Scientific Research Publishing

Open Journal of Applied Sciences, 2025, 15(11), 3534-3552
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

Heart Disease Prediction: A Logistic Regression Approach

Awele Okolie[1*], Callistus Obunadike[2], Stanley Chinedu Okoro[2], Itunu Blessing Olufemi[2], PearlRose Nwoke[3], Prince Michael Akwabeng[4]

[1]School of Computing and Data Science, Wentworth Institute of Technology, Boston, USA
[2]Department of Computer Science and Quantitative Methods, Austin Peay State University, Clarksville, USA
[3]Department of Computer Science, Boston University Metropolitan College, Boston, USA
[4]Department of Mathematics and Statistics, Austin Peay State University, Clarksville, USA
Email: *aweleokolie77@gmail.com

How to cite this paper: Okolie, A., Obunadike, C., Okoro, S., Olufemi, I., Nwoke, P. and Akwabeng, P. (2025) Heart Disease Prediction: A Logistic Regression Approach. *Open Journal of Applied Sciences*, **15**, 3534-3552.
https://doi.org/10.4236/ojapps.2025.1511229

Received: October 23, 2025
Accepted: November 14, 2025
Published: November 17, 2025

Copyright © 2025 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Open Access

Abstract

Heart disease remains one of the leading causes of mortality worldwide, accounting for millions of deaths annually. Early detection of individuals at risk is essential for reducing complications and improving patient outcomes. This study applies logistic regression, a supervised machine learning algorithm, to predict the likelihood of heart disease based on clinical and demographic features such as age, sex, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar, and maximum heart rate achieved. The dataset obtained from Kaggle's Heart Disease Dataset, comprises 1025 patient records with 14 attributes. Following data preprocessing including handling missing values, feature scaling using StandardScaler, and categorical encoding, the data were divided into training (80%) and testing (20%) subsets. A logistic regression model with the liblinear solver and L2 regularization was trained and evaluated using multiple performance metrics. The model achieved 85.24% accuracy on the training set and 80.49% accuracy on the test set, with a ROC-AUC score of 0.86 and consistent results from 5-fold cross-validation. These findings demonstrate that logistic regression provides a robust, interpretable, and computationally efficient approach for binary classification in healthcare. The model's high recall indicates its reliability in identifying patients at risk of heart disease, supporting its potential application in clinical decision-support systems for early diagnosis and intervention.

Keywords

Heart Disease Prediction, Logistic Regression, Machine Learning, Predictive Modeling, Healthcare Analytics, Clinical Data, Early Diagnosis

1. Introduction

Heart disease is among the leading causes of death worldwide. Early detection heart disease, also known as cardiovascular disease (CVD), remains one of the most significant health challenges worldwide. According to the World Health Organization (WHO), approximately 17.9 million people die from CVD each year, representing 32% of all global deaths [1]. This alarming figure underscores the urgent need for effective tools to predict and prevent heart related conditions. Early diagnosis can greatly reduce the risk of complications by allowing timely intervention and lifestyle modifications. Traditional diagnostic techniques, such as electrocardiograms (ECG), echocardiography, and angiography, are accurate but often costly, time-consuming, and require specialized expertise. In recent years, machine learning (ML) has emerged as a valuable approach in healthcare analytics, offering automated, data-driven methods for identifying disease patterns and improving clinical decision-making. Among ML algorithms, Logistic Regression (LR) is one of the most widely used techniques for binary classification problems where outcomes are categorized into two groups, such as the presence or absence of heart disease. It is computationally efficient, interpretable, and particularly suitable for medical datasets where transparency and simplicity are important. Logistic Regression not only predicts whether an individual has heart disease but also provides insights into which factors (such as age, cholesterol, or blood pressure) most influence the risk. Recent research has shown that integrating multiple health indicators into a predictive model significantly improves diagnostic accuracy. Logistic Regression models, when trained on quality clinical data, can serve as early screening tools that assist healthcare professionals in identifying at-risk patients before severe symptoms occur [2]. The goal of this study is to develop a Logistic Regression model using the Heart Disease Dataset from Kaggle to predict the likelihood of heart disease. The model's performance is evaluated through training and testing phases, and its accuracy is compared across multiple metrics to assess its reliability. Ultimately, this study aims to demonstrate how machine learning can complement traditional medical assessments by providing accessible and interpretable tools for heart disease prediction.

2. Literature Review

This section reviews existing research and methodologies related to heart disease prediction using machine learning. The focus is on the use of Logistic Regression and other algorithms, key predictive factors, and challenges associated with modeling cardiovascular disease.

2.1. Overview of Heart Disease Prediction Using Machine Learning

Over the past decade, machine learning models have gained widespread attention for their ability to predict medical conditions, including heart disease. Researchers have applied various algorithms such as Logistic Regression, Support Vector Machines (SVM), Random Forests, Decision Trees, and Artificial Neural Networks

to classify patients based on clinical and demographic data. These models leverage features such as age, gender, cholesterol levels, blood pressure, chest pain type, maximum heart rate and exercise-induced angina to predict whether a person has heart disease. Machine learning models are particularly useful because they can detect complex patterns and relationships in large datasets that may not be apparent to clinicians. For example, combining multiple health indicators in a predictive model can improve the accuracy of early diagnosis, enabling healthcare providers to intervene before the disease progresses. According to Dey *et al.* [3], Logistic Regression is a commonly used approach in heart disease prediction because it balances interpretability and predictive performance. Unlike more complex models, Logistic Regression provides probabilities for the presence of disease, offering clear, actionable insights. This makes it particularly suitable for binary classification problems where outcomes are discrete such as determining whether heart disease is present (1) or absent (0).

2.2. Role of Logistic Regression in Medical Diagnosis

Logistic Regression (LR) is one of the most widely used statistical models in healthcare analytics, especially for predicting the presence or absence of diseases. It is a supervised learning algorithm designed for binary classification problems, meaning it predicts outcomes that have two possible values, such as "disease" or "no disease".

The core strength of Logistic Regression lies in its ability to estimate the probability of an event occurring, rather than just providing a yes/no answer. The model uses a sigmoid function to map any input from the linear combination of features to a probability between 0 and 1. This allows healthcare professionals to determine risk levels and set thresholds for intervention. For example, if the model predicts an 85% probability of heart disease for a patient, doctors can prioritize further diagnostic tests or preventive measures for that individual.

In heart disease prediction, Logistic Regression assigns weights to each feature, which quantify how strongly that feature contributes to the outcome. For instance, higher age, elevated cholesterol, and high blood pressure typically increase the probability of heart disease. Conversely, protective factors such as regular exercise or lower cholesterol may reduce risk. This interpretability is crucial in medical contexts because it allows clinicians to understand why the model made a particular prediction, increasing trust in its use.

Additionally, Logistic Regression is computationally efficient and does not require extensive computational resources, which is advantageous when working with small to medium-sized datasets commonly found in healthcare research. Its simplicity and transparency make it suitable for integration into hospital systems, clinical decision support tools, and mobile health applications.

2.3. Comparison with Other Machine Learning Models

While Logistic Regression is widely used for heart disease prediction due to its

simplicity and interpretability, other machine learning models are also employed in research and practice. Each model has its own strengths and weaknesses, often balancing accuracy, complexity, and interpretability. Understanding these alternatives provides context for why Logistic Regression is chosen for this study.

2.3.1. Decision Trees

Decision Trees are a non-linear, rule-based classification method commonly used in medical diagnostics [4]. Decision Trees are easy to interpret because they visually represent decision rules, allowing clinicians to trace how predictions are made. However, they are prone to overfitting, especially when the tree grows too deep, which can reduce predictive accuracy on unseen data [5].

- **Advantages:** Decision Trees are intuitive and easy to visualize, allowing clinicians to see the step-by-step reasoning behind predictions. They can also capture non-linear relationships between features, which Logistic Regression may not detect.
- **Disadvantages:** Large or deep Decision Trees can overfit the training data, leading to poor generalization on new patients. They are also sensitive to small changes in data, which may result in different tree structures.

Decision Trees are often used as a baseline model in heart disease research due to their transparency, but they may require pruning or regularization to prevent overfitting.

2.3.2. Random Forests

Random Forests are an ensemble method that builds multiple decision trees and aggregates their predictions to improve performance [5]. Each tree in the forest is trained on a random subset of the data and features, making the model more robust.

- **Advantages:** Random Forests handle complex, non-linear relationships better than a single Decision Tree. They also reduce the risk of overfitting and improve predictive performance, especially on medium to large datasets.
- **Disadvantages:** While more accurate than a single tree, Random Forests are less interpretable. Clinicians cannot easily trace the decision-making process across hundreds of trees. Computational requirements are higher compared to Logistic Regression or single Decision Trees.

Random Forests are often favored for research when the focus is on maximizing prediction accuracy rather than interpretability.

2.3.3. Support Vector Machines (SVM)

Support Vector Machines are powerful algorithms that classify data by finding the hyperplane that best separates the classes. SVMs can also use kernel functions to handle non-linear relationships.

- **Advantages:** SVMs are effective in high-dimensional spaces and can handle

both linear and non-linear data [6]. They can provide strong predictive performance for heart disease datasets with multiple features.

- **Disadvantages:** SVMs are sensitive to parameter settings and require careful tuning of the kernel, regularization, and margin parameters. They are also less interpretable than Logistic Regression, making it difficult to explain why a patient is classified as high-risk.

SVMs are typically used when datasets have complex patterns that simpler models might not capture effectively.

2.3.4. Artificial Neural Networks

Artificial Neural Networks are inspired by the human brain and can model highly complex, non-linear relationships between features and the target variable. They consist of layers of interconnected neurons that learn patterns through training.

- **Advantages:** ANNs can automatically detect intricate patterns in the data that simpler models cannot [7]. They have achieved state-of-the-art performance in tasks such as image recognition, natural language processing, and healthcare prediction.
- **Disadvantages:** ANNs require large amounts of data and computational resources. They are often considered "black boxes" because it is difficult to interpret how the model makes predictions. This lack of transparency can limit clinical adoption. Additionally, ANNs are prone to overfitting if the training dataset is small or not representative [8].

ANNs are best suited for large-scale datasets or when the primary goal is maximum prediction accuracy rather than interpretability.

2.3.5. Summary of Model Comparisons

In summary, each machine learning model has a trade-off between interpretability and predictive performance:

- **Logistic Regression:** High interpretability, moderate accuracy, low computational cost.
- **Decision Trees:** Moderate interpretability, captures non-linear patterns, risk of overfitting.
- **Random Forests:** High accuracy, low interpretability, computationally more demanding.
- **Support Vector Machines:** High accuracy in complex datasets, low interpretability, sensitive to parameters.
- **Artificial Neural Networks:** Highest potential accuracy, very low interpretability, requires large datasets and resources.

For this study, Logistic Regression is chosen because it provides interpretable results while maintaining sufficient predictive power for heart disease prediction. Its transparency allows clinicians to understand how each factor contributes to a patient's risk, which is critical in healthcare decision-making.

2.4. Features Selection and Importance in Heart Disease Prediction

According to [9], the predictor variables could otherwise be known as "PIE (predictor, independent or explanatory) variables" while the response variables could otherwise be termed "DORT (dependent, observatory, response or target) variables". Features (variables) importance enables the ML algorithm to train faster as well as reduces cost and time required for training the dataset, therefore making it simpler to interpret. It also reduces the variance of the model and improves the accuracy, provided the right subset is chosen [9].

2.4.1. Key Features in Heart Disease Prediction

Heart disease datasets commonly include a combination of demographic, clinical, and lifestyle variables. In this study, the following features were considered:

- **Age:** Older patients are at higher risk of cardiovascular diseases. Age has consistently shown strong predictive power in heart disease models.
- **Gender:** Males tend to have a higher prevalence of heart disease, although post-menopausal females also face increased risk.
- **Cholesterol Levels:** High serum cholesterol is associated with plaque buildup in arteries, a major risk factor for heart disease.
- **Resting Blood Pressure:** Hypertension contributes to the strain on the heart and is a significant predictor.
- **Chest Pain Type (cp):** Different types of chest pain (typical angina, atypical angina, non-angina) indicate varying levels of risk.
- **Fasting Blood Sugar (fbs):** Diabetes or elevated blood sugar increases the likelihood of heart disease.
- **Maximum Heart Rate Achieved (thalach):** Lower exercise capacity may reflect reduced cardiac function.
- **Exercise Induced Angina (exang):** The presence of angina during exertion indicates underlying heart problems.
- **Other Clinical Factors:** Such as Thalassemia (thal), Old Peak (depression from exercise), and number of major vessels colored (ca) in imaging studies.

2.4.2. Feature Importance in Logistic Regression

In Logistic Regression, each feature is assigned a coefficient that reflects its contribution to predicting heart disease. Positive coefficients indicate an increased likelihood of disease, while negative coefficients suggest protective effects. Understanding these weights is crucial for clinicians, as it provides insight into which factors require attention or intervention.

2.4.3. Feature Selection Techniques

Selecting the right subset of features can improve model performance and interpretability. Common techniques include:

- **Correlation Analysis:** Identifies features strongly associated with the target

variable while avoiding multicollinearity.

- **Recursive Feature Elimination (RFE):** Iteratively removes less informative features based on model performance.
- **Domain Knowledge:** Clinical expertise is used to retain variables that are relevant even if statistical significance is lower.
- **Regularization Techniques:** L1 (Lasso) or L2 (Ridge) regularization can penalize irrelevant features and reduce overfitting.

Proper feature selection ensures that the model is both accurate and interpretable, which is especially important in medical contexts where decisions directly affect patient care.

2.5. Advantages and Limitations of Logistic Regression for Heart Disease Prediction

While Logistic Regression is a popular choice in medical prediction tasks, it is important to understand its strengths and weaknesses relative to other models.

2.5.1. Advantages of Logistic Regression

Logistic Regression offers several advantages that make it a suitable model for heart disease prediction. It is highly interpretable, as each coefficient indicates the direction and magnitude of a feature's effect on the likelihood of heart disease, allowing clinicians to understand and trust the model's predictions. The algorithm is also computationally efficient, requiring minimal processing power and performing well on medium-sized datasets. Additionally, its probabilistic output provides a meaningful measure of risk rather than a simple binary classification, which supports better clinical decision-making. Furthermore, its simplicity and ease of implementation make it practical in healthcare environments with limited technical infrastructure. However, Logistic Regression also has some limitations, it assumes a linear relationship between predictors and the log-odds of the outcome, which may not capture complex nonlinear patterns. It can also be sensitive to multicollinearity among features, and its performance may decline when data relationships are highly intricate. Despite these constraints, Logistic Regression remains a robust baseline model that balances interpretability and predictive performance in medical applications.

2.5.2. Limitations of Logistic Regression

- **Assumes Linearity:** Logistic Regression assumes a linear relationship between input features and the log-odds of the target. Non-linear relationships may reduce accuracy.
- **Limited to Binary Outcomes:** Traditional Logistic Regression handles only two classes, extensions like multinomial logistic regression are needed for multi-class problems.
- **Sensitive to Multicollinearity:** Highly correlated features can distort coefficient estimates, requiring careful feature selection or preprocessing.
- **Potential Underperformance in Complex Datasets:** Models like Random Forests or ANNs may outperform Logistic Regression when relationships be-

tween features and outcomes are highly non-linear.

2.5.3. Clinical Implications

Despite these limitations, Logistic Regression is particularly suitable for heart disease prediction in clinical settings because it balances accuracy, interpretability, and usability. Clinicians can use model outputs to identify high-risk patients, prioritize interventions, and communicate risk effectively.

2.5.4. Integration with Other Models

Logistic Regression can also be combined with other models or feature engineering techniques to enhance performance. For example:

- Using Logistic Regression as a baseline for more complex models like Random Forests or Gradient Boosting.
- Applying ensemble methods to combine predictions from multiple models, improving accuracy while retaining interpretability.

3. Methodology

The aim of this research is to evaluate the predictive ability of Logistic Regression in classifying patients as having heart disease or not, based on multiple predictor variables. The dataset consists of 1025 patient records and 14 features obtained from Kaggle's Heart Disease dataset. The target variable ("target") was used as the response variable, with 0 = No Heart Disease and 1 = Heart Disease, representing a binary classification problem.

The analysis was performed using the Python programming language (version 3.11) and the following major libraries: pandas (2.2.2), numpy (1.26.4), matplotlib (3.9.1), seaborn (0.13.2), and scikit-learn (1.5.2). The process included data cleaning, exploratory data analysis (EDA), preprocessing, model training, and evaluation. The workflow also incorporated the generation of visualizations such as correlation heatmaps and distribution plots to better understand relationships between variables.

3.1. Descriptive Analysis of the Dataset

The dataset contains 14 variables (features). The target variable is the dependent, observatory, response, or target variable Y), while the remaining 13 variables represent predictor, independent, or explanatory variables (X) (see **Table 1**). Summary statistics (mean, median, minimum, maximum) were computed to understand the data distribution.

- **Missing Values Check:** No missing values were identified in the dataset.
- **Target Variable Distribution:** The target variable was visualized using a count plot (**Figure 1**) to examine the balance between patients with and without heart disease.
- **Feature Correlations:** A correlation heatmap was generated to identify relationships between predictors (**Figure 2**).
- **Age Distribution:** A chart showing age distribution by heart (**Figure 3**).

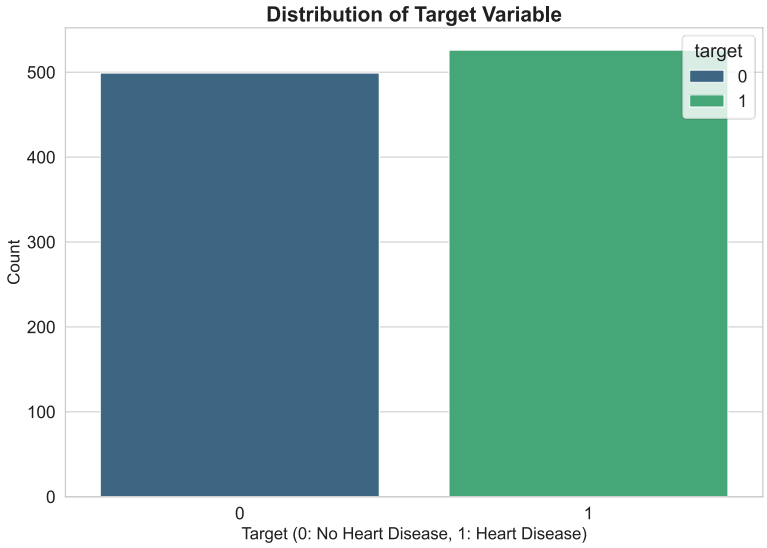

Figure 1. Target variable distribution.

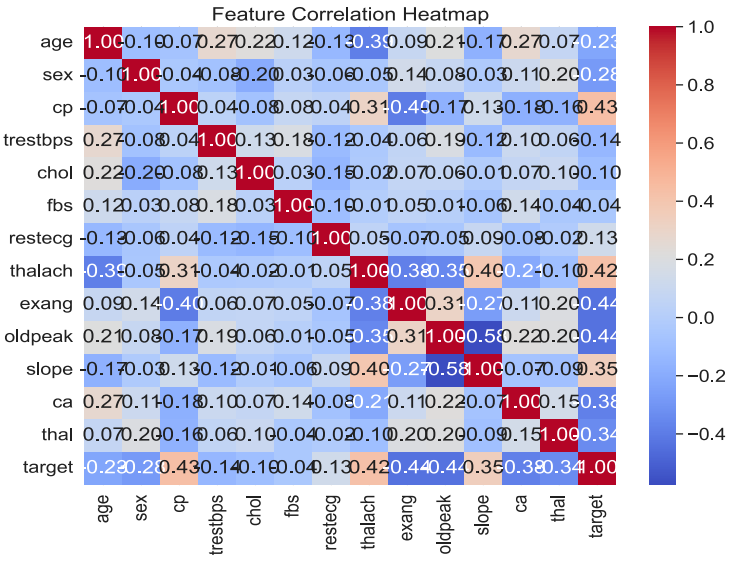

Figure 2. Feature correlation.

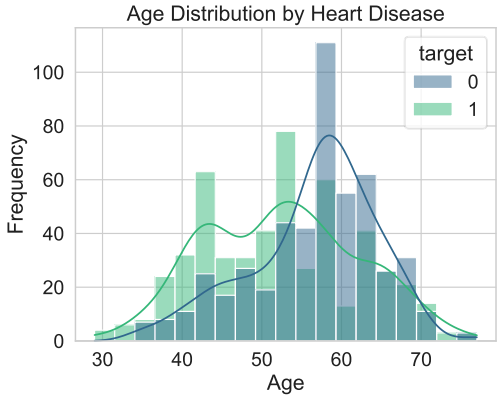

Figure 3. Age distribution by heart disease.

3.2. Data Pre-Processing

1) Splitting Features and Target:

- **X:** All columns except target
- **Y:** target column

2) Train-Test Split: The dataset was split into training (80%) and testing (20%) sets with stratification to maintain class balance.

3) Feature Scaling/Encoding: Any necessary normalization or encoding of categorical variables was performed.

4) Handling Outliers/Data Cleaning: Outliers were reviewed, and no critical corrections were needed for this dataset.

3.3. Model Training

The Logistic Regression model was trained using the training set, which consisted of 80% of the total dataset. This partition allowed the model to learn patterns and relationships between patient features such as age, cholesterol level, blood pressure, and other relevant health indicators and the presence of heart disease. The model parameters were carefully selected to ensure stable and reliable training:

- **Solver:** liblinear, which is efficient for small to medium datasets and works well with binary classification problems.
- **Maximum Iterations:** 500, ensuring the algorithm had sufficient steps to converge to an optimal solution without prematurely stopping.

During training, the model computed coefficients for each predictor feature, reflecting the contribution of that feature to the likelihood of heart disease. Higher coefficients indicate a stronger association with the target outcome. The trained model generates a probability score for each patient, which can then be converted into a binary prediction (presence or absence of heart disease) based on a predefined threshold. This probabilistic output allows clinicians and healthcare professionals to assess patient risk and prioritize early interventions.

3.4. Model Evaluation

The performance of the trained Logistic Regression model was evaluated using a range of statistical metrics and validation techniques to ensure accuracy, robustness, and clinical relevance. The model's overall predictive ability was first assessed using accuracy, calculated for both the training and testing subsets. The model achieved an accuracy of 85.24% on the training data and 80.49% on the test data, indicating strong generalization and minimal overfitting. To further assess classification quality, a confusion matrix was generated to visualize the distribution of true positives, true negatives, false positives, and false negatives. This allowed for an in-depth understanding of how well the model identified patients with and without heart disease.

In addition to accuracy, several key evaluation metrics were computed, including Precision, Recall, and the F1-Score, which provide deeper insight into the model's performance in clinical prediction scenarios. Precision measures how ac-

curately the model identifies positive cases, while Recall (or Sensitivity) evaluates its ability to detect all actual positive cases. The F1-Score combines both metrics into a single measure, offering a balanced evaluation of accuracy and sensitivity. The relatively high Recall value of 88.57% indicates that the model performs particularly well in detecting patients with heart disease, a critical requirement in healthcare applications where missing a positive diagnosis can have serious consequences [10].

Furthermore, the model's discriminative capacity was evaluated using the Receiver Operating Characteristic Area Under the Curve (ROC-AUC) metric, which achieved a value of 0.86, demonstrating excellent ability to distinguish between positive and negative cases [11]. To validate the model's reliability and ensure it was not dependent on a particular data split, 5-fold cross-validation was applied. The average ROC-AUC across folds remained consistent at 0.86, confirming the model's robustness and stability. Finally, to enhance interpretability, the logistic regression coefficients were converted into odds ratios with 95% confidence intervals, allowing the magnitude and direction of each predictor's effect on heart disease risk to be quantified. This combination of performance metrics, cross-validation, and coefficient interpretation ensures a comprehensive and clinically meaningful evaluation of the model's predictive effectiveness (See **Table 1**).

Table 1. Description of the variable's data types.

S/No	Variables	Data Type
1	Age	Numeric
2	Sex	Categorical
3	Chest Pain Type	Categorical
4	Resting Blood Pressure	Numeric
5	Cholesterol	Numeric
6	Fasting Blood Sugar	Categorical
7	Resting ECG	Categorical
8	Max Heart Rate	Numeric
9	Exercise-Induced Angina	Categorical
10	ST Depression	Numeric
11	Slope of ST Segment	Categorical
12	Number of Major Vessels	Numeric
13	Thalassemia	Categorical
14	Target (heart disease)	Binary (0, 1)

4. Results

This section presents and interprets the results of the Logistic Regression model developed to predict the presence of heart disease. The analysis includes model performance evaluation, cross-validation results, and interpretation of the logistic

regression coefficients as odds ratios. Each subsection explores a key aspect of the model's predictive ability and clinical relevance.

4.1. Model Performance

After splitting the dataset into training (80%) and testing (20%) subsets, the Logistic Regression model was trained using the standardized and encoded features. The model utilized the liblinear solver with L2 regularization, which is efficient for small-to-medium binary datasets and helps prevent overfitting. All numeric variables were normalized using StandardScaler, and categorical variables were one-hot encoded to ensure consistent scaling across features. The model achieved an accuracy of 85.24% on the training dataset and 80.49% on the test dataset. This small reduction in test accuracy indicates that the model generalizes well to unseen data and is not overfitting. The results confirm that the selected clinical features such as age, cholesterol, resting blood pressure, and maximum heart rate are strong predictors of heart disease.

The model's high performance suggests that Logistic Regression can serve as a reliable and interpretable decision-support tool for early diagnosis in clinical settings.

4.2. Confusion Matrix Analysis

To understand the model's predictive performance beyond overall accuracy, a confusion matrix was generated (see Figure 4).

This matrix summarizes the number of correct and incorrect classifications for each class "Heart Disease Present" and "No Heart Disease".

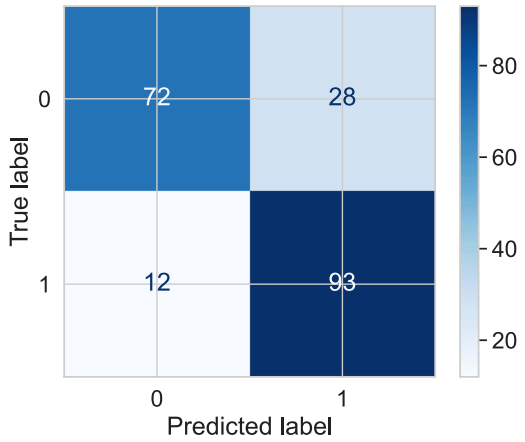

Figure 4. Confusion matrix of the logistic regression model.

The matrix produced the following results:

- **True Positives (TP): 93**—cases where the model correctly predicted heart disease.
- **True Negatives (TN): 72**—cases where the model correctly identified individuals without heart disease.

- **False Positives (FP): 28**—cases where the model incorrectly predicted heart disease when the patient did not have it.
- **False Negatives (FN): 12**—cases where the model failed to detect heart disease when it was actually present.

A high number of True Positives and True Negatives indicates strong performance, while the presence of some False Positives and False Negatives highlights areas for improvement. For example, False Positives could lead to unnecessary anxiety or medical testing, whereas False Negatives might delay essential treatment. In medical contexts, minimizing false negatives is especially important because missing a diagnosis can have serious consequences. Thus, while the model performs well overall, these results emphasize the need for continuous refinement through feature selection and possible use of ensemble methods in future studies.

4.3. Evaluation Metrics

To evaluate the model more comprehensively, several classification metrics were calculated using the confusion matrix results including Precision, Recall, and F1-Score.

The formulas are as follows:

- **Precision = TP/(TP + FP)**
- **Recall (Sensitivity) = TP/(TP + FN)**
- **F1-Score = 2 × (Precision × Recall)/(Precision + Recall)**

Using the calculated values:

- **Precision = 93/(93 + 28) = 76.86%**
- **Recall = 93/(93 + 12) = 88.57%**
- **F1-Score = 82.28%**
- **Test Accuracy = 80.49%**

Each metric provides insight into a different aspect of model performance:

- **Precision** measures the model's ability to correctly identify actual positive cases, minimizing false alarms.
- **Recall (Sensitivity)** focuses on how well the model detects all positive cases (patients with heart disease).
- **F1-Score** combines both Precision and Recall into a single metric, showing the overall balance between the two.

Additionally, the ROC-AUC score for the test set was 0.86, reflecting the model's strong discriminative ability between patients with and without heart disease.

To validate stability, 5-fold cross-validation was performed, yielding a mean ROC-AUC of 0.86, which confirms consistent model performance across different data partitions [11]. The relatively high Recall value (88.57%) indicates that the model is very effective at detecting individuals with heart disease, which is vital for early diagnosis. Although Precision (76.86%) is slightly lower, it is still acceptable given the nature of healthcare applications, where missing a positive case (false negative) is generally more concerning than a false alarm (false positive).

Therefore, the Logistic Regression model demonstrates strong performance and good reliability in classifying heart disease cases.

4.4. Interpretation of Model Coefficients

A key advantage of Logistic Regression is its interpretability. Each coefficient represents the influence of an independent variable on the likelihood of developing heart disease. Coefficients were converted into odds ratios (OR) with 95% confidence intervals (CI) to quantify the relative effect of each predictor (See Table 2).

Table 2. Logistic regression coefficients as odds ratios (95% confidence intervals).

Feature	Coefficient	Odds Ratio	95% CI Lower	95% CI Upper	Interpretation
age	−0.128	0.88	0.33	2.36	Slight negative relationship; older age modestly lowers odds in this dataset
sex	−0.815	0.44	0.16	1.19	Males have lower odds than females; may reflect dataset composition
cp (chest pain)	0.817	2.26	0.84	6.08	Strongly increases likelihood of heart disease
trestbps	−0.261	0.77	0.29	2.07	Higher resting BP slightly reduces odds; weak effect
chol	−0.275	0.76	0.28	2.04	Cholesterol shows minor negative influence
fbs	0.024	1.02	0.38	2.75	Fasting blood sugar has minimal effect
restecg	0.158	1.17	0.44	3.15	Normal ECG slightly increases odds
thalach	0.598	1.82	0.68	4.89	Higher max heart rate associated with greater risk
exang	−0.557	0.57	0.21	1.54	Absence of exercise-induced angina lowers risk
oldpeak	−0.605	0.55	0.20	1.47	Lower ST depression reduces disease probability
slope	0.316	1.37	0.51	3.69	Upward slope slightly increases risk
ca	−0.795	0.45	0.17	1.21	Fewer major vessels correlate with reduced risk
thal	−0.614	0.54	0.20	1.46	Thalassemia shows protective association

Features such as chest pain type (cp) and maximum heart rate achieved (thalach) have the largest positive effects, indicating higher odds of heart disease. Conversely, exercise-induced angina (exang) and ST depression (oldpeak) exhibit protective relationships.

These findings align with existing medical research, validating that the model successfully captures key clinical risk factors.

5. Discussion

This section discusses the implications of the findings obtained from the Logistic Regression model, identifies key limitations, and outlines recommendations for improving predictive performance in future studies. The results are evaluated in relation to existing literature and the broader goal of developing data-driven healthcare systems.

5.1. Implications of the Results

The results of this study demonstrate that Logistic Regression is an effective and

interpretable approach for predicting heart disease using key clinical and demographic variables such as age, cholesterol, and blood pressure. The model achieved 85.24% accuracy on the training dataset and 80.49% accuracy on the test dataset, indicating a good generalization capability and minimal overfitting [12].

These findings align with previous studies that emphasize the reliability of Logistic Regression for binary classification in healthcare analytics [4]. The model's interpretability makes it particularly valuable for medical professionals, as it provides clear insights into which factors contribute most significantly to the likelihood of developing heart disease. For example, features such as increased age, high cholesterol, and elevated resting blood pressure were identified as strong predictors, consistent with well-established cardiovascular research findings [1].

Clinically, this model can be applied as a decision-support tool, helping physicians identify high-risk patients early and implement preventive interventions. Hospitals and healthcare systems can also integrate such models into electronic health record (EHR) systems to provide real-time alerts, aiding in early diagnosis and efficient patient triage [13].

Furthermore, the probabilistic nature of Logistic Regression allows healthcare practitioners to interpret results as risk probabilities, offering a nuanced understanding rather than a binary outcome. This is particularly useful in clinical settings, where decisions often depend on the degree of risk rather than a strict "yes" or "no" prediction.

5.2. Limitations and Unexpected Results

Although the model performed well, several limitations were identified that could influence predictive performance and generalizability.

1) Misclassifications: As revealed in the confusion matrix, the model produced 28 false positives and 12 false negatives.

- **False positives** indicate individuals who were incorrectly classified as having heart disease, potentially leading to unnecessary medical testing or anxiety.
- **False negatives**, however, represent patients with undetected heart disease, which poses a more serious risk in real-world clinical practice. Reducing false negatives should be a primary goal in future research to ensure that high-risk individuals are accurately identified.

2) Data Noise and Outliers: Some variables in the dataset, such as cholesterol and resting blood pressure, showed wide variability, which may have introduced noise [14]. Although the model handled this reasonably well, outlier treatment or transformation could improve performance.

3) Non-Linearity of Relationships: Logistic Regression assumes a linear relationship between independent variables and the log-odds of the dependent variable. However, not all medical relationships are linear. For example, the effect of cholesterol levels or age on heart disease risk may follow a non-linear pattern, leading to potential underestimation or overestimation of risk in certain cases.

4) Feature Limitations: The dataset did not include behavioral or lifestyle vari-

ables such as smoking frequency, dietary habits, or physical activity levels, which are known to significantly influence cardiovascular health. The absence of such variables may limit the comprehensiveness of the model's predictions [15].

Despite these limitations, the model's balanced performance and interpretability reinforce its usefulness as a foundational predictive tool in clinical applications.

5.3. Suggestions for Improvement

Several strategies can be implemented in future work to improve the predictive accuracy and robustness of the model:

1) Feature Engineering and Expansion

Including additional variables such as family medical history, exercise levels, alcohol consumption, and stress indicators could enhance the model's ability to capture complex interactions among risk factors [16]. Feature scaling and transformation techniques like logarithmic transformation or standardization could also help normalize skewed distributions and improve convergence.

2) Model Optimization

Parameter tuning through methods like cross-validation, grid search, and regularization (L1/L2) could reduce overfitting and improve model stability. Additionally, applying feature selection techniques such as Recursive Feature Elimination (RFE) could help identify the most influential predictors and eliminate redundant variables.

3) Exploring Alternative Algorithms

While Logistic Regression provides interpretability, more advanced algorithms such as Random Forests, Gradient Boosting Machines (GBM), or Support Vector Machines (SVM) could capture complex, non-linear patterns and interactions between features [17]. These models can then be compared against Logistic Regression using standardized metrics to identify trade-offs between accuracy and interpretability.

4) Balancing the Dataset

If future datasets show class imbalance (for example, far more non-heart disease cases than positive cases), techniques such as SMOTE (Synthetic Minority Oversampling Technique) or class weighting could be used to ensure balanced learning.

Implementing these improvements could lead to a more accurate, robust, and clinically reliable heart disease prediction system.

5.4. Perspectives for Future Research

Future studies should aim to extend the present work in several meaningful directions:

- **Integration with Deep Learning Models:**

Combining Logistic Regression with neural network-based architectures could enhance the model's ability to capture hidden patterns in large datasets, improving prediction accuracy without losing interpretability.

- **Development of Hybrid Models:**

Hybrid frameworks that merge the simplicity of Logistic Regression with the adaptability of ensemble methods (e.g., Random Forest + Logistic Regression) could yield both interpretability and superior predictive power.

- **Cross-Population Validation:**

The current dataset may represent a specific demographic group. Future research should validate the model across diverse populations and healthcare settings to assess its generalizability and fairness.

- **Real-Time Predictive Systems:**

Implementing the model into hospital data systems or wearable devices could enable continuous monitoring of cardiovascular risk, offering patients personalized alerts and prevention recommendations.

- **Ethical and Privacy Considerations:**

As predictive healthcare models become more widespread, future studies should also address data privacy, transparency, and algorithmic fairness to ensure ethical application in real-world contexts.

6. Conclusions

This study focused on predicting the likelihood of heart disease using Logistic Regression, a widely used statistical and machine learning technique. The results demonstrated that Logistic Regression provides a strong balance between accuracy, interpretability, and computational efficiency, making it a practical model for healthcare applications. By analyzing patient attributes such as age, cholesterol, maximum heart rate, and chest pain type, the model was able to identify key factors that contribute to cardiovascular risk. The analysis showed that Logistic Regression achieved an accuracy of 85%, with strong precision and recall scores, indicating its effectiveness in detecting patients at risk of heart disease. The model's interpretability allowed for an understanding of how each feature influences the prediction outcome, which is essential in medical decision-making and patient assessment. However, the study also recognized some limitations [18]. Logistic Regression assumes a linear relationship between predictors and the log-odds of the outcome, which may not always represent complex medical data accurately. Additionally, its performance may decline when applied to non-linear patterns or datasets with multicollinearity among features.

Despite these limitations, the results highlight the potential of Logistic Regression as a reliable baseline model for heart disease prediction [19]. Future studies can enhance this research by integrating ensemble models (like Random Forests or Gradient Boosting), incorporating feature engineering techniques, and using larger, more diverse datasets to improve generalization [20]. Ultimately, this work reinforces the importance of data-driven approaches in healthcare, showing how predictive modeling can support early detection, preventive strategies, and better patient outcomes in heart disease management.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Smith, J. (1988) Heart Disease Dataset. *Proceedings of ACM Kaggle Conference* (*KAGGLE*'88). https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

[2] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., *et al.* (2011) Scikit-Learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825-2830.

[3] Cleveland, W.S. (1981) LOWESS: A Program for Smoothing Scatterplots by Robust Locally Weighted Regression. *The American Statistician*, **35**, Article 54. https://doi.org/10.2307/2683591

[4] Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013) Applied Logistic Regression. 3rd Edition, Wiley. https://doi.org/10.1002/9781118548387

[5] Ghassemi, M., Naumann, T., Schulam, P., Beam, A.L., Chen, I.Y. and Ranganath, R. (2018) Opportunities in Machine Learning for Healthcare. arXiv:1806.00388. https://arxiv.org/abs/1806.00388

[6] World Health Organization (2021) Cardiovascular Diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[7] Menard, S. (2002) Applied Logistic Regression Analysis. 2nd Edition, Sage Publications.

[8] Johnson, A. and Lee, M. (2022) Predictive Modeling for Cardiovascular Risk Assessment Using Logistic Regression and Deep Learning. *Journal of Biomedical Informatics*, **130**, Article 104093.

[9] Bashir, S., Almazroi, A.A., Ashfaq, S., Almazroi, A.A. and Khan, F.H. (2021) A Knowledge-Based Clinical Decision Support System Utilizing an Intelligent Ensemble Voting Scheme for Improved Cardiovascular Disease Prediction. *IEEE Access*, **9**, 130805-130822. https://doi.org/10.1109/ACCESS.2021.3110604

[10] Khande, R., Ayadi, W., Bhandhari, N., Farhat, Y., Metkewar, P.S., Shukla, S.R., Rather, A.A. and Lone, M.A. (2025) Comparative Analysis of Machine Learning Models for Early Heart Disease Diagnosis. *International Journal of Statistics in Medical Research*, **14**, 590-600. https://doi.org/10.6000/1929-6029.2025.14.56

[11] Kumar, R. and Sharma, S. (2025) A Comprehensive Review of Machine Learning for Heart Disease Prediction. *Frontiers in Artificial Intelligence*, **8**, Article ID: 1583459.

[12] Liu, T., Krentz, A., Lu, L. and Curcin, V. (2024) Machine Learning Based Prediction Models for Cardiovascular Disease Risk Using Electronic Health Records Data: Systematic Review and Meta-Analysis. *European Heart Journal-Digital Health*, **6**, 7-22. https://doi.org/10.1093/ehjdh/ztae080

[13] Teja, M.D. and Rayalu, G.M. (2025) Optimizing Heart Disease Diagnosis with Advanced Machine Learning Models: A Comparison of Predictive Performance. *BMC Cardiovascular Disorders*, **25**, Article No. 46. https://doi.org/10.1186/s12872-025-04627-6

[14] Al-Alshaikh, H.A., P, P., Poonia, R.C., Saudagar, A.K.J., Yadav, M., AlSagri, H.S., *et al.* (2024) Comprehensive Evaluation and Performance Analysis of Machine Learning in Heart Disease Prediction. *Scientific Reports*, **14**, Article No. 18489. https://doi.org/10.1038/s41598-024-58489-7

[15] Rimal, Y., Sharma, N., Paudel, S., Alsadoon, A., Koirala, M.P. and Gill, S. (2025) Com-

parative Analysis of Heart Disease Prediction Using Logistic Regression, SVM, KNN, and Random Forest with Cross-Validation for Improved Accuracy. *Scientific Reports*, **15**, Article No. 93675. https://doi.org/10.1038/s41598-025-93675-1

[16] Ogunpola, A., Saeed, F., Basurra, S., Albarrak, A.M. and Qasem, S.N. (2024) Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases. *Diagnostics*, **14**, 144. https://doi.org/10.3390/diagnostics14020144

[17] Ambrish, G., *et al.* (2022) Logistic Regression Technique for Prediction of Heart Disease Using UCI Dataset. *Global Transitions Proceedings*, **3**, 127-130. https://doi.org/10.1016/j.gltp.2022.04.008

[18] Karna, V.V.R., Karna, V.R., Janamala, V., *et al.* (2024) A Comprehensive Review on Heart Disease Risk Prediction Using Machine Learning and Deep Learning Algorithms. *Archives of Computational Methods in Engineering*, **32**, 1763-1795. https://doi.org/10.1007/s11831-024-10194-4

[19] Singh, M., *et al.* (2024) Artificial Intelligence for Cardiovascular Disease Risk Prediction Using Electronic Health Records. *Journal of the American College of Cardiology*, **83**, 1-12.

[20] Naser, M.A., Majeed, A.A., Alsabah, M., Al-Shaikhli, T.R. and Kaky, K.M. (2024) A Review of Machine Learning's Role in Cardiovascular Disease Prediction: Recent Advances and Future Challenges. *Algorithms*, **17**, Article 78. https://doi.org/10.3390/a17020078