Interpretable Linear Models for Heart Disease Prediction: A Comparative Study

Journal Title
XX(X):1–12
©The Author(s) 2016
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Abstract

Heart disease remains a leading cause of mortality worldwide, underscoring the importance of accurate and transparent methods for early diagnosis. While many machine learning and artificial intelligence models have demonstrated strong predictive performance, their limited interpretability poses challenges for clinical adoption. In this study, we evaluate three interpretable linear classification models—Generalized Linear Model (GLM) logistic regression, L1-regularized (Lasso) logistic regression, and Linear Discriminant Analysis (LDA)—for heart disease prediction using the Cleveland Heart Disease dataset. Following comprehensive data preprocessing, the models are assessed on a held-out test set using standard evaluation metrics, including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (ROC-AUC). The results show that all three models achieve strong discriminative performance. Among them, Lasso logistic regression attains the highest accuracy and F1-score, reflecting a favorable balance between sensitivity and specificity, while GLM and LDA exhibit comparable performance with slightly lower recall. Importantly, the GLM framework enables identification of clinically meaningful predictors, reinforcing its interpretability and relevance for medical decision-making. These findings demonstrate that interpretable linear models can provide reliable and transparent tools for heart disease prediction, offering a practical alternative to more complex black-box approaches in clinical settings.

Keywords

Heart disease, CDC, GLM, Lasso Logistic Regression, Linear Discriminant Analysis, Interpretability.

Introduction

Heart disease or cardiovascular disease (CVD) remains one of the most significant threats to human health and is the leading cause of mortality worldwide. According to the report of the World Health Organization (WHO), cardiovascular diseases (CVDs) account for over 17.9 million deaths annually, WHO (2024), Shah et al. (2020). In 2021, CVD accounted for approximately 20.5 million deaths globally, representing nearly 32% of all deaths. Among these, ischemic heart disease was responsible for 9.1 million deaths and stroke accounted for 6.6 million, together comprising 85% of all CVD-related mortality, World Heart Federation (2023). The burden is particularly severe in low- and middle-income countries, where more than 80% of CVD deaths occur and access to early detection and care remains limited, Di Cesare et al. (2024). The global burden of cardiovascular disease (CVD) has continued to escalate, with an estimated 523 million people living with CVD in 2022. The rising prevalence is largely driven by aging populations and modifiable risk factors such as high systolic blood pressure, obesity, and diabetes, Mensah et al. (2023), Savarese et al. (2022).

In the United States, the situation is equally dire: heart disease continues to be the leading cause of death among men, women, and most racial and ethnic groups. According to recent data from the Centers for Disease Control and Prevention (CDC), an estimated 702,880 individuals died from heart disease in 2022, accounting for approximately 1 in every 5 deaths nationwide. On average, one person dies every 33 seconds due to cardiovascular-related complications. Coronary artery disease (CAD), the most common form of heart disease, alone accounted for more than 371,500 deaths. Moreover, about 1 in 20 adults aged 20 or older—roughly 5% of the U.S. adult population—are living with CAD, and nearly 20% of all cardiovascular deaths occur in individuals under the age of 65, highlighting its impact across age groups, CDC (2024).

These statistics emphasize the urgent need for effective, accessible, and proactive diagnostic tools for early heart disease detection. Early identification of at-risk individuals significantly improves treatment outcomes, reduces long-term healthcare costs, and enhances patient quality of life. However, traditional diagnostic methods for heart disease, such as electrocardiograms (ECG), stress testing, coronary angiography, and blood biomarker analysis, often involve time-consuming, invasive, and expensive procedures that are not always feasible in low-resource settings or for routine screening.

To address these limitations, there has been growing interest in adopting data-driven predictive modeling techniques that leverage statistical learning and artificial intelligence (AI). The increasing availability of clinical datasets, along with advancements in computing power and algorithm development, has enabled the use of machine learning (ML) and AI models for risk prediction in cardiovascular medicine. These approaches can automatically learn complex patterns from patient data, such as age, cholesterol levels, blood pressure, and family history, to identify individuals at high risk for heart disease without requiring invasive testing.

By doing so, ML/AI models offer a promising pathway to improve diagnostic accuracy, reduce diagnostic delays, and support more timely clinical decision-making, Ahsan and Siddique (2022), Azmi et al. (2022), Zhou et al. (2024).

Several recent studies have explored the effectiveness of various ML algorithms for heart disease classification. For instance, Shah et al. (2020) conducted a comparative analysis using K-nearest neighbors (KNN), decision trees, and random forests on the Cleveland Heart Disease dataset, reporting encouraging results for KNN in terms of predictive accuracy. Yadav et al. (2023) extended this work by examining decision trees, logistic regression, random forests, and ensemble techniques, highlighting the performance benefits of tree-based models. Furthermore, Jha et al. (2025) incorporated artificial neural networks (ANNs) into their analysis, demonstrating that deep learning models can outperform traditional classifiers across multiple evaluation metrics such as accuracy, F1-score, and ROC-AUC.

While these findings underscore the predictive strength of complex ML models, one of the most significant challenges to their deployment in clinical practice is interpretability. Many of the high-performing models—such as neural networks, support vector machines, and ensemble methods—are considered ”black-box” models, meaning that their internal decision logic is not easily understandable by human experts. In the context of healthcare, this opacity limits the trust clinicians can place in model outputs, especially when these outputs inform critical diagnostic or treatment decisions. Moreover, regulatory standards and ethical considerations increasingly demand transparency and accountability in AI-driven medical tools, Mamun and Alouani (2022).

Motivated by the need for interpretable yet effective predictive models, this study focuses on developing and evaluating interpretable machine learning models for heart disease prediction. We specifically investigate three linear classification approaches—Generalized Linear Models (GLMs), L1-regularized (Lasso) logistic regression, and Linear Discriminant Analysis (LDA)—that offer the dual benefits of transparency and performance. These models are designed to be both statistically robust and clinically explainable, allowing healthcare providers to understand how individual features contribute to the model's predictions. Using the widely recognized Cleveland Heart Disease dataset, we assess each model's effectiveness across multiple evaluation metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. Our overarching goal is to contribute a framework that balances predictive power with interpretability, enabling practical, cost-effective, and ethically sound applications in real-world medical environments.

For interpretable, accurate, and clinically actionable models for heart disease prediction, the present study is designed with the following objectives:

- To evaluate the predictive performance of three transparent linear classification models—Generalized Linear Models (GLMs), L1-regularized (Lasso) logistic regression, and Linear Discriminant Analysis (LDA)—using the Cleveland Heart Disease dataset.
- To ensure model interpretability by identifying the most significant clinical features contributing to heart disease prediction and examining their associations with patient outcomes.
- To assess and compare the models across standard evaluation metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, with particular attention to sensitivity and false-positive rates.
- To highlight the practical value of interpretable models for real-world clinical integration, especially in resource-constrained environments where black-box models may be less viable.

Data

The dataset used for this study is the Cleveland Heart Disease dataset, obtained from the UCI Machine Learning Repository, Dua and Taniskidou (2017). Table 1 summarizes the dataset features used in this study. The dataset consists of 303 instances and includes both continuous and categorical variables relevant to cardiovascular health assessment. Continuous features, such as `age`, `trestbps` (resting blood pressure), `chol` (serum cholesterol), `thalach` (maximum heart rate achieved), and `oldpeak` (ST depression induced by exercise), capture various physiological measurements.

Table 1. Dataset overview.

Feature	Feature Type	Count	Description
age	Continuous	303	Age in years
sex	Categorical	303	1 = male; 0 = female
cp	Categorical	303	Chest pain type (0–3)
trestbps	Continuous	303	Resting blood pressure (mm Hg)
chol	Continuous	303	Serum cholesterol (mg/dl)
fbs	Categorical	303	Fasting blood sugar, > 120 mg/dl (1 = true)
restecg	Categorical	303	Resting electrocardiographic result (0–2)
thalach	Continuous	303	Maximum heart rate achieved
exang	Categorical	303	Exercise-induced angina (1 = yes)
oldpeak	Continuous	303	ST depression induced by exercise
slope	Categorical	303	Slope of peak exercise ST segment (0–2)
ca	Categorical	299	Number of major vessels colored by fluor.
thal	Categorical	301	Thalassemia (1 = normal; 2 = fixed def.)
condition	Categorical (Target Variable)	303	0 = no disease; 1 = disease

Categorical variables capture clinical indicators and demographic factors. The `sex` feature encodes gender as 1 for male and 0 for female. The `cp` feature categorizes the type of chest pain into four classes: typical angina, atypical angina, non-anginal pain, and asymptomatic. Fasting blood sugar levels, indicated by the `fbs` feature, represent whether the patient's fasting blood sugar exceeds 120 mg/dl (1 = true, 0 = false). Resting electrocardiographic results are captured by the `restecg` feature, with values indicating normal findings, ST-T wave abnormalities, or left ventricular hypertrophy based on Estes' criteria. The `exang` feature reflects the presence of exercise-induced angina (1 = yes, 0 = no), while `slope` describes the slope of the peak exercise ST segment as upsloping, flat, or downsloping. The `ca` feature records the number of major vessels (ranging

from 0 to 3) visualized by fluoroscopy. The `thal` feature indicates thalassemia status, differentiating between normal blood flow, fixed defects (regions with no blood flow), and reversible defects (abnormal blood flow that improves with rest). Missing values originally present in `ca` and `thal` have been addressed during preprocessing.

The target variable, labeled as `condition`, is binary, indicating the presence (1) or absence (0) of heart disease. This combination of demographic, clinical, and physiological features enables comprehensive modeling and analysis of heart disease prediction.

Methodology

Figure 1 provides an overview of the step-by-step workflow implemented in this study. The dataset will undergo preprocessing to address missing values, encode categorical variables, and normalize features as required. The binary classification task will involve splitting the data into training and test sets using stratified sampling to preserve class proportions. The first model to be implemented is a generalized linear model using logistic regression with a logit link function. This model serves as the baseline due to its statistical rigor and interpretability. The second model, Lasso Logistic Regression, introduces L1 regularization to the logistic regression framework. This approach helps in selecting the most relevant features by shrinking the coefficients of less informative predictors to zero, thereby offering a more parsimonious model. The third model, Linear Discriminant Analysis (LDA), classifies instances based on the assumption that each class follows a Gaussian distribution with equal covariance matrices.

Let's

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}, \quad x_i \in \mathbb{R}^p, \; y_i \in \{0, 1\}$$

denote our Cleveland Heart Disease dataset, where $p = 13$ clinical features and $n = 303$ patients. We describe below the preprocessing, model formulations, training procedures, and evaluation metrics.

Data Preprocessing

The following preprocessing steps have been applied to prepare the dataset for model training:

- **Missing-value:** To assess the presence of missing values in the dataset, a heatmap visualization has been generated, as shown in Figure 2. This initial analysis has identified a few missing entries, primarily concentrated in the `ca` and `thal` features. To address this issue, samples containing missing values have been removed from the dataset. Following this removal process, a second heatmap has been produced (Figure 3), confirming that the resulting dataset is free of missing values. After removing missing values, the number of samples in this dataset is 297.
- **Categorical encoding:** Categorical features `cp`, `restecg`, `slope`, `thal`, and `ca` have been transformed using one-hot encoding.
- **Feature scaling:** Continuous variables have been standardized to zero mean and unit variance using the `standard scalar` package from `scikit-learn` library to ensure uniform scaling across features.

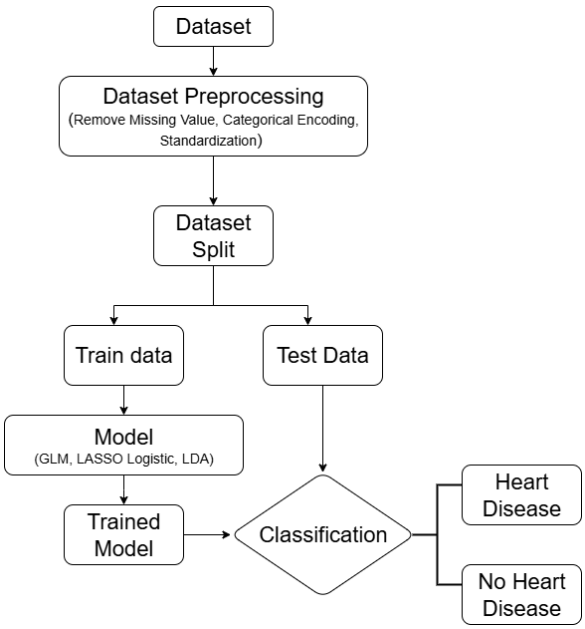


Figure 1. Model Flowchart

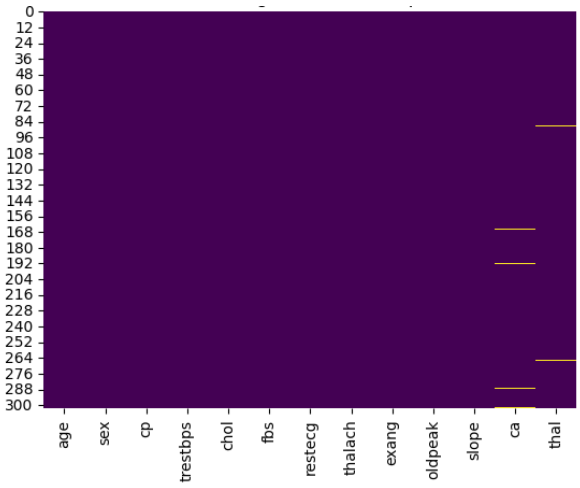


Figure 2. Missing values visualization heatmap.

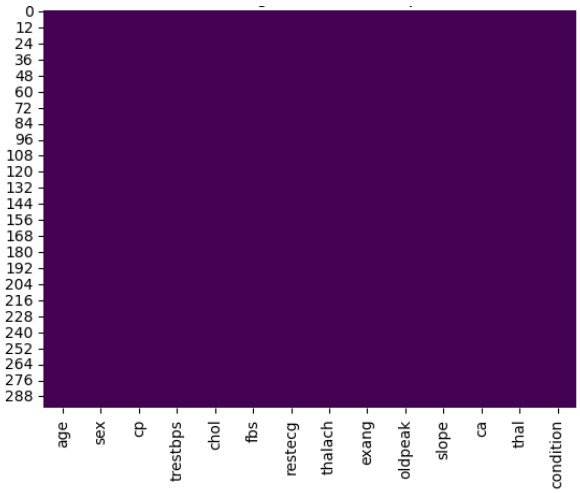


Figure 3. Heatmap visualization after removing missing values.

- **Train/test split:** The dataset has been partitioned into training (80%) and test (20%) subsets using stratified sampling, preserving the proportion of positive-class instances in each split. In the training data, there are 237 samples, and in the test data, there are 60 samples.

Models

Generalized Linear Model (GLM): The generalized linear model (GLM) provides a unified framework for modeling response variables that follow a distribution from the exponential family, McCullagh and Nelder (1989), Agresti (2015). For binary classification tasks, where the response variable $Y \in \{0, 1\}$, the GLM with a binomial family and a logistic (logit) link is appropriate.

The GLM comprises three components:

1. **Random Component:** The response variable Y follows a Bernoulli distribution:

$$Y \sim \text{Bernoulli}(p), \quad \text{with} \quad \mathbb{P}(Y = 1 \mid \mathbf{X}) = p \tag{1}$$

2. **Systematic Component:** The linear predictor is defined as:

$$\eta = \mathbf{X}\boldsymbol{\beta} \tag{2}$$

where $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the matrix of predictors, and $\boldsymbol{\beta} \in \mathbb{R}^p$ is the coefficient vector.

3. **Link Function:** The logistic link function relates the expected value $\mu = \mathbb{E}[Y \mid \mathbf{X}]$ to the linear predictor:

$$\log\left(\frac{\mu}{1-\mu}\right) = \mathbf{X}\boldsymbol{\beta} \tag{3}$$

Solving for μ gives the probability of success:

$$\mu = \frac{1}{1 + \exp(-\mathbf{X}\boldsymbol{\beta})} \tag{4}$$

Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ denote the dataset, where $y_i = Y \in \{0, 1\}$. The likelihood function for logistic regression is given by:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i} \tag{5}$$

where

$$p_i = \frac{1}{1 + \exp(-\mathbf{x}_i^\top \boldsymbol{\beta})} \tag{6}$$

The corresponding log-likelihood function is:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \tag{7}$$

This function is concave in $\boldsymbol{\beta}$ and is typically maximized using iterative numerical optimization techniques, such as Newton-Raphson or Iteratively Reweighted Least Squares (IRLS).

Lasso Logistic Regression (L1 Regularization): Lasso logistic regression extends the GLM by introducing an L1 penalty on the coefficients to perform variable selection and regularization, Tibshirani (1996). The optimization problem is:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left[-\sum_{i=1}^{n} (y_i \log p_i + (1 - y_i) \log(1 - p_i)) + \lambda \|\boldsymbol{\beta}\|_1 \right] \tag{8}$$

where the predicted probability is given by:

$$p_i = \frac{1}{1 + e^{-\mathbf{x}_i^\top \boldsymbol{\beta}}} \tag{9}$$

and the L1 penalty term is:

$$\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j| \tag{10}$$

Here, λ controls the strength of the regularization.

Linear Discriminant Analysis (LDA): Linear discriminant analysis (LDA) is a generative classification method that models the joint distribution of the predictors and the response, Hastie et al. (2009). It assumes that the feature vectors $\mathbf{x} \in \mathbb{R}^p$ for each class are drawn from a multivariate normal distribution with class-specific means $\boldsymbol{\mu}_k$, but a shared covariance matrix $\boldsymbol{\Sigma}$ across all classes.

Under this assumption, LDA predicts the class label $\hat{y}$ for a given observation $\mathbf{x}$ by maximizing the following discriminant function:

$$\hat{y} = \arg\max_{k \in \{0,1\}} \left(\mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \log \pi_k \right) \tag{11}$$

where:

- $\hat{y}$: Predicted class label (0 or 1),
- k: Class index (0 or 1),
- $\boldsymbol{\mu}_k$: Mean vector of predictors for class k,
- $\boldsymbol{\Sigma}$: Common covariance matrix,
- π_k: Prior probability of class k.

The discriminant function in Equation (11) arises from taking the log of the posterior probability (via Bayes' theorem) under the Gaussian assumption. The quadratic terms cancel due to the shared covariance matrix, resulting in a linear decision boundary.

LDA performs well when the Gaussian assumption holds and the classes are linearly separable. It is particularly effective when p (number of features) is not large compared to the number of observations and when class covariances are reasonably similar.

Model Training and Hyperparameter Tuning

Three classification models have been utilized in this study: Generalized Linear Model (GLM), Lasso Logistic Regression, and Linear Discriminant Analysis (LDA). The GLM has been implemented using the `statsmodels` library, Seabold and Perktold (2010), specifying a binomial family with a logistic link function. This setup models the log-odds of the binary outcome as a linear function of the predictors, without involving additional hyperparameter tuning. In contrast, Lasso Logistic Regression and LDA have been implemented using the `scikit-learn` library, Pedregosa et al. (2011). The

Lasso model has employed the `LogisticRegression` class with `penalty='l1'` and the `liblinear` solver to apply ℓ_1-norm regularization, encouraging sparsity in the coefficient estimates. LDA has been performed using the `LinearDiscriminantAnalysis` class with default settings, which assume equal covariance structures across classes. No further hyperparameter tuning has been accomplished beyond these standard configurations, ensuring comparability among the models.

Evaluation Metrics

The performance of the classification models has been evaluated using several standard metrics: accuracy, precision, recall, F1-score, and the area under the Receiver Operating Characteristic (ROC) curve (ROC-AUC). For each model, predicted class labels and predicted probabilities have been obtained on the test set. Model predictions and probability estimates have been generated using the respective `predict` and `predict_proba` functions in `scikit-learn`. For GLM, the probability estimates from the logistic regression output have been thresholded at 0.5 to generate class predictions. Subsequently, the following metrics have been computed using the test set. ROC curves have been plotted for each model, and the ROC AUC scores have been reported to evaluate each model's performance.

- **Accuracy:** The proportion of correct predictions among the total number of instances, defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

  where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.
- **Precision:** The proportion of correctly predicted positive instances among all predicted positives:

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall (Sensitivity):** The proportion of correctly predicted positive instances among all actual positives:

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-score:** The harmonic mean of precision and recall, providing a balance between the two:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **ROC-AUC:** The area under the ROC curve, which quantifies the trade-off between the true positive rate and false positive rate across different classification thresholds.
- **ROC Curve:** A graphical plot that illustrates the diagnostic ability of a binary classifier as its discrimination threshold is varied. It plots the true positive rate against the false positive rate.

We also present confusion matrices to visualize trade-offs between sensitivity and specificity.

Results and Discussion

Data Visualization

At the beginning of the data visualization stage, first, we visualized the response variable `"condition"`. Figure 4 illustrates the distribution of patients by heart disease status. The non-diseased group (condition = 0) contains slightly more patients than the diseased group (condition = 1). This figure demonstrates that the dataset maintains a balanced representation of both classes.

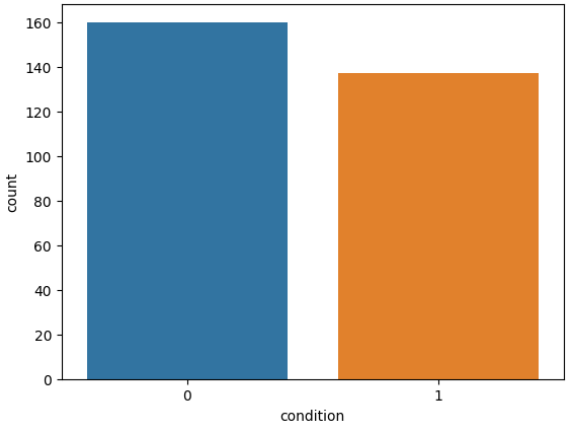


Figure 4. Class distribution of patients with and without heart disease. (Condition = 0, No Heart Disease; Condition = 1, Heart Disease).

Figure 5 presents the distribution of the key continuous variables in the Cleveland Heart Disease dataset. The age distribution (Figure 5a) is approximately normal, with most patients falling between 45 and 65 years, highlighting the higher prevalence of heart disease among middle-aged and older adults. Resting blood pressure (trestbps), Figure 5b, shows a unimodal pattern centered around 120–140 mm Hg, indicating that many patients exhibit borderline to moderately elevated systolic blood pressure. Serum cholesterol (chol) graph, Figure 5c, displays a right-skewed distribution, with the majority of patients clustered between 200 and 300 mg/dl, though a small subset demonstrates extreme hypercholesterolemia with values exceeding 400 mg/dl. The maximum heart rate achieved (thalach), Figure 5d, follows a slightly left-skewed distribution, with most individuals achieving 140–170 beats per minute, whereas lower values are more frequent among patients with heart disease, consistent with reduced exercise capacity. These patterns provide clinically relevant insights into the population, as elevated cholesterol, higher blood pressure, and lower maximum heart rate are well-recognized risk factors for cardiovascular disease.

Figure 6 presents the correlation heatmap depicting pairwise Pearson correlation coefficients among the dataset features, including the target variable `condition`.

Among the features, `thal` (thalassemia status) and `condition` exhibit the strongest positive correlation (0.52), followed closely by `ca` (number of major vessels colored by fluoroscopy) with a correlation of 0.46, and `cp` (chest pain type) at 0.41. These associations suggest that these features are potentially important indicators for

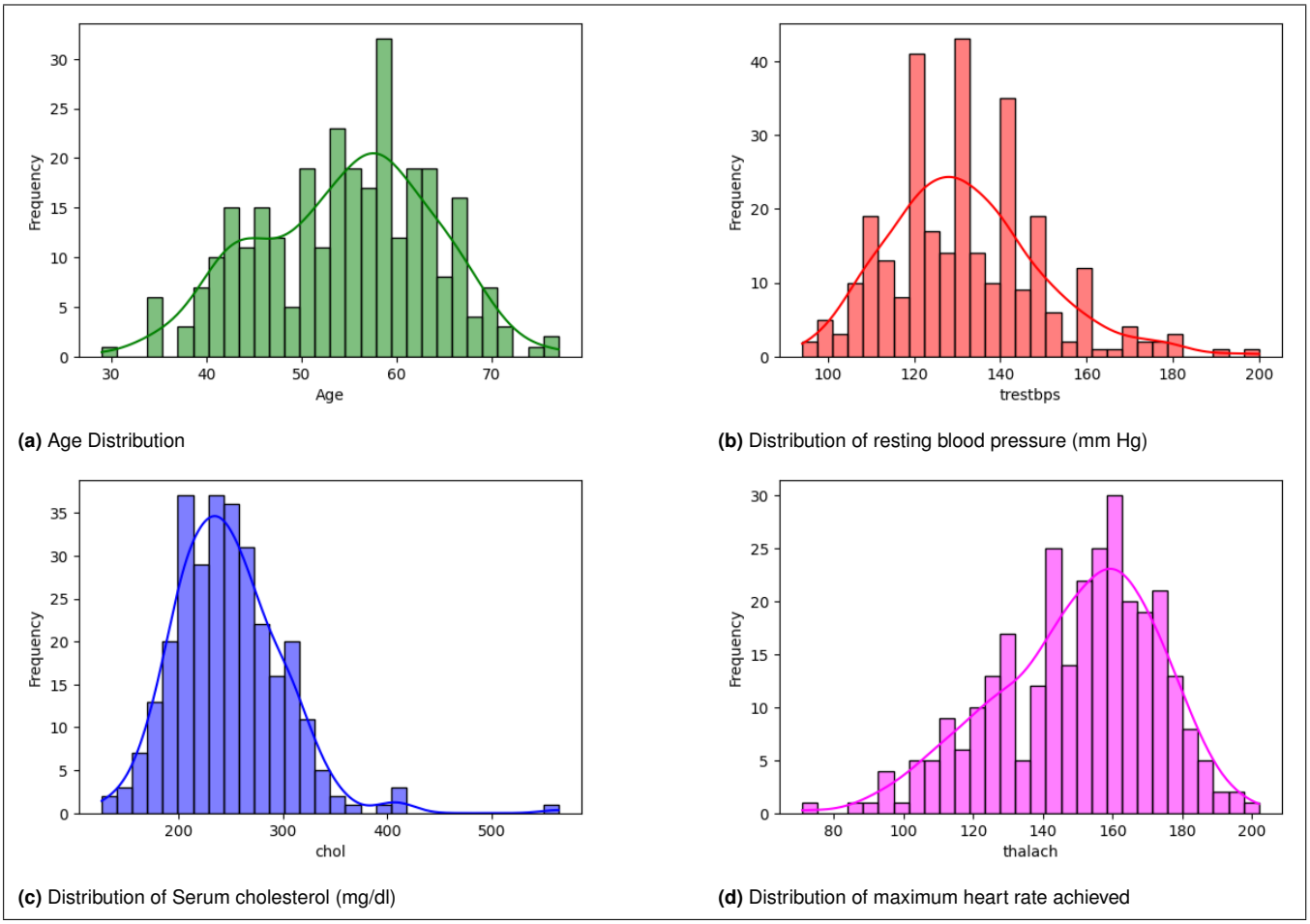


Figure 5. Distribution of continuous variables in the Cleveland Heart Disease dataset: (a) age, (b) resting blood pressure (trestbps), (c) serum cholesterol (chol), and (d) maximum heart rate achieved (thalach). The plots highlight typical ranges in the study population, with elevated cholesterol, higher blood pressure, and reduced maximum heart rate aligning with established cardiovascular risk factors.

predicting heart disease. Additionally, `exang` (exercise-induced angina) and `oldpeak` (ST depression) also show moderate positive correlations with `condition`, with coefficients of 0.42 and 0.33, respectively.

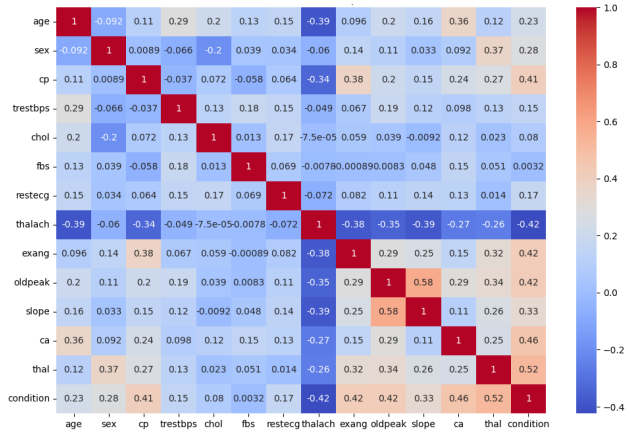
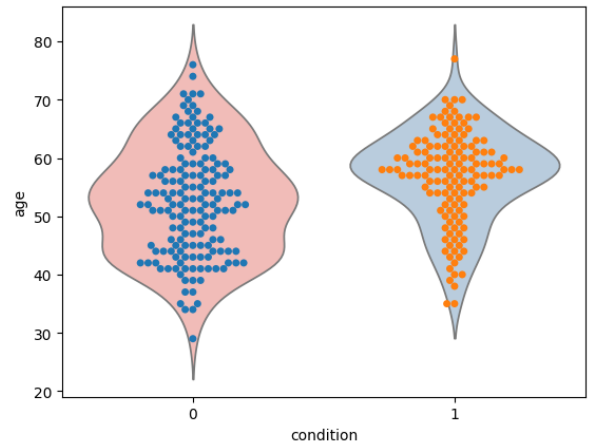


Figure 6. Correlation heatmap of each feature.

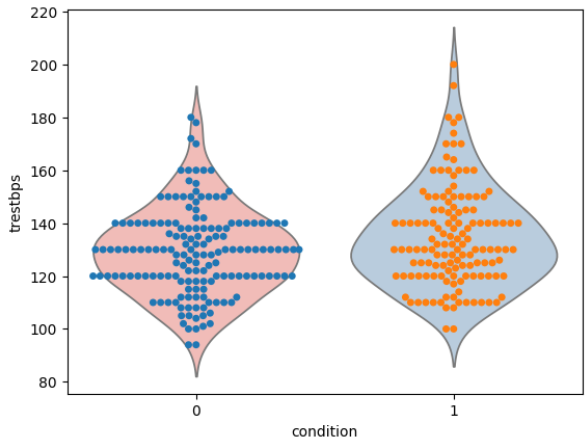
Conversely, `thalach` (maximum heart rate achieved) demonstrates the strongest negative correlation with `condition` at -0.42, implying that lower heart rates are associated with the presence of heart disease. Similarly, `age` and `trestbps` (resting blood pressure) exhibit weaker positive correlations with `condition`, at 0.23 and 0.15, respectively.

Overall, the correlation analysis highlights several features with moderate associations with heart disease status, which may contribute meaningfully to predictive modeling.

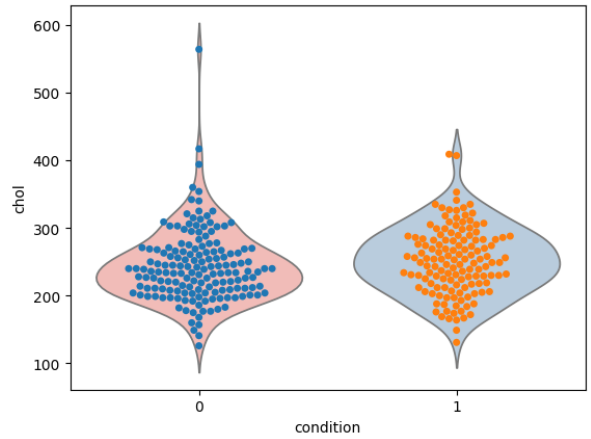
Figure 7 illustrates the distribution of key continuous variables in the Cleveland Heart Disease dataset using a swarm-violin plot, stratified by disease status. The age distribution (Figure 7a) illustrates that patients diagnosed with heart disease (condition = 1) exhibit an age distribution concentrated primarily between 50 and 65 years, with the central tendency near 60 years of age. In contrast, patients without heart disease (condition = 0) demonstrate a broader age range, spanning approximately from 35 to 75 years, with a greater proportion of younger individuals and a central tendency around 50 years. The density of younger individuals, particularly those below 35 years, is notably higher in the no-disease group, whereas very few patients in this age bracket present with heart disease. This visualization suggests that while heart disease occurs across a broad age range, it is relatively more concentrated among older individuals, supporting the well-established association between advancing age and increased cardiovascular risk.



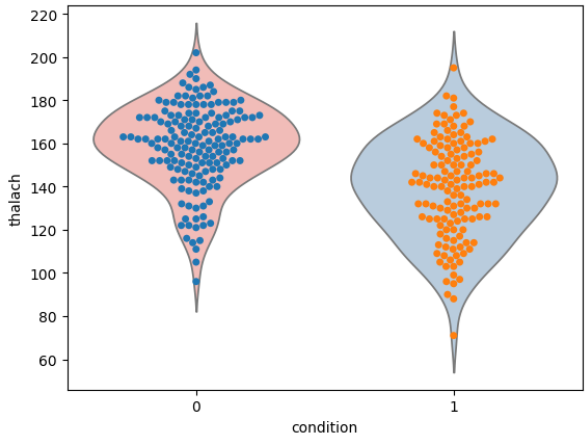
(a) Age distribution of patients with heart disease vs no heart disease (0 = No heart disease, 1 = Heart disease).



(b) Resting blood pressure (mm Hg) distribution for patients with and without heart disease.



(c) Serum cholesterol levels (mg/dL) in patients categorized by heart disease presence.



(d) Maximum heart rate achieved by patients, grouped by heart disease status.

Figure 7. The distribution of key continuous variables in the heart disease dataset, grouped by presence (1) or absence (0) of heart disease. Each plot visualizes the spread and central tendency of the data for: (a) age, (b) resting blood pressure, (c) serum cholesterol, and (d) maximum heart rate achieved.

Resting blood pressure (Figure 7b) shows a modest upward shift among patients with heart disease, with noticeably higher median values relative to those without disease, consistent with hypertension as a well-established cardiovascular risk factor. Among the diseased group (condition = 1), values cluster primarily between 120 and 150 mm Hg, but the distribution also exhibits a broader spread and an extended upper tail exceeding 200 mm Hg. In contrast, the non-diseased group (condition = 0) displays a more compact distribution, concentrated around 120 to 140 mm Hg, with relatively fewer extreme cases. This wider variability and heavier upper tail in the diseased cohort suggest that elevated blood pressure is more common and more severe in patients with heart disease. Although some overlap exists between the two groups, the greater density of readings above 140 mm Hg in the diseased group reinforces the strong association between higher resting blood pressure and cardiovascular risk.

Figure 7c compares the distribution of serum cholesterol levels (`chol`) between patients with and without heart disease. For both groups, cholesterol values are predominantly clustered around 200 to 300 mg/dl. However, the non-diseased group (condition = 0) exhibits a slightly wider spread, with cholesterol levels extending beyond 500 mg/dl in a few cases, which indicates an outlier. The diseased group (condition = 1) shows a comparatively narrower distribution, with most values concentrated between 180 and 300 mg/dl and fewer extreme outliers.

The maximum heart rate achieved (`thalach`, Figure 7d reveals a clear distinction between patients with and without heart disease. The non-diseased group (condition = 0) generally attains higher maximum heart rates, with values clustering between 140 and 170 beats per minute (bpm) and some extending beyond 190 bpm. By contrast, the diseased group (condition = 1) exhibits a lower distribution, concentrated around 120 to 150 bpm, with relatively few observations exceeding 170 bpm. This divergence suggests that patients without heart disease possess greater cardiovascular fitness and fewer physiological restrictions during exercise testing, whereas lower maximum heart rates in the diseased group likely reflect underlying cardiac limitations. Overall, the inverse relationship between `thalach` and disease status underscores its value as a

discriminative feature in distinguishing between groups, supporting its role as a meaningful predictor in classification models.

Collectively, these comparisons emphasize that age, blood pressure, cholesterol, and heart rate all exhibit clinically meaningful differences between the two groups, aligning with established cardiovascular risk profiles.

Model Performance and Discussion

The three linear classifiers—generalized linear model (GLM), Lasso logistic regression, and linear discriminant analysis (LDA)—all demonstrated strong discriminative ability on the held-out test set (Table 2).Among them, the Lasso logistic regression model achieved the highest accuracy (0.90), while GLM logistic and LDA attained accuracies of 0.87 and 0.88, respectively. In terms of precision, recall, and F1-score, the Lasso Logistic Regression model also achieved the best performance compared to the other two models. As shown in Figure 12, all three models exhibit steep ROC curves with high AUCs, confirming their strong discriminative capability. Notably, each model achieves a true positive rate exceeding 0.80 at a false positive rate below 0.10, which is a desirable property for clinical screening applications where minimizing false alarms is critical.

Table 2. Performance metrics for each model.

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
GLM Logistic	0.87	0.83	0.83	0.83	0.94
Lasso Logistic	0.90	0.88	0.88	0.88	0.94
LDA	0.88	0.87	0.83	0.85	0.95

Importantly, as illustrated in Figure 8–10, all three classifiers produced a small but non-negligible number of false positives and false negatives, indicating a balanced trade-off between sensitivity and specificity rather than perfect classification. The GLM logistic regression model yielded four false positives and four false negatives, resulting in a recall of 0.83 and an F1-score of 0.83. The Lasso logistic regression model showed slightly improved performance, with three false positives and three false negatives, achieving higher precision and recall (both 0.88) and the highest F1-score among the three models. In contrast, the LDA model produced three false positives and four false negatives, leading to a recall of 0.83 and a marginally lower F1-score of 0.85. Overall, the confusion matrices confirm the quantitative results in Table 2, with Lasso logistic regression offering the most balanced classification performance, while GLM and LDA exhibit comparable but slightly reduced sensitivity.

The results of the Generalized Linear Model (GLM) analysis (Figure 11, Table 3) provide insight into the key clinical features associated with heart disease. Among the thirteen predictors considered, five variables emerged as statistically significant contributors to the model's performance (Table 3), highlighting factors with meaningful associations after adjusting for all covariates.

The number of major vessels colored by fluoroscopy (`ca`) exhibited the strongest positive association with heart disease, with a highly significant p-value ($p < 0.001$), indicating that patients with a greater number of affected

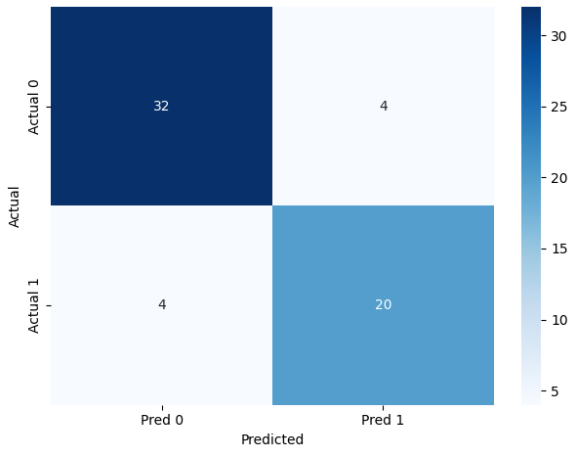


Figure 8. Confusion metrics of Generalized Linear Model.

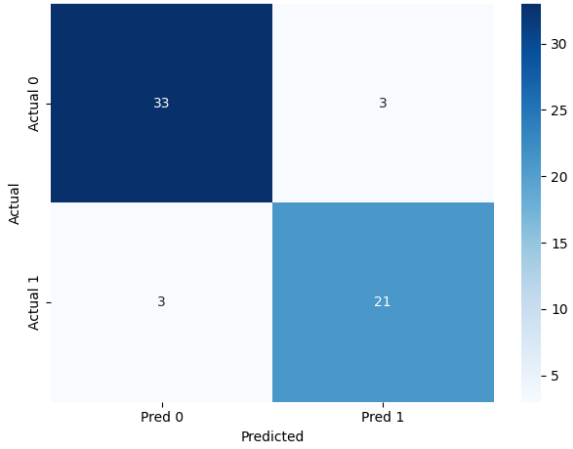


Figure 9. Confusion metrics of Lasso Logistic Model.

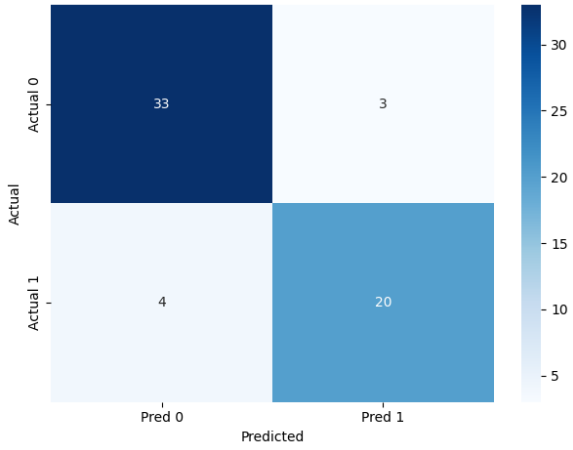


Figure 10. Confusion metrics of Linear Discriminant Analysis Model.

vessels are at substantially higher risk. Sex (`sex`) was also significantly associated with disease status ($p = 0.005$), with male patients showing an increased likelihood of heart disease. Resting blood pressure (`trestbps`) demonstrated a significant positive association ($p = 0.012$), reinforcing the well-established role of hypertension as a cardiovascular risk factor. In addition, thalassemia status (`thal`) was positively

associated with heart disease ($p = 0.011$), underscoring its importance as a clinically relevant biomarker.

Fasting blood sugar (`fbs`) exhibited a statistically significant negative association with heart disease ($p = 0.023$). This inverse relationship reflects the conditional effect of `fbs` within the multivariable model and should be interpreted in the context of interactions with other clinical covariates rather than as an isolated physiological effect.

Overall, these findings underscore the clinical relevance of the identified predictors in cardiovascular risk stratification. The interpretability of the GLM framework enables clear identification of influential variables, supporting transparent clinical decision-making. The significance of factors such as `ca`, `trestbps`, and `thal` aligns with established medical knowledge, further validating the reliability and clinical applicability of the proposed model.

Generalized Linear Model Regression Results

Dep. Variable:	condition	No. Observations:	237
Model:	GLM	Df Residuals:	223
Model Family:	Binomial	Df Model:	13
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-84.751
Date:	Fri, 19 Dec 2025	Deviance:	169.50
Time:	06:56:56	Pearson chi2:	236.
No. Iterations:	6	Pseudo R-squ. (CS):	0.4877
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0727	0.206	0.353	0.724	-0.331	0.477
x1	-0.0988	0.234	-0.423	0.672	-0.557	0.359
x2	0.6903	0.247	2.795	0.005	0.206	1.174
x3	0.3725	0.201	1.853	0.064	-0.022	0.766
x4	0.5155	0.206	2.502	0.012	0.112	0.919
x5	0.3619	0.215	1.686	0.092	-0.059	0.783
x6	-0.4936	0.217	-2.272	0.023	-0.920	-0.068
x7	0.2105	0.202	1.045	0.296	-0.184	0.606
x8	-0.5054	0.256	-1.974	0.048	-1.007	-0.004
x9	0.4514	0.217	2.080	0.038	0.026	0.877
x10	0.3317	0.269	1.233	0.218	-0.196	0.859
x11	0.2675	0.242	1.106	0.269	-0.207	0.742
x12	1.2448	0.271	4.594	0.000	0.714	1.776
x13	0.5378	0.213	2.530	0.011	0.121	0.954

Figure 11. Generalized Linear Model performance

Table 3. Most significant variables identified by the generalized linear model.

Variable	Interpretation
ca	Number of major vessels colored by fluoroscopy is strongly and positively associated with heart disease.
sex	Male sex is positively associated with the presence of heart disease.
trestbps	Higher resting blood pressure increases the likelihood of heart disease.
thal	Thalassemia status is positively associated with heart disease.
fbs	Fasting blood sugar level is negatively associated with heart disease.

Beyond GLM performance, the Lasso logistic regression model enforces sparsity—retaining a subset of key predictors (e.g., chest pain type, maximum heart rate, ST depression) and shrinking others to zero—which reduces measurement burden and highlights the most relevant biomarkers. LDA, though not explicitly sparse, projects patients onto a single continuous risk axis that can be readily visualized and thresholded in practice. The GLM model strikes a balance between full-feature modeling and straightforward odds-ratio interpretation.

Comparison with existing literature

Due to the high stakes of early diagnosis and treatment, heart disease prediction has been a prominent application area for machine learning. Numerous studies have evaluated a

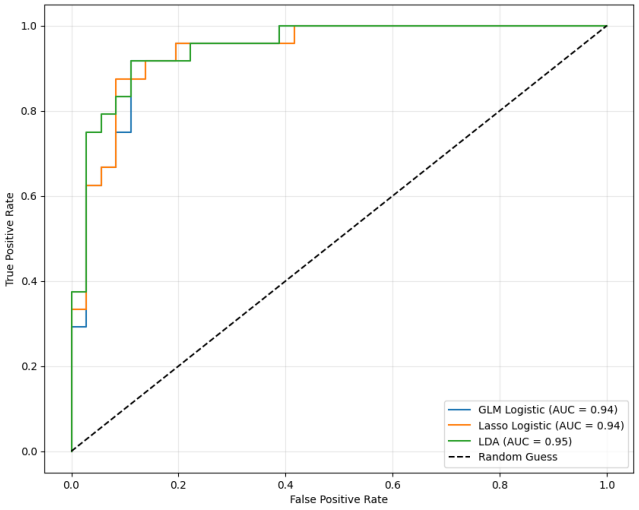


Figure 12. ROC curve comparison for each model.

wide variety of classifiers, ranging from traditional statistical methods to advanced ensemble techniques, each offering unique advantages in terms of accuracy, interpretability, and generalization. Table 4 presents a comprehensive summary of the relevant literature and compares the performance of our proposed models with those from previous studies.

Jha et al. (2025) explored a series of machine learning models including Decision Trees, Support Vector Machines (SVM), Random Forests, and Artificial Neural Networks (ANN) for heart disease classification. While ANN achieved the best overall performance with an accuracy of 0.86 and a balanced F1-score of 0.83, the SVM model showed a significant drop in recall (0.42) despite a similar accuracy, indicating a tendency toward false negatives—a critical concern in medical diagnostics. Decision Trees, while interpretable, yielded moderate results with an F1-score of 0.80, reflecting their limitations when applied without ensemble enhancements.

Yadav et al. (2023) demonstrated high accuracies for Decision Tree and Random Forest classifiers (0.97), but their study lacked additional evaluation metrics such as precision, recall, and ROC-AUC. This omission makes it challenging to fully assess the models' reliability, especially in imbalanced datasets where accuracy alone can be misleading. Similarly, Logistic Regression and KNN achieved accuracies of 0.81 and 0.70, respectively, suggesting that simpler linear models may require more feature engineering or regularization to perform competitively.

Bhatt et al. (2023) conducted a more comprehensive evaluation, incorporating metrics like precision, recall, F1-score, and ROC-AUC. Their use of Multilayer Perceptron, Random Forest, Decision Tree, and XGBoost models yielded consistent and high performance, with XGBoost achieving the best F1-score (0.86) and ROC-AUC (0.95). This emphasizes the strength of ensemble-based learning in capturing complex feature interactions, particularly in health data where nonlinear relationships are prevalent.

Shah et al. (2020) and Choudhary and Singh (2020) also employed traditional classifiers, with Naïve Bayes reaching an accuracy of 0.88 and Decision Trees up to 0.97. Choudhary's study is notable for reporting all key

Table 4. Performance comparison of classification models from literature.

Author name	Model	Accuracy	Precision	Recall	F1_Score	RoC-AUC
Jha et al. (2025)	Decision Tree (DT)	0.78	0.77	0.83	0.80	
	Support Vector Machine (SVM)	0.78	0.69	0.42	–	
	Random Forest (RF)	0.67	0.64	0.63	0.66	
	Artificial Neural Networks (ANN)	0.86	0.86	0.84	0.83	
Yadav et al. (2023)	Logistic Regression (LR)	0.81				
	Decision Tree (DT)	0.97				
	Random Forest (RF)	0.97				
	K-Nearest Neighbor (KNN)	0.70				
Bhatt et al. (2023)	Multilayer Perceptron (MLP)	0.87	0.89	0.83	0.86	0.95
	Random Forest (RF)	0.87	0.89	0.83	0.86	0.95
	Decision Tree (DT)	0.87	0.90	0.81	0.85	0.94
	Xgboost (XGB)	0.87	0.90	0.82	0.86	0.95
Shah et al. (2020)	Naïve Bayes (NB)	0.88				
	K-Nearest Neighbor (KNN)	0.79				
	Decision Tree (DT)	0.74				
	Random Forest (RF)	0.84				
Choudhary and Singh (2020)	Decision Tree (DT)	0.97				
	Ada-boost (AB)	0.89	0.91	0.89	0.90	
Jindal et al. (2021)	K-Nearest Neighbor (KNN)	0.89				
	Logistic Regression (LR)	0.89				
	Random Forest (RF)	0.85				
Saboor et al. (2022)	Multinomial Naïve Bayes (MNB)	0.93				
	Support Vector Machine (SVM)	0.97				
	Logistic Regression (LR)	0.87				
	CART	0.84				
	Linear Discriminant Analysis (LDA)	0.95				
	Ada-boost (AB)	0.93				
	Random Forest (RF)	0.90				
	Extra Tree (ET)	0.95				
	Xgboost (XGB)	0.92				
Liu and Fu (2014)	CS-PSO-SVM	0.85	0.82	0.90	0.86	
Yang et al. (2015)	PCA with quasi linear SVM	0.87	0.86	0.84	0.85	
Liu et al. (2017)	Ensemble classifier (k = 50)	0.89				
	Ensemble classifier (k = 100)	0.93				
	Ensemble classifier (k = 150)	0.91				
	C 4.5 tree	0.87				
	Naïve Bayes (NB)	0.83				
	Bayesian Neural Network (BNN)	0.85				
Nashif et al. (2018)	Naïve Bayes (NB)	0.86	0.86	0.86	0.86	
	Support Vector Machine (SVM)	0.98	0.96	0.98	0.97	
	Random Forest (RF)	0.96	0.96	0.96	0.96	
	Simple Logistic	0.95	0.95	0.95	0.95	
	Artificial Neural Networks (ANN)	0.77	0.78	0.77	0.77	
Ayatollahi et al. (2019)	Artificial Neural Networks (ANN)				0.88	
	Support Vector Machine (SVM)				0.92	
Our model	**Generalized Linear Model (Logistic)**	**0.87**	**0.83**	**0.83**	**0.83**	**0.94**
	Lasso Logistic Regression	**0.90**	**0.88**	**0.88**	**0.88**	**0.94**
	Linear Discriminant Analysis (LDA)	**0.88**	**0.87**	**0.83**	**0.85**	**0.95**

metrics for AdaBoost (e.g., F1-score of 0.90), highlighting how ensemble strategies can elevate even basic learners' performance. However, the lack of full metric coverage in some of these works reduces transparency in model comparison.

Jindal et al. (2021) and Saboor et al. (2022) expanded on this by benchmarking a wide range of classifiers. Saboor's study stands out with nine algorithms, among which SVM and LDA achieved the highest accuracies (up to 0.97 and 0.95, respectively). These results reinforce the robustness of margin-based classifiers and discriminant analysis in medical classification tasks. However, simpler models like CART and Logistic Regression showed slightly lower performance (0.84–0.87), suggesting that they might be better suited as baselines rather than final models.

More novel approaches were explored by Liu and Fu (2014), who implemented a hybrid PSO-SVM model, and Yang et al. (2015), who applied PCA-based clustering with

quasi-linear SVM. Both studies reported strong, balanced performance across key metrics, highlighting the benefit of combining optimization techniques and dimensionality reduction for medical data.

Ensemble classifiers were further studied by Liu et al. (2017), who varied ensemble sizes (k=50 to 150), noting that larger ensembles (k=100) improved classification outcomes (accuracy up to 0.93). Nashif et al. (2018) evaluated five different classifiers on real-time cardiovascular data, reporting very high accuracy for SVM (0.98) and Random Forest (0.96), with comprehensive metric coverage. These findings suggest that models with higher capacity and regularization perform better under noisy, real-world conditions.

Ayatollahi et al. (2019) focused on F1-scores for SVM and ANN, reporting values of 0.92 and 0.88, respectively. The preference toward SVM aligns with previous findings, especially in datasets with well-separated classes.

In summary, the literature highlights that ensemble and neural network-based models generally outperform traditional classifiers in heart disease prediction, especially when evaluated across multiple metrics. However, many works still emphasize accuracy alone, neglecting critical measures like recall and ROC-AUC that are vital in clinical settings. Additionally, some studies lack reproducibility due to limited reporting of experimental setups and metrics. These gaps motivate the development of more interpretable yet high-performing models, which our work aims to address by combining generalized linear and regularized classifiers with full performance evaluation.

Conclusion

This study has systematically evaluated the predictive capabilities of three interpretable linear classifiers—Generalized Linear Model (GLM), L1-regularized (Lasso) logistic regression, and Linear Discriminant Analysis (LDA)—on the Cleveland Heart Disease dataset. Addressing the clinical demand for models that balance accuracy with transparency, the results demonstrate that relatively simple linear approaches can achieve strong discriminative performance while remaining interpretable and clinically meaningful.

Across all models, ROC-AUC values exceeded 0.94, indicating robust class separation. Among the three approaches, Lasso logistic regression achieved the highest overall accuracy and F1-score, reflecting the most balanced trade-off between precision and recall. The GLM and LDA models produced comparable results, with slightly lower sensitivity but similarly strong discrimination. Analysis of the confusion matrices revealed that all models incurred small but non-negligible numbers of false positives and false negatives, emphasizing realistic performance trade-offs rather than idealized classification outcomes.

Beyond predictive accuracy, the GLM analysis provided valuable clinical insights by identifying significant predictors such as chest pain type, thalassemia status, number of major vessels visualized by fluoroscopy, exercise-induced angina, and maximum heart rate achieved. These variables align well with established cardiovascular risk factors, reinforcing the clinical credibility and interpretability of the model. While Lasso regression further enhances interpretability through coefficient sparsity and feature selection, LDA offers a compact one-dimensional risk score that can be flexibly thresholded in practice.

In summary, this work demonstrates that interpretable linear models—particularly Lasso-regularized and generalized linear approaches—can serve as effective and transparent tools for heart disease prediction. Their combination of competitive performance, interpretability, and ease of implementation makes them well suited for real-world clinical decision support. Future research should extend this analysis to larger and more diverse cohorts and explore hybrid approaches that preserve interpretability while capturing more complex data relationships.

References

World Health Organization (2024) *Cardiovascular diseases (CVDs)*. Available at: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 (Accessed: 15 May 2025).

Shah D, Patel S and Bharti SK (2020) *Heart disease prediction using machine learning techniques*. *SN Comput. Sci.* 1(6): 345.

World Heart Federation (2023) *World Heart Report 2023: Confronting the World's Number One Killer*. Geneva, Switzerland: World Heart Federation. Available at: https://world-heart-federation.org/resource/world-heart-report-2023 (Accessed: 15 May 2025).

Di Cesare M, Perel P, Taylor S, Kabudula C, Bixby H, Gaziano TA, McGhie DV, Mwangi J, Pervan B, Narula J, Pineiro D and Pinto FJ (2024) *The Heart of the World*. *Global Heart* 19(1): 11. doi: 10.5334/gh.1288. PMID: 38273998; PMCID: PMC10809869.

Mensah GA, Fuster V, Murray CJ, Roth GA and Global Burden of Cardiovascular Diseases and Risks Collaborators (2023) *Global burden of cardiovascular diseases and risks, 1990–2022*. *J. Am. Coll. Cardiol.* 82(25): 2350–2473. doi: 10.1016/j.jacc.2023.11.007.

Savarese G, Becher PM, Lund LH, Seferovic P, Rosano GM and Coats AJ (2022) *Global burden of heart failure: a comprehensive and updated review of epidemiology*. *Cardiovasc. Res.* 118(17): 3272–3287. doi: 10.1093/cvr/cvac013.

Centers for Disease Control and Prevention (2024) *Heart Disease Facts and Statistics*. Available at: https://www.cdc.gov/heart-disease/data-research/facts-stats/ (Accessed: 15 May 2025).

Ahsan MM and Siddique Z (2022) *Machine learning-based heart disease diagnosis: A systematic literature review*. *Artif. Intell. Med.* 128: 102289. doi: 10.1016/j.artmed.2022.102289.

Azmi J, Arif M, Nafis MT, Alam MA, Tanweer S and Wang G (2022) *A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data*. *Med. Eng. Phys.* 105: 103825. doi: 10.1016/j.medengphy.2022.103825.

Zhou C, Dai P, Hou A, et al. (2024) *A comprehensive review of deep learning-based models for heart disease prediction*. *Artif. Intell. Rev.* 57: 263. doi: 10.1007/s10462-024-10899-9.

Jha KM, Velaga V, Routhu KK, Sadaram G and Boppana SB *et al.* (2025) *Evaluating the effectiveness of machine learning for*

heart disease prediction in healthcare sector. *J. Cardiobiol.* 9(1): 1.

Yadav AL, Soni K and Khare S (2023) *Heart diseases prediction using machine learning*. In: *Proc. 14th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)*, pp. 1–7.

Mamun MMR and Alouani A (2022) *Automatic detection of heart diseases using biomedical signals: A literature review of current status and limitations*. In: Arai K (ed) *Advances in Information and Communication. FICC 2022*. Lecture Notes in Networks and Systems, vol. 439. Cham: Springer. doi: 10.1007/978-3-030-98015-3_29.

Dua D and Taniskidou EK (2017) *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences. [Online]. Available: https://archive.ics.uci.edu/ml

McCullagh P and Nelder JA (1989) *Generalized Linear Models*. 2nd ed. London: Chapman and Hall/CRC.

Agresti A (2015) *Foundations of Linear and Generalized Linear Models*. 2nd ed. Hoboken, NJ: Wiley.

Tibshirani R (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58(1): 267–288.

Hastie T, Tibshirani R and Friedman J (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B *et al.* (2011) *Scikit-learn: Machine learning in Python*. *J. Mach. Learn. Res.* 12: 2825–2830. [Online]. Available: https://scikit-learn.org

Seabold S and Perktold J (2010) *Statsmodels: Econometric and statistical modeling with Python*. In: *Proc. 9th Python Sci. Conf.*, pp. 92–96. [Online]. Available: https://www.statsmodels.org

Bhatt CM, Patel P, Ghetia T and Mazzeo PL (2023) *Effective heart disease prediction using machine learning techniques*. *Algorithms* 16(2): 88.

Choudhary G and Singh SN (2020) *Prediction of heart disease using machine learning algorithms*. In: *Proc. Int. Conf. Smart Technol. Comput. Electr. Electron. (ICSTCEE)*, Bengaluru, India, pp. 197–202. doi: 10.1109/ICSTCEE49637.2020.9276802.

Jindal H, Agrawal S, Khera R, Jain R and Nagrath P (2021) *Heart disease prediction using machine learning algorithms*. In: *IOP Conf. Ser. Mater. Sci. Eng.* 1022(1): 012072. IOP Publishing.

Saboor A, Usman M, Ali S, Samad A, Abrar MF and Ullah N (2022) *A method for improving prediction of human heart disease using machine learning algorithms*. *Mob. Inf. Syst.* 2022(1): 1410169.

Liu X and Fu H (2014) *PSO-based support vector machine with cuckoo search technique for clinical disease diagnoses*. *Sci. World J.* 2014(1): 548483.

Yang C, Yang K and Zhou B (2015) *A hierarchical clustering method based on PCA-clusters merging for quasi-linear SVM*. In: *Proc. Int. Conf. Autom. Mech. Control Comput. Eng.*, pp. 1022–1028. Atlantis Press.

Liu X, Wang X, Su Q, Zhang M, Zhu Y, Wang Q and Wang Q (2017) *A hybrid classification system for heart disease diagnosis based on the RFRS method*. *Comput. Math. Methods Med.* 2017(1): 8272091.

Nashif S, Raihan MR, Islam MR and Imam MH (2018) *Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system*. *World J. Eng. Technol.* 6(4): 854–873.

Ayatollahi H, Gholamhosseini L and Salehi M (2019) *Predicting coronary artery disease: a comparison between two data mining algorithms*. *BMC Public Health* 19: 1–9.