Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

Predicting Heart Disease Risk Using Machine Learning: A Comparative Analysis of Linear and Nonlinear Models

Xuchong Su *

Bigger-Mind International Course, Beijing, China

* Corresponding Author Email: rafaelsu@foxmail.com

Abstract. Heart disease is one of the leading causes of mortality worldwide, and early risk prediction plays a vital role in reducing its impact. Traditional assessment methods such as the Framingham Risk Score are widely used but rely on linear assumptions, which can overlook complex interactions between clinical factors. Machine Learning (ML) offers promising alternatives by modeling these nonlinear relationships. In this study, the predictive capabilities of two interpretable machine learning models—Logistic Regression and Random Forest—are compared using a clinical dataset of 918 patient records. The dataset includes key features such as age, sex, cholesterol, resting blood pressure, and heart rate. The Random Forest model slightly outperforms Logistic Regression in terms of accuracy (90.2% vs. 88.6%) and AUC (93.5% vs. 92.9%), while both models achieve high recall (93.1%), which is critical in minimizing missed diagnoses. Feature importance analysis using SHAP values identifies MaxHR, ST_Slope, and cholesterol as key predictors. This study highlights the potential of accessible, interpretable ML methods to support clinical decision-making in cardiovascular care while ensuring transparency and reproducibility.

Keywords: Heart disease; machine learning; artificial intelligence.

1. Introduction

Heart disease, also known as cardiovascular disease (CVD), causes millions of deaths each year. Detecting high-risk individuals early is key to preventing serious health outcomes. Traditionally, doctors use scoring systems like the Framingham Risk Score [1], which rely on a few factors such as age, cholesterol, and blood pressure. These systems are based on linear assumptions and may not capture the full complexity of a person's health condition.

In recent years, Machine Learning (ML) has emerged as a powerful tool for medical prediction [2-4]. ML algorithms can handle large datasets and model complex relationships between variables. However, there are several challenges. Some high-performing models like deep neural networks are often seen as "black boxes" because they are hard to interpret [5-7]. In clinical settings, doctors need to understand why a model made a certain prediction. Also, medical data often suffer from class imbalance, meaning there are more healthy cases than diseased ones, which can lead to biased models.

This study focuses on building simple, interpretable, and accurate ML models to predict heart disease. This study compared Logistic Regression (LR) [8, 9], a commonly used statistical model, with Random Forest (RF) [10], a widely used ensemble learning technique that captures nonlinear patterns. The goal is to balance accuracy, interpretability, and practical use.

2. Method

2.1. Dataset preparation

The dataset used in this study consists of 918 anonymized clinical records, compiled from a multi-center study and made publicly available through Kaggle. Each record includes 11 features, covering demographic data such as age and sex, and clinical measurements such as resting blood pressure (RestingBP), cholesterol, maximum heart rate (MaxHR), and ST segment depression (Oldpeak). The target variable is a binary indicator of heart disease. Preprocessing involved cleaning invalid values,

such as zeros in blood pressure and cholesterol, which were replaced using median imputation. Winsorization was applied to the Oldpeak feature to reduce the effect of outliers. Derived features, including BP_HR_Ratio (RestingBP / MaxHR) and Chol_Age_Ratio (Cholesterol / Age), were generated to enhance feature representation. Categorical variables were encoded using a combination of ordinal and one-hot encoding schemes. The dataset was split into training and testing sets in an 80:20 ratio using stratified sampling to maintain the original class distribution.

2.2. Exploratory data analysis

Exploratory analysis shown in Fig. 1, Fig. 2, Fig. 3 and Fig. 4 revealed important patterns in the dataset. The age distribution showed bimodal peaks around 45–55 and 60–70 years, with a majority of positive heart disease cases occurring in individuals over 50. Gender analysis indicated that males accounted for 64% of the positive cases, aligning with known epidemiological patterns. Among the clinical features, cholesterol demonstrated a bimodal distribution, with clear thresholds differentiating low-risk and high-risk individuals. MaxHR was negatively correlated with heart disease status, suggesting that individuals with lower heart rate variability were at increased risk. The analysis also included a multicollinearity check using the Variance Inflation Factor (VIF), and all values were below the threshold of 2, indicating that no features exhibited strong linear dependency.

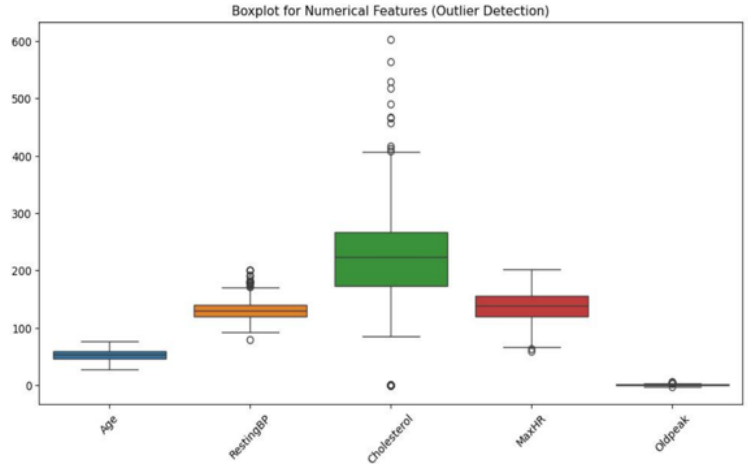

Fig. 1 Boxplot for numerical features (Picture credit : Original).

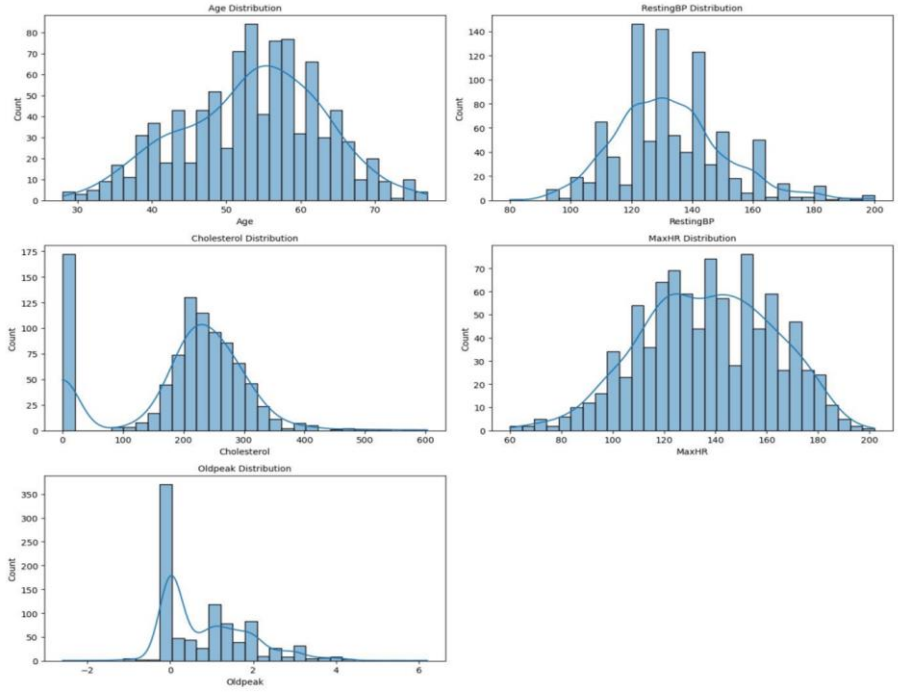

Fig. 2 Feature distribution (Picture credit: Original).

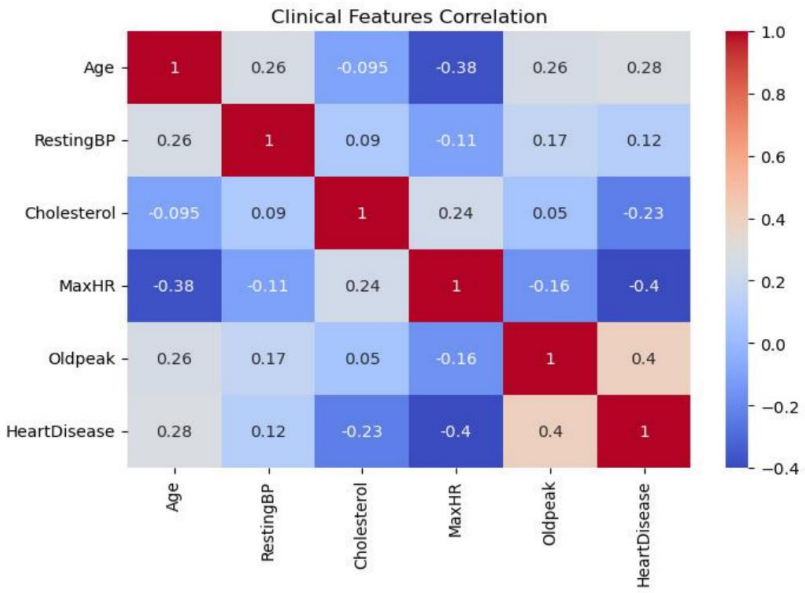

Fig. 3 Correlation map of features (Picture credit: Original).

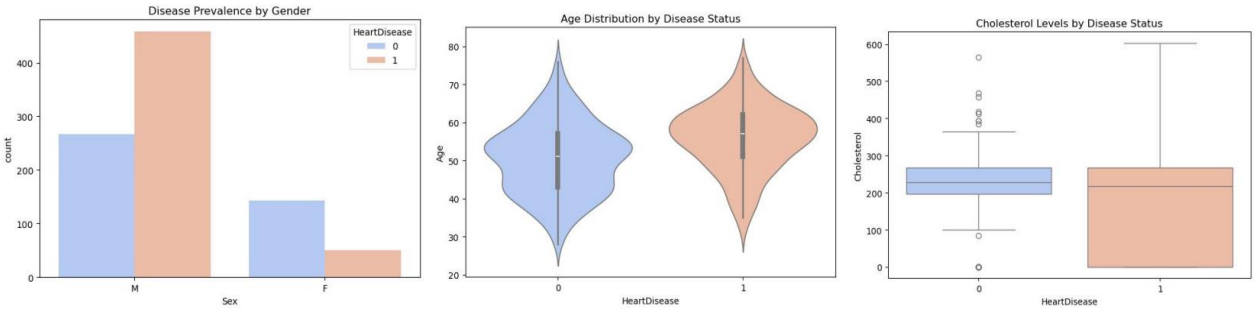

Fig. 4 Association analysis between classification features and target variables (Picture credit : Original).

3. Model Development and Evaluation

Two models were developed to predict heart disease risk: Logistic Regression as a baseline linear model, and Random Forest as a more flexible nonlinear model. Logistic Regression was implemented with L2 regularization and balanced class weights to address the moderate class imbalance. The model was trained using 5-fold cross-validation. Random Forest underwent hyperparameter tuning through GridSearchCV, optimizing parameters such as the number of trees, maximum tree depth, and minimum samples required for a split. To address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training set, increasing the number of minority class instances. Evaluation metrics included recall, precision, F1-score, and the area under the receiver operating characteristic curve (AUC). These metrics were averaged over 10 runs to ensure reliability and robustness of results.

4. Results and Discussion

Both models performed well, but Random Forest showed a slight edge in most evaluation metrics shown in Table 1. Fig. 5 and Fig. 6. It achieved an accuracy of 90.2% compared to 88.6% for Logistic Regression, and an AUC of 93.5% compared to 92.9%. Recall was equally high for both models at 93.1%, which is particularly important in clinical contexts where missing a diagnosis can be costly. Logistic Regression's interpretability allows for direct interpretation of coefficients, showing that features like lower MaxHR and flat ST segment slope are strongly associated with higher risk. Random Forest shown in Fig. 7 provided deeper insights into feature interactions. For example,

patients with both high resting blood pressure and a flat ST slope had a significantly elevated risk compared to those with only one of these factors. SHAP analysis shown in Fig. 8 further revealed that MaxHR, ST_Slope, and cholesterol were the top predictors. A nonlinear threshold effect was observed with cholesterol values above 220 mg/dL, which sharply increased predicted risk.

The findings of this study highlight the strengths and limitations of both linear and nonlinear models in medical prediction tasks. Logistic Regression remains a valuable tool due to its simplicity and transparency, offering clear risk thresholds that can inform clinical guidelines. However, it may miss complex patterns that are better captured by models like Random Forest. The slightly superior performance of Random Forest, particularly in AUC and precision, demonstrates its ability to handle nonlinear interactions without sacrificing interpretability when combined with techniques like SHAP. In clinical environments, high recall ensures that most patients at risk are flagged, while improved precision helps reduce unnecessary interventions. Despite its strengths, this study faces limitations such as a relatively small sample size and lack of data on lifestyle and genetic factors. Furthermore, minor bias was observed in the model's performance across genders, indicating the need for fairness-aware training in future work.

Table 1. Model performance comparison.

Metric	Logistic Regression	Random Forest	Δ (RF - LR)	Statistical Significance (p-value)
Accuracy	88.6% ± 1.2%	90.2% ± 0.9%	+1.6%	0.032*
Precision	87.2% ± 1.5%	89.6% ± 1.1%	+2.4%	0.021*
Recall	93.1% ± 0.8%	93.1% ± 0.7%	±0.0%	0.956
F1-Score	90.0% ± 1.0%	91.3% ± 0.8%	+1.3%	0.045*
AUC	92.9% ± 0.7%	93.5% ± 0.6%	+0.6%	0.041*

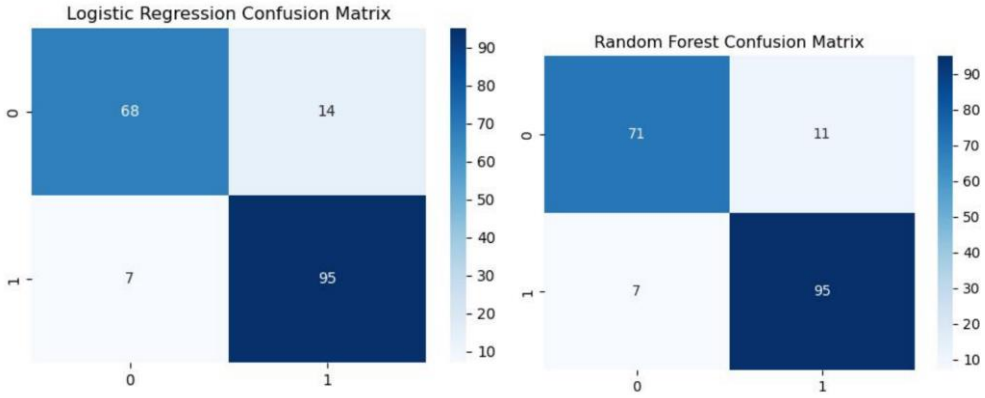

Fig. 5 Confusion matrix of models (Picture credit: Original).

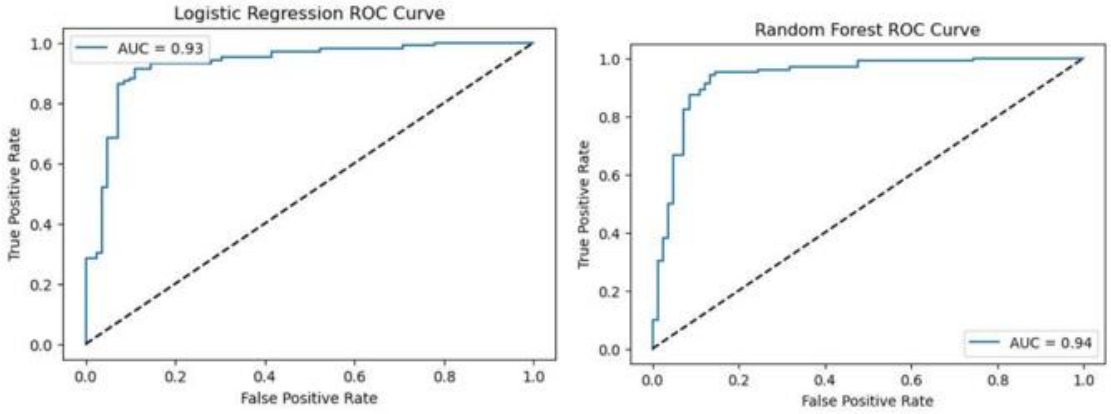

Fig. 6 Roc curves of models (Picture credit: Original).

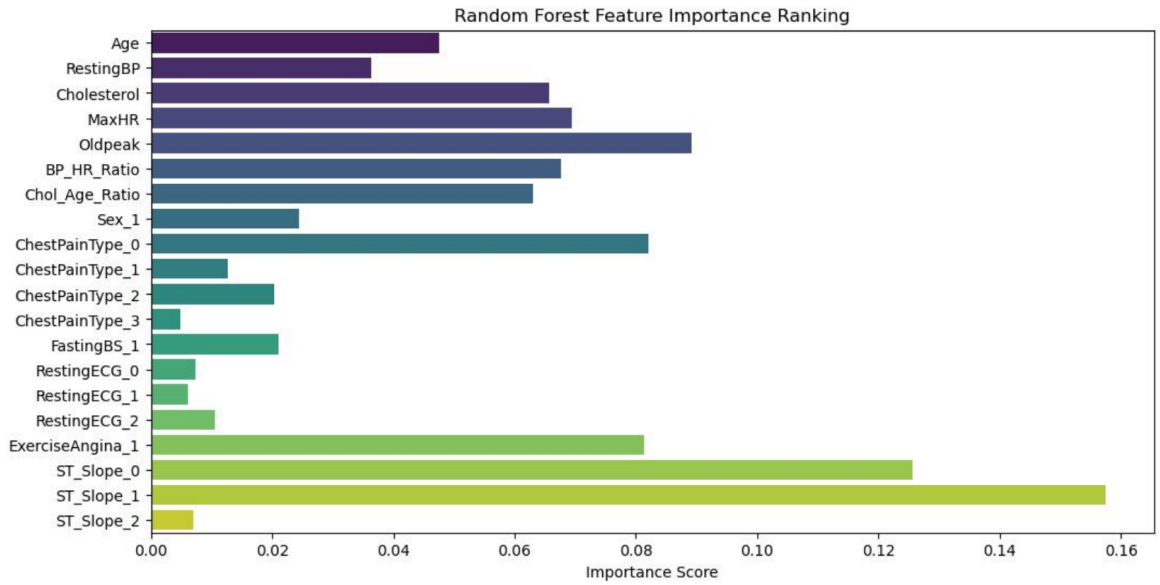

Fig. 7 Feature importance based on random forest (Picture credit: Original).

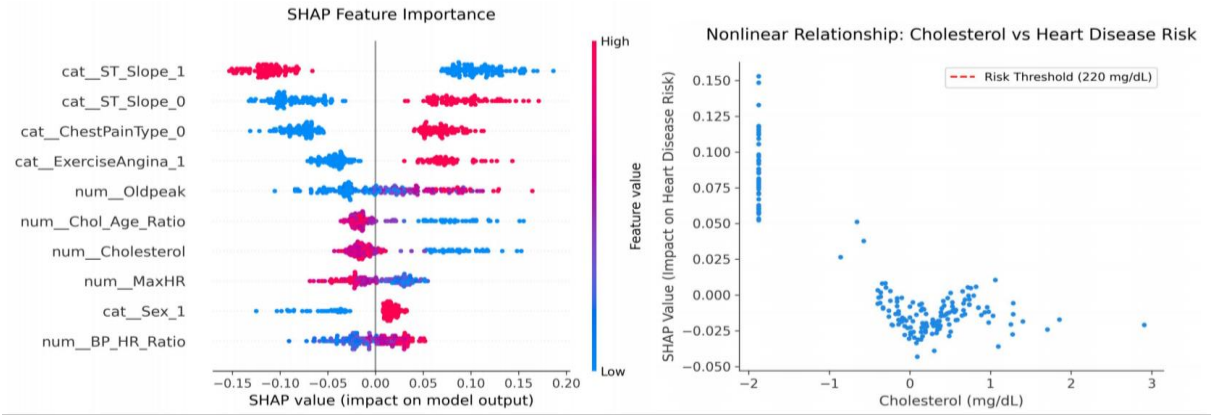

Fig. 8 Shap analysis results (Picture credit: Original).

5. Conclusion

This comparative study confirms that both Logistic Regression and Random Forest can effectively predict heart disease risk with high recall and reasonable precision. Random Forest offers slightly better overall performance and captures complex feature interactions that linear models may miss. The results support the use of interpretable machine learning methods in clinical settings, where both accuracy and transparency are essential. Future work should focus on expanding the dataset to include more diverse patient populations and additional features such as smoking habits, diet, and family history. Integrating time-series data from electronic health records could enable dynamic modeling of disease progression using models like LSTM. Additionally, incorporating fairness optimization techniques will help ensure that predictions are equitable across different demographic groups. By advancing these areas, machine learning can play a more integral and trustworthy role in supporting cardiovascular risk assessment and prevention.

References

[1] Dehghan A, Jahangiry L, Khezri R, Jafari A, Pezeshki B, Rezaei F, Aune D. Framingham risk scores for determination the 10-year risk of cardiovascular disease in participants with and without the metabolic syndrome: results of the Fasa Persian cohort study. BMC Endocrine Disorders. 2024 Jun 24; 24(1):95.

[2] Asif S, Wenhui Y, ur-Rehman S, ul-ain Q, Amjad K, Yueyang Y, Jinhai S, Awais M. Advancements and prospects of machine learning in medical diagnostics: unveiling the future of diagnostic precision. Archives of Computational Methods in Engineering. 2024 Jun 26:1-31.

[3] Jones C, Castro DC, De Sousa Ribeiro F, Oktay O, McCradden M, Glocker B. A causal perspective on dataset bias in machine learning for medical imaging. Nature Machine Intelligence. 2024 Feb; 6(2):138-46.

[4] Gayap HT, Akhloufi MA. Deep machine learning for medical diagnosis, application to lung cancer detection: a review. BioMedInformatics. 2024 Jan 18; 4(1):236-84.

[5] Cheng H, Zhang M, Shi JQ. A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2024 Aug 21.

[6] Ahmadilivani MH, Taheri M, Raik J, Daneshtalab M, Jenihhin M. A systematic literature review on hardware reliability assessment methods for deep neural networks. ACM Computing Surveys. 2024 Jan 22; 56(6):1-39.

[7] Antamis T, Drosou A, Vafeiadis T, Nizamis A, Ioannidis D, Tzovaras D. Interpretability of deep neural networks: A review of methods, classification and hardware. Neurocomputing. 2024 Jul 17:128204.

[8] Elkahwagy DM, Kiriacos CJ, Mansour M. Logistic regression and other statistical tools in diagnostic biomarker studies. Clinical and Translational Oncology. 2024 Sep; 26(9):2172-80.

[9] Sunarya PA, Rahardja U, Chen SC, Lic YM, Hardini M. Deciphering digital social dynamics: A comparative study of logistic regression and random forest in predicting e-commerce customer behavior. Journal of Applied Data Sciences. 2024 Jan 29; 5(1):100-13.

[10] Iranzad R, Liu X. A review of random forest-based feature selection methods for data science education and applications. International Journal of Data Science and Analytics. 2024 Feb 3:1-5.