Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU33293

Predicting Heart Disease Risk: A Machine Learning-Based Analysis

Kaijie Tian

School of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University, Taicang, China
tiankaijie4@gmail.com

Abstract. Heart disease continues to be a significant global health concern that requires accurate risk assessment tools to assist researchers and clinicians in decision-making. This study employs various mechanical learning models to evaluate heart disease risk using a dataset comprising 1026 records and 14 predictor variables. This study employs Gradient Boosting, Support Vector Machine, and Logistic Regression, with evaluation based on accuracy and AUC to explore key clinical factors of heart disease. The findings demonstrate that the Gradient Boosting model has the advantage over alternative models, with an accuracy of almost 98% and an AUC of 0.988. The significant predictors identified are as follows: chest pain type (cp), count of major vessels (ca), and ST depression (oldpeak), which are consistent with clinical medical literature. This search offers a robust machine learning-based methodology for predicting heart disease risk, which serves as a valuable clinical reference for early intervention and disease management.

Keywords: Heart Disease Prediction, Support Vector Machine, Gradient Boosting, Clinical Data Analysis

1. Introduction

Cardiovascular diseases, specifically heart diseases, are among the leading causes of death or disability worldwide, which cause significant strain on health systems and global public health [1]. The ability to assess risk accurately and quickly, therefore, is essential; it supports early diagnosis, the formulation of personalized treatment plans, and helps to reduce negative patient outcomes. Key risk factors include demographic traits (age, gender) and clinical indicators (blood pressure, cholesterol, chest pain characteristics) [1]. However, traditional risk assessment methods often rely on linear correlations and limited variables, leading to suboptimal predictive performance in complex clinical scenarios. Machine learning (ML) has emerged as a powerful tool for extracting nonlinear patterns from high-dimensional data [2]. Various ML algorithms have been applied to heart disease prediction, including logistic regression, SVM, and gradient boosting [3,4]. However, few studies systematically compare these algorithms using a comprehensive dataset or have identified the most influential predictive features through rigorous feature importance analysis.

This paper employs a heart disease dataset to systematically evaluate and compare the performance of three machine learning models: logistic regression, SVM, and Gradient Boosting. This study, by providing a transparent and reproducible framework for heart disease prediction,

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GU33293

validates the clinical importance of insights derived from machine learning. It provides a practical, strong tool for healthcare providers to improve risk stratification, enable targeted intervention in high-risk populations, and, overall, improve the efficiency of healthcare service delivery.

2. Literature review

Numerous studies have identified key risk factors for heart disease. Age is a major non-modifiable risk factor, and gender differences also contribute to disease vulnerability [5,6]. Common modifiable clinical indicators include elevated resting blood pressure and serum cholesterol, which are closely associated with atherosclerosis [7]. In addition, chest pain type, exercise-induced angina, and ST depression are important markers of myocardial ischemia in clinical diagnosis [8].

Machine learning tools are increasingly used to predict heart disease for picking up on the complicated, non-linear connections between different factors that affect heart health. Logistic regression, a traditional statistical learning method, is favored for its simplicity and interpretability, making it a benchmark in clinical prediction models [9]. HSVM projects data into higher-dimensional space to handle non-linear separability, achieving 80-90% accuracy for heart disease prediction depending on the dataset [10,11]. Gradient boosting combines weak learners (decision trees) through iterative error reduction [12]. Recent studies report its accuracy often exceeds 95%, outperforming logistic regression and SVM [13].

Despite promising results, several gaps remain. Few studies directly compare algorithms with thorough feature importance analysis. Many studies use small or heterogeneous datasets with inconsistent formats, limiting generalizability. Feature selection often lacks systematic rigor and clinical validation. Moreover, comparisons among ML algorithms frequently omit evaluations of model stability, generalization on imbalanced data, and AUC metrics.

3. Methodology

3.1. Dataset

This paper uses a dataset containing 1026 patient records, with 14 attributes per record. Each record corresponds to a single patient [14]. The attributes are separated into 4 main groups: demographic data (age, sex), clinical measurements (resting systolic blood pressure, cholesterol levels, fasting blood sugar), physiological measurements (maximum heart rate achieved, exercise-induced angina, ST depression), and a target variable. The target variable indicates whether the patient had heart disease (coded as 1) or not (coded as 0).

3.2. Data preprocessing

Data preprocessing included handling missing values, outlier clipping, categorical encoding, and standardization. Missing values were identified and processed. Records with missing key features were removed, and remaining numerical missing values were imputed by the median. Regarding outlier detection, the interquartile range (IQR) method was applied. Values falling outside the range of $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$ were treated as outliers and subsequently clipped to the respective lower or upper bound to mitigate their influence on model training. Categorical variables, such as sex and chest pain type, were encoded using one-hot encoding to ensure compatibility with machine learning algorithms that require numerical inputs. All numerical features were standardized via z-score normalization to meet the requirements of SVM and gradient boosting.

DOI: 10.54254/2755-2721/2026.GU33293

3.3. Exploratory Data Analysis (EDA)

EDA was performed to analyze feature distributions, relationships, and correlations with heart disease. Univariate analysis included bar charts for the target variable and histograms for age. Bivariate analysis used grouped bar charts, scatter plots, and a correlation heatmap to explore associations among gender, blood pressure, cholesterol, and heart disease, and to identify strongly predictive features

3.4. Model selection and evaluation

3.4.1. Model selection

Three models were selected for comparison based on clinical prediction suitability: 1. Logistic Regression: A generalized linear model suitable for binary classification; it was implemented as a baseline due to its interpretability and computational efficiency. It allows direct examination of feature coefficients, providing insight into each variable's contribution to the prediction.

Support Vector Machine (SVM): A kernel-based classifier capable of handling high-dimensional data through margin maximization. Its ability to capture nonlinear relationships makes it suitable for datasets with correlated features and complex decision boundaries.

Gradient Boosting: An ensemble method that sequentially builds weak learners (decision trees) to minimize residual errors. This algorithm is particularly effective at modeling intricate nonlinear patterns and interactions within the data.

3.4.2. Hyperparameter tuning and model configuration

A systematic hyperparameter optimization was conducted for each algorithm using 5-fold cross-validation on the training set:

Logistic Regression: L2 regularization with C values searched logarithmically in $[10^{-3}10^{3}]$.

SVM: Radial basis function kernel with C in $[0.110]$ and γ in $[0.0010.1]$.

Gradient Boosting: Learning rate in $[0.010.2]$, n_estimators in $[50200]$, and max_depth in $[36]$. The dataset was partitioned into training, validation, and test sets using stratified random sampling to preserve class distribution: Training set (70%): Model training and tuning. Validation set (15%): model selection and early stopping. Test set (15%): Unbiased final performance evaluation.

3.5. Evaluation

Using the confusion matrix of the Gradient Boosting model, SHAP enhances interpretability from global and local perspectives. Globally, SHAP confirms that cp, ca, and oldpeak are the core predictive features, changes in these indicators significantly influence the model's classification of disease presence. Locally, taking a correctly diagnosed heart disease patient as an example, SHAP identifies that severe chest pain with 2 stained vessels and significant ST depression are key to the model's decision, while age has a weaker impact. This interpretation transforms the model from a "black box" to a transparent tool, clarifying both dataset-level predictors and individual case predictions, thereby meeting clinical needs for understanding how and why a decision is made.

DOI: 10.54254/2755-2721/2026.GU33293

4. Results

4.1. Exploratory data analysis findings

4.1.1. Heart disease and age distribution

As shown in Figure 1, the bar chart presents a balanced distribution of heart disease across age groups, with roughly equal numbers of affected and unaffected individuals. The age histogram indicates that middle-aged and elderly people are more likely to have heart disease.

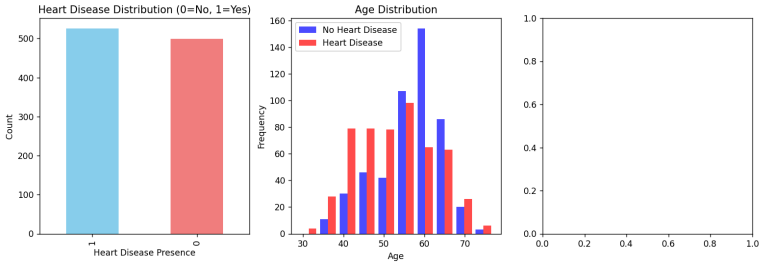

Figure 1. Heart disease distribution

4.1.2. Gender, blood pressure, and cholesterol analysis

Figure 2 illustrates the relationships between gender, resting systolic blood pressure, cholesterol and the presence of heart disease. For females, approximately 75 have no heart disease and 20 have heart disease. Thus, heart disease is much more common in males. The left subplot shows that resting blood pressure distributions for both groups (blue: no disease; red: disease) overlap heavily, ranging from 100 to 200 mmHg with peaks around 120-140 mmHg. The right subplot shows a similar pattern for cholesterol, with both groups peaking at 200-250 mg/dL. These high overlaps indicate that blood pressure and cholesterol alone have limited discriminative power for heart disease.

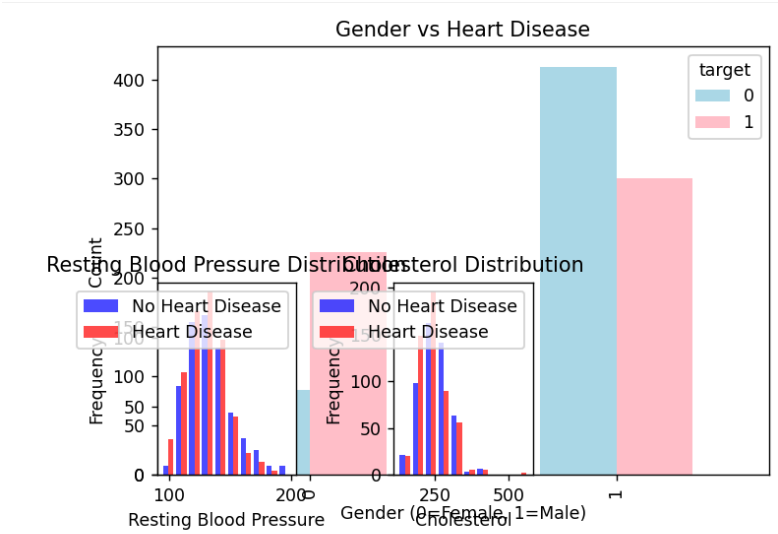

Figure 2. Relationship between gender, resting systolic blood pressure, cholesterol level, andthe presence of heart disease

DOI: 10.54254/2755-2721/2026.GU33293

4.1.3. Feature correlation

The relevant heat map (Figure 3) indicates the core features closely related to heart disease: cp: 0.43, thalach: 0.42, ca: -0.39, and old peak: -0.42. These correlation patterns are consistent with established medical knowledge. In addition, it can be seen from their relatively high correlation coefficient of 0.58 that there is redundancy between the old peak and slope variables.

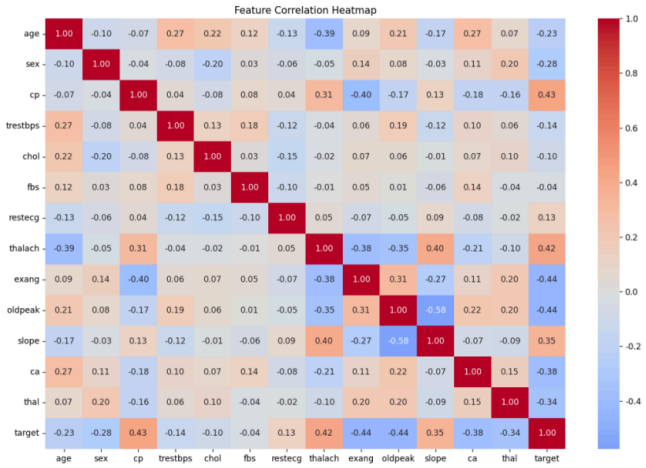

Figure 3. Feature correlation heatmap of the dataset

4.2. Model performance comparison

Table 1. Performance comparison of three classification models (Gradient Boosting, SVM and Logistic Regression) on the test set

Model	Accuracy	AUC	CV Accuracy	CV Std
Gradient Boosting	0.9756	0.9876	0.9634	0.0102
SVM	0.9219	0.9771	0.9121	0.0195
Logistic Regression	0.8097	0.9297	0.8439	0.0239

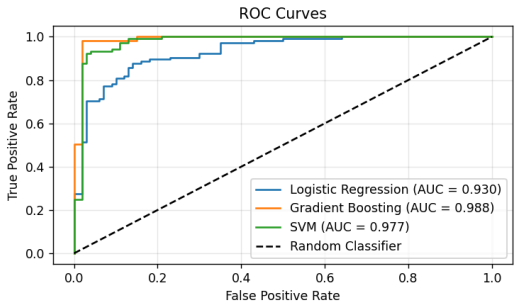

Figure 4. ROC curves of the three models

Table 1 summarizes the test set performance of three models. Gradient Boosting achieves the highest accuracy (0.9756) and AUC (0.9876), with a cross-validation standard deviation of 0.0102, indicating good stability. SVM performs well (in terms of accuracy), while logistic regression lags behind. Figure 4 shows the ROC curves of the three models, visually confirming that Gradient

DOI: 10.54254/2755-2721/2026.GU33293

Boosting has the best classification performance. Logistic Regression performs worst on both metrics, with scores lower than those of the other two models.

4.3. Key predictive features

In Figure 5, the horizontal axis shows feature importance scores (0.00–0.30) from the Gradient Boosting model, and the vertical axis lists the corresponding features. Feature cp is the most important (close to 0.30), followed by ca (≈0.15) and oldpeak (≈0.12). Traditional risk factors such as age (0.10) and sex (0.05) have much lower importance, while restecg and fbs show scores near 0.00. These results confirm that cp, ca, and oldpeak are the strongest predictors of heart disease in the Gradient Boosting model, outperforming conventional demographic factors. Clinically, prioritizing these three indicators can improve the accuracy of heart disease risk assessment.

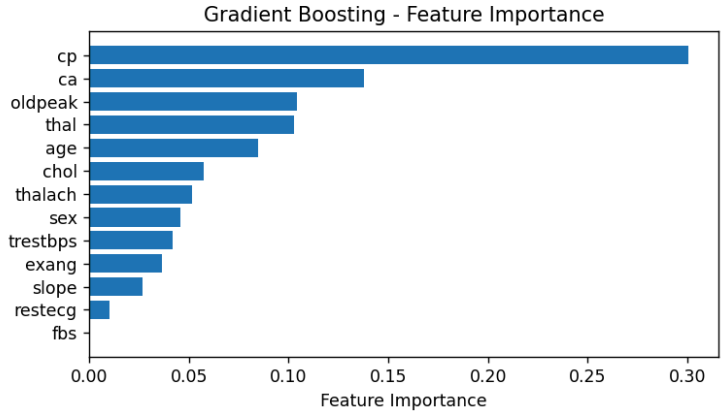

Figure 5. Feature importance scores of the Gradient Boosting model for heart disease prediction

SHAP enhances the interpretability of the Gradient Boosting model from both global and local perspectives. At the global level, SHAP verifies that cp, ca, and oldpeak are the core predictive features, and changes in these indicators significantly influence disease classification. At the local level, for a correctly diagnosed patient, SHAP identifies severe chest pain, two stained vessels, and significant ST depression as key factors for positive prediction, while age has a weaker impact.

5. Discussion

5.1. Model performance interpretation and clinical relevance

The results show that machine learning models, especially Gradient Boosting, achieve outstanding performance in heart disease prediction, with accuracy approaching 98% and an AUC of 0.988. Its superiority over SVM and logistic regression arises from its strong capability in capturing complex nonlinear relationships and feature interactions in clinical data. SVM performs well in high-dimensional data but is slightly less accurate due to sensitivity to hyperparameters and noise. Logistic regression remains useful for its high interpretability. The key predictors cp, ca, and oldpeak are consistent with clinical knowledge. They reflect typical symptoms, coronary artery disease severity, and myocardial ischemia, respectively [15]. The relatively lower importance of conventional factors such as age and sex suggests that modern machine learning models capture more integrated and subtle patterns, rather than diminishing the value of traditional indicators.

This study has several limitations. First, dataset constraints such as potential biases and standardization issues may limit the generalizability of the conclusions. Second, the black-box

DOI: 10.54254/2755-2721/2026.GU33293

nature of ensemble models like Gradient Boosting reduces transparency and hinders clinical acceptance.

5.2. Practical implications

The proposed models can serve as effective auxiliary tools for initial risk screening and clinical resource allocation. For clinical practice, the Gradient Boosting model can help clinicians evaluate individual risk, enable early intervention, and support personalized treatment. For public health, accurate risk stratification helps optimize medical resource allocation and reduce the overall burden of heart disease. Future work should adopt more explainable AI techniques and validate the models in larger, multi-center, prospective cohorts to improve reliability and promote clinical translation.

6. Conclusion

This study conducted empirical analysis on machine learning-based heart disease risk prediction using a comprehensive dataset of 1026 patients. The gradient boosting model outperforms logistic regression and support vector machines, with an accuracy of 97.56% and an AUC of 0.988, making it the best model for predicting heart disease risk. Key predictive features include chest pain type (cp), number of major blood vessels (ca), and ST segment depression (oldpeak), which are consistent with clinical medical knowledge.

This study has several limitations: the data may limit its generalizability to different populations (e.g., races, geographic regions). This study used cross-sectional data and excluded assessments of the relationship between variables and heart disease risk. Preserving outliers may introduce errors in the model, thereby affecting performance. Future studies should use diverse, longitudinal data and advanced feature engineering to improve performance and interpretability.

References

[1] World Health Organization. (2023). Cardiovascular Diseases (CVDs). https: //www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
[2] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press, Cambridge.
[3] Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer, New York.
[4] Zhang, Y., Li, J., & Wang, Z. (2020). Heart Disease Prediction Using Support Vector Machines. Journal of Medical Systems, 44(5), 78.
[5] Benjamin, E. J., Muntner, P., Alonso, A., et al. (2019). Heart disease and stroke statistics—2019 update: A report from the American Heart Association. Circulation, 139(10), e56-e528.
[6] Mosca, L., Barrett-Connor, E., & Wenger, N. K. (2011). Cardiovascular disease in women: A statement for healthcare professionals from the American Heart Association. Circulation, 123(11), 1243-1262.
[7] Grundy, S. M., Stone, N. J., Bailey, A. L., et al. (2018). 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA guideline on the management of blood cholesterol. Circulation, 139(25), e1082-e1143.
[8] Thygesen, K., Alpert, J. S., Jaffe, A. S., et al. (2018). Fourth universal definition of myocardial infarction (2018). Circulation, 138(20), e618-e651.
[9] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd ed.). Wiley, Hoboken.
[10] Vapnik, V. N. (1998). Statistical Learning Theory. Wiley, New York.
[11] Dua, D., & Graff, C. (2017). UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences.
[12] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.

DOI: 10.54254/2755-2721/2026.GU33293

[13] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM, San Francisco.
[14] Kaggle. (2025). https: //www.kaggle.com/datasets/johnsmith88/heart-disease-dataset.
[15] Libby, P., Mann, D. L., & Zipes, D. P. (2018). Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine (11th ed.). Elsevier, Philadelphia.