Heart Disease Risk Prediction — Machine Learning Analysis

Dataset: raw_merged_heart_dataset.csv (Public Heart Disease Cohort, Kaggle)
Population: n = 1,744 adult patients (retrospective cohort)
Outcome: Binary heart disease indicator (presence vs absence)
Modeling Approach: Supervised classification with explainable ML
Prepared by: Mo Mo Ko Win Myint, global-ad-snap
Date: October 2025

Executive Summary

Problem Statement

Cardiovascular disease remains one of the leading global causes of illness and death. Early identification of at-risk individuals can significantly improve treatment outcomes and reduce healthcare costs. This project developed a predictive model to assess heart disease risk using patient health data, providing clinicians with an evidence-based decision support tool.

Key Findings

- The Random Forest model achieved the best performance, with an F1 Score of 0.959 (95.9%), indicating a strong balance between precision and recall.
- LightGBM (94.8%) and XGBoost (94.1%) also performed strongly, offering alternative options depending on resource constraints.
- Model interpretation using SHAP analysis highlighted clinically meaningful predictors:
  - Maximum heart rate (thalachh) – higher values increase predicted risk
  - Chest pain type (cp) – strong diagnostic relevance
  - ST depression (oldpeak) – elevated stress marker linked to heart disease
- The dataset was balanced between positive and negative cases (862 positive, 882 negative), ensuring reliable training and minimizing bias.

Visual Highlights

Key supporting visuals are provided in the Technical Appendix (see Figures 1, 6, 14 & 15), which illustrate the balanced dataset, model performance comparison, and the interpretability of key predictors through feature importance and SHAP analysis.

Recommendations

Random Forest is recommended for deployment given its top F1 Score (0.959) and strong interpretability. LightGBM is recommended as an alternative when computational efficiency is prioritized, offering competitive performance with lower resource usage. To maximize adoption and trust:

- Integrate SHAP-based explanations into the interface, allowing clinicians to see which factors drive predictions.
- Validate the model on external datasets before clinical deployment.
- Explore ensemble approaches to further enhance robustness.
- Develop a Streamlit demo for stakeholder review and usability testing.

Strategic Value

This predictive model enables earlier screening of high-risk patients, reduces clinician workload through automation, and can be scaled to hospital systems with minimal cost. Built-in interpretability ensures transparency and supports clinician trust and regulatory compliance.

Potential Clinical & Operational Applications

This model is designed as a **population-level risk stratification and decision-support prototype**, rather than an individual diagnostic tool.

Potential applications include:

- **Preventive cardiology screening**
  Identifying individuals at elevated cardiovascular risk to prioritize lifestyle interventions, further diagnostic testing, or specialist referral.

- **Risk-based patient stratification**
  Supporting clinicians and care managers in categorizing patients into low- and high-risk groups for targeted follow-up programs.

- **Clinical decision-support prototyping**
  Demonstrating how machine learning–based risk scores and explainability tools (SHAP) can augment clinician understanding of key risk drivers.

- **Healthcare analytics and education**
  Serving as a teaching and demonstration tool for medical trainees, data scientists, and healthcare stakeholders exploring interpretable ML in clinical contexts.

- **Health-tech product development**
  Acting as a proof-of-concept model for startups or research teams developing early-stage cardiovascular risk assessment tools.

These applications assume **population-level analysis and retrospective data exploration** and are **not intended for real-time clinical decision-making** without prospective validation and regulatory approval.

Technical Appendix

Dataset Description

- Source: raw_merged_heart_dataset.csv
- Size: 1,744 samples (after cleaning)
- Target Variable: Binary indicator of heart disease (1 = Yes, 0 = No)
- Key Features: age, cholesterol (chol), resting blood pressure (trestbps), ST depression (oldpeak), maximum heart rate (thalachh), chest pain type (cp), resting ECG (restecg), slope, thal

Preprocessing Steps

- Replaced ambiguous values (?, unknown, 999, -1) with NaN
- Imputed missing numeric values using median strategy
- Scaled numeric features with StandardScaler

- One-hot encoded categorical variables (restecg, slope, thal)

Target Class Distribution

- Balanced dataset:
  - 862 cases with heart disease
  - 882 cases without heart disease

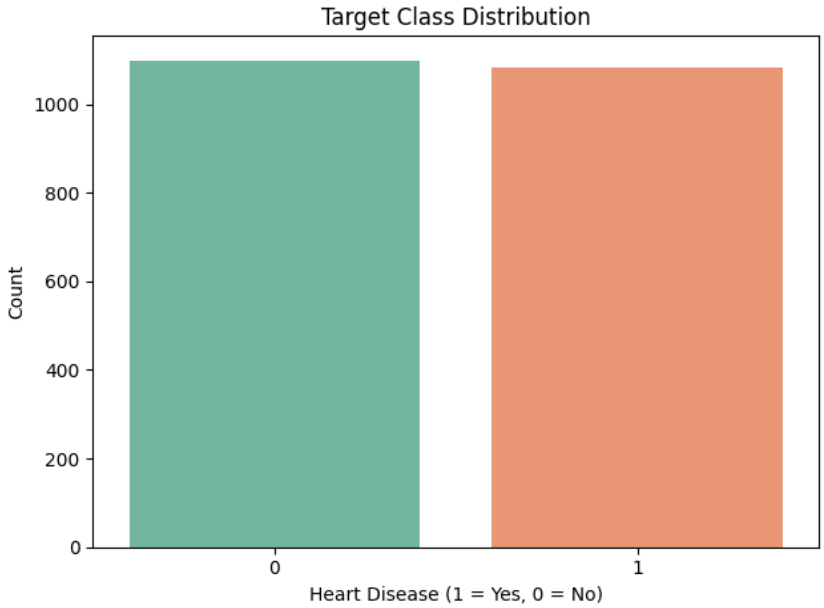
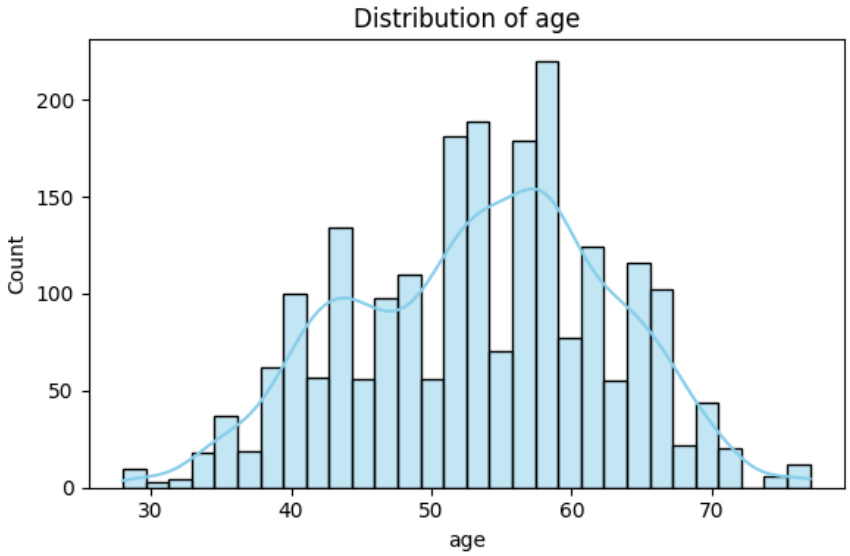
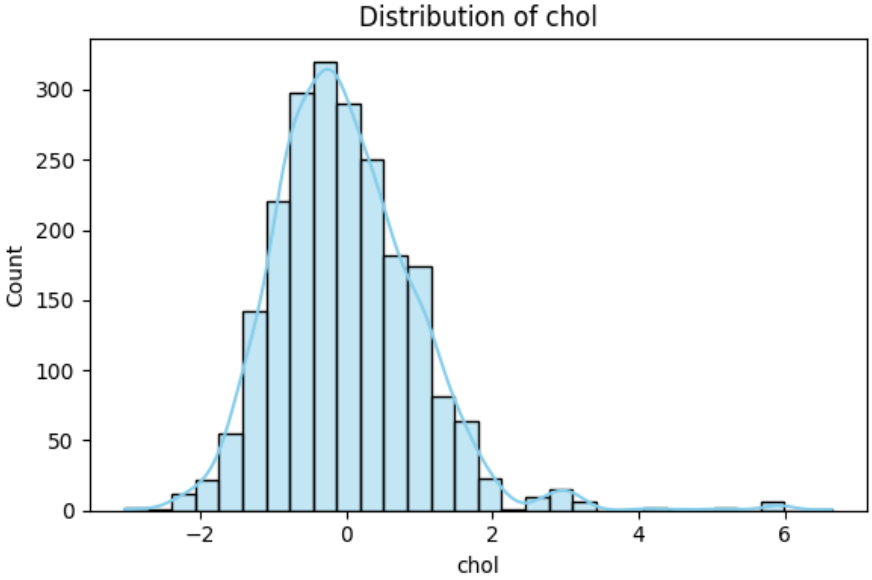


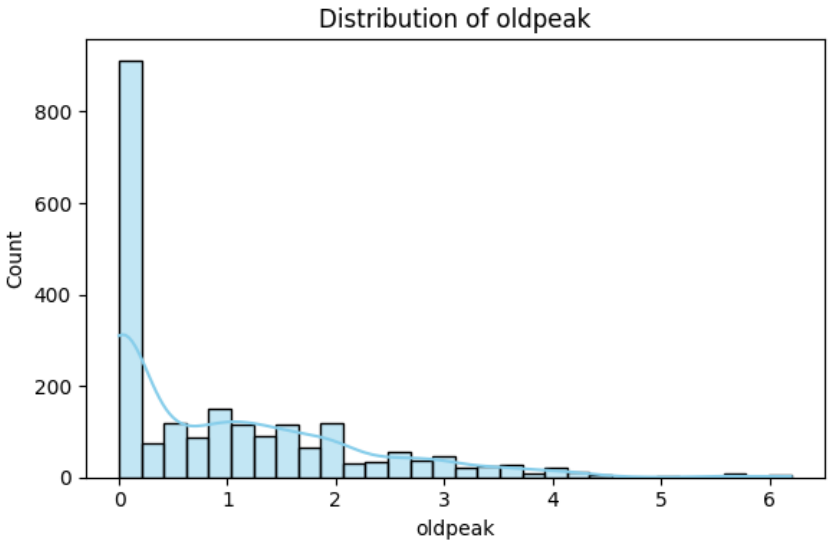
Figure 1: Target class distribution — shows near-equal balance between classes

Feature Distributions

- Age: Centered around 55–60 years
- Cholesterol (chol): Normalized after scaling
- ST depression (oldpeak): Right-skewed, indicating elevated cardiac stress in a subset of patients


Distribution of age
Count
age
0
50
100
150
200
30
40
50
60
70


Distribution of chol
Count
chol
0
50
100
150
200
250
300
−2
0
2
4
6



Figures 2–4: Feature distribution plots

Feature Correlation

- Positive correlation: Maximum heart rate (thalachh) with heart disease
- Negative correlation: Exercise-induced angina (exang) with heart disease
- Minimal multicollinearity: Supports model stability

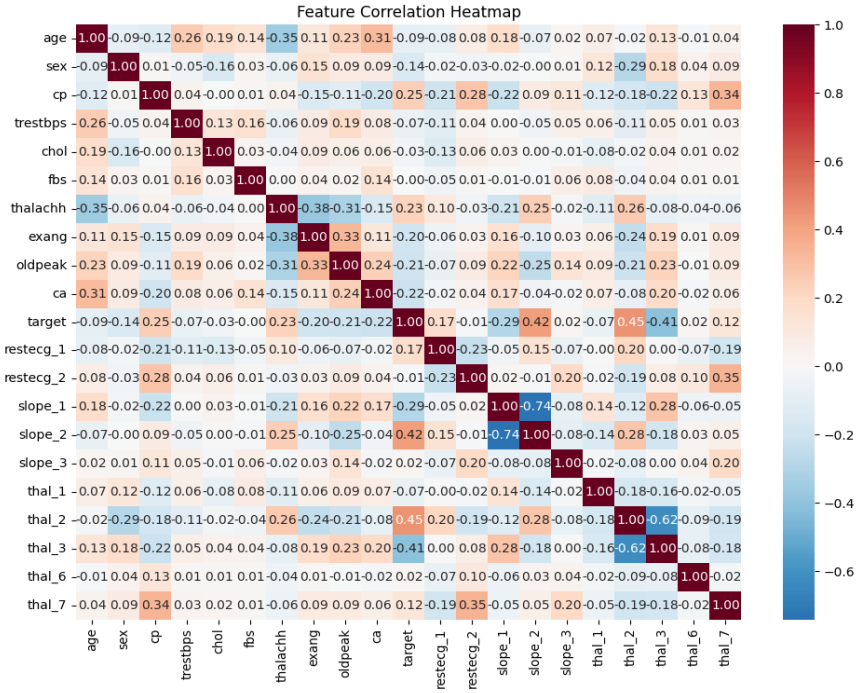


Figure 5: Correlation heatmap

Model Comparison

Model	Accuracy	Precision	Recall	F1 Score	ROC–AUC
Random Forest	0.959	0.951	0.968	0.959	0.995
LightGBM	0.947	0.942	0.955	0.948	0.991
XGBoost	0.941	0.937	0.945	0.941	0.983
KNN	0.870	0.856	0.891	0.873	0.920
Logistic Regression	0.792	0.795	0.791	0.793	0.868
Naive Bayes	0.753	0.757	0.750	0.753	0.808
SVM	0.714	0.669	0.855	0.750	0.763

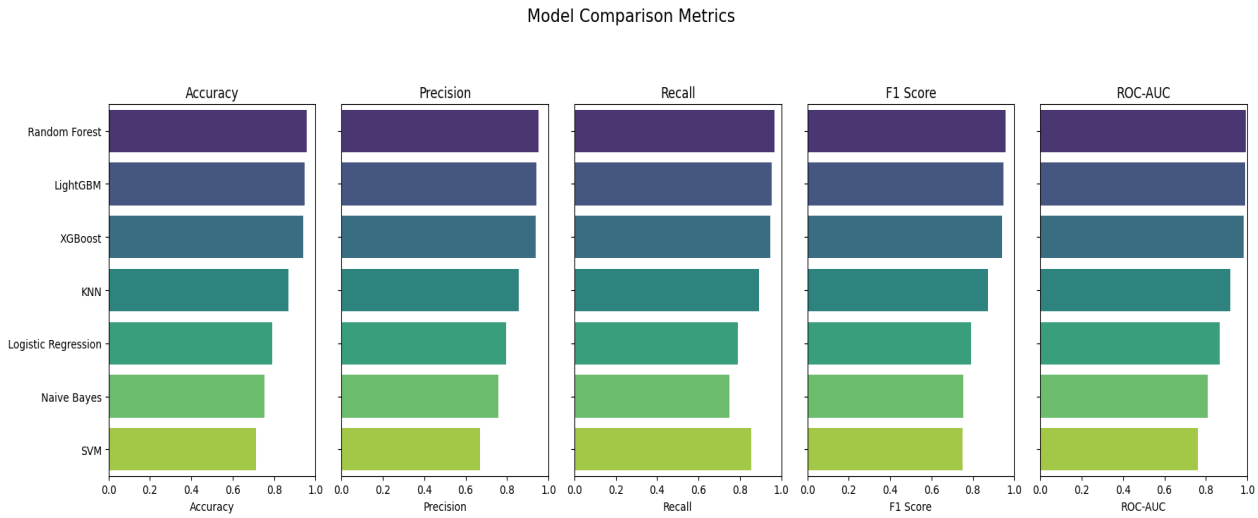
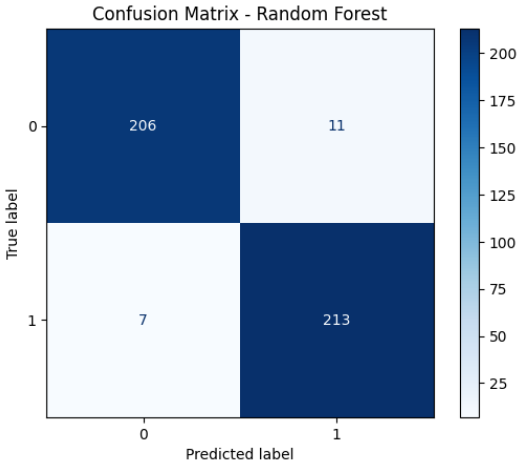
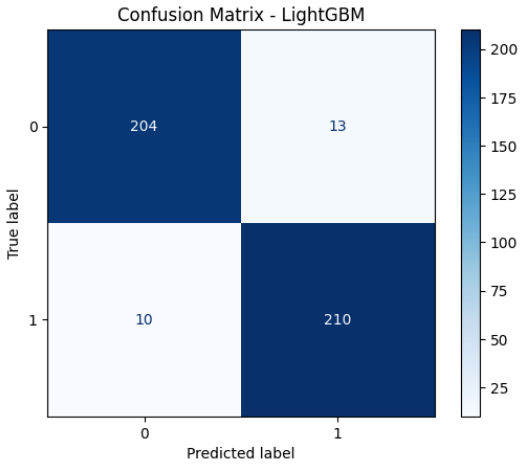


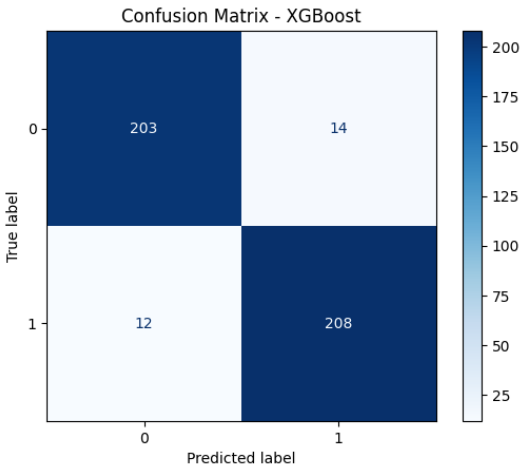
Figure 6: Bar chart comparison — Random Forest leads across all metrics

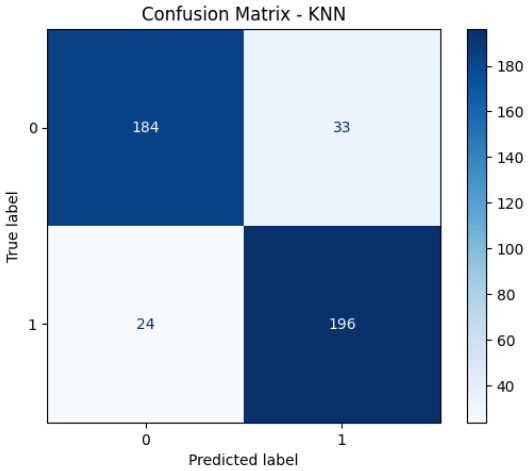
Evaluation Metrics

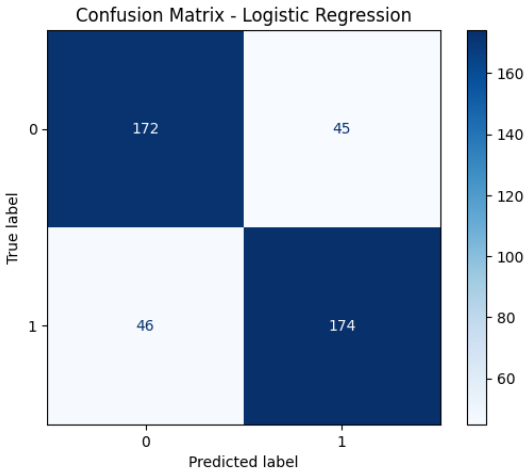
- Metrics used: Accuracy, Precision, Recall, F1 Score, and ROC–AUC
- Confusion matrices for all models are included in the Technical Appendix (Figures 7–13) to provide additional insight into classification errors.
- Evaluation performed on 20% test split (random_state = 42)

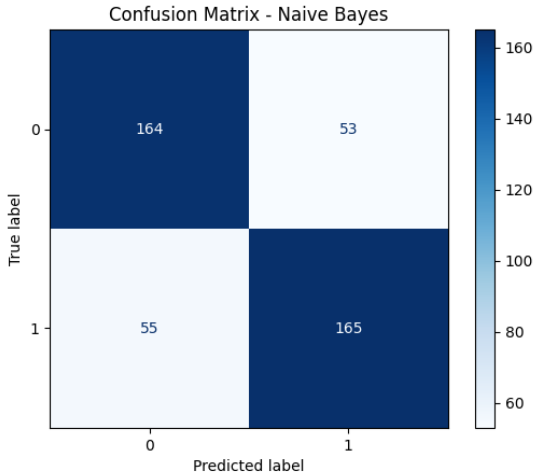

Confusion Matrix - Random Forest
206
11
7
213
True label
Predicted label
0
1
200
175
150
125
100
75
50
25


Confusion Matrix - LightGBM
204
13
10
210
True label
Predicted label
0
1
200
175
150
125
100
75
50
25


Confusion Matrix - XGBoost
203
14
12
208
True label
Predicted label
0
1
200
175
150
125
100
75
50
25


Confusion Matrix - KNN
184
33
24
196
True label
Predicted label
0
1
180
160
140
120
100
80
60
40


Confusion Matrix - Logistic Regression
172
45
46
174
True label
Predicted label
0
1
160
140
120
100
80
60


Confusion Matrix - Naive Bayes
164
53
55
165
True label
Predicted label
0
1
160
140
120
100
80
60

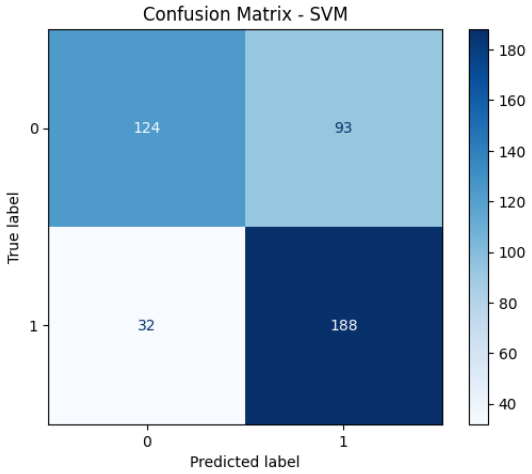


Figure 7-13: Confusion Matrices

Feature Importance

- LightGBM identified top predictors:
  - Maximum heart rate (thalachh)
  - Chest pain type (cp)
  - ST depression (oldpeak)
  - Age

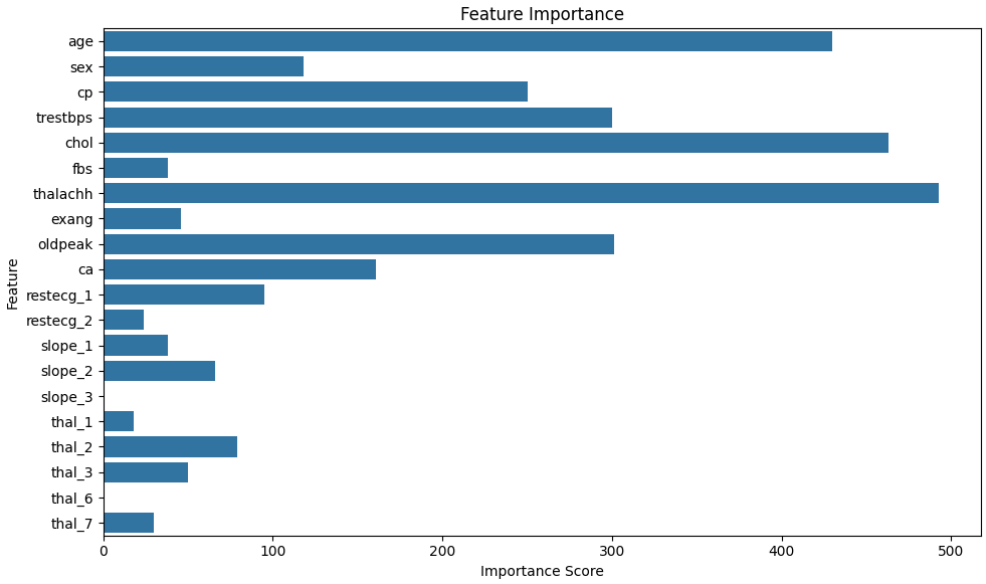


Figure 14: Feature importance plot

SHAP Interpretability Insights

- High thalachh → Increases predicted risk
- Low oldpeak → Lowers predicted risk
- Chest pain type → Strong predictor of outcomes
- SHAP confirms both directional impact and clinical relevance

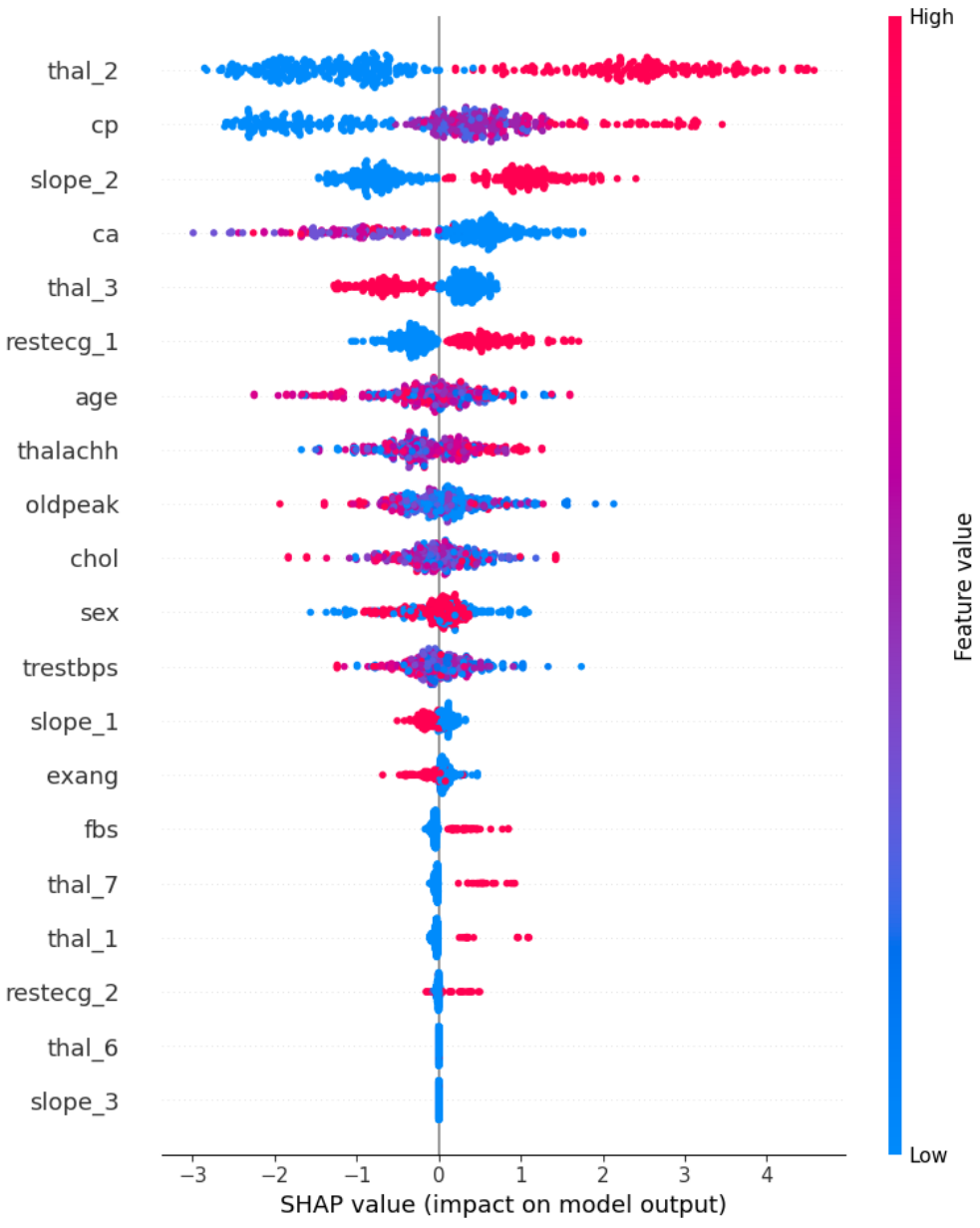


Figure 15: SHAP summary plot