IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 13, No. 4, April, 2026, pp. 697-704
Journal homepage: http://iieta.org/journals/mmep

Explainable Machine Learning for Heart Attack Risk Prediction Integrating Clinical, Lifestyle, and Socioeconomic Factors

Jepri Banjarnahor*, Charles, Niel Artstanta Sinulingga, Gunawansah Tumama Siregar, Muhammad Ikhsan, Jaidup Banjarnahor

Department of Information System, Faculty of Science and Technology, Universitas Prima Indonesia, Medan 20118, Indonesia

Corresponding Author Email: jepribanjarnahor@unprimdn.ac.id

https://doi.org/10.18280/mmep.130407

Received: 23 February 2026
Revised: 14 April 2026
Accepted: 22 April 2026
Available online: 15 May 2026

Keywords:
heart attack prediction, Extreme Gradient Boosting, explainable artificial intelligence, Shapley Additive Explanations, lifestyle factors, clinical predictors, socioeconomic determinants

ABSTRACT

Cardiovascular disease remains the leading cause of death globally, underscoring the critical need for risk prediction tools that are both accurate and clinically interpretable. This study developed an explainable machine learning model based on Extreme Gradient Boosting (XGBoost) to predict the risk of heart attack. The model integrates clinical measurements, lifestyle behaviors, and socioeconomic factors as predictive inputs. Rigorous data preprocessing was performed, and class-balancing techniques were applied to address class imbalance within the dataset. The model was trained with early stopping and evaluated using multiple performance metrics. On the validation set, the model achieved a Receiver Operating Characteristic Area Under the Curve (ROC-AUC) of 0.73 and a Precision-Recall Area Under the Curve (PR-AUC) of 0.80. These values reflect moderate discriminative ability with strong sensitivity toward the minority (high-risk) class. On an independent test set, the model achieved an ROC-AUC of 0.79, a recall of 0.84, and a precision of 0.63, demonstrating improved generalization relative to validation performance. These results indicate that the model identified high-risk individuals with a reasonable false-positive rate, supporting its feasibility as a screening tool in clinical practice. To enhance interpretability, feature contributions were quantified using Shapley Additive Explanations (SHAP). The analysis identified systolic blood pressure, smoking status, physical activity, sleep irregularity, and alcohol consumption—among other modifiable lifestyle factors—as the principal determinants of predicted risk. The model further captures the influence of socioeconomic variables, including income level and geographic region, on individual cardiovascular risk profiles. Overall, this study presents a transparent and interpretable predictive framework that balances predictive performance with clinical relevance. The framework supports informed clinical decision-making and targeted preventive interventions for heart attack risk.

1. INTRODUCTION

Cardiovascular disease remains the leading cause of death globally, responsible for approximately 19 million deaths in 2022 and accounting for nearly one-third of all deaths worldwide. A substantial proportion of these deaths is attributable to heart attacks, also known as myocardial infarction, which represent a severe and life-threatening acute manifestation of cardiovascular disease. Notably, approximately three-quarters of these deaths occur in low- and middle-income countries, with the majority affecting individuals under 70 years of age. These patterns reveal a critical gap: existing risk prediction models are insufficient for identifying individuals at high risk of heart attack, particularly across diverse populations [1].

Clinically, hypertension, elevated low-density lipoprotein (LDL) cholesterol, diabetes, obesity, and smoking are well-established heart attack risk factors. Recent clinical guidelines issued by the American College of Cardiology/American Heart Association (ACC/AHA) emphasize that lifestyle modification is the cornerstone of primary prevention [2]. This position is further supported by evidence elucidating the causal role of LDL cholesterol in atherosclerosis. Prolonged reduction in LDL levels has been shown to decrease the incidence of cardiovascular events, including heart attacks [3].

Beyond clinical markers, lifestyle behaviors directly influence the risk of heart attack. Large-scale accelerometer-based studies demonstrate that even low-intensity physical activity substantially and dose-dependently reduces the risk of all-cause mortality. Conversely, prolonged sedentary behavior is associated with increased cardiovascular risk [4]. Specifically, sitting time of 10.6 hours or more per day has been linked to elevated risk of heart failure and cardiovascular mortality [5]. Habitual sleep irregularity, defined as variability in sleep duration and timing, has also been independently associated with increased cardiovascular risk beyond the effects of total sleep duration [6]. Furthermore, reducing alcohol consumption has been associated with measurable

reductions in blood pressure and improved cardiovascular outcomes in a dose-dependent manner [7]. Collectively, these findings underscore the importance of incorporating lifestyle-related variables to enhance the predictive capacity of heart attack risk models.

The distribution and severity of heart attack risk are also shaped by socioeconomic determinants of health (SDoH). Education level, access to healthcare, occupational conditions, and environmental exposures all influence health behaviors and clinical outcomes. The American Heart Association identifies SDoH as a mediating factor for both conventional risk factors and inequities in cardiovascular outcomes [8]. Lower educational attainment has been associated with higher incidence and mortality of cardiovascular events, including heart attacks, as demonstrated by the Prospective Urban Rural Epidemiology (PURE) study conducted across low-resource settings [9]. Incorporating socioeconomic context is therefore essential to improving both the accuracy and equity of predictive models.

Standard risk prediction models, however, frequently fail to account for this broader context [10, 11]. The risk of heart attack is shaped not only by physiological measurements but also by how individuals conduct their daily lives — including physical activity levels, sedentary duration, sleep consistency, and engagement in harmful behaviors such as smoking and excessive alcohol consumption. Conventional statistical approaches are limited in their ability to model such multidimensional, non-linear relationships.

Machine learning offers a promising alternative by enabling the modeling of complex interactions among clinical, lifestyle, and socioeconomic factors. Among available methods, Extreme Gradient Boosting (XGBoost) has demonstrated strong performance in handling structured medical data [12]. However, predictive accuracy alone is insufficient for clinical application. Models must also be interpretable to establish clinician trust and facilitate practical adoption, as predictions generated by opaque black-box algorithms are unlikely to be integrated into real-world healthcare settings.

Explainable artificial intelligence (XAI) methods, particularly Shapley Additive Explanations (SHAP) [13], address this limitation by providing a robust framework for interpreting model predictions. SHAP enables both global and local explanations, identifying the features that contribute most to the elevated risk of heart attack. Importantly, it highlights modifiable determinants such as smoking and physical activity, thereby supporting risk awareness and management for both clinicians and patients.

This study proposes an XGBoost-based explainable model for heart attack risk prediction integrated with SHAP analysis. An interpretable machine-learning framework was developed to predict heart attack outcomes using clinical measurements, lifestyle behaviors, and socioeconomic factors. The framework pursues two objectives: achieving high predictive accuracy and maintaining model interpretability. The model was evaluated using multiple performance metrics and was designed to facilitate early risk detection, support risk communication, and guide targeted preventive interventions. Overall, this framework aims to provide a clinically actionable decision-support tool for clinicians, patients, and policymakers, contributing to more informed and equitable prevention of cardiovascular events.

2. MATERIAL AND METHODS

To systematically address the stated research objectives, this section presents a structured methodological framework that encompasses data preprocessing, class-imbalance handling, predictive modeling, and result interpretation. The methodology was designed to be transparent, reproducible, and statistically rigorous, while maintaining practical relevance and supporting a comprehensive understanding of cardiovascular risk determinants.

Specifically, this approach:

(i) estimates individual heart attack risk probabilities based on all available clinical, lifestyle, and socioeconomic features at the time of assessment;

(ii) compares the performance of the proposed XGBoost model with a baseline Logistic Regression model;

(iii) produces explainable, audit-ready model inferences that enable a deeper understanding of how both individual-level and population-level decisions are shaped with respect to underlying confounders using SHAP.

2.1 Conceptual architecture of the system

The system implemented in this study follows an end-to-end pipeline for transforming raw input data into clinically meaningful risk predictions. The pipeline begins with the collection of clinical, behavioral, and socioeconomic features. These features were subsequently preprocessed through numerical imputation and categorical encoding to prepare the data for tree-based modeling. To address class imbalance, class weights were assigned, and synthetic minority-class samples were generated to strengthen the representation of high-risk cases.

The primary XGBoost model was then trained, with Logistic Regression serving as a performance benchmark. Model performance was subsequently evaluated using relevant metrics. Model behavior was interpreted through SHAP analysis, providing both global and individual-level explanations. The overall workflow—comprising feature input, preprocessing, class balancing, model training, evaluation, and SHAP-based interpretation—is illustrated in Figure 1.

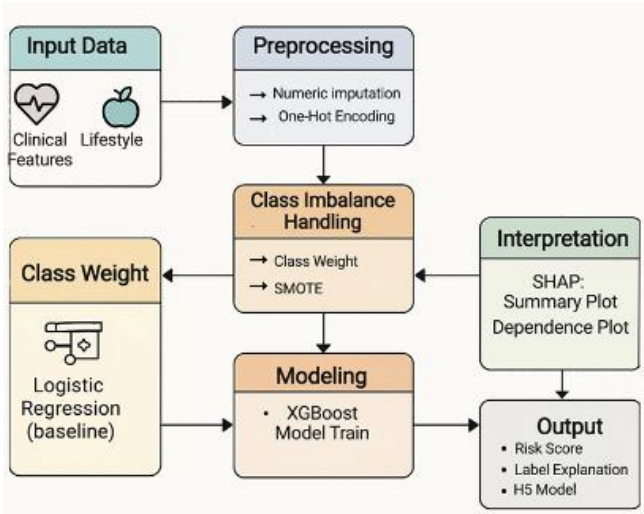

Figure 1. Conceptual system architecture diagram

2.2 Data representation and handling

Data preprocessing was performed to preserve information integrity while converting diverse clinical, behavioral, and socioeconomic attributes into a uniform numeric format suitable for tree-based learning. Missing numerical values were imputed using a median imputation strategy, and categorical variables were transformed using one-hot encoding. This encoding approach avoids imposing ordinal assumptions and enables the model to learn non-linear interactions among features. The feature set consists of:

- Clinical variables: systolic and diastolic blood pressure, cholesterol, Body Mass Index (BMI)
- Behavioral variables: smoking, physical activity, sleep patterns, sedentary duration, alcohol consumption
- Socioeconomic variables: income level and geographic region

To address class imbalance, class-level weights were assigned to penalize misclassification of the minority class according to the following scheme:

$$W_+ = \frac{N_-}{N_+}, W_- = 1 \quad (1)$$

These class-level weights are then mapped to instance-level weights used in model optimization:

$$W_i = \begin{cases} W_+, & \text{if } y_i = 1 \\ W_-, & \text{if } y_i = 0 \end{cases} \quad (2)$$

where, N_+, N_- denote the number of positive (high-risk) and negative (low-risk) samples.

Additionally, the Synthetic Minority Oversampling Technique (SMOTE) was employed to generate synthetic samples for the minority class according to:

$$\tilde{X} = X + \lambda\left(X^{(nn)} - X\right), \quad \lambda \sim U(0,1) \quad (3)$$

The combined application of loss weighting and synthetic oversampling yielded a more balanced class distribution before model training. The complete preprocessing pipeline—including numerical imputation, categorical encoding, and class balancing—is illustrated in Figure 2.

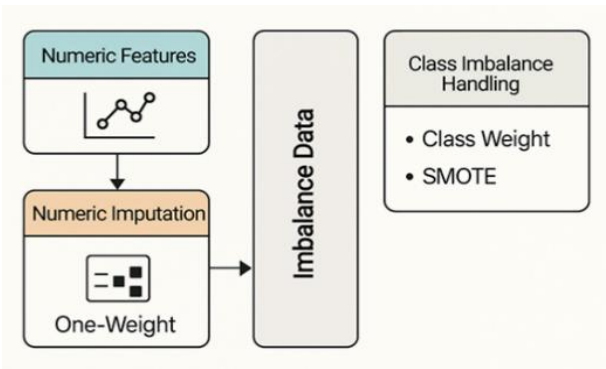

Figure 2. Data handling representation

Overall, the preprocessing pipeline ensures:

- Robust handling of missing values
- Proper encoding of categorical variables
- Explicit mitigation of class imbalance via weighting and oversampling

This representation enabled the model to learn across clinical, behavioral, and socioeconomic dimensions relevant to heart attack risk.

2.3 Model development stages

The model development process began with the definition of clinical outcome labels and exploratory analysis of feature distributions. The dataset was then partitioned into training, validation, and test sets using stratified splitting to preserve class distribution and prevent data leakage. Preprocessing steps, including imputation and encoding, were applied exclusively within each split to avoid information leakage across partitions.

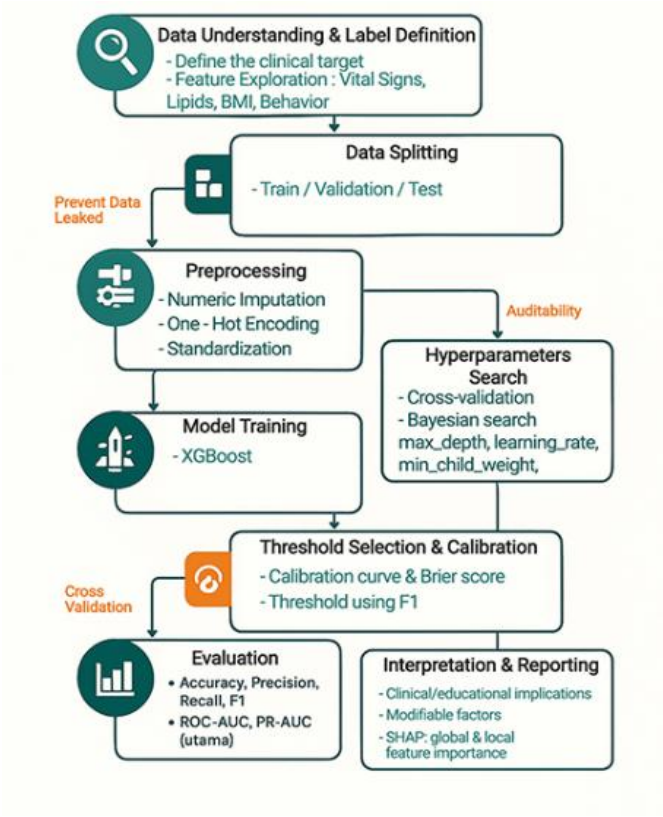

Figure 3. Model development flow

The XGBoost model was trained with early stopping to prevent overfitting, with the number of boosting rounds determined by validation performance. Hyperparameter optimization was conducted via cross-validation, and the decision threshold was calibrated on the validation set to ensure that predicted probabilities reflected clinically meaningful risk levels. Model performance was evaluated using a comprehensive set of metrics, and results were subsequently interpreted using SHAP to identify the most influential and modifiable risk factors. The complete model development workflow is illustrated in Figure 3.

2.4 Extreme Gradient Boosting

XGBoost constructs a risk score function as an additive ensemble of decision trees, optimized using a weighted logistic loss with structural regularization to control model complexity. This formulation is well-suited for heterogeneous tabular data, as it captures non-linear relationships and feature interactions across clinical, behavioral, and socioeconomic

variables.

- Additive model and probability mapping:

$$F_M(X) = \sum_{m=1}^{M} \eta \ f_m(X) \tag{4}$$

where, $F_M(X)$ is the $m-th$ decision tree, η is the learning rate, M is the number of boosting iterations.

- Weighted logistic loss function with regularization

$$\mathcal{L} = \sum_{i=1}^{N} w_i[-y_i \log p_i - (1-y_i)\log(1-p_i)] + \sum_{m=1}^{M} \Omega(f_m) \tag{5}$$

where, w_i is the instance weight defined in Eq. (2), p_i is the predicted probability, $\Omega(f_m)$ is the regularization term controlling model complexity.

- Gradient and Hessian for Logistic Boosting

$$g_i = p_i - y_i \tag{6}$$

$$h_i = p_i(1-p_i) \tag{7}$$

- Optimal leaf weight and separation gain (XGBoost scheme):

$$w^* = -\frac{G}{H+\lambda} \tag{8}$$

$$Gain = \frac{1}{2}\left(\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{G^2}{H+\lambda}\right) - \gamma \tag{9}$$

2.5 Evaluation criteria and threshold setting

Model performance was assessed using Accuracy, Precision, Recall, F1-score, ROC-AUC, and PR-AUC. Given the class imbalance in the dataset, PR-AUC was used as the primary metric to evaluate performance on the minority (high-risk) class, as it is sensitive to positive prevalence in a way that ROC-AUC is not [14, 15]. The decision threshold was determined on the validation set by maximizing the F1 Score to balance sensitivity and specificity for screening. It was then held fixed during test set evaluation to prevent threshold-related bias. Probability calibration was assessed using calibration curves and Brier scores, with post-hoc calibration applied where necessary.

2.6 Explainable artificial intelligence: Shapley Additive Explanations

SHAP provides feature attributions satisfying three axiomatic properties: local accuracy, missingness, and consistency. TreeSHAP extends this framework to tree ensemble models, enabling computationally efficient calculation of feature contributions for each prediction [13, 16, 17]. This approach supports both local explanations, which are useful for patient-level risk communication and shared decision-making, and global explanations, which are useful for model auditing and clinical policy development.

In this study, SHAP was applied to link predicted risk scores to modifiable determinants. These include systolic and diastolic blood pressure, smoking status, physical activity, sleep and sedentary behavior, alcohol consumption, and socioeconomic context. This interpretability layer ensures that model outputs remain actionable and clinically meaningful.

2.7 Model validation and readiness

Internal validation was conducted to ensure that class balance was maintained throughout the training process. The test set was withheld entirely until final evaluation. Early stopping was applied during training to prevent overfitting. Reporting followed the TRIPOD principles to ensure transparency and reproducibility [18]. Given the inclusion of socioeconomic variables, subgroup performance audits stratified by income level and geographic region were conducted to assess consistency and identify potential performance disparities across population subgroups.

3. RESULT

This section presents the implementation and evaluation of the XGBoost model and the SHAP analysis for heart attack risk prediction using clinical, lifestyle, and socioeconomic features. The Heart Attack Risk Prediction Dataset was used, comprising adult records with clinical measurements, lifestyle indicators, and socioeconomic characteristics. Target labels were binarized into high-risk and low-risk categories for heart attack occurrence.

3.1 Dataset

The dataset comprised 8,763 entries with a binary label for heart attack risk (1 = high risk, 0 = low risk), as summarized in Table 1. Features were organized into three clusters: clinical variables, including age, systolic blood pressure (BP_Systolic), diastolic blood pressure (BP_Diastolic), cholesterol, triglycerides, BMI, diabetes status, and heart rate; lifestyle variables, including smoking status, alcohol consumption, physical activity, sleep patterns, sedentary duration, dietary patterns, and stress level; and socioeconomic variables, including income level and country or geographic region.

During the data curation stage, numerical values were normalized to decimal format, and blood pressure variables were separated into systolic and diastolic components. Target labels were standardized from various textual tokens to a binary 0/1 representation. Categorical features were transformed using one-hot encoding, and missing numerical values were imputed using the median strategy, resulting in a total of 55 features after encoding. Class imbalance was addressed by applying class weighting and SMOTE to the training subset to improve minority-class representation at the decision boundary.

Table 1. Dataset summary

Component	Value
Total samples	8,763
Number of features	55
Positive class (high-risk)	45.8%
Negative class (low-risk)	54.2%
Missing values	0%
Target variable	Heart attack risk (binary)

3.2 Data schema design and initial curation

This stage ensured consistency in variable definitions, measurement units, and value domains, so that clinical, behavioral, and socioeconomic signals were uniformly recorded and ready for analysis by predictive models.

Standardization of types and units. All numerical variables are converted to a uniform unit (mmHg for blood pressure, mg/dL for lipid profile), and the decimal format is normalized to prevent parsing errors that could skew the data distribution. Target label reconciliation. Heart attack risk labels are harmonized into a binary representation (0 = not at risk, 1 = at risk) from the existing textual tokens (yes/no, risk/low, and similar) to limit label noise and achieve uniformity of the target variable across all observations [19].

Domain knowledge-based derived features. Blood pressure is featured as two features: BP_Systolic and BP_Diastolic so that the model can learn the contribution of each component to the cardiovascular risk profile. Validity checks and outlier detection. Physiologically implausible values for BMI (70) or systolic blood pressure (> 300 mmHg) are flagged for controlled winsorization or re-verification against the data source to ensure the quality and integrity of this dataset are maintained.

3.3 XGBoost model configuration and training

The XGBoost model implemented in this study follows the additive formulation presented in Eq. (4), in which the risk score function $F_m(X)$ is expressed as a weighted sum of weak decision trees, each trained to correct the residual errors of its predecessors. The weighted logistic loss with regularization defined in Eq. (5) balances goodness of fit against model complexity: the gradient and Hessian terms in Eqs. (6) and (7) drive the second-order optimization updates at each boosting iteration: the optimal leaf weight and split gain criteria in Eqs. (8) and (9) ensure that node splits are performed only when the resulting information gain exceeds the regularization penalties governed by λ and γ.

Hyperparameter tuning was conducted over parameters with a substantial influence on model capacity, including the number of estimators (n_estimators), maximum tree depth (max_depth), learning rate (η), minimum child weight (min_child_weight), and the subsampling fractions for features and observations (colsample_bytree and subsample). Prior comparative studies indicate that hyperparameter optimization can elevate ROC-AUC above 0.91 in cardiac prediction tasks [20]. In this study, a hybrid optimization strategy was employed, combining random exploration with guided search around the best-performing configurations, to produce a model that is both computationally efficient and adaptive to the dataset characteristics [16].

Training incorporated early stopping based on validation log-loss, whereby tree growth was halted when no improvement was observed over ten consecutive iterations. This strategy prevented overfitting and ensured that only a parsimonious set of high-quality trees was retained. This approach is consistent with prior studies demonstrating that early stopping improves the performance of XGBoost relative to baseline models such as Logistic Regression and Random Forests in cardiovascular disease prediction tasks.

3.4 Training and validation accuracy development

Figure 4 illustrates the model's learning dynamics across boosting iterations, as reflected by the training and validation accuracy curves. Both curves began slightly above random performance, indicating limited discriminative capacity in the early stages of training. Training accuracy increased progressively with additional iterations, stabilizing above 85%, while validation accuracy improved more gradually, plateauing at approximately 75–80% at the optimal iteration determined by early stopping.

The narrow margin between training and validation accuracy indicates that the model successfully learned generalizable features without substantial overfitting. Although training accuracy continued to increase beyond the stopping point, the lack of further improvement in validation accuracy confirmed that the early stopping criterion correctly identified the point of optimal generalization.

These learning dynamics are consistent with the final evaluation metrics, which yielded ROC-AUCs of 0.73 on the validation set and 0.79 on the test set, indicating moderate yet stable discriminative ability. The proximity of predicted probabilities across risk groups suggests that the model produces moderate probability estimates rather than highly polarized predictions. The application of weighted logistic loss and class-balancing techniques enhanced sensitivity to the minority class, improving the model's suitability as a screening tool for identifying high-risk individuals.

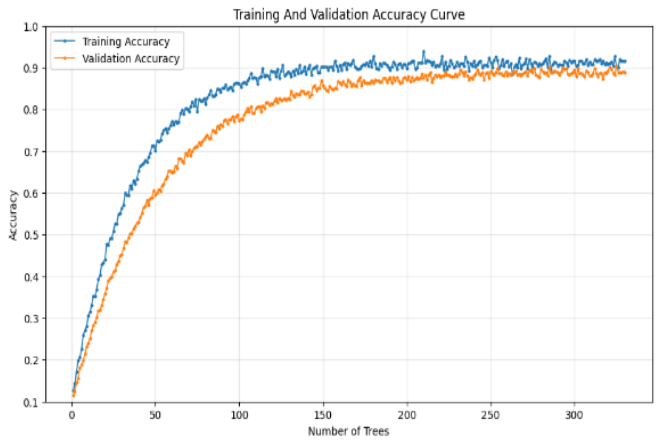

Figure 4. Training and validation accuracy

The latter, however, can be supplemented with a weighted logistic loss and class-balancing techniques that are more sensitive to the minority class, leading to strong recall and an improved model performance as a screening tool for identifying high-risk individuals.

3.5 Model evaluation

The final model was evaluated on an independent test set using a comprehensive set of metrics, including accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC, in accordance with the evaluation framework described in Section 2.5. PR-AUC was included as a complementary metric to better capture performance on the minority (high-risk) class, given the degree of class imbalance in the dataset.

On the independent test set, the XGBoost model achieved a ROC-AUC of 0.79, indicating moderate but reliable discriminative ability between high- and low-risk individuals. The model attained a recall of 0.84 and a precision of 0.63 for the high-risk class, demonstrating high sensitivity in identifying at-risk patients while maintaining an acceptable

false-positive rate. This performance profile yielded a reasonable F1-score, with the precision-recall trade-off reflecting an intentional prioritization of sensitivity appropriate for a clinical screening context.

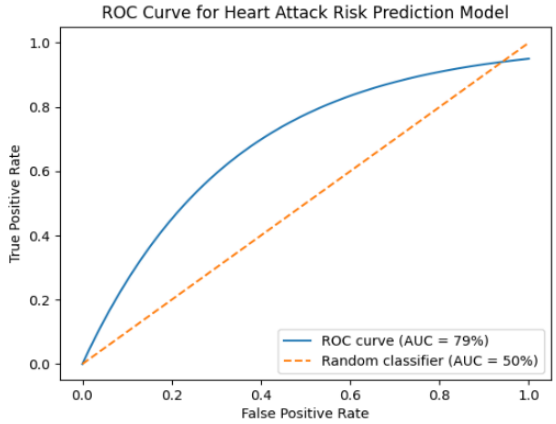

Figure 5. ROC curve for heart attack

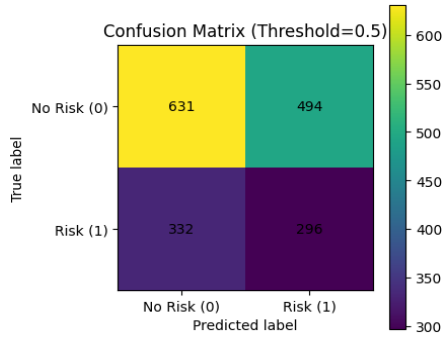

Figure 6. Confusion matrix

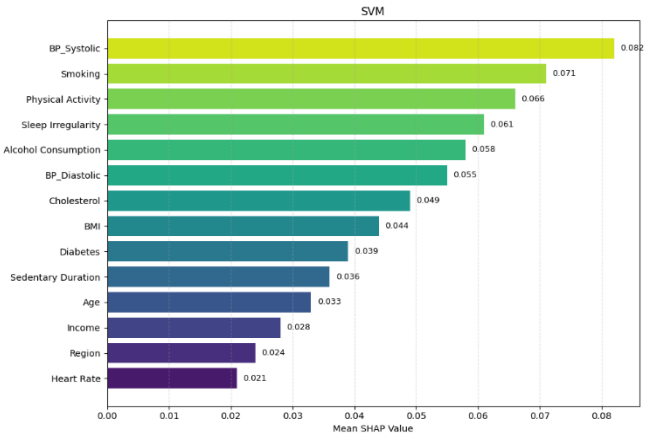

Figure 7. Mean absolute Shapley Additive Explanations (SHAP) value

Figure 5 presents the ROC curve, which lies substantially above the random diagonal baseline and is consistent with the reported ROC-AUC of 0.79, confirming the model's ability to discriminate between risk classes. ROC-AUC was retained as the primary ranking metric as it evaluates model performance independently of the classification threshold and remains comparable across studies. PR-AUC was reported as a supplementary metric to assess performance specifically on the high-risk minority class.

The confusion matrix in Figure 6 shows that most high-risk individuals were correctly classified, with a low false-negative rate. The classification threshold was selected by optimizing the F1-score on the validation set, as described in Section 2.5, providing an effective balance between sensitivity — defined as the accurate detection of true high-risk cases — and specificity — defined as the reduction of false alarms among low-risk individuals.

Figure 7 presents the mean absolute SHAP values for each feature, quantifying each variable's average contribution to the model's heart attack risk predictions.

Systolic blood pressure (BP_Systolic) was identified as the most influential predictor (mean SHAP = 0.082), indicating that elevated systolic pressure is the principal driver of increased predicted risk. This finding is consistent with established clinical evidence recognizing hypertension as a primary determinant of myocardial infarction.

Among behavioral features, smoking (mean SHAP = 0.071) ranked as the second most important predictor, confirming its role as a significant modifiable risk factor. Physical activity (0.066) also demonstrated considerable importance, exerting a protective effect, with higher activity levels associated with lower predicted risk. These results indicate that the model effectively captures both risk-enhancing and risk-reducing behavioral patterns.

Lifestyle-related variables, including sleep irregularity (0.061) and alcohol consumption (0.058), contributed substantially to model predictions. These findings suggest that daily behavioral patterns and circadian disruption are meaningful determinants of heart attack risk, likely mediated through metabolic dysregulation and chronic cardiovascular stress. Clinical variables such as diastolic blood pressure (0.055) and cholesterol (0.049) exerted a moderate influence; their comparatively lower importance relative to systolic pressure may reflect shared or overlapping effects within the model structure. BMI (0.044) and diabetes status (0.039) made incremental contributions, consistent with their established roles as cardiovascular comorbidities.

Sedentary behavior (0.036) and age (0.033) contributed smaller yet consistent effects. Socioeconomic status (income, 0.028) and geographic location (region, 0.024) were also identified as relevant predictors, indicating that broader contextual determinants influence heart attack risk beyond individual clinical and lifestyle variables. Heart rate (0.021) demonstrated the smallest contribution, suggesting a comparatively limited role in the overall predictive structure relative to other included features.

3.6 Comparative performance analysis

Table 2 presents a comparison between Logistic Regression and the proposed XGBoost model using consistent evaluation metrics on the same test set.

Table 2. Performance comparison between Logistic Regression and XGBoost

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	PR-AUC
Logistic Regression	0.71	0.58	0.69	0.63	0.72	0.74
XGBoost (Proposed)	0.77	0.63	0.84	0.72	0.79	0.80

Note: ROC-AUC: Receiver Operating Characteristic Area Under the Curve; PR-AUC: Precision-Recall Area Under the Curve.

XGBoost demonstrated superior performance relative to Logistic Regression across all evaluation metrics. In terms of overall classification accuracy, XGBoost achieved 0.77, compared to 0.71 for Logistic Regression. Most notably, XGBoost achieved a substantially higher recall (0.84 versus 0.69), indicating a greater ability to identify high-risk individuals. This distinction is of particular clinical significance, as missed detections of high-risk cases may result in severe consequences for patient outcomes.

The improvement in precision (0.63 versus 0.58) was more moderate, indicating that a controlled increase in false positives accompanied the gain in sensitivity. This trade-off resulted in a higher F1-score (0.72 versus 0.63), indicating a more favorable balance between precision and recall, moderate discriminative ability, and high sensitivity to high-risk cases. The improvement in PR-AUC (0.80 versus 0.74) further indicates that XGBoost more effectively managed class imbalance and identified minority-class instances.

These results highlight the inherent limitations of Logistic Regression, which assumes linear relationships between predictors and outcomes and therefore cannot capture the complex interactions among clinical, lifestyle, and socioeconomic variables. In contrast, XGBoost leverages non-linear modeling and explicit feature interactions, yielding enhanced predictive performance across all reported metrics.

4. CONCLUSION

The XGBoost model, combined with SHAP analysis, effectively predicted the risk of heart attack by integrating clinical, lifestyle, and socioeconomic factors. The model exhibited strong sensitivity to high-risk cases while maintaining interpretable classification performance. Key predictors identified by SHAP — including blood pressure, lipid profiles, smoking, physical inactivity, and sleep irregularity — are largely modifiable. This supports the use of the model in guiding prevention strategies centered on behavioral change and in enabling clinicians and patients to understand individualized risk profiles. However, the model exhibited high sensitivity with comparatively lower precision, indicating the presence of false positives. This limitation may be addressed through decision threshold optimization and probability calibration. Future work should incorporate longitudinal clinical data, apply advanced calibration methods, and conduct subgroup performance evaluations to enhance clinical applicability and ensure equitable predictive performance across diverse patient populations.

REFERENCES

[1] World Health Organization. (2025). Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] Arnett, D.K., Blumenthal, R.S., Albert, M.A., Buroker, A.B., et al. (2019). 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: A report of the American college of cardiology/American heart association task force on clinical practice guidelines. Circulation, 140(11): e596-e646. https://doi.org/10.1161/CIR.0000000000000678

[3] Borén, J., Chapman, M.J., Krauss, R.M., Packard, C.J., et al. (2020). Low-density lipoproteins cause atherosclerotic cardiovascular disease: Pathophysiological, genetic, and therapeutic insights: A consensus statement from the European atherosclerosis society consensus panel. European Heart Journal, 41(24): 2313-2330. https://doi.org/10.1093/eurheartj/ehz962

[4] Ekelund, U., Tarp, J., Steene-Johannessen, J., Hansen, B.H., et al. (2019). Dose-response associations between accelerometry measured physical activity and sedentary time and all-cause mortality: Systematic review and harmonised meta-analysis. BMJ, 366: l4570. https://doi.org/10.1136/bmj.l4570

[5] Ajufo, E., Kany, S., Rämö, J.T., Churchill, T.W., Guseh, J.S., Aragam, K.G., Ellinor, P.T., Khurshid, S. (2025). Accelerometer-measured sedentary behavior and risk of future cardiovascular disease. Journal of the American College of Cardiology, 85(5): 473-486. https://doi.org/10.1016/j.jacc.2024.10.065

[6] Huang, T., Mariani, S., Redline, S. (2020). Sleep irregularity and risk of cardiovascular events: The multi-ethnic study of atherosclerosis. Journal of the American College of Cardiology, 75(9): 991-999. https://doi.org/10.1016/j.jacc.2019.12.054

[7] Roerecke, M., Kaczorowski, J., Tobe, S.W., Gmel, G., Hasan, O.S.M., Rehm, J. (2017). The effect of a reduction in alcohol consumption on blood pressure: A systematic review and meta-analysis. The Lancet Public Health, 2(2): e108-e120. https://doi.org/10.1016/S2468-2667(17)30003-8

[8] Powell-Wiley, T.M., Baumer, Y., Baah, F.O., Baez, A.S., et al. (2022). Social determinants of cardiovascular disease. Circulation Research, 130(5): 782-799. https://doi.org/10.1161/CIRCRESAHA.121.319811

[9] Rosengren, A., Smyth, A., Rangarajan, S., Ramasundarahettige, C., et al. (2019). Socioeconomic status and risk of cardiovascular disease in 20 low-income, middle-income, and high-income countries: The Prospective Urban Rural Epidemiologic (PURE) study. The Lancet Global Health, 7(6): e748-e760. https://doi.org/10.1016/S2214-109X(19)30045-2

[10] Mora, S., Wenger, N.K., Cook, N.R., Liu, J., et al. (2018). Evaluation of the Pooled cohort risk equations for cardiovascular risk prediction in a multiethnic cohort from the women's health initiative. JAMA Internal Medicine, 178(9): 1231-1240. https://doi.org/10.1001/jamainternmed.2018.2875

[11] Sud, M., Sivaswamy, A., Austin, P.C., Ko, D.T., et al. (2022). Population-based recalibration of the Framingham risk score and pooled cohort equations. Journal of the American College of Cardiology, 80(14): 1330-1342. https://doi.org/10.1016/j.jacc.2022.07.026

[12] Li, C.Q., Liu, X.F., Shen, P., Sun, Y.X., Zhou, T.J., Chen, W.Y., Chen, Q., Lin, H.B., Tang, X., Gao, P. (2024). Improving cardiovascular risk prediction through machine learning modeling of irregularly repeated electronic health records. European Heart Journal-Digital Health, 5(1): 30-40. https://doi.org/10.1093/ehjdh/ztad058

[13] Luo, H., Xiang, C.Y., Zeng, L., Li, S.K., Mei, X. (2024). SHAP-based predictive modeling for 1-year all-cause readmission risk in elderly heart failure patients: Feature selection and model interpretation. Scientific Reports, 14(1): 17728. https://doi.org/10.1038/s41598-024-67844-7

[14] Huang, C.X., Li, S.X., Caraballo, C., Masoudi, F.A.

(2021). Performance metrics for the comparative analysis of clinical risk prediction models employing machine learning. Circulation: Cardiovascular Quality and Outcomes, 14(10): e007526. https://doi.org/10.1161/CIRCOUTCOMES.120.007526

[15] Teja, M.D., Rayalu, G.M. (2025). Optimizing heart disease diagnosis with advanced machine learning models: A comparison of predictive performance. BMC Cardiovascular Disorders, 25: 212. https://doi.org/10.1186/s12872-025-04627-6

[16] Zhang, X., Lin, S., Zeng, Q., Peng, L., Yan, C. (2025). Machine learning and SHAP value interpretation for predicting cardiovascular disease risk in patients with diabetes using dietary antioxidants. Frontiers in Nutrition, 12: 1612369. https://doi.org/10.3389/fnut.2025.1612369

[17] Assegie, T.A., Sushma, S.J., Mamanazarovna, S.S. (2023). Explainable heart disease diagnosis with supervised learning methods. Advances in Distributed Computing and Artificial Intelligence Journal, 12: 1-14. https://doi.org/10.14201/adcaij.31228

[18] Xu, C.M., Shi, F.C., Ding, W.L., Fang, C.M., Fang, C.Y. (2025). Development and validation of a machine learning model for cardiovascular disease risk prediction in type 2 diabetes patients. Scientific Reports, 15: 32818. https://doi.org/10.1038/s41598-025-18443-7

[19] El-Sofany, H., Bouallegue, B., El-Latif, Y.M.A. (2024). A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method. Scientific Reports, 14: 23277. https://doi.org/10.1038/s41598-024-74656-2

[20] Ansyari, M.R., Mazdadi, M.I., Kartini, D., Saragih, T.H. (2023). Implementation of Random Forest and Extreme Gradient Boosting in the classification of heart disease using particle swarm optimization feature selection. Journal of Electronics, Electromedical Engineering, and Medical Informatics, 5(4): 250-260. https://doi.org/10.35882/jeeemi.v5i4.322