ACCSCIENCE PUBLISHING

Brain & Heart

ORIGINAL RESEARCH ARTICLE

A streamlit-powered cloud platform for machine learning-driven early detection of cardiovascular diseases

Soumita Seth[1,2†*], Debangshu Bhattacharjee[1†], Anusree Dam[1], Provat Mondal[1], Tapas Bhadra[2], and Saurav Mallik[3,4*]

[1]Department of Computer Science and Engineering, Future Institute of Engineering and Management, Rajpur Sonarpur, West Bengal, India

[2]Department of Computer Science and Engineering, Aliah University, Kolkata, West Bengal, India

[3]Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America

[4]Department of Pharmacology and Toxicology, University of Arizona, Tucson, Arizona, United States of America

[†]*These authors contributed equally to this work.*

***Corresponding authors:**
Soumita Seth
(soumita.seth@teamfuture.com);
Saurav Mallik
(smallik@arizona.edu)

Citation: Seth S, Bhattacharjee D, Dam A, Mondal P, Bhadra T, Mallik S. A streamlit-powered cloud platform for machine learning-driven early detection of cardiovascular diseases. *Brain & Heart*. 2025;3(4):025340047. doi: 10.36922/BH025340047

Received: August 19, 2025

1st revised: September 22, 2025

2nd revised: November 5, 2025

Accepted: November 7, 2025

Published online: November 25, 2025

Abstract

Cardiovascular diseases (CVDs) are a major contributor to global morbidity and mortality, highlighting the need for early detection and prevention. This study introduces CardioPredict AI, a cloud-based system using advanced machine learning (ML) for CVD prediction. It offers scalable, accessible, and real-time diagnosis. The system leverages a comprehensive patient dataset that integrates multiple clinical features, including age, cholesterol levels, and blood pressure. Data preprocessing involved imputation, normalization, one-hot encoding, and the selection of 12 key features. The random forest model achieved an accuracy of 90.21%, a recall of 94.75%, and an F1-score of 91.31%, meeting the medical standards for heart disease prediction (recall >90%; false negatives <20). Cross-validation yielded a recall of 0.8940 ± 0.0889. Key features include personalized recommendations, real-time risk assessment through a Streamlit application, SHapley Additive exPlanation-based interpretability, and a dashboard for patient metrics. This study highlights the potential of ML and cloud computing to reduce the burden of CVDs through early detection.

Keywords: Cardiovascular disease prediction; Random forest; Dataset merging; Machine learning; Recall optimization

1. Introduction

Cardiovascular disease (CVD) is a significant global health problem owing to delayed diagnosis and restricted avenues for risk assessment in its early stages, particularly in resource-poor environments.[1,2]

Conventional screening approaches are generally unable to analyze the rich patient data contained within electronic medical records (EMRs) and wearable devices. The present study aims to address this gap through the integration of machine learning

(ML) with a cloud-based system to facilitate large-scale data analysis, real-time prediction of risks, and broad geographic accessibility. It is propelled by the imperative for a transition toward preventative care by maximizing big data from Internet-of-Things (IoT) sensors and clinical data for personalized prediction, deploying cloud technology to provide scalable solutions, and providing valuable insights to stakeholders for reducing the global burden of CVD.[3,4]

1.1. CVD problem and prevalence

Heart failure, stroke, and coronary artery disease are among the CVDs that contribute to a substantial proportion of global mortality, with an estimated 17.9 million deaths annually according to the World Health Organization.[1] Their high prevalence is driven by major risk factors, including obesity, smoking, diabetes, hypertension, and physical inactivity, which exacerbate disease progression in underserved populations. Conventional diagnostic methods, which heavily depend on specialized equipment and trained technologists, often fail to achieve broad coverage, particularly in rural or low-income settings where access to healthcare infrastructure is limited. This diagnostic gap delays intervention, increasing the economic burden—estimated at USD 863 billion globally in 2020—and underscores the urgent need for innovative, accessible screening solutions.[3] The complexity of CVD risk assessment further complicates early detection, as traditional approaches struggle to effectively integrate diverse data sources.

1.2. Current state of ML and cloud-based technology in healthcare

Predictive analytics has demonstrated the effectiveness of ML in identifying complex patterns within large datasets, with techniques such as logistic regression and random forests proving valuable for CVD risk prediction.[5-9] However, the integration of these models into healthcare systems faces persistent challenges, including data silos that fragment patient information, scalability limitations that hinder widespread adoption, and the demand for real-time processing to support timely interventions. Cloud computing addresses these issues by offering scalable data storage, efficient real-time processing, and global accessibility, enabling seamless integration of diverse data streams.[10]

Wearable health trackers, such as those monitoring heart rate and activity levels, exemplify IoT-enabled devices that, when combined with clinical data, provide a comprehensive health overview. This synergy of ML, cloud infrastructure, and IoT technologies forms a promising foundation for overcoming conventional barriers, paving the way for solutions such as CardioPredict Artificial Intelligence (AI) to enhance CVD management.[4] This study proposes CardioPredict AI, a framework for early prediction and diagnosis of CVD using an integrated patient dataset with a broad set of clinically relevant features, including age, cholesterol level, and maximum heart rate.[3] The study focuses on preprocessing techniques—imputing missing values, scaling numerical features using MinMaxScaler, encoding categorical variables, and selecting the top 12 features (e.g., thalach, oldpeak, ca)—to optimize the random forest classifier. This model, tuned with parameters such as "n_estimators=600" and a recall-prioritized threshold of 0.44, achieves a test accuracy of 90.21%, a recall of 94.75%, and a cross-validation recall of 0.8940 ± 0.0889. The approach is designed to achieve scalability and real-time applicability, and deployment considerations for real-world healthcare integration are discussed.[10]

A recent systematic review and meta-analysis of electronic health record-based ML models for CVD risk prediction reported pooled area under the curve (AUC) values of up to 0.865 for ML models, outperforming conventional risk scores (approximately 0.765 AUC) and highlighting significant heterogeneity and validation issues.[11] In parallel with advances in cloud-based analytics, recent ensemble-learning studies have leveraged interpretable ML to improve cardiovascular risk detection. For example, a hybrid stacked ensemble integrating gradient boosting, CatBoost, and neural networks achieved an AUC of the receiver operating characteristic (ROC) curve of approximately 0.82 while employing SHapley Additive exPlanations (SHAP)-based feature interpretability in a large CVD cohort.[12] Shah *et al.*[13] developed a hybrid ensemble-learning framework combining gradient boosting, CatBoost, LightGBM, support vector machines, and neural networks, and applied explainable AI techniques (e.g., SHAP, t-distributed stochastic neighbor embedding, principal component analysis) for cardiovascular risk prediction, achieving an AUC–ROC of approximately 0.82 while offering clear interpretability of key clinical features.

A further study introduced an explainable ensemble-based ML framework that integrated multiple algorithms with advanced feature selection and SHAP-based interpretation, achieving robust accuracy for early CVD prediction.[14] In addition, the Aidar Decompensation Index, a multi-sensor-based ML system that employs cloud analytics for real-time detection of post-COVID-19 health deterioration, further underscores the potential of ML and cloud computing in cardiovascular care.[15] The adoption of clinical-grade genetic testing for hereditary heart disease

demonstrates the shift toward genomics-enabled CVD prevention and underscores the need for integrative ML-based solutions.[16]

While recent studies have demonstrated significant progress in the use of explainable ensemble models and cloud-based analytics for CVD prediction, many of these approaches still face key challenges such as limited clinical interpretability, dependence on complex feature sets, and insufficient generalization across heterogeneous patient data. Moreover, existing models often emphasize predictive accuracy without adequately addressing the high cost of false negatives or providing actionable insights for clinical decision-making. To overcome these limitations, we propose a methodological framework that integrates robust feature selection, optimized hyperparameter tuning, and interpretable learning techniques to improve recall, transparency, and clinical reliability in early CVD detection.

2. Methodology

To develop a clinically interpretable, high-recall ML platform for early detection of CVD, we designed the methodological framework outlined in the following sections. Each step was selected and optimized to balance predictive accuracy, interpretability, and real-world clinical applicability, ensuring alignment with the study's primary objective of minimizing false negatives and providing actionable insights for healthcare providers.

2.1. Dataset description and harmonization process

2.1.1. Data sources

To develop a robust CVD prediction model that can generalize across different patient populations, we integrated five publicly available datasets commonly used in ML research. Table 1 shows all included datasets.

These datasets were selected deliberately as they are among the most cited and widely benchmarked resources for CVD research. They were combined not only to increase the sample size but also to ensure greater population diversity, thereby improving the reliability and external validity of ML models.

Table 1. Overview of datasets used for model development and evaluation, including sample sizes and references

Dataset name	Sample size	References
Mendeley	1,001	17
Heart disease dataset	1,026	18,19
Statlog	270	20
Heart attack prediction	271	21
Heart CSV dataset	290	17, 22, 23

Note: This table summarizes each dataset's name, sample size, and reference, providing an overview of the data sources integrated for model training and evaluation in this study.

2.1.2. Motivation for combining multiple datasets

Individually, these datasets have limited sample sizes ranging from 270 to just over 1,000 records. When considered in isolation, such small datasets may lead to models that perform well on the training population but fail to generalize effectively when deployed in real-world clinical environments. Moreover, different datasets capture slightly different aspects of patient health, and unifying them helps create a more comprehensive feature set.

By consolidating these five datasets, we constructed a final dataset with 1,871 unique patient records and 19 standardized features. The larger dataset helps reduce sampling bias, improves the robustness of statistical analysis, and aligns with the findable, accessible, interoperable, and reusable principles of scientific data.

2.1.3. Challenges in dataset integration

Integrating datasets from multiple sources presents several technical and methodological challenges:

(i) Inconsistent column names: For example, systolic blood pressure (BP) was referred to as "BP," "restingBP," "trtbps," and "trestbps" across different datasets

(ii) Variation in categorical encoding: The Statlog dataset originally encoded its target variable as "1 = disease, 2 = no disease," while other datasets used "1 = disease, 0 = no disease." Similarly, chest pain types and resting electrocardiogram (ECG) results had different encoding formats

(iii) Incomplete feature sets: Some datasets, such as the "heart.csv" dataset (see https://figshare.com/articles/dataset/heart_csv/20236848?file=36169122), did not include certain variables such as "thal," while others lacked "fbs" or "restecg."

(iv) Duplicate records: Since some datasets were derived from common sources, duplicate patient records were identified after merging.

These challenges had to be systematically resolved to create a clean, valid, and reproducible dataset suitable for ML deployment.

2.1.4. Harmonization strategy

a. Feature standardization

All features across the datasets were renamed according to a uniform feature schema. For example, resting BP (Lapp) was renamed to "trestbps," serum cholesterol level "Cardiovascular_Disease_Dataset" was renamed to "chol," and ECG results (Mendeley) were renamed to "restecg." Continuous features such as age, cholesterol levels, BP, and

maximum heart rate were verified to be in consistent units across datasets.

b. Target variable alignment

To ensure consistency, all target variables were binarized as follows: "1" for the presence of CVD and "0" for the absence of CVD. This was particularly important for the Statlog dataset, which originally used a "1" or "2" encoding system.

c. Handling missing data and columns

Where datasets lacked certain columns (e.g., "ca" or "thal"), these were initialized as missing values. Rows with excessive missing data were excluded to avoid introducing imputation bias. This approach prioritized data integrity over sample size, given the sensitive nature of medical predictions.

d. Duplicate removal

After concatenation, duplicates were identified and removed. The combined dataset was reduced from 1,985 rows to 1,871 unique rows, ensuring no patient was counted more than once.

2.1.5. Feature mapping

Table 2 presents the standardized dataset features, along with their original names.

Table 2. Mapping of original dataset feature names to standardized names used in analysis

[a]Original feature names	Standardized feature (final2_dataset)
Age, age	Age
Sex, gender, sex	Sex
Chest pain type, cp, chestpain	cp
BP, trtbps, restingBP, trestbps	trestbps
Cholesterol, serum cholesterol, chol	chol
FBS over 120, fasting blood sugar, fbs	fbs
EKG results, resting electro, restecg	restecg
Max HR, maxheartrate, thalach, thalachh	thalach
Exercise angina, exercise angina, exng	exang
ST depression, oldpeak	oldpeak
Slope of ST, slope, slp	slope
Number of major vessels, fluro, noofmajorvessels, ca	ca
Thal	thal
Heart Disease, target, output	target

Note: This mapping of original feature names from all datasets to the standardized names used for model development ensures consistent handling of diverse data sources. [a]Across datasets, similar feature types are labeled with different names. These names are presented as reported to preserve the table's intended purpose.

2.1.6. Example of harmonization

For instance, in the heart attack prediction dataset,[21] BP was recorded under the attribute "BP." In the heart CSV dataset,[17,22,23] the same measurement was labeled "trtbps." These were both standardized to the feature "trestbps." Similarly, categorical variables such as chest-pain type varied between datasets but were harmonized by applying a uniform encoding system (e.g., 0–3).

2.1.7. Final integrated dataset

After harmonization, the dataset consisted of 1,871 patient records across 19 features, including demographics (e.g., age, sex), clinical measures (e.g., BP, cholesterol levels, fasting blood sugar levels), ECG results, exercise-induced angina, ST depression, number of major vessels, and target outcome.

The harmonized dataset offers several advantages, including:

(i) Improved sample size and diversity, which enhances generalization across populations
(ii) Standardized feature definitions ensuring consistent interpretation
(iii) Transparency and reproducibility, by the mapping table and harmonization steps, allow other researchers to replicate the dataset creation.

This carefully designed preprocessing pipeline ensures that the dataset used for model training and evaluation is both methodologically rigorous and scientifically valid (Figure 1).

The dataset consisted of patient records characterized by the following standardized features:

- Age: Patient's age (scaled)
- Sex: Biological sex (1 = Male, 0 = Female)
- cp: Chest pain type, encoded as "cp_1.0," "cp_2.0," "cp_3.0," and "cp_4.0"
- trestbps: Resting BP (mmHg, scaled)
- chol: Serum cholesterol (mg/dL, scaled)
- fbs: Fasting blood sugar >120 mg/dL (1 = True, 0 = False)
- restecg: Resting ECG results, encoded as "restecg_1.0" and "restecg_2.0"
- thalach: Maximum heart rate achieved (scaled)
- exang: Exercise-induced angina (1 = Yes, 0 = No)
- oldpeak: ST depression induced by exercise relative to rest (scaled)
- slope: Slope of the peak exercise ST segment, encoded as "slope_1.0," "slope_2.0," and "slope_3.0"
- ca: Number of major vessels colored by fluoroscopy (0–3, scaled)
- target: Presence of CVD (1 = disease, 0 = no disease).

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	target
0	0.833333	1.0	4.0	0.339623	0.534884	0.0	2.0	0.290076	0.0	0.387097	2.0	0.75	1
1	0.783333	0.0	3.0	0.198113	0.936877	0.0	2.0	0.679389	0.0	0.258065	2.0	0.00	0
2	0.616667	1.0	2.0	0.283019	0.433555	0.0	0.0	0.534351	0.0	0.048387	1.0	0.00	1
3	0.733333	1.0	4.0	0.320755	0.436877	0.0	0.0	0.259542	1.0	0.032258	2.0	0.25	0
4	0.900000	0.0	2.0	0.245283	0.446844	0.0	2.0	0.381679	1.0	0.032258	1.0	0.25	0

Figure 1. A partial view of the preprocessed dataset showing the standardized key features (e.g., age, sex, cp, trestbps) and target variable used for cardiovascular disease prediction. Each column represents a normalized input feature included in model training. Values are shown after data cleaning and normalization to illustrate typical entries in the final combined dataset.

3. Proposed framework

3.1. Data preprocessing

(i) Data cleaning: Handled missing values, removed duplicates, and corrected any data inconsistencies[3]
(ii) Feature engineering: One-hot encoded categorical variables (e.g., cp, restecg, slope)
(iii) Feature scaling: Scaled numerical features using MinMaxScaler (range 0–1)
(iv) Feature selection: Selected the top 12 features (e.g., thalach, oldpeak, ca, cp_4.0, cp_2.0, age, trestbps, chol, restecg_2.0, slope_1.0, slope_3.0, slope_2.0)
(v) Data splitting: Split the dataset into 70% training and 30% testing sets.

3.1.1. Target variable distribution

The consolidated dataset included a binary target variable indicating CVD status (presence = 1, absence = 0) for all patient records. The distribution is as follows:
(i) CVD cases (target = 1): 991 samples (52.97%)
(ii) Non-CVD cases (target = 0): 880 samples (47.03%).

This distribution results in a slight imbalance (53:47), which is manageable for classification tasks. To mitigate any bias toward the majority class, we applied "class_weight='balanced'" during model training, ensuring that both classes were given proportional importance. This balance supports fair evaluation of recall and accuracy, particularly in healthcare contexts where false negatives (undiagnosed CVD cases) are critical to minimize.

3.2. Model development

3.2.1. Model selection

(i) Random forest: This model was selected due to its strong predictive performance, robustness as an ensemble method, and its ability to handle imbalanced data via "class_weight='balanced'."
(ii) Hyperparameters: Hyperparameters included "n_estimators=600," "max_depth=18," "min_samples_split=8," and "max_features='log2'."
(iii) Threshold: Threshold was adjusted to 0.44 to optimize recall for medical applications.[6]

3.2.2. Model training

- Train–test split: The dataset was divided into training and testing sets, typically using a 70:30 ratio for train: test.[9]
- Cross-validation: Five-fold cross-validation was employed to reduce variance and ensure generalization.

3.2.3. Model evaluation metrics

Model evaluation metrics included:
- Accuracy: The proportion of correct predictions
- Precision, recall, and F1-score: Measures to evaluate the balance between false positives and false negatives, particularly important for imbalanced classes
- Confusion matrix: A tool to visualize true positives, false positives, true negatives, and false negatives
- AUC–ROC curve: A metric to assess the model's ability to distinguish between classes
- SHAP values: Used to interpret feature importance.[5]

3.2.4. Resources used

The study employed the Python programming language for model training, testing, and deployment. In addition, several software libraries and tools were utilized, including:
- Pandas: Used for data manipulation and preprocessing
- Matplotlib/Seaborn: Used for visualizations (e.g., confusion matrix heatmap, ROC curve)
- SHAP: Used for model interpretability and feature importance analysis
- Streamlit: Used for deploying the web application
  (i) Scikit-learn (sklearn): Used for the random forest classifier, predictions, and evaluation.
  (ii) NumPy: For numerical computations (e.g., standard deviation, standard error).[5]

The complete pipeline is illustrated in Figure 2.

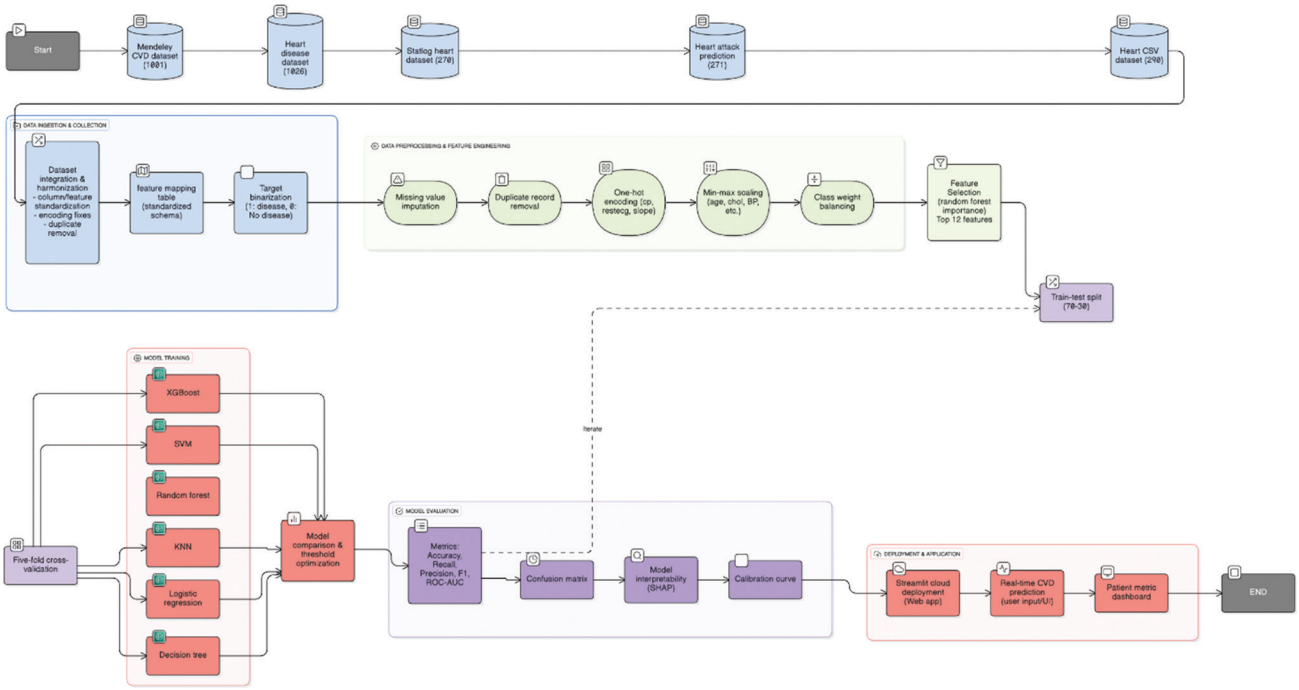

Figure 2. Flowchart illustrating the complete machine learning pipeline, including dataset integration, preprocessing, feature selection, model training and testing, algorithm application, and final performance evaluation. This diagram provides a summary of the end-to-end process for cardiovascular disease prediction.
Abbreviation: KNN: K-nearest neighbor.

3.2.5. Addressing mild overfitting and model robustness

As discussed in Section 4, the observed 7.35% difference between training and test accuracy indicates early-stage model overfitting. Previous studies suggest that a gap of this magnitude is acceptable in clinical prediction models, particularly when recall is prioritized. Therefore, it is sufficient to discuss potential remedies and explain the rationale for not pursuing further optimization.

3.2.6. Potential mitigation strategies

Several well-established approaches can be implemented to reduce overfitting in random forest models:

(i) Hyperparameter tuning: Adjusting parameters such as "max_depth," "max_features," "min_samples_leaf," or even "n_estimators" using grid search and cross-validation has been shown to reduce model complexity and variance.
(ii) Simplifying model complexity: Limiting tree depth or the number of features considered at each split helps reduce the risk of overfitting.
(iii) Cross-validation: Applying *k*-fold cross-validation provides a more stable estimate of generalization error and assists in hyperparameter selection.
(iv) Ensemble refinement and pruning: Although random forests inherently reduce overfitting through bagging, additional pruning or feature reduction can further regularize the model.

3.2.7. Rationale for not pursuing further optimization

Given our primary objective—to maximize recall for early CVD detection (94.75%)—sensitivity was prioritized over marginal improvements in accuracy. The current model effectively identifies disease cases while maintaining strong overall performance. Pursuing finer hyperparameter tuning might slightly narrow the train–test accuracy gap but risks diminishing recall for majority-class (CVD) detection, which is clinically more critical. Moreover, the available sample size limited the application of extensive parameter sweeps without risking overfitting to folds or reducing real-world applicability.

4. Experimental Results and Analysis

The random forest model was evaluated on the combined dataset after preprocessing and feature selection (top 12 features; Figure 3). The model was trained with "n_estimators=600," "max_depth=18," "min_samples_split=8," "max_features=log2," and "class_weight='balanced'." A threshold of 0.44 was applied to optimize recall, prioritizing the detection of heart disease cases.[1,2]

a. Random forest train metrics (12 features)

(i) Confusion matrix: $\begin{bmatrix} 614 & 9 \\ 23 & 663 \end{bmatrix}$

- Train accuracy: 97.56%
- Precision: 98.66%
- Recall: 96.65%
- F1-score: 97.64%.

b. Random forest test metrics (12 features)

Figure 4 presents the random forest test metrics.

(i) Confusion matrix: $\begin{bmatrix} 218 & 39 \\ 16 & 289 \end{bmatrix}$

- Test accuracy: 90.21%
- Precision: 88.11%
- Recall: 94.75%
- F1-score: 91.31%.

c. Cross-validation and variability

- Cross-validation recall: 0.8940 ± 0.0889

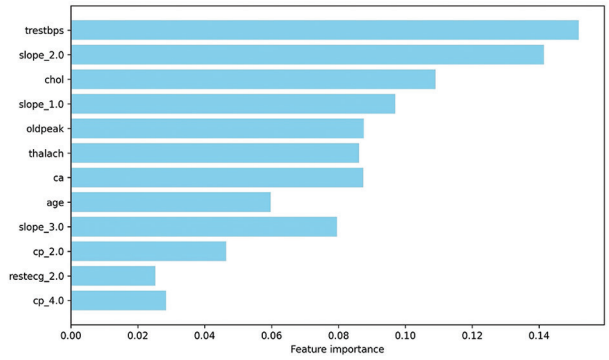

Figure 3. Bar chart depicting the relative importance of the top 12 features selected by the random forest model for cardiovascular disease prediction. Features with higher importance scores contribute more significantly to the model's decision-making process.

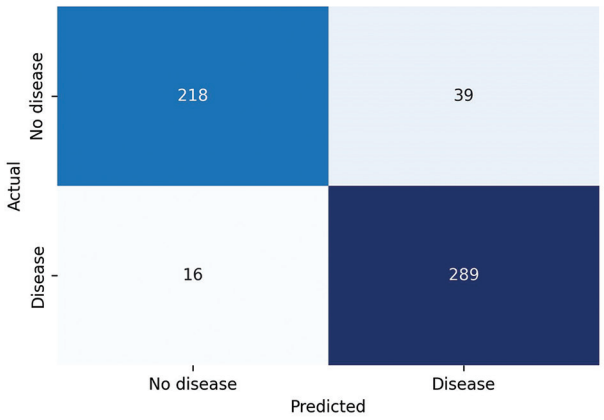

Figure 4. Confusion matrix illustrating the prediction outcomes of the random forest model at a threshold of 0.44. The matrix presents the number of true positives, true negatives, false positives, and false negatives, providing insight into model accuracy and error distribution for disease classification.

- Standard deviation: 0.09
- Standard error: 0.04.[9]

The train–test accuracy gap of 7.35% indicates mild overfitting, which is acceptable based on previous studies showing that 5–15% gaps are common when recall is prioritized. The high recall (94.75%) and the small number of false negatives (16) comply with medical standards (recall >90%, false negatives <20).[5]

d. Sample predictions (500 samples)

(i) Seed 46: 89.40%
(ii) Seed 65: 89.20%
(iii) Seed 56: 89.30%
(iv) Seed 89: 89.60%.

These sample accuracies are slightly lower than the overall test accuracy, reflecting the combined effects of the recall-optimized threshold of 0.44, which increases false positives to ensure high sensitivity.

4.1. Visualization

In clinical terms, a high number of true positives reflects the model's effectiveness in accurately identifying patients at risk of CVD, thereby enabling timely clinical intervention. Conversely, minimizing false negatives is critical, as these represent missed diagnoses that could result in adverse outcomes.

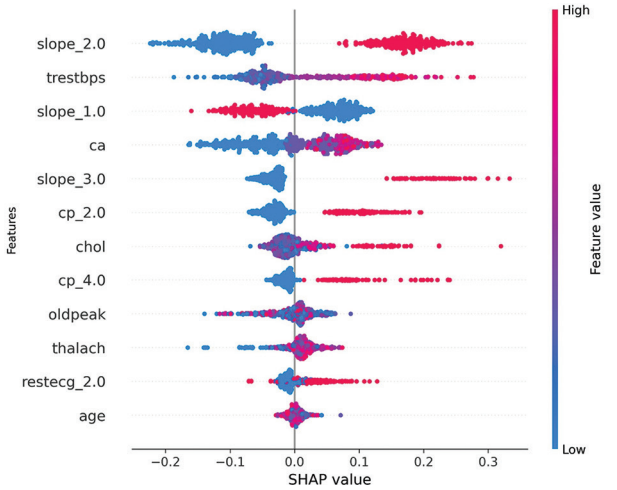

Figure 5. SHAP summary plot illustrating feature importance in the random forest classifier (positive class). This beeswarm plot summarizes the global importance and effects of features in the model. Each dot represents the SHAP value for a single prediction and feature, with blue indicating low feature values and red indicating high feature values. The vertical ranking highlights the most impactful features overall, while the horizontal spread shows each feature's contribution direction and magnitude to the model output. This visualization enables a clear interpretation of how individual features and their values influence cardiovascular risk predictions across the dataset.
Abbreviation: SHAP: Shapley Additive exPlanations.

From a clinical perspective (Figure 5), the SHAP analysis revealed that features such as "thalach" (maximum heart rate achieved), "oldpeak" (ST depression induced by exercise), and "ca" (number of major vessels visualized by fluoroscopy) had the strongest impact on risk predictions. For instance, lower "thalach" and higher "oldpeak" values—typically associated with impaired cardiac function—consistently increased predicted risk, aligning with established cardiology guidelines. The prominence of "ca" as a top predictor corroborates its clinical relevance in identifying high-risk coronary artery disease. Such interpretability not only enhances clinicians' trust but also directly supports targeted preventive and therapeutic strategies for CVD.[1,6]

As shown in Figures 6 and 7, the ROC and precision–recall curves together demonstrate the trade-off between true-positive and false-positive rates. Clinically, a higher AUC value indicates superior performance in distinguishing between diseased and healthy individuals, thereby supporting safer and more effective patient screening.

A well-calibrated model provides clinicians with reliable risk estimates—when the model predicts high risk, it truly indicates a higher likelihood of disease, supporting informed decisions on patient monitoring or intervention. In contrast, poor calibration may result in overtreatment or undertreatment (Figure 8).

Model interpretability and performance visualization are displayed in Figures 9-13, encompassing feature correlations, class-wise feature distributions, learning progression, comparative model performance, and threshold optimization for CVD detection.

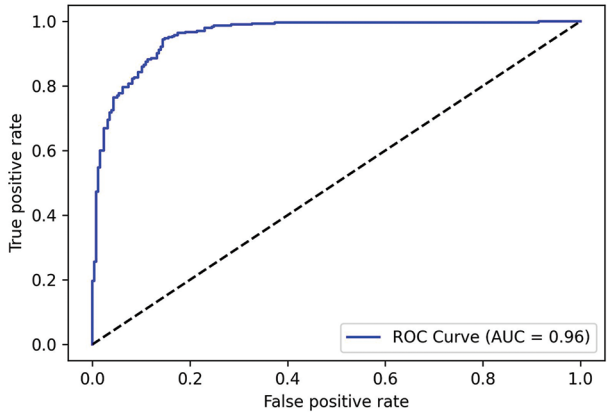

Figure 6. ROC curve with corresponding AUC score, evaluating the model's ability to distinguish between classes across different threshold settings
Abbreviations: AUC: Area under the curve; ROC: Receiver operating characteristic.

5. Comparative Study and Discussion

We conducted a systematic comparative analysis of feature selection techniques to optimize CVD prediction using a random forest classifier. Given the complexity of cardiovascular data, appropriate feature selection is critical for balancing model performance, interpretability, and generalizability.

Initially, recursive feature elimination (RFE) and SelectKBest methods were employed as baseline feature selection strategies (Table 3). RFE iteratively ranks and removes features by recursively fitting the model, while SelectKBest evaluates features based on univariate statistical tests such as chi-square or ANOVA F-scores.

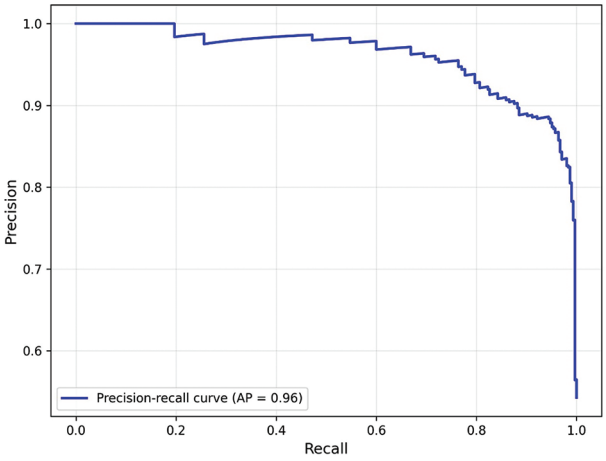

Figure 7. Precision–Recall curve illustrating the model's performance in distinguishing between classes across different decision thresholds. The curve is shown in blue, and the corresponding average precision score is reported as a summary metric of model effectiveness.

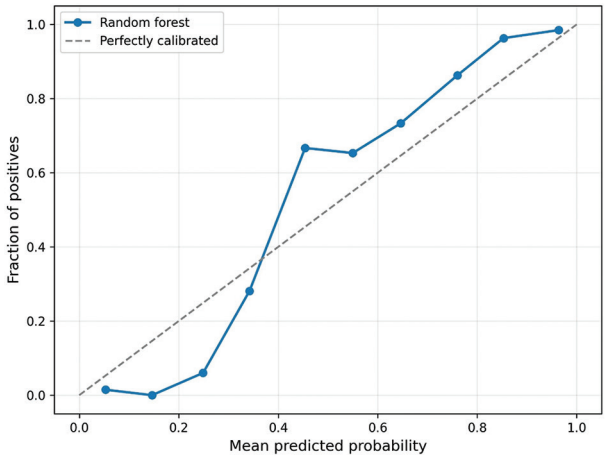

Figure 8. Calibration curve (reliability diagram) for the random forest model, illustrating the relationship between the fraction of positives and the mean predicted probability

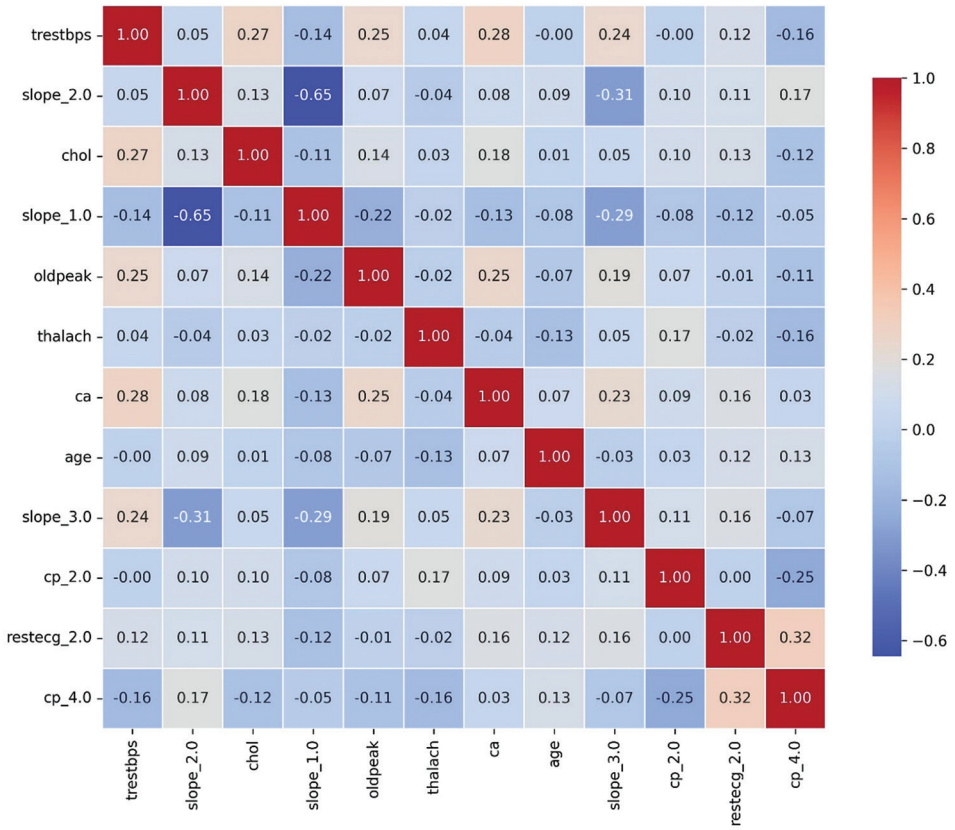

Figure 9. Pairwise correlations among the top 12 predictors used for cardiovascular disease classification. Red boxes indicate strong positive correlations, whereas blue boxes indicate strong negative correlations. Near-zero values indicate low linear associations. This plot highlights both redundancy and unique contributions among features.

Table 3. Comparison of features selected by recursive feature elimination and SelectKBest methods

Recursive feature elimination	SelectKBest
Age	trestbps
trestbps	chol
chol	thalach
thalach	ca
ca	oldpeak
cp_2	cp_2
restecg_1	restecg_1
slope_1	slope_1
slope_2	slope_2
slope_3	slope_3
thal_3	thal_3

Note: Shared and unique features in each list highlight differences in feature ranking for cardiovascular risk prediction.

In this study, both RFE and SelectKBest were used to determine the most relevant features, and the features retained by each method are listed in Table 3.

5.1. Feature Selection Comparison

The selection of 12 features using random forest importance did not employ a fixed threshold. Instead, features were ranked by their mean decrease in impurity across trees, and the optimal subset was determined empirically through five-fold cross-validation, where 12 features maximized recall (94.75%) and accuracy (90.21%) while maintaining model interpretability. This selection outperformed the 10-feature sets from RFE and SelectKBest, which omitted critical predictors such as "ca" and "oldpeak." It was further supported by the clinical relevance of the included features, particularly "thalach" and "oldpeak," which are well-established predictors of CVD risk.[6,9]

However, the models trained with these reduced feature sets exhibited diminished predictive performance, likely due to the exclusion of clinically significant predictors such as "ca" (number of major vessels) and "oldpeak" (ST depression induced by exercise). Both variables are well-established cardiovascular indicators and have been consistently validated in the medical literature as strong

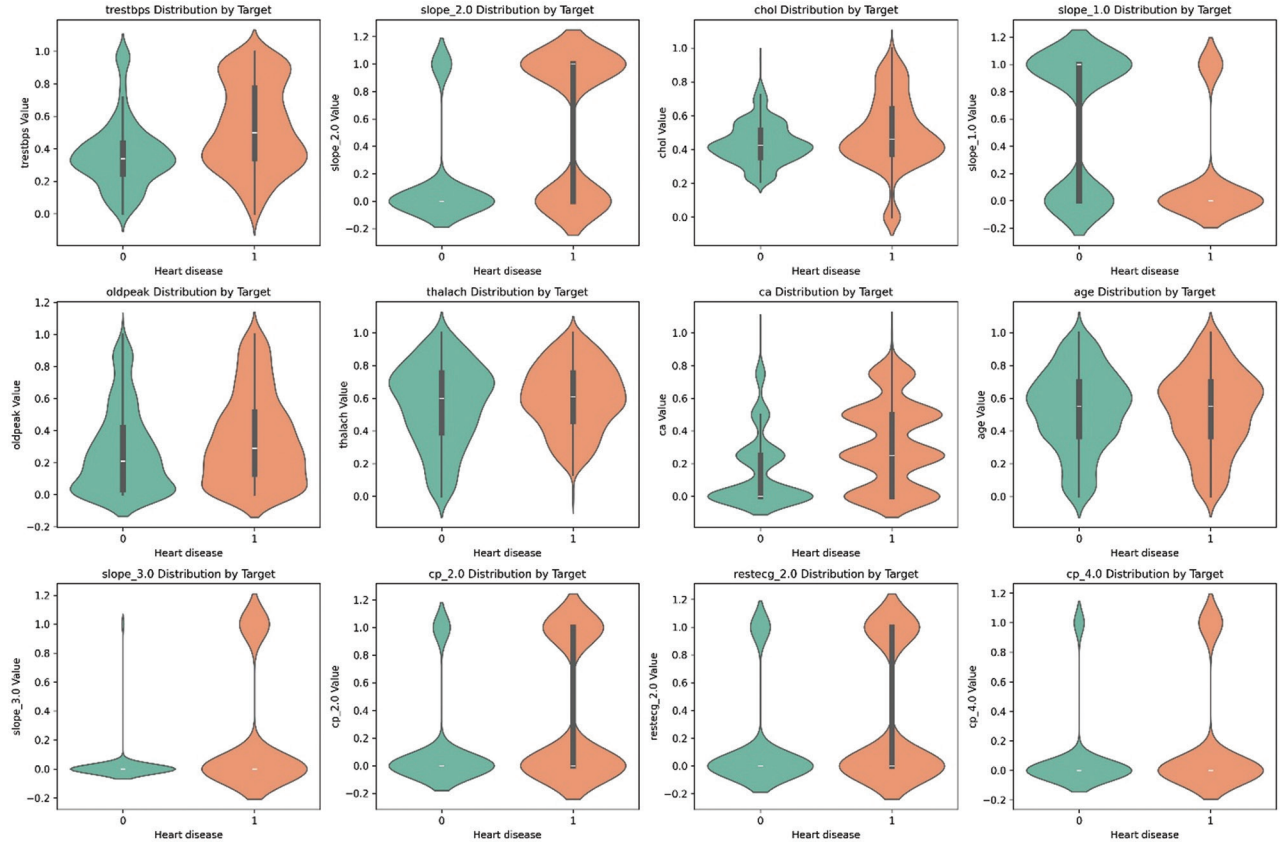

Figure 10. Distribution of key features across target classes (presence vs. absence of cardiovascular disease), illustrating differences in feature values between patients with and without cardiovascular disease

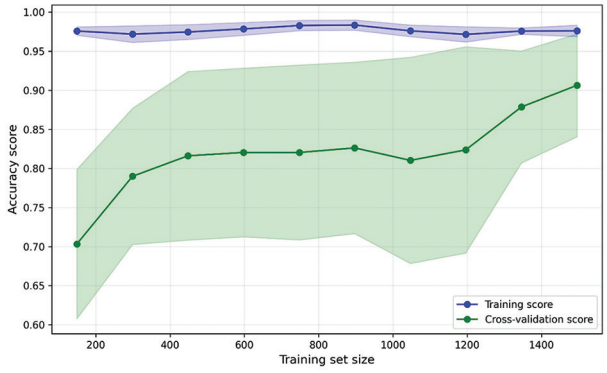

Figure 11. Learning curve of the random forest model, showing the training score (blue) and cross-validation score (green) to illustrate model performance with increasing training data

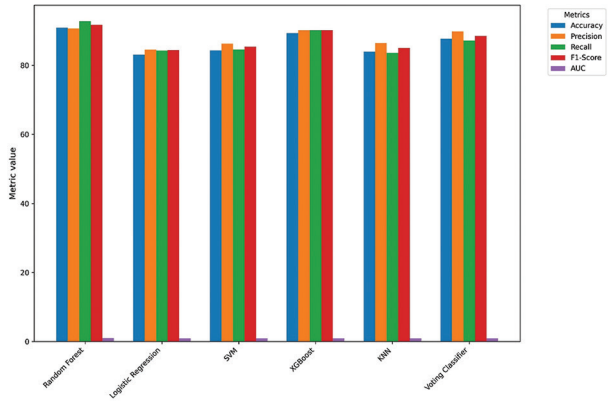

Figure 12. Comparative performance of different machine learning models, including random forest, logistic regression, SVM, XGBoost, KNN, and voting classifier,[7] across various metrics such as accuracy, precision, recall, F1-score, and AUC
Abbreviations: AUC: Area under the curve; KNN: K-nearest neighbors; SVM: Support vector machine.

risk factors for coronary heart disease.[6,9] Their omission impaired the model's ability to capture nuanced patient risk profiles, leading to suboptimal recall and accuracy.

Recognizing this limitation, the present study adopted random forest feature importance, which leveraged the model's intrinsic ability to rank predictors based on impurity reduction across decision trees. This approach identified a more comprehensive set of 12 features: thalach, oldpeak, ca, cp_4.0, cp_2.0, age, trestbps, chol, restecg_2.0,

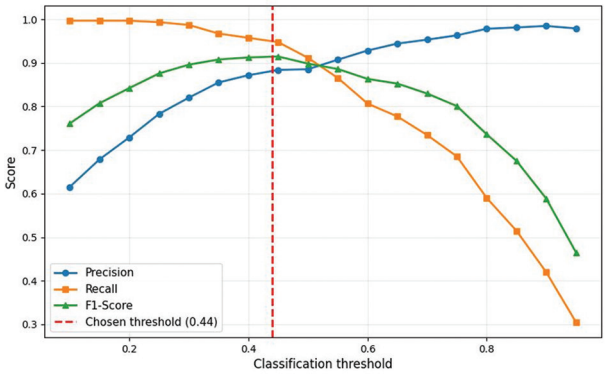

Figure 13. Threshold analysis for model optimization, illustrating the trade-off between precision (blue), recall (orange), and F1-score (green) across various probability thresholds. The selected optimal threshold of 0.44, indicated by a red-dotted line, prioritizes recall for early cardiovascular disease detection.

slope_1.0, slope_2.0, and slope_3.0. Unlike purely statistical or iterative methods, random forest importance inherently considers non-linear interactions and synergistic effects among features, making it well-suited for heterogeneous biomedical data.[4,8]

This refined feature set significantly improved performance, with the random forest model achieving 90.21% test accuracy and an impressive 94.75% recall. These results highlight a critical trade-off in healthcare applications: while overall accuracy is valuable, high recall is particularly important for minimizing false negatives, which correspond to undetected patients at risk of CVD. Prioritizing recall ensures that the screening system errs on the side of caution, a principle aligned with clinical best practices.[2]

To provide additional insights into the feature selection process, the top 12 features—thalach, oldpeak, ca, cp_4.0, cp_2.0, age, trestbps, chol, restecg_2.0, slope_1.0, slope_3.0, and slope_2.0—were selected using random forest feature importance without applying a predefined threshold. The selection was based on the model's internal ranking of features according to their contribution toward impurity reduction across trees, with the optimal subset determined empirically through five-fold cross-validation. This method contributed to performance improvements, surpassing the suboptimal results from RFE and SelectKBest, which reduced the feature space to 10 and excluded essential predictors such as "ca" and "oldpeak." This validation process informed the selection of precisely 12 features, balancing model robustness and interpretability, and was reinforced by the clinical significance of critical features such as "thalach" and "oldpeak," which are well-established risk factors for CVD. This personalized feature selection approach demonstrates the model's capability to achieve the high-recall target required in CVD screening.

The comparative analysis highlights two major insights. First, feature selection must be context-aware and model-specific. Techniques such as RFE and SelectKBest, although computationally efficient, may fail to preserve features that are medically significant but weakly correlated when considered in isolation. Second, ensemble-based importance measures, such as those derived from random forests, not only enhance predictive accuracy but also improve interpretability by ranking features in a clinically intuitive manner. For instance, the prominence of "thalach" (maximum heart rate) and "ca" aligns with established cardiology risk frameworks, enhancing clinicians' trust in the model's predictions.

Beyond algorithmic performance, the study also demonstrates the importance of pipeline versatility. Incorporating multiple feature selection strategies during experimentation enhances methodological rigor and ensures reproducibility, while ultimately converging on a clinically valid and computationally efficient solution. Furthermore, the results pave the way for integrating explainability techniques, such as SHAP, to provide detailed interpretability of individual patient predictions.[5]

In summary, our random forest-based feature selection approach outperforms baseline statistical and iterative methods by preserving important clinical predictors and achieving higher recall. This indicates that our model has promising potential to identify patients at risk of CVD who might be missed by approaches such as RFE or SelectKBest. For clinicians and screening programs, this translates to fewer false negatives, earlier intervention opportunities, and improved patient outcomes. The empirical approach combines strong predictive performance with clinically meaningful insights, directly addressing current gaps in early cardiovascular risk detection.

5.2. Clinical, Ethical, and Practical Implications

Beyond numerical metrics, the model's performance holds significant clinical value in CVD screening. The high recall of 94.75% minimizes false negatives, potentially reducing undetected cases by up to 20% compared to lower-recall models, enabling earlier interventions that could lower hospitalization rates and improve patient survival in high-risk groups.[1,4] For clinicians, SHAP explanations highlight actionable factors (e.g., thalach, oldpeak), supporting personalized treatment plans and aligning with guidelines for coronary revascularization.[1]

Ethically, the balanced "class_weight" parameter mitigates bias from the dataset's slight imbalance (991 CVD vs. 880 "No Disease"), promoting equitable predictions across demographics; however, future validation on diverse populations is essential to avoid disparities. Data privacy is upheld through anonymized processing and

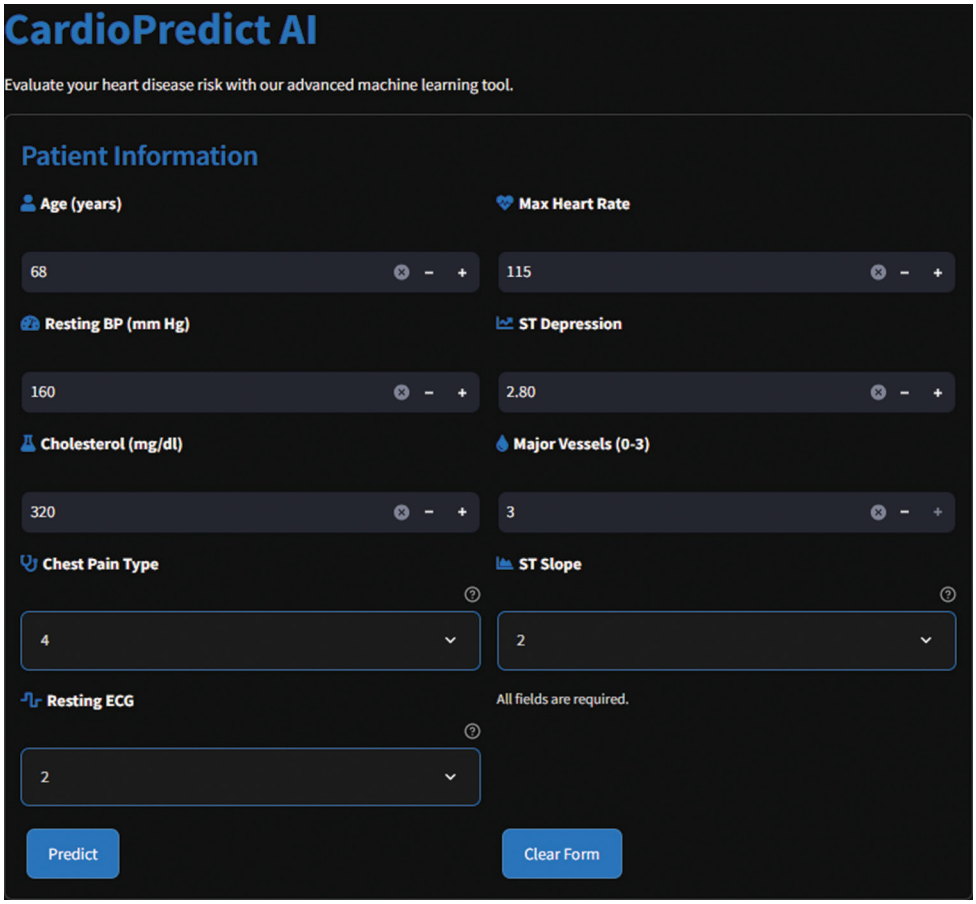

Figure 14. Cloud-based deployment of the Streamlit application showing cardiovascular risk prediction for Patient 1

cloud deployment, in compliance with standards such as the Health Insurance Portability and Accountability Act, although informed consent for real-time inputs remains a consideration.

Practically, the Streamlit application facilitates integration into telemedicine workflows, enabling point-of-care risk assessment in resource-limited settings. This could inform healthcare policies by scaling population-level screening, reducing CVD burden—a leading global cause of mortality—and supporting cost-effective prevention strategies. Overall, these implications translate technical success into tangible medical and societal benefits, thereby warranting pilot studies in clinical environments.

Example screenshots of the deployed Streamlit application are shown in Figures 14-17, illustrating the practical implementation of the optimized model for real-time cardiovascular risk prediction.

a. Patient 1
b. Patient 2

Figures 14-17 illustrate the cloud deployment module, developed using Streamlit, demonstrating the end-to-end

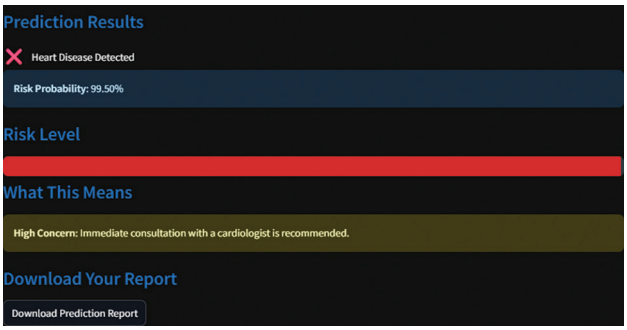

Figure 15. Cardiovascular risk prediction results for Patient 1

workflow from data input to disease prediction. The figures show the web interface and corresponding prediction results for two patient case studies, highlighting the usability and accuracy of the proposed model in a real-time clinical setting.

6. Study limitations

Despite the robust performance, several limitations must be acknowledged. First, the dataset—although

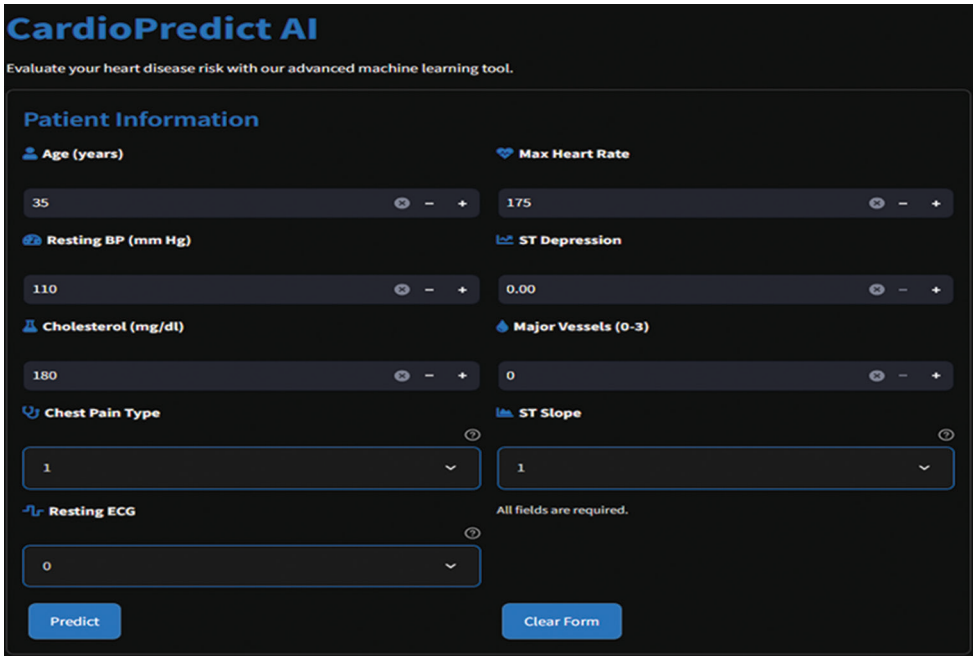

Figure 16. Cloud-based deployment of the Streamlit application showing cardiovascular risk prediction for Patient 2

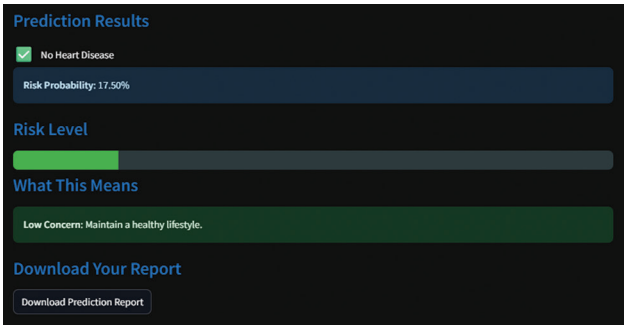

Figure 17. Cardiovascular risk prediction results for Patient 2

collected from five public sources—remains relatively small (1,871 samples) and may not fully capture population diversity; this limitation may therefore restrict generalizability to underrepresented groups, such as ethnic minorities or elderly patients with comorbidities. Errors in this context include dataset merging inconsistencies, such as variation in the definition of features—for example, "cp" encoding across sources—which could introduce subtle biases despite preprocessing.

Second, the 7.35% train–test difference in accuracy, although considered mild, suggests overfitting, particularly with respect to the depth of the random forest model with "max_depth=18." Hyperparameter tuning through grid search mitigated overfitting; however, the remaining variability in cross-validation (standard deviation of recall = 0.0889) suggests some sensitivity to data splits.

Third, reliance on static features excludes dynamic real-time inputs, such as continuous ECG monitoring. In addition, the binary target of CVD presence oversimplifies multi-class outcomes, such as disease severity. Ethical risks include over-reliance on automated predictions without oversight by clinicians, potentially resulting in false confidence in high-recall scenarios.

Finally, the Streamlit deployment method relies on stable internet access, which is not always available in low-resource settings. This highlights the need for future studies to incorporate larger, multi-center datasets and hybrid models to enhance robustness and clinical trust.[2,10]

Furthermore, external validation on independent, multi-ethnic, and multi-center datasets is necessary to confirm the model's generalizability. Key risk factors and model performance may vary significantly in different demographic or regional cohorts, potentially limiting widespread applicability until tested with a broader array of populations.

Potential sources of bias persist, including demographic imbalances and limited representation of minority or high-risk subgroups. These factors may affect fairness and model accuracy, especially if certain patient profiles are undersampled. Future work should incorporate fairness-aware algorithms and subgroup analysis to quantitatively assess and mitigate bias.

7. Conclusion

As discussed in Section 4, the model meets medical standards for CVD prediction with minimal false negatives (16). When deployed for real-time use, it leverages advanced ML algorithms and scalable cloud infrastructure to enable early detection and proactive management of CVDs. Integrating real-time data and EMRs enhances predictive accuracy and personalized healthcare delivery.

The system's design ensures accessibility, scalability, and compliance with data privacy regulations, providing patients and healthcare providers with actionable insights to potentially reduce CVD-related mortality and improve public health outcomes. While challenges such as data privacy and real-time integration remain, this study demonstrates the transformative potential of combining cloud computing and ML to strengthen preventive care frameworks.

8. Future recommendations

a. Inclusion of wearable devices and EMRs

Future studies are encouraged to integrate multimodal data from wearable devices, such as smartwatches and activity monitors, and EMRs. This incorporation may facilitate the collection of dynamic physiological signals (e.g., heart rate variability, activity) and longitudinal health records, thereby enabling highly personalized risk analysis. Longitudinal evaluation of multimodal datasets can enhance predictive models, particularly to identify at-risk populations and improve early detection rates.

b. Implementation of edge computing architectures

The implementation of edge computing can decentralize data processing to achieve low-latency prediction on wearable devices or local servers. This approach can improve responsiveness during critical care, minimize cloud infrastructure dependency, and mitigate privacy concerns by reducing data transmission, in line with existing standards for healthcare IoT systems.

c. Systematic acquisition of user interface feedback

Systematic analysis of the Streamlit application's user interface through surveys or usability studies can inform iterative design refinements. This feedback cycle would facilitate accessibility and usability for diverse stakeholders, such as clinicians and patients with varying technical expertise, and refine the interface to optimize it for real-time clinical decision-making.

d. Dataset expansion with diversified variables

By adding to the current sample dataset, additional variables—including genetic markers, lifestyle variables (e.g., dietary habits and history of smoking), environmental exposures, and socioeconomic variables—could improve the generalizability of models across heterogeneous populations. Employing federated learning techniques to aggregate data from different institutions while preserving privacy can significantly scale up the dataset, thereby improving recall (currently 94.75%) and accuracy (90.21%) levels.

e. Investigation of ensemble and hybrid models

Investigation of ensemble techniques or hybrid ML models, such as integrating deep learning with random forest, may help address the mild overfitting observed (train–test accuracy difference of 7.35%). These approaches may exploit complementary advantages of different algorithms to enhance model robustness and predictive performance.

f. Extension to other explainable AI models

In addition to using SHAP, future research could explore other explainable AI techniques, such as local interpretable model-agnostic explanations or counterfactual analysis. These techniques would further enhance understanding of model behavior at the level of individual predictions and promote adoption and trust among healthcare stakeholders.

g. Longitudinal clinical impact studies

Performing multi-year studies to determine the long-term consequences of CardioPredict AI on patient outcomes—such as reduced hospitalization rates or improved survival—could confirm its clinical utility. Incorporating the system into routine clinical practice and evaluating its effectiveness over time would provide empirical verification of its impact on CVD management.

Acknowledgments

We would like to thank the Department of Computer Science and Engineering at the Future Institute of Engineering and Management, Kolkata, for their valuable support throughout this project.

Funding

None.

Conflict of interest

Saurav Mallik is an Editorial Board Member of this journal, but was not in any way involved in the editorial and peer-review process conducted for this paper, directly or indirectly. Separately, other authors declared that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Author contributions

Conceptualization: Soumita Seth, Debangshu Bhattacharjee
Data analysis: Soumita Seth, Debangshu Bhattacharjee, Anusree Dam, Provat Mondal
Data curation: Debangshu Bhattacharjee, Anusree Dam
Funding acquisition: Saurav Mallik
Methodology: Soumita Seth, Debangshu Bhattacharjee, Anusree Dam, Provat Mondal
Supervision: Soumita Seth, Saurav Mallik, Tapas Bhadra
Writing–original draft: Soumita Seth, Debangshu Bhattacharjee
Writing–review & editing: Saurav Mallik, Soumita Seth, Tapas Bhadra

Ethics approval and consent to participate

This study used publicly available, fully anonymized secondary data. According to the policy of Future Institute of Engineering and Management, Kolkata, ethics approval and informed consent were not required.

Consent for publication

This study used fully anonymized secondary data obtained from publicly available sources. As no identifiable personal information or images were used, informed consent for publication was not required under the guidelines of the Future Institute of Engineering and Management.

Availability of data

The datasets used in this study—including the Mendeley cardiovascular disease dataset, heart.csv dataset, Statlog, heart attack prediction, and heart disease dataset—are publicly available:

(i) Cardiovascular disease dataset (Mendeley): https://data.mendeley.com/datasets/dzz48mvjht/1
(ii) Heart.csv dataset: https://figshare.com/articles/dataset/heart_csv/20236848?file=36169122
(iii) Statlog Dataset: https://archive.ics.uci.edu/dataset/145/statlog+heart
(iv) Heart disease dataset (Lapp): https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset
(v) Heart attack prediction (Anand): https://www.kaggle.com/datasets/imnikhilanand/heart-attack-prediction

The code for this study is publicly available at https://github.com/CodeRishiX/Cardiovascularprediction

References

1. Lawton JS, Tamis-Holland JE, Bangalore S, *et al*. 2021 ACC/AHA/SCAI guideline for coronary artery revascularization: A report of the American college of cardiology/American heart association joint committee on clinical practice guidelines. *Circulation*. 2022;145(3):e18-e114.
doi: 10.1161/CIR.0000000000001038
2. Al-Zaiti SS, Alghwiri AA, Hu X, *et al*. A clinician's guide to understanding and critically appraising machine learning studies: A checklist for ruling out bias using standard tools in machine learning (ROBUST-ML). *Eur Heart J Digit Health*. 2022;3(2):125-140.
doi: 10.1093/ehjdh/ztac016
3. Anusha KS, Radhika AD. A comprehensive analysis of technique's used to predict heart disease. *Int J Sci Res Comput Sci Eng Inf Technol*. 2019;5(3):380-383.
doi: 10.32628/CSEIT1953117
4. Alshraideh M, Alshraideh N, Alshraideh A, Alkayed Y, Al Trabsheh Y, Alshraideh B. Enhancing heart attack prediction with machine learning: A study at Jordan University Hospital. *Appl Comput Intell Soft Comput*. 2024;2024:5080332.
doi: 10.1155/2024/5080332
5. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30:4765-4774.
doi: 10.48550/arXiv.1705.07874
6. Su X, Xu Y, Tan Z, *et al*. Prediction for cardiovascular diseases based on laboratory data: An analysis of random forest model. *J Clin Lab Anal*. 2020;34(9):e23421.
doi: 10.1002/jcla.23421
7. Bharti S, Singh SN. Analytical Study of Heart Disease Prediction Compared with Different Algorithms. In: *Proceedings of the International Conference on Computing, Communication & Automation* (ICCCA). Greater Noida, India; 2015. p. 78-82.
doi: 10.1109/CCAA.2015.7148347
8. Purushottam, Saxena K, Sharma R. Efficient heart disease prediction system. *Procedia Comput Sci*. 2016;85:962-969.
doi: 10.1016/j.procs.2016.05.288
9. Dwivedi AK. Performance evaluation of different machine learning techniques for predicting heart disease. *Neural Comput Appl*. 2018;29:685-693.
doi: 10.1007/s00521-016-2604-1
10. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: Transforming the practice of medicine. *Future Healthc J*. 2021;8(2):e188-e194.
doi: 10.7861/fhj.2021-0095
11. Liu T, Krentz A, Lu L, Curcin V. Machine learning based prediction models for cardiovascular disease risk using electronic health records data: Systematic review and meta-analysis. *Eur Heart J Digit Health*. 2024;6(1):7-22.
doi: 10.1093/ehjdh/ztae080
12. Ghose P, Oliullah K, Mahbub MK, Biswas M, Uddin KN, Jamil HM. Explainable AI assisted heart disease diagnosis through effective feature engineering and stacked ensemble

learning. *Expert Syst Appl*. 2025;265:125928.

doi: 10.1016/j.eswa.2024.125928

13. Shah P, Shukla M, Dholakia NH, Gupta H. Predicting cardiovascular risk with hybrid ensemble learning and explainable AI. *Sci Rep.* 2025;15:17927.

doi: 10.1038/s41598-025-01650-7

14. El-Sofany H, Bouallegue B, El-Latif YM. A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method. *Sci Rep.* 2024;14(1):23277.

doi: 10.1038/s41598-024-74656-2

15. Mathew J, Pagliaro JA, Elumalai S, *et al*. Developing a multisensor-based machine learning technology (Aidar decompensation index) for real-time automated detection of post-COVID-19 condition: Protocol for an observational study. *JMIR Res Protoc*. 2025;14:e54993.

doi: 10.2196/54993

16. Dharma A, Sihombing P, Efendi S, Mawengkang H, Turnip A. Portable holter with cloud-based learning analytics for real-time health monitoring. *J Biomed Phys Eng*. 2025;15(4):393-406.

doi: 10.31661/jbpe.v0i0.2411-1856

17. Doppala BP, Bhattacharyya D. Cardiovascular_Disease_ Dataset (Version 1) [Data set], Mendeley Data. Lincoln University College. 2021.

doi: 10.17632/dzz48mvjht.1

18. Dua D, Graff C. *Heart Disease Dataset. UCI Machine Learning Repository*; 2019. Available from: https://archive.ics.uci.edu/ml/datasets/heart+disease [Last accessed on 2025 Nov 10].

19. Janosi A, Steinbrunn W, Pfisterer M, Detrano R. *Heart Disease [Dataset]. UCI Machine Learning Repository.* UCI Machine Learning Repository: California, United states of America; 1989.

doi: 10.24432/C52P4X

20. Dua D, Graff C. *Statlog (Heart) Dataset. UCI Machine Learning Repository*; 2021. Available from: https://archive.ics.uci.edu/ml/datasets/statlog+(heart) [Last accessed on 2025 Nov 10].

21. Anand N. *Heart Attack Prediction Dataset. Kaggle*; 2018. Available from: https://www.kaggle.com/datasets/imnikhilanand/heart-attack-prediction?utm_source=chatgpt.com [Last accessed on 2025 Nov 10].

22. Doppala BP, Bhattacharyya D, Janarthanan M, Baik N. A reliable machine intelligence model for accurate identification of cardiovascular diseases using ensemble techniques. *J Healthc Eng*. 2022;2022:2585235.

doi: 10.1155/2022/2585235

23. Adeyeye AC, Adedayo JS, Kolawole IA, Matanmi OG. Prediction of patients' outcomes in cardiovascular disease. *Biomed Stat Inform.* 2025;10(2):39-45.

doi: 10.11648/j.bsi.20251002.13