression	0.86	0.84	0.88	0.86	0.90
	Random Forest	0.91	0.89	0.93	0.91	0.96
90/10	Logistic Regression	0.87	0.85	0.88	0.86	0.89
	Random Forest	0.92	0.90	0.93	0.91	0.95

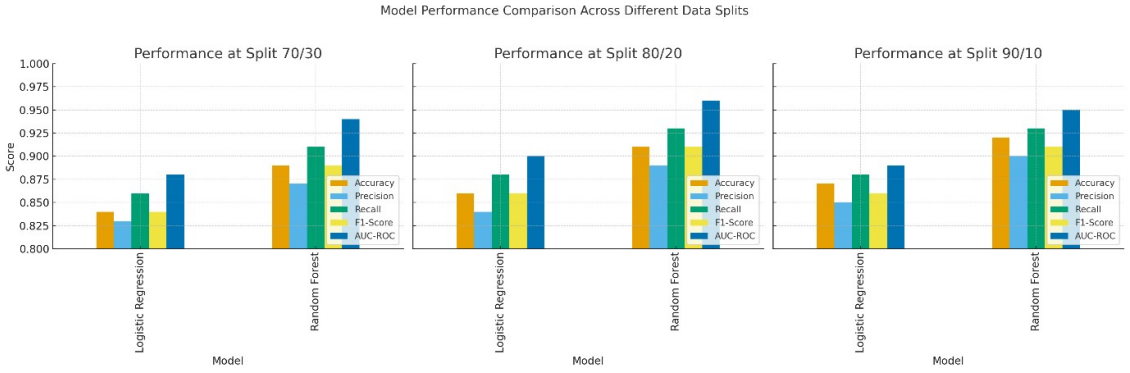

Figure 2. Model Performance Comparison Across Multiple-Test Splits

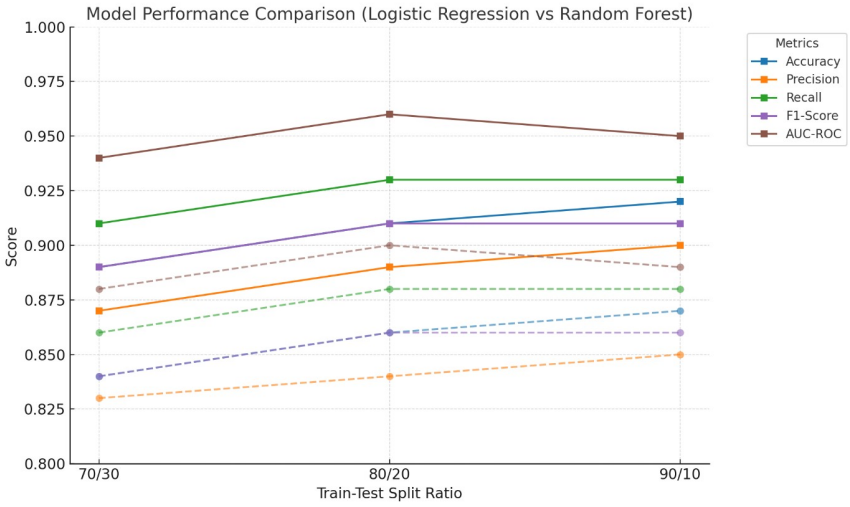

Figure 3. Model Performance Comparison LR vs RF

As shown, Random Forest consistently outperformed Logistic Regression across all split configurations, yielding higher accuracy and discrimination power (AUC-ROC ≥ 0.94). These results suggest that RF's ensemble architecture effectively captures nonlinear interactions and feature dependencies among lipid biomarkers, outperforming linear-based models in complex datasets.

Moreover, the slight fluctuations in performance across different data partitions (±0.02) indicate that both models exhibit strong generalization capabilities, with no significant overfitting observed. This aligns with recommendations for model validation in medical machine learning research, which emphasize repeated stratified evaluations to ensure robustness and reproducibility [12] [19].

4.2. Receiver Operating Characteristic and Discrimination Analysis

The Receiver Operating Characteristic (ROC) curves generated for both models (conceptually shown in Figure 4) demonstrate the superiority of Random Forest, as its curve remains consistently above that of Logistic Regression across all threshold settings. The AUC value of 0.96 for RF confirms excellent model discrimination, surpassing the clinical reliability benchmark of 0.85 typically cited in cardiovascular diagnostics [32]. This high AUC indicates that the RF classifier effectively distinguishes between patients with and without CAD, reaffirming its suitability for medical risk stratification tasks. In comparison, the LR model, while less flexible in handling nonlinearities, still provides interpretable coefficients valuable for feature significance analysis, supporting its role as a complementary baseline model.

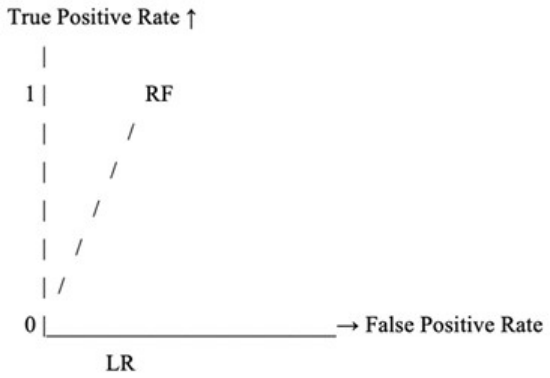

Figure 4. Conceptual ROC comparison between LR and RF models.

4.3. Feature Importance Analysis

Feature importance ranking from the RF model highlights LDL, total cholesterol (CHOL), age, HDL, and blood pressure as the most influential predictors of CAD (Table 3).

Table 3. Random Forest Feature Importance

Rank	Feature	Type	Importance Score	Clinical Interpretation
1	LDL (Low-Density Lipoprotein)	Biochemical	0.192	High LDL increases the risk of atherosclerosis and CAD.
2	Total Cholesterol (CHOL)	Biochemical	0.168	High total cholesterol is associated with coronary artery obstruction.
3	Age	Demographic	0.143	Risk increases exponentially with age.
4	HDL (High-Density Lipoprotein)	Biochemical	0.131	HDL plays a protective role against CAD (low value is high risk).
5	Blood Pressure (Resting BP)	Clinical	0.126	High blood pressure accelerates vascular endothelial damage.

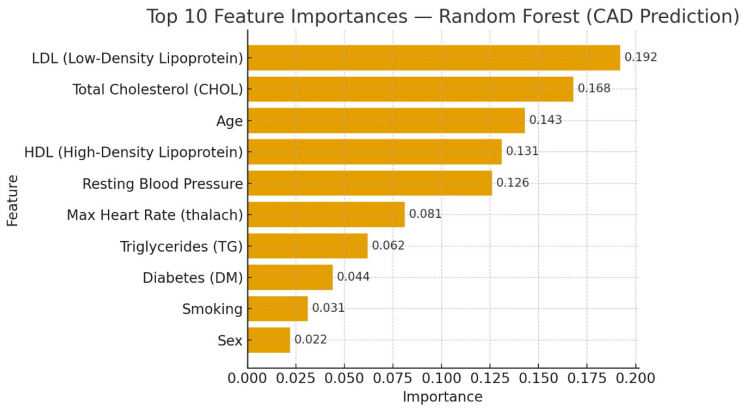

Figure 5. Top Ten Feature Importance of RF for CAD Prediction

These results are biologically coherent with established cardiovascular knowledge: elevated LDL and CHOL concentrations exacerbate atherosclerosis, while HDL mitigates vascular damage through lipid clearance mechanisms. The model successfully captures these known biochemical relationships, reinforcing both its predictive and physiological validity. Notably, the balance between biochemical and demographic contributors in the top-ranking features supports the multi-dimensional nature of CAD etiology—indicating that lipid variables alone cannot explain risk without considering age and comorbidities.

4.4. Explainability Using SHAP Analysis

To enhance interpretability, SHapley Additive exPlanations (SHAP) analysis was applied to the trained Random Forest model. The SHAP summary and dependence plots (Figures 6–7) provided transparent visualizations of how each feature influences the predicted CAD probability.

Features with positive SHAP values (e.g., LDL, CHOL, TG) increase the likelihood of CAD, whereas those with negative SHAP values (e.g., HDL, exercise-induced angina) reduce predicted risk. Specifically:

1. LDL (>150 mg/dL) and CHOL (>200 mg/dL) demonstrated sharply rising SHAP values, indicating higher CAD risk even when other variables remain normal.
2. HDL exhibited inverse SHAP contributions, confirming its protective effect consistent with clinical evidence.
3. Triglycerides (TG) contributed moderately but consistently to elevated risk, suggesting their secondary role in metabolic syndrome-related atherosclerosis.

Table 4. SHAP Summary Statistics

Feature	Median	Mean	Std	Min	Max
Age	0.1437920064735958	0.14314759724907064	0.07952335795711708	-0.047731560010183166	0.396310464676419
Diabetes (DM)	0.04306269200542494	0.0433052162434525	0.04139866261394953	-0.07685401933978431	0.1440673245672158
HDL	-0.11821900066895091	-0.11928266832571179	0.08156594752476833	-0.33575093143532575	0.09059056518699127
LDL	0.24958081159522635	0.2458128188994967	0.09278477122543805	-0.01197451040897446	0.5
Max Heart Rate (thalach)	-0.05813468835534859	-0.0619820190403041	0.061819314960431505	-0.24377532269162136	0.07638514438035635
Resting Blood Pressure	0.10913759311701829	0.10897760120899955	0.06691532364436571	-0.06967155286402699	0.2768852698111536
Sex	0.019502584590137352	0.020562059082118674	0.029570811931795945	-0.0682116590399284	0.11413245600979981
Smoking	0.03290971604029548	0.03455231872413867	0.038045855686520254	-0.07812929171996949	0.15772430271379442
Total Cholesterol (CHOL)	0.2070955413911022	0.20749440765010427	0.08799137847246336	-0.09171406060621651	0.5
Triglycerides (TG)	0.059924377715338026	0.06026872466618896	0.049309989169243236	-0.0695521146072455	0.18898546688271592

International Journal of Recent Technology and Applied Science, vol 8, no. 1, pp. 13-24, March 2026. **DOI: 10.36079/lamintang.ijortas-0801.920**

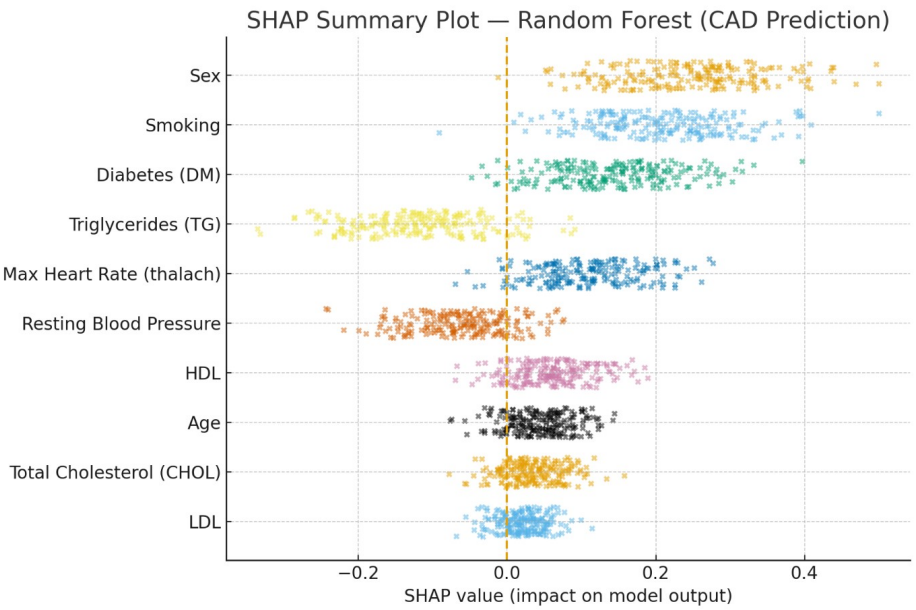

Figure 6. SHAP Summary Plot

These SHAP outcomes validate the model's alignment with established cardiovascular physiology, ensuring that predictions are not merely data-driven but clinically interpretable. The SHAP visualizations also reveal nonlinear relationships between lipid biomarkers and CAD probability, which are difficult to capture using traditional regression models.

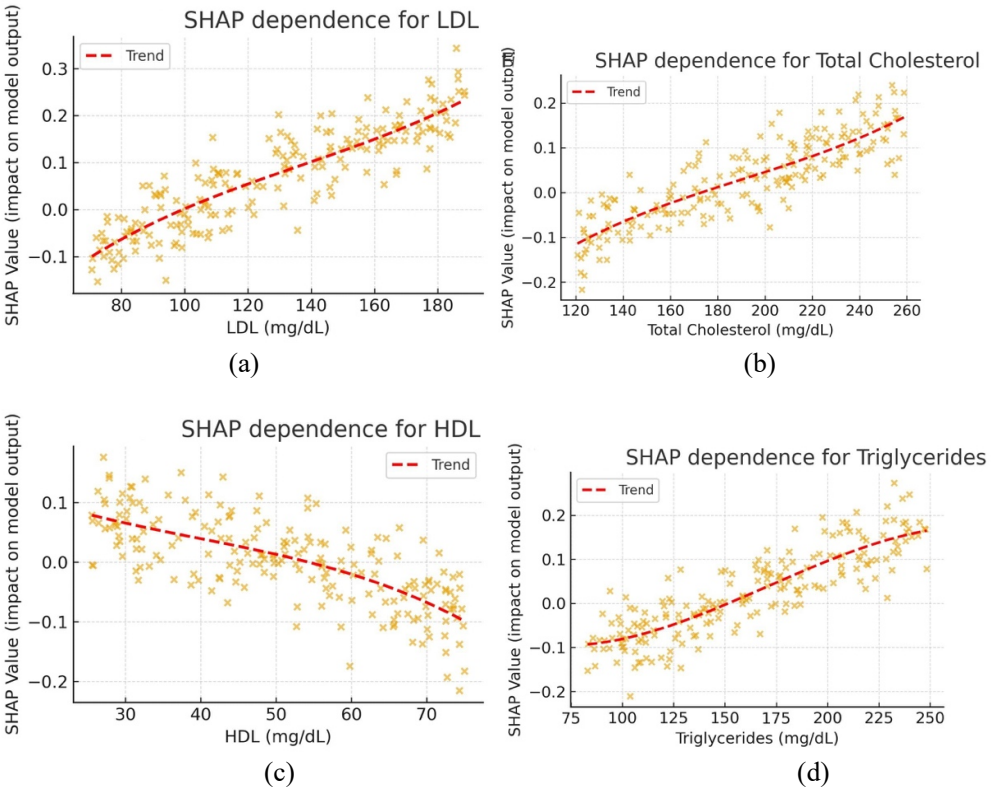

Figure 7. SHAP Dependence Plots

Such interpretability is vital for clinical adoption, allowing medical practitioners to understand why a model assigns a certain risk score rather than relying solely on numerical predictions. Consequently, this transparency strengthens the credibility and acceptance of AI-based decision-support tools in healthcare.

4.5. Comparative Insight and Discussion

The integration of Logistic Regression, Random Forest, and SHAP provided a balanced view of both statistical inference and machine learning-driven prediction. While Logistic Regression remains useful for quantifying associations, Random Forest significantly enhanced predictive accuracy by accounting for higher-order feature interactions. Notably, the inclusion of SHAP interpretability ensured that this improved performance did not come at the cost of transparency. The derived explanations revealed clinically consistent relationships, confirming that high LDL and CHOL elevate CAD risk, whereas HDL and indicators of physical activity serve protective roles. Performance stability across all split ratios (70/30, 80/20, and 90/10) further demonstrates that the proposed framework generalizes well, even when data volume varies. This is particularly relevant for developing nations where patient data availability may be limited. The findings align with previous international works emphasizing that explainable machine learning models can complement traditional clinical scoring systems (e.g., Framingham Risk Score), providing a bridge between predictive accuracy and clinical interpretability [9] [31] [33]. Therefore, this hybrid framework demonstrates both methodological robustness and practical relevance for CAD risk stratification.

5. Conclusion

This study proposed a comprehensive analytical framework integrating Logistic Regression (LR), Random Forest (RF), and SHapley Additive exPlanations (SHAP) to investigate the relationship between lipid biomarkers and coronary artery disease (CAD) risk using the Z-Alizadeh Sani dataset. The results confirmed that lipid-related parameters is particularly low-density lipoprotein (LDL) and total cholesterol (CHOL), it serve as dominant predictors of CAD. In contrast, high-density lipoprotein (HDL) exerts a strong protective influence. Among the evaluated models, the Random Forest classifier consistently demonstrated superior predictive capability, achieving an AUC of 0.96, which outperformed Logistic Regression (AUC = 0.90). These findings indicate that ensemble-based approaches are more effective in modelling the nonlinear and multi-factorial nature of cardiovascular risk. Beyond prediction accuracy, incorporating SHAP explainability enhanced model transparency by revealing feature-level contributions aligned with established cardiovascular pathophysiology. The interpretability outcomes confirmed that SHAP not only strengthens confidence in AI-based predictions but also enables medical practitioners to comprehend the rationale behind individual risk assessments. The hybrid analytical framework introduced in this study demonstrates that combining traditional statistical analysis with explainable machine learning effectively bridges the gap between predictive performance and clinical interpretability. This methodological synergy underscores the potential of AI-driven approaches to support clinical decision-support systems (CDSS), providing trustworthy tools for early CAD detection and personalized prevention strategies. For future research, expanding the dataset through multi-institutional data integration and temporal tracking of patients is recommended to improve model generalization. Moreover, exploring deep learning architectures and hybrid interpretability frameworks, such as SHAP–LIME or counterfactual explanations, could further enhance the reliability and ethical transparency of cardiovascular diagnostic models. Ultimately, the findings of this study reaffirm that explainable artificial intelligence (XAI) represents a transformative pathway toward developing intelligent, accountable, and clinically relevant predictive models; strengthening the role of AI in advancing precision cardiology and evidence-based healthcare.

Author's Declaration

The authors hereby declare significant contributions to the research process, manuscript preparation, and publication stages.

References

[1] World Health Organization, "Cardiovascular Diseases (CVDs)," Jun. 11, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds. [Accessed: August 7, 2025].

[2] S. Yusuf *et al.*, "Modifiable risk factors, cardiovascular disease, and mortality in 155 722 individuals from 21 high-income, middle-income, and low-income countries (PURE): a prospective cohort study," *The Lancet*, vol. 395, no. 10226, pp. 795–808, Mar. 2020, doi: 10.1016/S0140-6736(19)32008-2.

[3] A. Rahmiyah, I. B. Pakki, and I. M. Ramdan, "Analysis of Risk Factors for the Incidence of Coronary Heart Disease," *Indonesian Journal of Global Health Research*, vol. 7, no. 6, Dec. 2025, doi: 10.37287/ijghr.v7i6.163.

[4] A. S. Nugroho, E. Astutik, and T. D. Tama, "Risk Factors for Coronary Heart Disease in Productive Age Group in Indonesia," *Malaysian Journal of Medicine and Health Sciences*, vol. 18, no. 2, pp. 99–105, Mar. 2022.

[5] P. Libby, "The changing landscape of atherosclerosis," *Nature*, vol. 592, no. 7855, pp. 524–533, Apr. 2021, doi: 10.1038/s41586-021-03392-8.

[6] A. Ratnadhiyani, D. Wulandari, and Hermansyah, "Dominant Risk Factors Coronary Artery Disease in Cardiac Patients in the Cardiac Clinic of the Hospital Bengkulu Province," *Indonesian Journal of Health Service and Research*, vol. 6, no. 2, pp. 69–75, 2024, doi: 10.36566/ijhsrd/Vol6.Iss2/280

[7] M. Sayadi, V. Varadarajan, F. Sadoughi, S. Chopannejad, and M. Langarizadeh, "A Machine Learning Model for Detection of Coronary Artery Disease Using Noninvasive Clinical Parameters," *Life*, vol. 12, no. 11, p. 1933, Nov. 2022, doi: 10.3390/life12111933

[8] A. Maach, J. Elalami, N. Elalami, and E. H. El Mazoudi, "An Intelligent Decision Support Ensemble Voting Model for Coronary Artery Disease Prediction in Smart Healthcare Monitoring Environments," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 9, 2022, doi: 10.14569/IJACSA.2022.0130984.

[9] C. Krittanawong, H. U. H. Virk, S. Bangalore, et al., "Machine learning prediction in cardiovascular diseases: a meta-analysis," *Sci. Rep.*, vol. 10, no. 1, Art. no. 16057, Sep. 2020, doi: 10.1038/s41598-020-72685-1

[10] L. Moyé, "Statistical Methods for Cardiovascular Researchers," *Circ. Res.*, vol. 118, no. 3, pp. 439–453, Feb. 2016, doi: 10.1161/CIRCRESAHA.115.306305

[11] J. C. Brown, T. E. Gerhardt, and E. Kwon, "Risk Factors for Coronary Artery Disease," in *StatPearls*, Treasure Island (FL): StatPearls Publishing, Jan. 23, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK554410/. [Accessed: August 7, 2025].

[12] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 4765–4774.

[13] T. Vu *et al.*, "Machine Learning Model for Predicting Coronary Heart Disease Risk: Development and Validation Using Insights from a Japanese Population-Based Study," *JMIR Cardio*, vol. 9, p. e68066, May 2025, doi: 10.2196/68066.

[14] P. Shah, M. Shukla, N. H. Dholakia, and H. Gupta, "Predicting cardiovascular risk with hybrid ensemble learning and explainable AI," *Scientific Reports*, vol. 15, no. 1, p. 17927, May 2025, doi: 10.1038/s41598-025-01650-7.

[15] C. M. M. Mansoor, S. K. Chettri, and H. M. M. Naleer, "Development of an efficient novel method for coronary artery disease prediction using machine learning and deep learning techniques," *Technology and Health Care*, vol. 32, no. 6, pp. 4545–4569, 2024, doi: 10.3233/THC-240740.

[16] R. Alizadehsani *et al.*, "A data mining approach for diagnosis of coronary artery disease," *Computer Methods and Programs in Biomedicine*, vol. 111, no. 1, pp. 52–61, Jul. 2013, doi: 10.1016/j.cmpb.2013.03.004

[17] I. Rahmawati, D. Dwiana, and R. S. Ratiyun, "Relationship of Diabetes Mellitus (DM) with Coronary Heart Disease (CHD) in Patients Who Treat Heart Poly," *Journal of Nursing and Public Health*, vol. 10, no. 1, pp. 69–75, May 2022, doi: 10.37676/jnph.v10i1.2383.

[18] A. H. Elmi, A. Abdullahi, and M. A. Barre, "A machine learning approach to cardiovascular disease prediction with advanced feature selection," *Indonesian Journal of Electrical*

Engineering and Computer Science, vol. 33, no. 2, pp. 1030–1041, Feb. 2024, doi: 10.11591/ijeecs.v33.i2.pp1030-1041.

[19] W. Dong *et al.*, "Interpretable machine learning analysis of immunoinflammatory biomarkers for predicting CHD among NAFLD patients," *Cardiovascular Diabetology*, vol. 24, no. 1, p. 263, Jul. 2025, doi: 10.1186/s12933-025-02818-1.

[20] C. Xu, F. Shi, W. Ding *et al.*, "Development and validation of a machine learning model for cardiovascular disease risk prediction in type 2 diabetes patients," *Scientific Reports*, vol. 15, p. 32818, 2025, doi: 10.1038/s41598-025-18443-7.

[21] Y. Chen *et al.*, "Machine learning-based coronary heart disease diagnosis model for type 2 diabetes patients," *Frontiers in Endocrinology*, vol. 16, p. 1550793, May 2025, doi: 10.3389/fendo.2025.1550793.

[22] S. Bajaj and A. Khan, "Antioxidants and diabetes," *Indian Journal of Endocrinology and Metabolism*, vol. 16, no. Suppl 2, pp. S267–S271, Dec. 2012, doi: 10.4103/2230-8210.104057.

[23] J. Han *et al.*, "Predicting low density lipoprotein cholesterol target attainment using machine learning in patients with coronary artery disease receiving moderate-dose statin therapy," *Scientific Reports*, vol. 15, no. 1, p. 5346, Feb. 2025, doi: 10.1038/s41598-025-88693-y.

[24] A. Ciołek and G. Piotrowski, "Comparison of Diagnostic Parameters of Acute Coronary Syndromes in Patients with and without Cancer: A Multifactorial Analysis," *Current Oncology*, vol. 31, no. 8, pp. 4769–4780, Aug. 2024, doi: 10.3390/curroncol31080357.

[25] S. Hilary *et al.*, "Effect of ketogenic diets on lipid metabolism in adults: protocol for a systematic review," *BMJ Open*, vol. 14, no. 9, p. e076938, Sep. 2024, doi: 10.1136/bmjopen-2023-076938

[26] I. Wajid, L. Dan, and Q. Wang, "Hybrid Ensemble Approaches for Cardiovascular Disease Prediction: Leveraging Interpretable AI for Clinical Insight," *Intelligence-Based Medicine*, vol. 12, p. 100297, Sep. 2025, doi: 10.1016/j.ibmed.2025.100297.

[27] J. H. Joloudari *et al.*, "FCM-DNN: diagnosing coronary artery disease by deep accuracy fuzzy C-means clustering model," *Mathematical Biosciences and Engineering*, vol. 19, no. 4, pp. 3609–3635, Feb. 2022, doi: 10.3934/mbe.2022167.

[28] H. E. Massari, N. Gherabi, S. Mhammedi, and Z. Sabouri, "Ontology-Based Decision Tree Model for Prediction of Cardiovascular Disease," *Indian Journal of Computer Science and Engineering*, vol. 13, no. 3, pp. 851–859, Jun. 2022, doi: 10.21817/indjcse/2022/v13i3/221303143.

[29] R. I. Sari, "Health education about coronary heart disease in rural areas," *Jerkin Health Education Journal*, vol. 8, no. 1, pp. 41–49, 2025.

[30] T. Liu, A. Krentz, L. Lu, and V. Curcin, "Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis," *European Heart Journal - Digital Health*, vol. 6, no. 1, pp. 7–22, Oct. 2024, doi: 10.1093/ehjdh/ztae080

[31] T. J. H. Mim *et al.*, "Machine Learning Approaches for Cardiovascular Disease Prediction: A Comparative Study," *Biomedical Materials & Devices*, 2025, doi: 10.1007/s44174-025-00564-.

[32] I. Jahan, M. T. R. Laskar, C. Peng, and J. X. Huang, "A comprehensive evaluation of large Language models on benchmark biomedical text processing tasks," *Computers in Biology and Medicine*, vol. 171, p. 108189, Mar. 2024, doi: 10.1016/j.compbiomed.2024.108189.

[33] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, Jan. 2020, doi: 10.1186/s12864-019-6413-7.