Heart Attack Risk Prediction Using Machine Learning and SHAP Explainability

Rahul Chaduvula

June 14, 2025

Abstract

Heart disease remains one of the leading causes of mortality worldwide. Early detection of heart attack risk is crucial for timely intervention. In this study, we employ Random Forest and XGBoost models for predictive analysis of patient health data. In addition, SHAP explainability enhances the interpretation of the model, identifying key risk factors such as cholesterol levels, age, and blood pressure. Our final XGBoost model achieves 87% accuracy, demonstrating reliable predictive capabilities.

1 Introduction

Cardiovascular diseases are a growing global health concern, making risk prediction models essential for preventive healthcare. Machine learning methods allow us to extract valuable information from patient data, improving the accuracy of early detection. This paper explores the effectiveness of the XGBoost and Random Forest classifiers, coupled with SHAP analysis, to improve interpretability and trustworthiness.

2 Related Work

Several machine learning approaches have been applied to the prediction of heart disease, including logistic regression, neural networks, and ensemble methods. Although deep learning models have shown promise, they often lack transparency. Our research bridges this gap by integrating SHAP-based feature importance to provide explainability in healthcare ML models.

3 Methodology

3.1 Dataset

We have used a structured data set consisting of patient characteristics such as **age, cholesterol levels, blood pressure, and symptoms indicative of the risk of heart attack**. The target variable represents the presence (1) or absence (0) of the risk of heart attack.
Credits for the data set are given to "Aftab, Rakin Sad (2024), "Prediction of Heart Attack", Mendeley Data, V1, doi: 10.17632/yrwd336rkz.1"

3.2 Feature Analysis

Feature analysis is a critical step in understanding which variables contribute the most to the prediction of heart attack risk. By examining the importance scores, distributions, and correlations, we can refine the model to improve accuracy and interpretability.

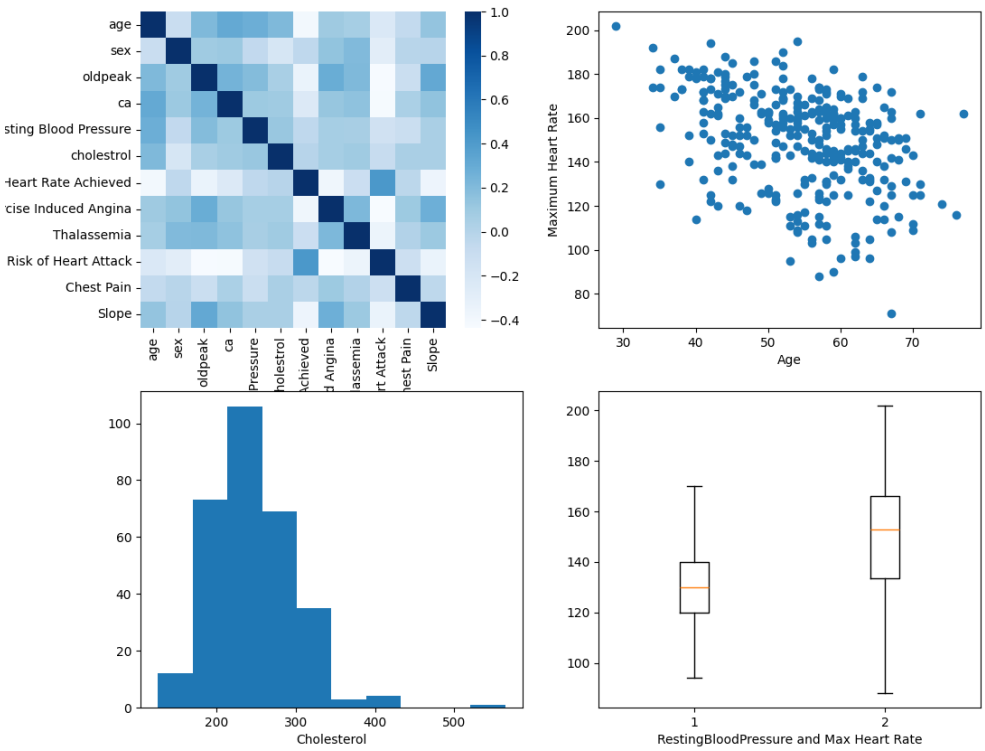


Figure 1: Subplot

3.3 Model Selection

To achieve optimal accuracy, we evaluate both ***Random Forest and XGBoost*** classifiers: - **Random Forest**: A tree-based ensemble method effective for structured data. - **XGBoost**: An optimized gradient boosting technique known for superior performance.

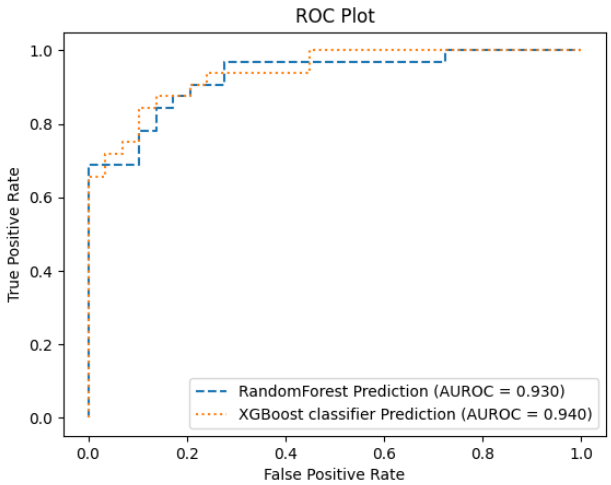


Figure 2: ROC plot for XGBoost and RandomForest Classifiers

Hyperparameter tuning is applied via ***GridSearchCV*** to optimize model parameters for maximum accuracy.

3.4 Explainability with SHAP

To ensure transparency, **SHAP (SHapley Additive exPlanations)** is used to analyze feature contributions. The SHAP bar plot highlights the most influential features:

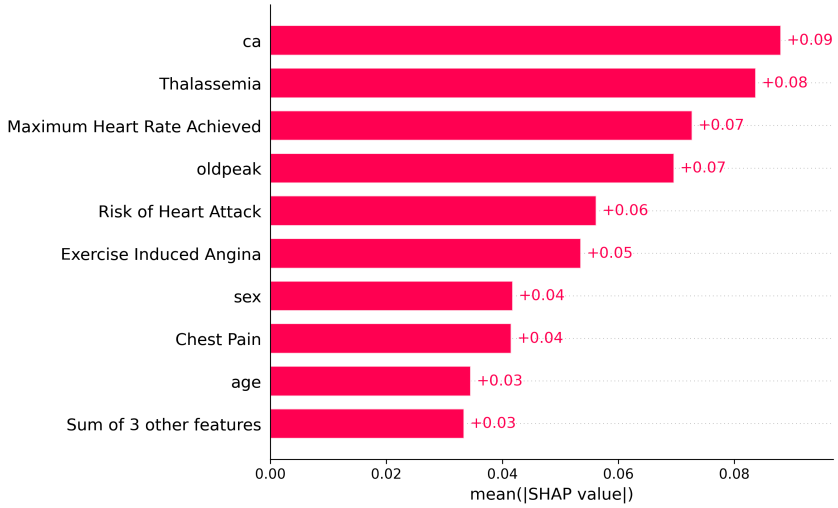


Figure 3: SHAP Bar Plot Showing Feature Contributions

In addition, a SHAP summary plot provides an in-depth look at the significance of features in patient predictions.

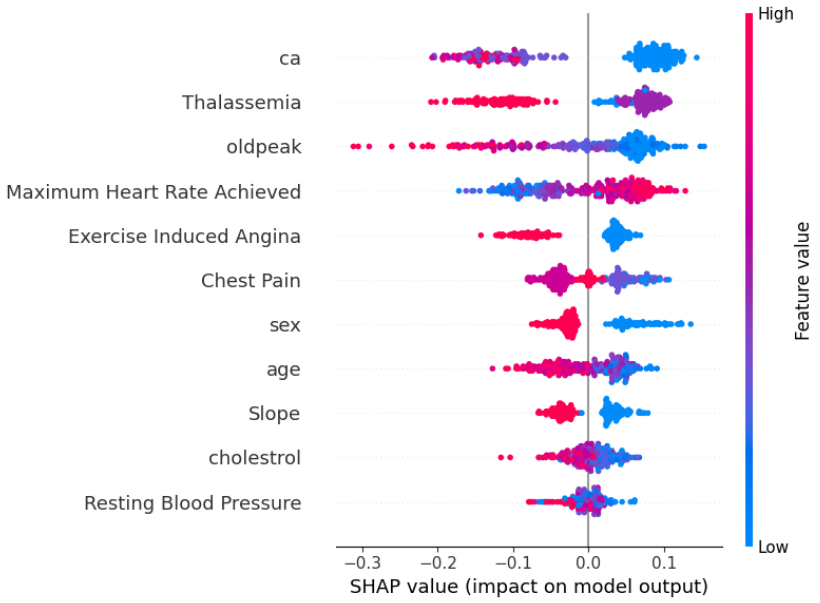


Figure 4: SHAP Summary Plot showing feature importance

This summary plot reveals how different features push predictions toward "high-risk" or "low-risk" outcomes.

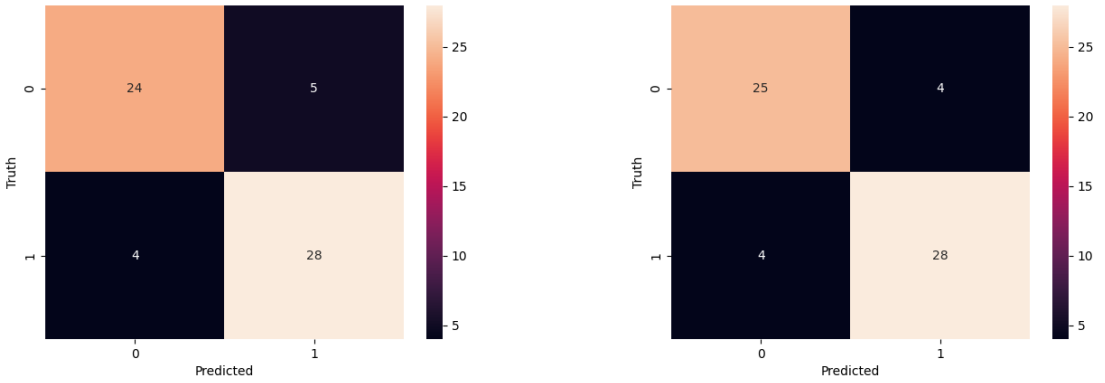
4 Results

The final model evaluation is based on standard classification metrics: accuracy, precision, recall, and F1 score. Below is a comparative analysis of our tuned models:

Metric	Random Forest	XGBoost
Accuracy	85%	87%
Precision	86%	88%
Recall	85%	88%
F1 Score	85%	88%

Table 1: Performance Comparison Between Models

In addition, the confusion matrices below highlight the classification results for each model.



(a) Confusion Matrix for Random Forest Model

(b) Confusion Matrix for XGBoost Model

Figure 5: Confusion Matrices for the Tuned Models

– XGBoost outperforms Random Forest in correctly classifying high-risk patients, with fewer false negatives.
– Misclassifications indicate that possible interactions of some low-risk patients are incorrectly flagged, probably due to marginal cholesterol levels.

5 Discussion

The XGBoost model exhibits better generalization, consistently outperforming Random Forest in precision and recall. Cholesterol levels and age prove to be the main contributors to the risk of heart attack, as revealed by SHAP analysis.

5.1 Limitations

1. **Data Quality and Bias**
   Many data sets used for the prediction of heart disease are unbalanced, which means that they contain more healthy patients than high-risk cases, leading to biased predictions.

2. **Feature Selection Challenges**
   Some important medical indicators (e.g., genetic predisposition, lifestyle factors) may not be included in data sets, limiting prediction accuracy

6 Key Insights

1. The explainability of the **SHAP** feature ensures that medical professionals can interpret the predictions.

2. Hyperparameter tuning significantly improves performance.

3. Future work could explore deep learning-based predictive modeling for heart attack detection.

7 Conclusion and Future Work

Our research validates the effectiveness of machine learning models in predicting the risk of heart attack. The integration of SHAP explainability enhances medical trustworthiness, paving the way for real-world applications in healthcare. Future efforts could involve developing deep learning models or integrating live patient data streams for real-time risk prediction. [1]

References

[1] A. S. Shaikh, R. M. Samant, K. S. Patil, N. R. Patil, and A. R. Mirkale, "Review on explainable ai by using lime and shap models for healthcare domain," *International Journal of Computer Applications*, vol. 185, no. 45, pp. 18–23, Nov 2023. [Online]. Available: https://ijcaonline.org/archives/volume185/number45/32992-2023923263/