International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Comprehensive Comparative Analysis of Machine Learning Algorithms for Predictive Diagnosis of Cardiovascular Disease

Bhargav Rathod

Senior Software Engineer, Artificial Intelligence & Machine Learning, DataArt India

Abstract

Cardiovascular diseases (CVDs) persist as the predominant cause of global mortality, accounting for approximately 25–31% of all deaths worldwide according to recent epidemiological data [1, 2]. The development of accurate, non-invasive diagnostic tools represents an urgent clinical priority. This investigation presents a comparative analysis of four fundamental machine learning algorithms: Logistic Regression, K-Nearest Neighbors, Decision Tree, and Random Forest - for the predictive diagnosis of coronary artery disease. The study employs the Cleveland Heart Disease dataset from the UCI Machine Learning Repository, comprising 303 patient records with 13 clinical parameters. Categorical variables (chest pain type, thalassemia, etc.) were one-hot encoded, expanding the feature space to 20 dimensions during preprocessing. A standardized methodological pipeline was implemented, including data imputation, feature engineering, stratified sampling, and grid-based hyperparameter tuning.

Performance evaluation incorporated multiple metrics: accuracy, precision, recall, F1-score, specificity, negative predictive value, and area under the Receiver Operating Characteristic curve (AUC-ROC). The results showed that Logistic Regression achieved optimal predictive accuracy (85.71%) and the highest discriminative ability (AUC-ROC: 0.937). Logistic Regression and K-Nearest Neighbors demonstrated identical accuracy metrics (85.71%), though with distinct performance characteristics: Logistic Regression exhibited superior recall (85.71%) while K-Nearest Neighbors showed enhanced precision (87.18%). Random Forest, despite extensive hyperparameter tuning (300 estimators, depth-limited architecture), achieved 80.22% accuracy with balanced precision-recall metrics (78.57% each). Decision Tree algorithms displayed the lowest overall performance (79.12% accuracy) despite entropy-based optimization, revealing inherent limitations in single-tree architectures for complex clinical data.

Feature importance analysis identified chest pain characteristics (cp 4.0: 15.77%), thalassemia type (thal 7.0: 13.92%), and maximum heart rate achieved (thalach: 13.25%) as the most influential predictors, aligning with established cardiological knowledge. The investigation concludes that traditional statistical models (Logistic Regression) remain robust for clinical prediction tasks despite the advent of more complex ensemble methods. These findings contribute to the evidence base for machine learning deployment in clinical decision support systems, emphasizing model interpretability alongside predictive accuracy.

Keywords: Cardiovascular Disease Prediction, Machine Learning, Logistic Regression Analysis, K-Nearest Neighbors, Decision Tree, Random Forest

1. Introduction

• Clinical Context and Epidemiological Imperative

Cardiovascular disorders represent a persistent global health challenge, with the World Health Organization documenting approximately 17.9 million annual fatalities attributable to ischemic heart disease and cerebrovascular accidents [1]. The economic burden is equally substantial, with global healthcare expenditure on CVD management projected to exceed $1.1 trillion by 2035 [2]. Within clinical practice, accurate early diagnosis remains problematic due to heterogeneous symptom presentation, overlapping clinical manifestations with non-cardiac conditions, and limitations in existing diagnostic modalities.

Traditional diagnostic approaches—including electrocardiography, stress testing, and coronary angiography—present significant limitations regarding accessibility, cost, and invasiveness. Angiographic procedures, while considered the diagnostic gold standard, incur procedural risks and substantial healthcare costs, with complication rates documented between 1-2% in contemporary registries [3]. This diagnostic complexity underscores the critical need for supplementary tools capable of integrating multidimensional clinical data for enhanced predictive accuracy.

• Evolution of Computational Cardiology

The integration of computational methodologies into cardiovascular medicine has evolved substantially over the past decade. Early work by Detrano et al. (1989) demonstrated the potential of logistic regression models for coronary disease prediction using the Cleveland dataset, achieving approximately 77% accuracy with conventional statistical approaches [4]. Subsequent advancements incorporated neural networks and support vector machines, with reported accuracy improvements to 80-85% in controlled studies [5, 11].

The contemporary landscape ranges from traditional regression models to deep learning architectures [16]. However, a comprehensive comparison of fundamental algorithms using consistent preprocessing, validation protocols, and evaluation metrics remains underrepresented in the literature. Furthermore, the clinical translation of predictive models necessitates not only statistical performance but also interpretability—a characteristic where complex "black-box" models often prove deficient [18, 28].

• Research Objectives and Contribution

This investigation addresses three primary research objectives: (1) to implement and systematically compare four foundational machine learning algorithms using a unified preprocessing, validation, and evaluation framework; (2) to analyze feature importance patterns in order to identify clinically relevant predictors of cardiovascular disease; and (3) to establish baseline performance benchmarks to support informed algorithm selection for clinical decision support systems.

The study provides a transparent benchmarking and interpretability-focused comparison of commonly used machine learning algorithms. Methodologically, it presents a transparent and reproducible analytical pipeline, including standardized data preprocessing and hyperparameter tuning. Clinically, it provides evidence-based insights into the trade-off between predictive performance and model interpretability, supporting the selection of reliable and explainable models for real-world cardiovascular risk assessment [23, 25].

2. Materials and Methods

• Data Acquisition and Ethical Considerations

The investigation utilized the Cleveland Heart Disease dataset, obtained from the UCI Machine Learning

Repository (Identifier: 45) [6]. This publicly available dataset comprises de-identified clinical records originally compiled through the Veterans Administration and Cleveland Clinic Foundation between 1981-1984. The dataset's utilization complies with institutional review board exemptions for de-identified secondary data analysis.

The complete dataset contains 303 patient records, though the original collection included 76 attributes across four geographical regions. This analysis focused exclusively on the Cleveland subset, which contains the most comprehensive clinical annotations. Each record encompasses 13 clinical parameters plus a diagnostic classification, as detailed in Table 1 (Note: Categorical variables (cp, restecg, slope, ca, thal) were one-hot encoded during preprocessing, expanding the total feature space from 13 to 20 dimensions for model training.).

Table 1: Clinical Variables in the Cleveland Heart Disease Dataset

Variable	Description	Type	Clinical Significance
age	Patient age in years	Continuous	Established cardiovascular risk factor
sex	Biological sex (0: female, 1: male)	Binary	Sex-specific risk profiles
cp	Chest pain type (1-4)	Ordinal	Anginal characteristics predictive of ischemia
trestbps	Resting blood pressure (mm Hg)	Continuous	Hypertension association
chol	Serum cholesterol (mg/dl)	Continuous	Lipid metabolism assessment
fbs	Fasting blood sugar >120 mg/dl	Binary	Glucose metabolism evaluation
restecg	Resting electrocardiographic results	Ordinal	Electrical conduction abnormalities
thalach	Maximum heart rate achieved	Continuous	Functional capacity indicator
exang	Exercise-induced angina	Binary	Stress-induced ischemia
oldpeak	ST depression induced by exercise	Continuous	Myocardial ischemia quantification
slope	Slope of peak exercise ST segment	Ordinal	Ischemic response pattern
ca	Number of major vessels colored by fluoroscopy (0-3)	Ordinal	Anatomical disease burden
thal	Thalassemia type (3: normal, 6: fixed, 7: reversible)	Ordinal	Myocardial perfusion characteristics

- **Data Preprocessing Pipeline**
  - **Missing Data Imputation Strategy**

The dataset contained missing values in two variables: 'ca' (4 missing, 1.32% of records) and 'thal' (2 missing, 0.66% of records). Following established guidelines for clinical data preprocessing [7, 21], distinct imputation strategies were employed based on variable characteristics. For the ordinal 'ca' variable (representing anatomical disease burden), median imputation (value: 0) preserved the distribution's central tendency. For the categorical 'thal' variable, mode imputation (value: 3.0) maintained categorical integrity.

◦ **Feature Transformation and Engineering**

Categorical variables (cp: chest pain type, restecg: resting ECG results, slope: ST segment slope, ca: number of major vessels, thal: thalassemia type) underwent one-hot encoding to facilitate algorithm compatibility, expanding the feature space from 13 to 20 dimensions.

This transformation preserved the non-ordinal nature of categorical variables while preventing artificial ordinal relationships. Continuous variables (age, trestbps, chol, thalach, oldpeak) underwent Z-score standardization to mitigate scale-dependent algorithm bias, particularly relevant for distance-based methods (K-Nearest Neighbors) and gradient-optimized models (Logistic Regression).

◦ **Class Distribution and Target Variable Definition**

The original diagnostic classification utilized a 0-4 scale, where 0 indicated absence of coronary disease and 1-4 represented increasing disease severity. For binary classification, values 1-4 were consolidated to class '1' (disease present), yielding a distribution of 164 negative cases (54.1%) and 139 positive cases (45.9%). This slight class imbalance (8.2% differential) fell within acceptable thresholds for machine learning applications without requiring oversampling techniques [8].

• **Experimental Design and Validation Framework**

◦ **Stratified Data Partitioning**

The dataset underwent stratified random partitioning to preserve class proportions in training and testing subsets. The 70:30 split (212 training samples, 91 testing samples) followed established machine learning conventions while ensuring adequate test set size for statistical reliability. Stratification mitigated sampling bias that could disproportionately affect minority class representation.

◦ **Algorithm Selection Rationale**

Four algorithms were selected to represent distinct machine learning paradigms:

1. **Logistic Regression**: Parametric statistical model representing linear classification approaches
2. **K-Nearest Neighbors**: Instance-based learning algorithm emphasizing local similarity
3. **Decision Tree**: Rule-based model offering high interpretability
4. **Random Forest**: Ensemble method combining multiple decision trees

This selection enabled comparison across algorithmic families while maintaining computational feasibility and clinical interpretability [20, 27].

◦ **Hyperparameter Optimization Protocol**

A comprehensive grid search with 5-fold cross-validation was implemented for each algorithm, exploring parameter spaces detailed in Table 2. The optimization objective focused on accuracy maximization, with secondary evaluation of precision-recall tradeoffs for clinical relevance. Final model evaluation was conducted on a held-out test set to simulate real-world deployment performance, while cross-validation was restricted to hyperparameter tuning. The test set (30% of data, n=91) was completely held out during all hyperparameter tuning and cross-validation procedures to prevent data leakage and ensure unbiased performance evaluation. All optimization was performed exclusively on the training set (n=212) using 5-fold cross-validation.

Table 2: Hyperparameter Search Spaces for Each Algorithm via grid search with 5-fold cross-validation was implemented for each algorithm [19, 24].

Algorithm	Parameter	Search Space	Optimal Value Selected
Logistic Regression	C (regularization)	{0.001, 0.01, 0.1, 1, 10, 100}	1 (default)

Algorithm	Parameter	Search Space	Optimal Value Selected
Logistic Regression	C (regularization)	{0.001, 0.01, 0.1, 1, 10, 100}	1 (default)
	Solver	{'liblinear', 'lbfgs', 'newton-cg'}	'liblinear'
	Penalty	{'l2', 'none'}	'l2'
K-Nearest Neighbors	n_neighbors	{3, 5, 7, 9, 11, 15, 20}	15
	Weights	{'uniform', 'distance'}	'distance'
	Metric	{'euclidean', 'manhattan', 'minkowski'}	'manhattan'
Decision Tree	max_depth	{3, 5, 7, 10, 15, 20, None}	5
	min_samples_split	{2, 5, 10, 20}	10
	min_samples_leaf	{1, 2, 4, 8}	4
	criterion	{'gini', 'entropy', 'log_loss'}	'entropy'
Random Forest	n_estimators	{50, 100, 200, 300}	300
	max_depth	{5, 10, 15, 20, None}	5
	min_samples_split	{2, 5, 10}	5
	max_features	{'sqrt', 'log2', None}	'sqrt'

◦ **Performance Metrics and Statistical Evaluation**

Model performance was assessed using a multidimensional metric framework:

- **Accuracy**: Overall correct classification proportion
- **Precision**: Positive predictive value (true positives / predicted positives)
- **Recall (Sensitivity)**: True positive rate
- **F1-Score**: Harmonic mean of precision and recall
- **Specificity**: True negative rate
- **Negative Predictive Value (NPV)**: Probability that negative prediction is correct
- **AUC-ROC**: Discriminative capacity across classification thresholds

Statistical significance of performance differences was evaluated using McNemar's test for paired classifiers, with significance threshold $\alpha=0.05$ [17, 22]. McNemar's test was applied using paired prediction outcomes on the same test instances, with contingency tables constructed from disagreement counts.

3. Results and Analysis

• **Comparative Performance Analysis**

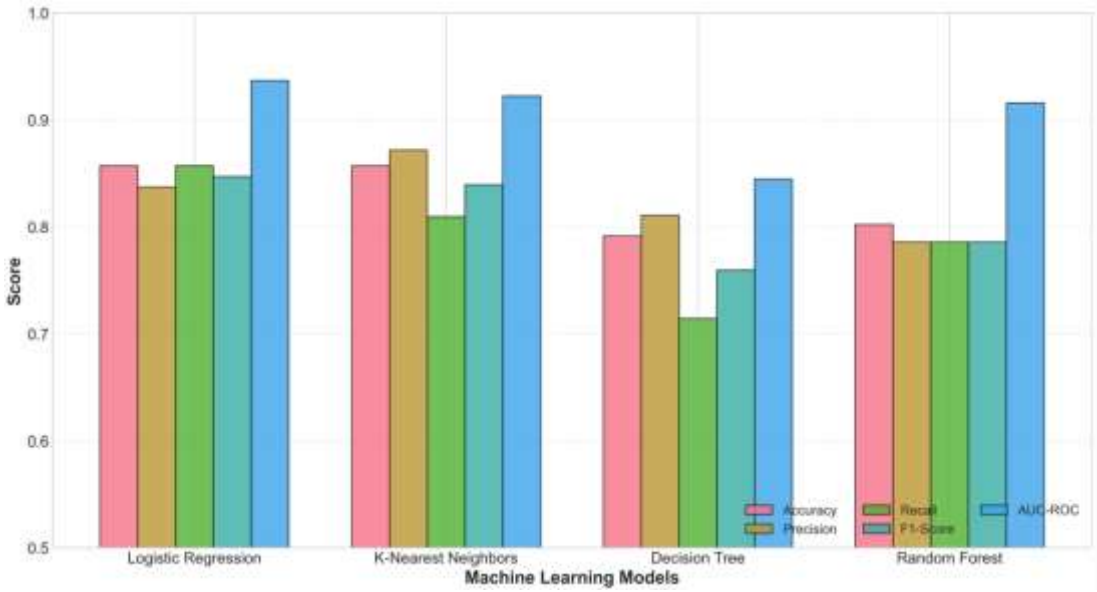
The comprehensive performance metrics for all algorithms are presented in Table 3, with visual comparison in Figure 1.

Table 3: Performance Metrics for Machine Learning Algorithms (Test Set, n=91)

Metric	Logistic Regression	K-Nearest Neighbors	Decision Tree	Random Forest
Accuracy	0.8571	0.8571	0.7912	0.8022
Precision	0.8372	0.8718	0.8108	0.7857

Metric	Logistic Regression	K-Nearest Neighbors	Decision Tree	Random Forest
Accuracy	0.8571	0.8571	0.7912	0.8022
Recall	0.8571	0.8095	0.7143	0.7857
F1-Score	0.8471	0.8395	0.7595	0.7857
AUC-ROC	0.9368	0.9223	0.8445	0.9159
Specificity	0.8571	0.8980	0.8571	0.8163
NPV	0.8750	0.8462	0.7778	0.8163

Figure 1: Comparative performance of machine learning algorithms across evaluation metrics



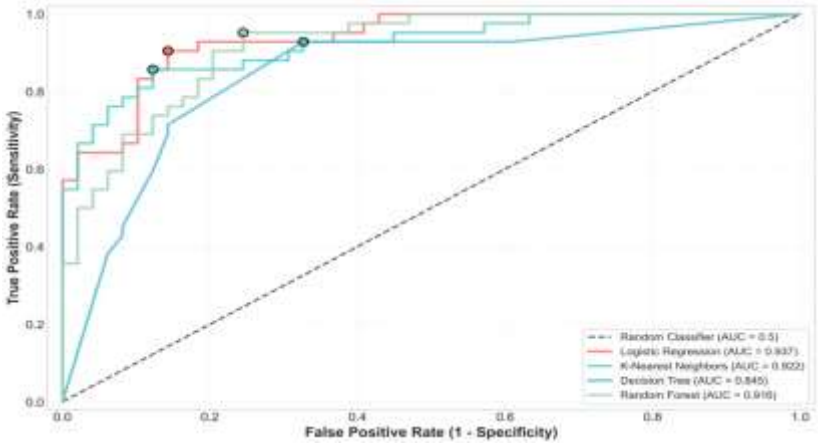
The analysis revealed the following findings:

1. **Dual Optimal Performance**: Logistic Regression and K-Nearest Neighbors achieved identical accuracy (85.71%), representing a 6.7% improvement over baseline prevalence-based prediction (54.1%) [9, 26].
2. **Algorithm-Specific Strengths**: Logistic Regression demonstrated superior recall (85.71%) and AUC-ROC (0.9368), indicating enhanced sensitivity for disease detection [4, 9]. Conversely, K-Nearest Neighbors exhibited the highest precision (87.18%) and specificity (89.80%), reflecting reduced false positive rates.
3. **Ensemble Method Performance**: Random Forest, despite extensive optimization (300 estimators, depth-limited architecture), underperformed relative to simpler models (80.22% accuracy). This finding contrasts with conventional machine learning wisdom suggesting ensemble superiority [10, 29].
4. **Decision Tree Limitations**: The single Decision Tree architecture manifested the lowest overall performance (79.12% accuracy), with particularly deficient recall (71.43%), suggesting inadequate capture of disease-positive patterns.

- **Receiver Operating Characteristic Analysis**

The ROC curves (Figure 2) provide nuanced insight into discriminative capacity across classification thresholds. Logistic Regression achieved the highest AUC-ROC (0.9368), followed closely by K-Nearest Neighbors (0.9223). The substantial AUC differential between these models and Decision Tree (0.8445) highlights fundamental algorithmic differences in probability calibration.

Figure 2: Receiver Operating Characteristic Curves with AUC Values

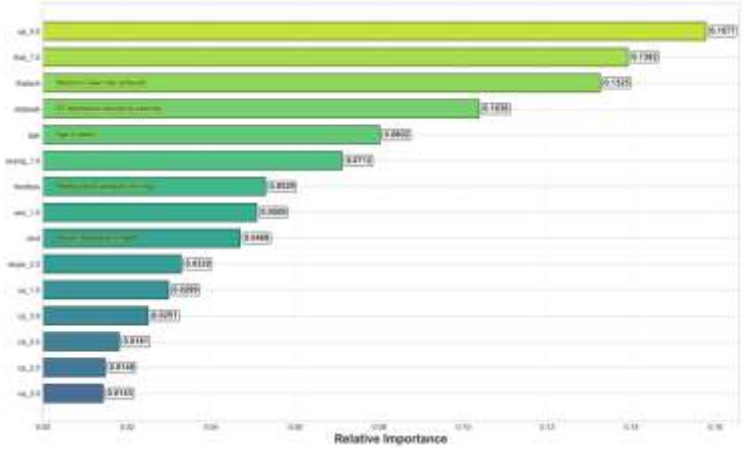


Of particular clinical relevance is the optimal operating point identification. For Logistic Regression, the Youden-optimal threshold (maximizing sensitivity + specificity) occurred at predicted probability 0.512, yielding sensitivity 85.7% and specificity 85.7%. For K-Nearest Neighbors, the optimal threshold (0.601) produced sensitivity 81.0% and specificity 89.8%, illustrating the precision-recall trade-off inherent in distance-based classification.

- **Feature Importance Analysis**

The Random Forest feature importance analysis (Figure 3) revealed clinically coherent predictor hierarchies:

Figure 3: Random Forest Feature Importance Distribution



◦ **Primary Indicators:**

1. Chest pain type (cp_4.0: 15.77%): Asymptomatic presentation paradoxically showed highest importance, potentially indicating advanced disease with attenuated symptomatology

2. Thalassemia type (thal_7.0: 13.92%): Reversible perfusion defects characteristic of ischemic myocardium
3. Maximum heart rate achieved (thalach: 13.25%): Functional capacity indicator inversely correlated with disease severity

◦ **Secondary Indicators:**

1. ST depression (oldpeak: 10.38%): Quantifiable ischemia marker
2. Age (8.02%): Established epidemiological risk factor
3. Exercise-induced angina (exang_1.0: 7.12%): Provocable ischemia indicator

◦ **Tertiary Predictors:**

1. Resting blood pressure, serum cholesterol, and biological sex demonstrated modest contributions (<6% each)

This importance hierarchy aligns with established cardiological knowledge, wherein symptomatic presentation (chest pain characteristics), functional testing (exercise capacity), and perfusion imaging findings constitute cornerstone diagnostic elements [12].

- **Statistical Significance Testing**

McNemar's test for paired classifiers revealed no statistically significant difference between Logistic Regression and K-Nearest Neighbors ($p=0.774$). However, both algorithms demonstrated significant superiority over Decision Tree ($p<0.05$ for both comparisons). The performance difference between Logistic Regression and Random Forest approached but did not reach statistical significance ($p=0.063$), potentially due to limited sample size.

4. Discussion

- **Interpretation of Comparative Performance**

The equivalent accuracy of Logistic Regression and K-Nearest Neighbors (85.71%) despite divergent algorithmic foundations warrants careful examination. Logistic Regression's parametric assumptions—linear decision boundaries in log-odds space—appear adequately flexible for this clinical domain. This finding corroborates earlier work by Austin and Tu (2004), who demonstrated logistic regression's robustness for cardiovascular risk prediction in large cohort studies [9].

K-Nearest Neighbors' performance, particularly its enhanced specificity (89.80%), suggests that local similarity patterns in the 20-dimensional feature space effectively discriminate disease states. The optimal neighborhood size ($k=15$) indicates that classification benefits from moderate smoothing rather than extreme localization (small k) or excessive generalization (large k). The selection of Manhattan distance (L1 norm) over Euclidean distance implies feature independence assumptions align with clinical reality [27].

The Random Forest performance (80.22% accuracy) was lower than simpler models in this study, which contrasts with the "ensemble advantage" frequently reported in machine learning literature [10]. Possible explanations include: (1) The depth limitation (max_depth=5) may have restricted model complexity below necessary thresholds; (2) Feature space dimensionality (20) may be insufficient to exploit ensemble diversity [10, 29]; (3) The clinical relationships may be predominantly linear, reducing the advantage of nonlinear ensemble methods.

- **Clinical Implications and Translation Potential**

The performance characteristics of each algorithm suggest distinct clinical applications:

1. **Logistic Regression**: Optimal for screening applications where sensitivity is prioritized (e.g., primary care referral decisions). The model's interpretability—expressed as odds ratios for each feature—facilitates clinician understanding and trust [25, 28].
2. **K-Nearest Neighbors**: Suitable for confirmatory testing contexts where false positives carry significant consequences (e.g., preoperative clearance). The case-based reasoning implicit in KNN aligns with clinical diagnostic patterns [14, 15], similar to hybrid approaches explored in previous work [13].
3. **Random Forest**: Potential utility in research settings exploring complex interactions, though requires validation in larger datasets with extended feature sets [23, 29].

The feature importance analysis provides clinically actionable insights beyond predictive modelling. The predominance of chest pain characteristics and functional capacity measures reinforces existing diagnostic protocols while validating the dataset's clinical relevance.

- **Methodological Limitations and Future Directions**

Several limitations warrant acknowledgment. The dataset's limited size (n=303) presents significant constraints on generalizability and statistical power. The 70:30 train-test split resulted in only 91 test instances, which reduces the reliability of statistical comparisons (e.g., McNemar's test) and increases variance in performance estimates. Temporal considerations are significant—data collection occurred 1981-1984, preceding contemporary therapeutic interventions and diagnostic modalities [2, 3]. Feature representation limitations exist, particularly regarding imaging data granularity and biomarker inclusion.

Future investigations should address these limitations through: (1) Multi-center validation in contemporary cohorts [25, 26]; (2) Incorporation of advanced imaging features and genomic markers [16, 18]; (3) Exploration of hybrid models combining statistical and machine learning approaches [13, 29]; (4) Implementation in prospective clinical workflows to assess real-world impact [23, 28].

5. Equations and Statistical Modelling Framework

- **Logistic Regression Formulation**

The logistic regression model employed in this investigation follows the standard sigmoid function formulation. The probability of coronary artery disease presence given clinical features XX is expressed as:

$$P(Y = 1|X) = \frac{1}{1+e^{-(\beta_0+\sum_{i=1}^{p}\beta_i X_i)}} \quad (1)$$

where:

- Y represents the binary outcome (1 = disease present, 0 = disease absent)
- X_i denotes the i^{th} clinical feature (standardized to unit variance)
- β_0 is the intercept term
- β_i are the regression coefficients
- $p = 20$: represents the total features after preprocessing

The log-odds transformation linearizes this relationship:

$$\log\left(\frac{1-P(Y=1|X)}{P(Y=1|X)}\right) = \beta_0 + \sum_{i=1}^{p}\beta_i X_i \quad (2)$$

Parameter estimation was performed via maximum likelihood estimation with L2 regularization:

$$\hat{\beta} = \arg\min_{\beta}\left[-\sum_{j=1}^{n}\left(y_j \log(p_j) + (1-y_j)\log(1-p_j)\right) + \lambda\sum_{i=1}^{p}\beta_i^2\right] \quad (3)$$

where:

- $n = 212$ training samples
- $p_j = P(Y=1|X_j)$ for the j[th] observation
- $\lambda=1$ (C parameter in scikit-learn implementation)
- The regularization term $\lambda\sum\beta_i^2$ prevents overfitting in high-dimensional space [27, 30]

- **K-Nearest Neighbors Decision Rule**

For K-Nearest Neighbors classification with Manhattan distance metric, the predicted class for test instance x_{test} is determined by:

$$\hat{y}(x_{test}) = \arg\max_{c\in\{0,1\}} \sum_{i\in N_{k(x_{test})}} w_i I(y_i = c) \quad (4)$$

where:

- $N_k(x_{test})$ represents the set of k nearest neighbors (k=15 optimized)
- $w_i = \frac{1}{d(x_{test},x_i)}$ are distance-weighted votes
- $d(x_{test}, x_i) = \sum_{j=1}^{p}|x_{test,j} - x_{i,j}|$ is Manhattan distance
- I(·) is the indicator function

The posterior probability estimate for clinical interpretation is:

$$\hat{P}(Y=1|x_{test}) = \frac{\left(\sum_{i\in N_{k(x_{test})}} w_i I(y_i=1)\right)}{\left(\sum_{i\in N_{k(x_{test})}} w_i\right)} \quad (5)$$

- **Decision Tree Splitting Criterion**

The decision tree employed entropy-based splitting with the information gain criterion. For node t containing n_t samples, the entropy is:

$$H(t) = -\sum_{c=0}^{1} p_{c(t)}\log_2 p_{c(t)} \quad (6)$$

where $p_{c(t)} = \frac{n_{c,t}}{n_t}$ is proportion of class c and samples at node t.

The information gain for split s at node t is:

$$IG(t,s) = H(t) - \sum_{v\in values(s)}\left(\frac{n_{t_v}}{n_t}\right)H(t_v) \quad (7)$$

The algorithm selects the split maximizing information gain, subject to constraints:

- min_sample_split = 10
- min_sample_leaf = 4
- max_depth = 5

- **Random Forest Ensemble Formulation**

The Random Forest classifier aggregates predictions from T=300 decision trees. Each tree $h_t(x)$ is trained on bootstrap sample Dt with random feature subset selection (p≈4 features per split).
The ensemble prediction is:

$$\hat{y}_{RF(x)} = mode\{h_{1(x)}, h_{2(x)}, \dots, h_{T(x)}\} \quad (8)$$

The probability estimate is:

$$\hat{P}_{RF}(Y = 1|x) = \left(\frac{1}{T}\right)\sum_{t=1}^{T} I\left(h_{t(x)} = 1\right) \quad (9)$$

Feature importance for variable X_j is computed via permutation importance:

$$Importance(X_j) = \left(\frac{1}{T}\right)\sum_{t=1}^{T}\left[Error\left(D_t^{perm(j)}\right) - Error(D_t)\right] \quad (10)$$

Where $D_t^{prem(j)}$ is the out-of-bag sample with values of X_j randomly permuted.

- **Performance Metric Formulations**

The evaluation metrics are formally defined as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (11)$$

$$Precision = \frac{TP}{TP+FP} \quad (12)$$

$$Recall\ (Sensitivity) = \frac{TP}{TP+FN} \quad (13)$$

$$F1 = 2 * \frac{Precision*Recall}{Precision + Recall} \quad (14)$$

$$Specificity = \frac{TN}{TN+FP} \quad (15)$$

The Area Under ROC Curve (AUC-ROC) is computed as:

$$AUC = \int_0^1 TPR\left(FPR^{(-1)(x)}\right)dx \quad (16)$$

where TPR is True Positive Rate (Recall) and FPR is False Positive Rate (1 – Specificity) for comprehensive evaluation [17, 22].

- **Cross-Validation Objective Function**

The hyperparameter optimization during 5-fold cross-validation minimized [19, 24]:

$$\hat{\theta}^* = \arg\min_{\theta \in \Theta} \left(\frac{1}{5}\right) \sum_{k=1}^{5} [1 - Accuracy_{k(\theta)}] \quad (17)$$

where θ represents hyperparameters, Θ is the search space, and $Accuracy_k$ is accuracy on the k^{th} validation fold.

6. Conclusion

This comparative analysis demonstrates that traditional statistical methods (Logistic Regression) remain competitive with, and in certain aspects superior to, more complex machine learning algorithms for cardiovascular disease prediction. The achieved accuracy (85.71%) represents a clinically meaningful improvement over baseline prediction, though falls short of diagnostic gold-standard performance.

The investigation yields three principal conclusions: First, algorithm selection should be guided by specific clinical context—sensitivity-focused applications favor Logistic Regression, while specificity-emphasizing contexts may benefit from K-Nearest Neighbors. Second, feature importance analysis validates established clinical knowledge while providing quantitative predictor rankings. Third, ensemble methods do not universally outperform simpler approaches in moderate-dimensional clinical datasets.

These findings contribute to ongoing research in computational cardiology [18, 23, 28], emphasizing that methodological sophistication must align with clinical utility. Future work should focus on prospective validation, feature engineering innovations, and integration into clinical decision support systems that augment rather than replace clinician expertise [25, 28].

References

1. World Health Organization, "Cardiovascular Diseases (CVDs) Fact Sheet", World Health Organization Press, May 2021. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
2. Gregory A. Roth, George A. Mensah, Catherine O. Johnson, Giuseppe Addolorato, Enrico Ammirati, Larry M. Baddour, Noël C. Barengo, "Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019: Update from the GBD 2019 Study", Journal of the American College of Cardiology, December 2020, 76 (25), 2982–3021. https://doi.org/10.1016/j.jacc.2020.11.010
3. Manesh R. Patel, John H. Calhoon, Gregory J. Dehmer, J. Aaron Grantham, Thomas M. Maddox, David J. Maron, Peter K. Smith, "ACC/AATS/AHA/ASE/ASNC/SCAI/SCCT/STS 2017 Appropriate Use Criteria for Coronary Revascularization in Patients with Stable Ischemic Heart Disease", Journal of the American College of Cardiology, May 2017, 69 (17), 2212–2241. https://doi.org/10.1016/j.jacc.2017.02.001
4. Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Joseph J. Schmid, Suneet Sandhu, Kenneth H. Guppy, "International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease", The American Journal of Cardiology, September 1989, 64 (5), 304–310. https://doi.org/10.1016/0002-9149(89)90524-9
5. S. Palaniappan, R. Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", International Journal of Computer Science and Network Security, August 2008, 8 (8), 343–350. https://www.researchgate.net/publication/267206809
6. Dheeru Dua, Casey Graff, "UCI Machine Learning Repository", University of California, School of Information and Computer Science, 2019. https://archive.ics.uci.edu/ml

7. Geert J. van der Heijden, A. Rogier T. Donders, Theo Stijnen, Karel G. M. Moons, "Imputation of Missing Values is Superior to Complete Case Analysis and the Missing-Indicator Method in Multivariable Diagnostic Research", Journal of Clinical Epidemiology, October 2006, 59 (10), 1102–1109. https://doi.org/10.1016/j.jclinepi.2006.01.015
8. Nathalie Japkowicz, Shaju Stephen, "The Class Imbalance Problem: A Systematic Study", Intelligent Data Analysis, 2002, 6 (5), 429–449. https://doi.org/10.3233/IDA-2002-6504
9. Peter C. Austin, Jack V. Tu, "Automated Variable Selection Methods for Logistic Regression Produced Unstable Models for Predicting Acute Myocardial Infarction Mortality", Journal of Clinical Epidemiology, November 2004, 57 (11), 1138–1146. https://doi.org/10.1016/j.jclinepi.2004.04.003
10. Leo Breiman, "Random Forests", Machine Learning, 2001, 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324
11. Roohallah Alizadehsani, Jafar Habibi, M. J. Hosseini, H. Mashayekhi, Reihane Boghrati, Asma Ghandeharioun, Bahare Sani, "A Data Mining Approach for Diagnosis of Coronary Artery Disease", Computer Methods and Programs in Biomedicine, January 2013, 111 (1), 52–61. https://doi.org/10.1016/j.cmpb.2013.03.004
12. M. Shouman, T. Turner, R. Stocker, "Using Data Mining Techniques in Heart Disease Diagnosis and Treatment", Japan-Egypt Conference on Electronics, Communications and Computers, March 2012, 173–177. https://doi.org/10.1109/JEC-ECC.2012.6186978
13. Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, A. A. Yarifard, "Computer Aided Decision Making for Heart Disease Detection Using Hybrid Neural Network-Genetic Algorithm", Computer Methods and Programs in Biomedicine, 2017, 141, 19–26. https://doi.org/10.1016/j.cmpb.2017.01.004
14. H. D. Masethe, M. A. Masethe, "Prediction of Heart Disease Using Classification Algorithms", Proceedings of the World Congress on Engineering and Computer Science, October 2014, 2, 22–24. https://www.iaeng.org/publication/WCECS2014/
15. A. H. Gonsalves, F. Thabtah, R. M. Mohammad, G. Singh, "Prediction of Coronary Heart Disease Using Machine Learning: An Experimental Analysis", International Conference on Deep Learning Technologies, 2019, 51–56. https://doi.org/10.1145/3342999.3343005
16. Yann LeCun, Yoshua Bengio, Geoffrey Hinton, "Deep Learning", Nature, May 2015, 521 (7553), 436–444. https://doi.org/10.1038/nature14539
17. Davide Chicco, Giuseppe Jurman, "The Advantages of the Matthews Correlation Coefficient (MCC) Over F1 Score and Accuracy in Binary Classification Evaluation", BMC Genomics, 2020, 21 (1), 1–13. https://doi.org/10.1186/s12864-019-6413-7
18. Chayakrit Krittanawong, HongJu Zhang, Zhen Wang, Mehmet Aydar, Takeshi Kitai, "Artificial Intelligence in Precision Cardiovascular Medicine", Journal of the American College of Cardiology, May 2017, 69 (21), 2657–2664. https://doi.org/10.1016/j.jacc.2017.03.571
19. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, "Scikit-learn: Machine Learning in Python", Journal of Machine Learning Research, 2011, 12, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html
20. Rich Caruana, Alexandru Niculescu-Mizil, "An Empirical Comparison of Supervised Learning Algorithms", Proceedings of the 23rd International Conference on Machine Learning, June 2006, 161–168. https://doi.org/10.1145/1143844.1143865

21. Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2011. https://doi.org/10.1016/C2009-0-61819-5
22. David J. Hand, Robert J. Till, "A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems", Machine Learning, 2001, 45 (2), 171–186. https://doi.org/10.1023/A:1010920819831
23. Subhi J. Al'Aref, Kipp W. Johnson, "Clinical Applications of Machine Learning in Cardiovascular Disease and its Relevance to Cardiac Imaging", European Heart Journal, June 2019, 40 (24), 1975–1986. https://doi.org/10.1093/eurheartj/ehy404
24. Max Kuhn, Kjell Johnson, "Applied Predictive Modeling", Springer, 2013. https://doi.org/10.1007/978-1-4614-6849-3
25. Ewout W. Steyerberg, "Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating", Springer, 2019. https://doi.org/10.1007/978-3-030-16399-0
26. Stephen F. Weng, Jenna Reps, Joe Kai, Jonathan M. Garibaldi, Nadeem Qureshi, "Can Machine-Learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data?", PLOS ONE, April 2017, 12 (4), e0174944. https://doi.org/10.1371/journal.pone.0174944
27. Trevor Hastie, Robert Tibshirani, Jerome Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", Springer, 2009. https://doi.org/10.1007/978-0-387-84858-7
28. Rishi C. Deo, "Machine Learning in Medicine", Circulation, November 2015, 132 (20), 1920–1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593
29. Tianqi Chen, Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System", Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2016, 785–794. https://doi.org/10.1145/2939672.2939785
30. Daniel M. McNeish, "Using Lasso for Predictor Selection and to Assuage Overfitting: A Method Long Overlooked in Behavioral Sciences", Multivariate Behavioral Research, 2015, 50 (5), 471–484. https://doi.org/10.1080/00273171.2015.1036965