

Heart Disease Prediction Using Logistic Regression: A Machine Learning Approach with Interactive Clinical Decision Support Dashboard

Murari Raja Krishna¹, Maddisetti Balu², Muppalla Ravi³, Ms. Sivasangari⁴

^{1,2,3}UG Student, Department of Artificial Intelligence and Data Science, Dhanalakshmi Srinivasan University, Samayapuram, Trichy, Tamil Nadu.

⁴Assistant Professor, Department of Artificial Intelligence and Data Science, Dhanalakshmi Srinivasan University, Samayapuram, Trichy, Tamil Nadu.

Abstract- Cardiovascular disease remains the leading cause of global mortality, claiming approximately 17.9 million lives annually. Early and accurate prediction of individual patient risk is essential for enabling timely clinical intervention and reducing preventable mortality. This paper presents a comprehensive machine learning-based clinical decision support system for heart disease prediction using Logistic Regression, deployed as an interactive Streamlit web application. The system is trained and evaluated on the UCI Cleveland Heart Disease dataset comprising 303 patient records and 13 clinical attributes. The Logistic Regression model achieves a test accuracy of 88.52%, F1-score of 0.89, and AUC-ROC of 0.9267 on a held-out 20% test partition. Rigorous comparative evaluation against five baseline classifiers—Naïve Bayes, K-Nearest Neighbors, Support Vector Machine, Decision Tree, and Random Forest—confirms that Logistic Regression provides the best overall balance of interpretability, predictive performance, and generalization stability. The Streamlit web application enables real-time risk stratification and probabilistic output visualization across all 13 clinical features, bridging the gap between algorithmic performance and practical clinical utility.

Keywords: Heart Disease Prediction; Logistic Regression; Machine Learning; UCI Cleveland Dataset; Streamlit; AUCROC; Clinical Decision Support; Cardiovascular Risk; Interactive Visualization.

I. INTRODUCTION

Cardiovascular diseases (CVDs) represent the foremost public health challenge of the 21st century, responsible for nearly one-third of all global deaths. The World Health Organization estimates that 17.9 million lives are claimed annually by heart attacks, strokes, and related conditions [1]. Despite remarkable advances in cardiac medicine, late-stage diagnosis continues to drive preventable mortality and enormous healthcare expenditure.

The proliferation of electronic health records has catalyzed a new generation of machine learning-based clinical decision support systems (CDSS). Unlike traditional statistical risk calculators such as the Framingham Risk Score, ML approaches can identify complex interactions among clinical risk factors without requiring explicit mechanistic assumptions, offering superior performance on heterogeneous patient populations [2].

This paper addresses heart disease prediction through three contributions: (i) a rigorously evaluated Logistic Regression pipeline on the UCI Cleveland dataset; (ii) systematic comparison against five baseline classifiers; and (iii) a production-ready Streamlit CDSS application enabling real-time probabilistic risk assessment across all 13 clinical features.

The remainder is organized as follows. Section II surveys related work. Section III describes the dataset. Section IV presents the architecture. Section V details methodology. Section VI analyzes results. Section VII describes the web application. Sections VIII–IX discuss limitations, future work, and conclusions.

II. RELATED WORK

A. Classical Statistical Approaches

The Framingham Heart Study [2] established multivariate cardiovascular risk scoring,

demonstrating that demographic and biochemical variables collectively predict coronary artery disease risk with clinically useful accuracy. However, the Framingham model assumes linear additive effects and was calibrated on a predominantly Caucasian cohort, limiting generalizability. Logistic Regression has been the statistical workhorse of medical prediction for decades, offering interpretable coefficients, calibrated probability outputs, and robustness to moderate class imbalance [3]. Diamond and Forrester [4] established the foundational probabilistic framework for coronary artery disease prediction.

B. Machine Learning Approaches

The UCI Cleveland dataset [5] catalyzed extensive ML research into cardiac risk classification. Detrano et al. [6] demonstrated 77% prediction accuracy using clinical features alone. Subsequent SVM studies achieved 84–86% accuracy [7]. Mohan et al. [9] proposed a hybrid random forest with linear model feature selection achieving 88.7% accuracy, highlighting the value of feature selection. Tomar and Agarwal [8] systematically reviewed ML approaches in healthcare, identifying interpretability as the primary barrier to clinical translation.

C. Deep Learning Methods

Recurrent neural networks and convolutional architectures achieve remarkable results in ECG-based arrhythmia detection [10]. However, for tabular clinical datasets of modest size ($n < 1000$), deep learning is consistently outperformed by well-tuned classical ML models due to excessive parameterization relative to available training data. Arabasadi et al. [11] demonstrated that hybrid neural network–genetic algorithm approaches can reach 93% accuracy on Cleveland data, though at the cost of interpretability—a critical requirement in clinical CDSS contexts.

D. Clinical Decision Support Tools

Streamlit [12] has emerged as the leading Python framework for deploying ML models as clinical decision tools, enabling web-based inference without front-end engineering expertise. This paper contributes to this literature by integrating a

validated Logistic Regression model with a full-featured Streamlit CDSS interface.

III. DATASET AND PROBLEM FORMULATION

A. Dataset Overview

The UCI Cleveland Heart Disease dataset [5] comprises 303 de-identified patient records from the Cleveland Clinic Foundation. Each record contains 13 clinical input features and a binary target (disease=1, healthy=0). Table I summarizes dataset characteristics.

TABLE I: Dataset Characteristics Summary

Disease Cases	165 (54.5%)
Healthy Cases	138 (45.5%)
Patient Age Range	29–77 years (mean 54.4)
Sex Distribution	Male: 207, Female: 96
Train/Test Split	80% / 20% (seed=42)

B. Feature Description

The 13 features span demographic, hemodynamic, biochemical, and stress-test domains: age (years); sex (1=male, 0=female); cp – chest pain type (0–3); trestbps – resting BP (mm Hg); chol – serum cholesterol (mg/dl); fbs – fasting blood sugar >120; restecg – resting ECG; thalach – max heart rate; exang – exercise angina; oldpeak – ST depression; slope – ST slope; ca – major vessels (0–3); thal – thalassemia type.

C. Problem Formulation

We formulate this as binary classification. Given $x \in \mathbb{R}^{13}$, Logistic Regression models the posterior probability as: $P(y=1|x) = \sigma(wTx + b) = 1 / (1 + \exp(-(wTx + b)))$ where w are learned weights with L2 regularization ($C=1.0$) to mitigate overfitting on the 303sample dataset. The decision threshold is set at $P(y=1|x) \geq 0.5$ for standard classification; the interactive application exposes the full probability distribution to support threshold adjustment.

IV. PROPOSED SYSTEM ARCHITECTURE

The system is a modular end-to-end CDSS pipeline with four components: Data Preprocessing, Feature Standardization, Model Training and Evaluation, and Interactive Web Deployment.

A. Data Preprocessing Module

Raw clinical data is loaded from CSV. The pipeline performs: (i) missing value removal for six records with missing ca values; (ii) data type validation; (iii)

Attribute	Value
Dataset Source	UCI Cleveland Heart Disease
Total Records	303 patients
Input Features	13 clinical attributes
Target Variable	Disease (1) / Healthy (0)

stratified 80/20 train-test splitting (random_state=42). The dataset's balanced class distribution (54.5%/45.5%) requires no oversampling adjustments.

B. Feature Standardization

All 13 features are standardized: $x' = (x - \mu) / \sigma$ using StandardScaler fit on training data only. The scaler is serialized (heart_scaler.pkl) alongside the model (heart_disease_model.pkl) for consistent production inference without data leakage. Logistic Regression (lbfgs solver, L2 penalty, C=1.0, max_iter=1000, random_state=42) is trained on 242 samples. Model and feature names are serialized via joblib. Table II details the hyperparameter configuration.

TABLE II: Logistic Regression Hyperparameter Configuration

Parameter	Value	Rationale
solver	lbfgs	Efficient, small datasets
penalty	L2	Ridge regularization
C	1.0	Balanced bias-variance
max_iter	1000	Convergence guarantee
random_state	42	Reproducibility

D. Evaluation Module

Performance is measured via: Accuracy, Precision, Recall (sensitivity), F1-Score, AUC-ROC, and 5-fold stratified cross-validation. Recall is prioritized clinically since false negatives (missed diagnoses) carry greater patient risk than false positives.

V. METHODOLOGY AND EXPERIMENTAL SETUP

A. Implementation Environment

All experiments use Python 3.10. Key dependencies: pandas (v1.5+) for data ingestion; scikit-learn (v1.1+) for training and evaluation; joblib for serialization; Streamlit (v1.20+) for deployment; Plotly for interactive visualization.

B. Baseline Classifiers

Five baselines are evaluated under identical conditions:

Naïve Bayes (Gaussian): Assumes conditional feature independence. Fast and interpretable but ignores clinically important feature correlations such as the interaction between chest pain type and exercise-induced angina.

K-Nearest Neighbors (k=5): Instance-based classifier using Euclidean distance on standardized features. Captures local data structure but computationally expensive at inference and sensitive to the curse of dimensionality.

Support Vector Machine (RBF kernel): Maximizes classification margin in a transformed feature space. Strong discriminator on standardized clinical data; C=1.0, γ ='scale'.

Decision Tree (CART): Single unpruned tree (random_state=42). Fully interpretable but prone to overfitting on 303-sample data without pruning, as confirmed by results.

Random Forest (100 trees): Bootstrap-aggregated ensemble (random_state=42). Reduces variance through tree averaging; provides stable cross-validation performance but sacrifices coefficient interpretability.

C. Cross-Validation Protocol

Stratified 5-fold cross-validation preserves the 54.5%/45.5% class ratio across folds, preventing fold-level imbalance from inflating variance estimates. Scores are reported as mean \pm standard deviation across the five folds, providing a more reliable generalization estimate than single-split accuracy for the small 303-sample dataset.

D. Pseudocode

The core classification pipeline is summarized as:

```
df = load_csv(heart_disease_data.csv)
X, y = df.drop('target'), df['target']
X_tr, X_te, y_tr, y_te = train_test_split(X, y, 0.2,
seed=42)
scaler = StandardScaler().fit(X_tr)
model = LogisticRegression(C=1.0).fit(scaler(X_tr), y_tr)
evaluate(model, scaler(X_te), y_te)
```

VI. RESULTS AND DISCUSSION

A. Classification Performance

Table III presents Logistic Regression performance on the 61 sample test set. The model achieves 88.52% accuracy with AUCROC 0.9267. The confusion matrix shows: 25 true negatives, 29 true positives, 4 false positives, 3 false negatives—a clinically favorable ratio.

TABLE III: Logistic Regression Test Performance (n=61)

Disease-class recall of 0.91 is the most critical clinical metric: 91% of actual heart disease patients are correctly identified for further workup. The 5-fold CV accuracy of 83.49% \pm 4.11% confirms reliable generalization with moderate variance. The 5-point gap between test (88.52%) and CV (83.49%) accuracy is within expected range for a 303-sample dataset.

B. Comparative Evaluation

Table IV compares all six classifiers. Logistic Regression (highlighted) provides the best balance of accuracy, AUC-ROC, and generalization.

KNN achieves the highest raw test accuracy (90.16%) but exhibits the largest gap to its CV score (81.19%), indicating test-partition overfitting (gap of 8.97 points vs. Logistic Regression's 5.03 points). SVM achieves the highest AUCROC (0.9310) but lower accuracy (86.89%). Random Forest demonstrates the most stable CV (83.82% \pm 2.88%) but lower test

accuracy. Decision Tree performs worst (75.41%), confirming that single trees overfit severely on small clinical coronary artery disease. Number of major vessels (ca) and thalassemia type (thal) rank second and third, reflecting their direct relationship with coronary anatomy and myocardial perfusion.

TABLE IV: Comparative Performance of All Classifiers datasets without pruning.

Metric	Healthy (0)	Disease (1)	Overall
Precision	0.89	0.88	0.89 (wtd)
Recall	0.86	0.91	0.89 (wtd)
F1-Score	0.88	0.89	0.89 (wtd)
Support	29	32	61 total
Accuracy	—	—	88.52%
AUC-ROC	—	—	0.9267
5-Fold CV Acc.	—	—	83.49% \pm 4.11%
Method	Test Acc.	AUC-ROC	5-Fold CV
Naïve Bayes	86.89%	0.8944	80.51% \pm 5.55%
KNN (k=5)	90.16%	0.9154	81.19% \pm 4.28%
SVM (RBF)	86.89%	0.9310	81.83% \pm 4.99%
Decision Tree	75.41%	0.7592	75.55% \pm 5.29%
Random Forest	83.61%	0.9224	83.82% \pm 2.88%
Logistic Reg. \checkmark	88.52%	0.9267	83.49%\pm4.11%

Logistic Regression delivers the best holistic profile: competitive accuracy, high AUC-ROC (0.9267), consistent cross-validation, and—critically—interpretable coefficients that clinical staff can validate against established cardiovascular domain knowledge.

C. Feature Importance

Standardized Logistic Regression coefficients reveal clinically meaningful feature rankings. Chest pain type (cp) carries the largest positive coefficient magnitude, consistent with literature identifying angina patterns as strongly predictive of Maximum heart rate (thalach) shows a negative coefficient—higher exercise capacity correlates with lower disease probability—consistent with its known cardioprotective role.

D. Clinical Significance

In a screening context, false negatives (disease patients classified as healthy) represent the most dangerous error type, potentially leading to delayed treatment and adverse outcomes. The model's sensitivity of 91% means 9 out of 10 disease-positive patients are correctly flagged for specialist referral—a performance level comparable to validated clinical risk scores in analogous cardiovascular screening applications.

The AUC-ROC of 0.9267 confirms strong discrimination across all classification thresholds, enabling clinicians to tune the decision boundary based on their specific sensitivity-specificity requirements. For high-sensitivity screening (e.g., population health programs), a lower threshold ($P > 0.35$) increases recall at the cost of specificity, while specialist referral contexts may prefer higher thresholds ($P > 0.60$).

VII. WEB APPLICATION AND DEPLOYMENT

A. Streamlit Application Design

The model is deployed as a Streamlit application (heart_app.py) with three functional zones: a sidebar panel, a two-column clinical input form, and a results visualization panel. The sidebar displays model metadata and key risk factors. The main panel presents all 13 input controls: continuous features use interactive sliders with clinically appropriate ranges; categorical features use dropdown selectors with human-readable labels. Upon prediction, the application renders: (i) a color-coded risk box (red: disease risk, green: healthy); (ii) confidence percentage; (iii) risk level (Low/Medium/High based on disease probability thresholds 0.4/0.7); and (iv) a Plotly interactive probability bar chart for both classes. Clinical guidance messages direct high-risk patients to seek specialist consultation.

B. Deployment and Performance

Local deployment: streamlit run heart_app.py at localhost:8501. Cloud deployment via Streamlit Community Cloud requires only a GitHub repository with heart_app.py, requirements.txt, and serialized model files. Model files are loaded once via

@st.cache_resource, enabling inference in <50 ms. Enterprise deployment is supported through Docker containerization for integration with hospital information systems. Total application memory footprint is under 200 MB.

C. Application Workflow

The complete user workflow is: (1) the clinician opens the browser-based interface; (2) enters patient demographics via age/sex sliders and selectors; (3) enters heart metrics including chest pain type, blood pressure, cholesterol, ECG results, and stress-test parameters; (4) clicks 'Predict Heart Disease'; (5) receives an immediate color-coded risk assessment with probability breakdown. The application's non-technical design requires no statistical or programming expertise, enabling direct use by clinicians, nurses, and health screening professionals. A key design decision is exposing the full class probability distribution (not just the binary prediction) via the Plotly bar chart. This enables clinicians to distinguish between borderline cases (e.g., 52% disease probability) and high-confidence predictions (e.g., 85% disease probability), supporting appropriately calibrated clinical decision-making. The three-tier risk stratification (Low: $P < 0.4$, Medium: 0.4–0.7, High: $P > 0.7$) further aids clinical triage.

The 'Input Summary' expander provides an audit trail of all entered feature values alongside the prediction, supporting documentation of AI-assisted clinical decisions. All predictions include a disclaimer that the tool is for educational and screening purposes and does not substitute professional medical diagnosis.

D. Comparison with Existing Tools

The proposed application advances beyond existing web-based cardiovascular risk calculators (e.g., ACC/AHA ASCVD Risk Estimator) in two key ways: (i) it leverages ML classification rather than parametric regression, capturing non-linear feature interactions; and (ii) it provides interactive probability visualization rather than only a point risk estimate. Compared to published ML-based CDSS tools, the Streamlit framework enables deployment without

dedicated infrastructure, lowering the barrier to adoption in resource-constrained clinical settings.

VIII. LIMITATIONS AND FUTURE WORK

A. Current Limitations

Several limitations merit acknowledgment. First, the Cleveland dataset comprises only 303 patients from a single US institution, constraining generalizability to diverse demographic and geographic populations. Multi-site, multi-ethnic validation is needed before clinical deployment.

Second, the 13-feature input set omits potentially informative variables including troponin levels, BNP, echocardiography findings, CT angiography results, and lifestyle factors (smoking history, physical activity). Integration of these variables is expected to substantially improve discriminative performance toward AUC-ROC > 0.95.

Third, the model produces only point predictions without confidence intervals. Fourth, the model is static and does not update as new patient data accumulates. Online learning pipelines are needed for deployed systems to maintain calibration over time.

B. Future Work

Feature enrichment with additional biomarkers represents the highest-priority improvement. SHAP (SHapley Additive exPlanations) values will enable patient-level feature attribution. Calibration via Platt scaling will improve probability reliability. EHR integration via HL7 FHIR APIs will automate feature extraction. Prospective clinical trials will provide ground-truth validation.

IX. CONCLUSION

This paper presented an end-to-end machine learning pipeline for binary heart disease classification using Logistic Regression, evaluated on the UCI Cleveland Heart Disease dataset and deployed as an interactive Streamlit clinical decision support application.

Logistic Regression achieves 88.52% test accuracy, 0.9267

AUC-ROC, 0.91 disease-class recall, and 83.49% ± 4.11% 5-fold CV accuracy. Comparative analysis against five baselines confirms the best overall balance of performance, generalization, and clinical interpretability. The Streamlit application bridges the gap between algorithmic performance and clinical utility, enabling real-time risk stratification.

ACKNOWLEDGMENT

The authors acknowledge Dhanalakshmi Srinivasan University

for computational resources, the UCI ML Repository for the dataset, and the scikit-learn, Streamlit, and Plotly communities.

REFERENCES

1. WHO, "Cardiovascular Diseases (CVDs)," WHO Fact Sheet, 2023.
2. W. B. Kannel et al., "Factors of Risk in Development of Coronary Heart Disease," *Ann. Intern. Med.*, vol. 55, pp. 33–50, 1961.
3. D. W. Hosmer, S. Lemeshow, R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Wiley, 2013.
4. G. A. Diamond and J. S. Forrester, "Analysis of Probability as Aid in Clinical Diagnosis of CAD," *N. Engl. J. Med.*, vol. 300, pp. 1350–1358, 1979.
5. D. Dua and C. Graff, *UCI Machine Learning Repository*, Univ. of California, Irvine, 2019. <http://archive.ics.uci.edu/ml>.
6. R. Detrano et al., "International Application of a New Probability
7. Algorithm for Diagnosis of CAD," *Am. J. Cardiol.*, vol. 64, pp. 304–310, 1989.
8. K. Polat and S. Güneş, "A Novel Hybrid Intelligent Method Based on C4.5 Decision Tree," *Expert Syst. Appl.*, vol. 36, pp. 1587–1592, 2009.
9. D. Tomar and S. Agarwal, "A Survey on Data Mining Approaches for Healthcare," *Int. J. Bio-Sci. Bio-Technol.*, vol. 5, pp. 241–266, 2013.
10. S. Mohan, C. Thirumalai, G. Srivastava, "Effective Heart Disease Prediction Using Hybrid ML Techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019.
11. A. Y. Hannun et al., "Cardiologist-Level Arrhythmia Detection Using Deep Neural Network," *Nat. Med.*, vol. 25, pp. 65–69, 2019.

12. Z. Arabasadi et al., "Computer Aided Decision Making for Heart Disease Detection Using Hybrid Neural Network–GA," *Comput. Methods Programs Biomed.*, vol. 141, pp. 19–26, 2017.
13. Streamlit Inc., "Streamlit: Build and Share Data Apps," <https://streamlit.io>. [Accessed: Mar. 3, 2026].
14. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.