Journal of Computer and Communications, 2025, 13(11), 163-179
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Comparative Study of Machine Learning Techniques for Early Detection of Heart Diseases

Chaibou Kadri*, Moussa Idi Bachir, Sidi Zakari Ibrahim, Naroua Harouna, Mamadou Fougou Mamadou

Département de Mathématiques et Informatique, Faculté des Sciences et Techniques, Université Abdou Moumouni, Niamey, Niger

Email: *kadrichaibou73@gmail.com, bachir.moussaidi@yahoo.fr, sidizakariibrahim@gmail.com, hnaroua@yahoo.com, mamadoufougoumamadou@gmail.com

How to cite this paper: Kadri, C., Bachir, M.I., Ibrahim, S.Z., Harouna, N. and Mamadou, M.F. (2025) Comparative Study of Machine Learning Techniques for Early Detection of Heart Diseases. *Journal of Computer and Communications*, **13**, 163-179.
https://doi.org/10.4236/jcc.2025.1311010

Received: September 25, 2025
Accepted: November 22, 2025
Published: November 25, 2025

Abstract

Cardiovascular diseases (CVDs) are the leading cause of death worldwide, accounting for millions of deaths each year according to the World Health Organization (WHO). Early detection of these diseases is essential to reduce mortality, improve preventive care, and alleviate the burden on healthcare systems. However, traditional diagnostic approaches based on clinical assessment or risk scores have several limitations, particularly in terms of sensitivity, generalizability, and their ability to capture complex interactions among risk factors. The rise of Artificial Intelligence (AI), and particularly Machine Learning (ML), offers new opportunities for developing more effective predictive systems. This study presents a comparative analysis of five supervised ML algorithms—Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Artificial Neural Network (ANN)—applied to the early detection of heart disease. Three benchmark datasets were used: the primary Kaggle Heart Disease dataset (1,025 records, 14 clinical variables), the UCI Cleveland Heart Disease dataset (303 records), and the Framingham Heart Study dataset (4,020 records). Models were evaluated using multiple performance metrics, including accuracy, F1-score, precision, recall, ROC-AUC, and the Matthews Correlation Coefficient (MCC). Experimental results revealed that the Random Forest classifier achieved the best overall performance on the Kaggle dataset (accuracy = 99.26%, F1 = 99.28%), followed by ANN and DT (accuracy = 98.78%). On the Cleveland dataset, RF also outperformed other models (accuracy = 90.34%), while ANN and SVM reached 83.78% and 86.07%, respectively. For the Framingham dataset, RF maintained strong results (accuracy = 86.34%), confirming its robustness across heterogeneous data sources. These results highlight the importance of selecting the ap-

propriate algorithm according to reliability and sensitivity requirements in medical contexts. The study also demonstrates the importance of selecting models according to data characteristics and clinical objectives, reinforcing the potential of AI-based approaches for early and reliable cardiovascular risk prediction.

Keywords

Cardiovascular Diseases, Artificial Intelligence, Early Detection, Machine Learning Techniques, Predictive Healthcare System

1. Introduction

Cardiovascular diseases (CVDs) represent a major global public health concern. According to the World Health Organization (WHO), they are responsible for approximately 17.9 million deaths each year, accounting for 32% of global mortality [1]. These conditions encompass a wide range of disorders affecting the heart and blood vessels, including ischemic heart disease, stroke, and hypertension. The majority of these deaths occur in low and middle-income countries, where healthcare systems often lack efficient tools for early screening and effective treatment [2]. Early detection is crucial to prevent severe complications and improve patient outcomes. However, traditional diagnostic approaches, such as clinical assessments, laboratory tests, or risk scores (e.g., the Framingham score), often show limitations in their ability to model complex, nonlinear, and multidimensional relationships among risk factors [3]. In this context, Artificial Intelligence technologies, particularly Machine Learning, provide powerful alternatives capable of processing large volumes of clinical data to extract predictive patterns [4] [5].

This study presents a comparative analysis of five supervised machine-learning algorithms Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Artificial Neural Network (ANN) for the early detection of cardiovascular diseases. To ensure robustness and generalizability, three benchmark datasets were employed: the Kaggle Heart Disease Dataset, the UCI Cleveland Heart Disease Dataset and the Framingham Heart Study Dataset. These datasets collectively provide a diverse representation of clinical and demographic features associated with cardiovascular risk. The performance of each model is evaluated using several well-established metrics, including accuracy, F1-score, precision, recall, ROC-AUC, and the Matthews Correlation Coefficient (MCC). The objective is to identify the most suitable and generalizable algorithm capable of supporting automated cardiovascular disease diagnosis across heterogeneous clinical datasets.

2. Literature Review

The application of machine learning in the early detection of cardiovascular dis-

eases (CVDs) has gained significant momentum in recent years. Numerous studies have demonstrated that intelligent algorithms can improve diagnostic accuracy and provide effective decision support, often outperforming traditional risk-based approaches.

In their work on improving heart disease diagnosis, Neeraja Joshi and Tejal Dave [6] applied four ML algorithms—Linear Regression, Neural Networks, Support Vector Machines, and K-Nearest Neighbors on the Standard-Cleveland dataset. Their preprocessing pipeline included standardization, normalization, and data splitting into training and testing sets. Among the tested models, SVM achieved the highest predictive performance, highlighting its robustness for medical signal classification tasks.

Similarly, K. Vembandasamy *et al*. [7] investigated the use of the Naïve Bayes algorithm for heart disease detection and pre-diction using a dataset from a diabetes research institute in Chennai. The study was conducted with the Weka software, which provides functionalities for preprocessing, classification, and visualization. After splitting the data into 70% for training and 30% for testing, Naïve Bayes achieved an accuracy of 86.41%, demonstrating efficiency with minimal computational cost.

Talukdar and Singh [8] developed an Artificial Neural Network based approach using real hospital data collected from five institutions in Assam, India. Their model incorporated eight clinically relevant variables, selected through correlation-based analysis, and trained using a Multi-Layer Perceptron with backpropagation. The model reached 81% accuracy with satisfactory sensitivity and specificity. This study underscored the importance of rigorous preprocessing and careful feature selection, while also demonstrating the feasibility of ANN-based approaches in localized clinical environments.

On a broader scale, Vishnu Vardhana Reddy *et al*. [9] published a systematic review covering a decade of research (2014-2024) on ML and deep learning (DL) methods for cardiovascular risk prediction. Their review identified SVM, Random Forest, ANN, and Long Short-Term Memory (LSTM) networks as the most widely adopted techniques. They also discussed the shortcomings of traditional diagnostic methods such as ECG and angiography, noting their high cost, invasiveness, and limited generalizability. Key challenges identified for ML/DL adoption included class imbalance, reproducibility issues, and lack of interoperability, which need to be addressed for clinical implementation.

Chong *et al*. [10] further provided a comprehensive review of ML strategies in cardiovascular disease detection, focusing particularly on medical imaging. Their analysis showed that Convolutional Neural Networks (CNNs) outperform other methods for image-based modalities like echocardiography and MRI, while RF and SVM remain strong candidates for structured clinical datasets. The authors emphasized the importance of explainability through XAI methods such as SHAP and LIME, as well as validation rigor and model transparency. Persistent issues such as data leakage, poor reproducibility, and weak generalizability were high-

lighted as barriers to clinical translation.

Complementing these reviews, Kamble *et al*. [11] tested five ML models—RF, KNN, SVM, Decision Tree, and Logistic Regression—on the UCI Heart Disease dataset. RF achieved the highest accuracy (90%), followed by Logistic Regression (88%) and SVM (86%). Despite the dataset's small size (303 records), the study emphasized the robustness of ensemble methods, the role of structured pipelines, and the necessity of proper preprocessing, including normalization and feature selection. However, the lack of real-world validation and XAI integration were noted as limitations.

Finally, Munmun *et al*. [12] proposed a comparative classification framework using Logistic Regression, RF, and SVM for coronary heart disease detection. By combining five heart disease repositories into a hybrid dataset, they applied rigorous pre-processing, including SMOTE balancing and hyperparameter tuning. Their findings showed RF as the most accurate model (93.5%) and SVM as the most sensitive (97.5%), emphasizing the importance of selecting models based on clinical objectives. The study further recommended multi-metric evaluation and future research into real-time deployment with wearable technologies.

Taken together, these studies demonstrate the rapid progress and diverse applications of ML in CVD prediction. While ensemble methods such as RF consistently achieve strong results, ANN and CNN show promise when dealing with nonlinear and image-based data. Nevertheless, challenges such as data imbalance, reproducibility, lack of generalization, and limited explainability remain central obstacles that future research must overcome to ensure reliable and clinically applicable AI-driven cardiovascular prediction systems.

3. Methodologies

This section presents the methodological framework adopted for the early detection of cardiovascular diseases using machine learning algorithms. It includes data sources and preprocessing steps.

3.1. Proposed System

The proposed system for cardiovascular disease detection is based on a multi-stage architecture, as illustrated in **Figure 1**. The process begins with dataset acquisition, followed by rigorous preprocessing. This step includes data cleaning, handling missing values, normalization, standardization, balancing classes, and splitting the dataset into two subsets: one for training and the other for testing. The training subset is used to develop and optimize different classification models. Once trained, these models are applied to the test data to predict the patient's condition, classifying each case as either "normal" or "diseased."

3.2. Data Collection

The dataset used in this study was obtained from a publicly available Kaggle repository and includes several medically relevant features for heart disease predic-

tion, such as age, sex, cholesterol level, and blood pressure. It contains 76 attributes, including the target variable; however, many published works commonly rely on a subset of 14 attributes. The target field indicates the presence of heart disease in the patient, represented as an integer value: 0 = no disease and 1 = disease [13]. All attribute are described and listed in Table 1.

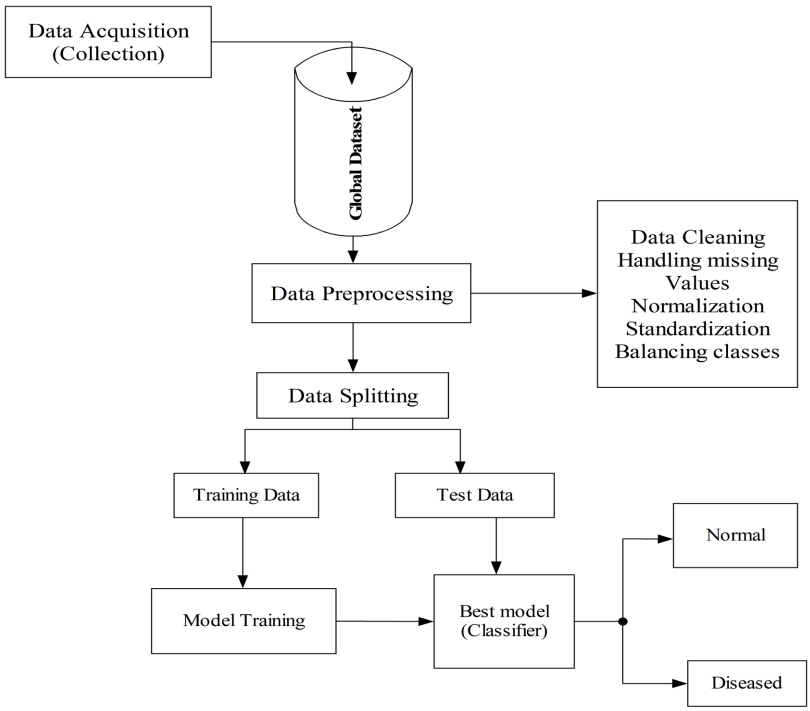

Figure 1. Architecture of the proposed system for cardiovascular disease detection.

Table 1. Description of dataset attributes.

Attribute	Description
Age	Patient's age in years.
Sex	Patient's sex (1 = male, 0 = female).
ChestPainType	Type of chest pain: (0 = Typical angina, 1 = Atypical angina, 2 = Non-anginal pain, 3 = Asymptomatic).
Resting Blood Pressure	Resting blood pressure (in mm Hg).
Cholesterol	Serum cholesterol level (in mg/dl).
Fasting Blood Sugar	Fasting blood sugar > 120 mg/dl (1 = true, 0 = false).
Resting Electrocardiogram	ECG results: (0 = Normal, 1 = ST-T wave abnormality, 2 = Left ventricular hypertrophy).
Maximum Heart Rate Achieved	Maximum heart rate achieved during exercise.
Exercise Induced Angina	Exercise-induced angina (1 = yes, 0 = no).
Oldpeak	ST segment depression induced by exercise relative to rest.
Slope of Peak Exercise ST Segment	Slope of the peak exercise ST segment: (0 = Upsloping, 1 = Flat, 2 = Downsloping).
Num Major Vessels	Number of major blood vessels colored by fluoroscopy (0 - 3).
Thal	Thalassemia: (1 = Normal, 2 = Fixed defect, 3 = Reversible defect).
Target	Heart disease diagnosis (1 = presence, 0 = absence).

In addition to the primary dataset, two other datasets were also employed for comparative purposes: the Framingham Heart Study dataset and the UCI Heart Disease dataset (heart_disease_uci), which is also available on Kaggle repository.

- ✓ Cleveland Heart Disease Dataset: A benchmark dataset from the UCI Machine Learning Repository, consisting of 303 records and 14 key attributes.
- ✓ Framingham Heart Study Dataset: A longitudinal dataset derived from an epidemiological study of more than 4,020 individuals, including a wide range of cardiovascular risk factors.

3.3. Data Preprocessing

Data preprocessing is a critical step in any machine learning pipeline, particularly in medical applications where the quality and consistency of the data directly impact the reliability of the results. In this study, several preprocessing techniques were systematically applied to ensure the robustness of the models.

3.3.1. Data Cleaning

Duplicate and invalid records were removed to prevent bias and redundancy. Outliers, identified using statistical thresholds or automatic detection methods, were either corrected or eliminated.

3.3.2. Handling Missing Values

For attributes with missing data, an appropriate imputation strategy was adopted. Mean or median imputation was used for numerical variables, while mode imputation was applied for categorical attributes.

3.3.3. Normalization

To ensure a common scale among numerical features, Min-Max normalization was applied, rescaling all values to the [0, 1] range. This is particularly beneficial for scale-sensitive algorithms such as K-Nearest Neighbors (KNN) and Artificial Neural Networks (ANN).

3.3.4. Standardization

In addition to normalization, certain variables were standardized to have a mean of zero and a standard deviation of one. This transformation improves the convergence of models such as Support Vector Machines (SVM).

3.3.5. Categorical Encoding

Non-numerical attributes were converted into numerical representations. One-hot encoding was applied to nominal variables with no intrinsic order, while ordinal encoding was used for hierarchically structured attributes.

3.3.6. Dimensionality Reduction

Correlation analysis was performed to eliminate highly redundant features. Feature selection techniques such as Random Forest Feature Importance and SelectK-Best were also employed in some cases to retain only the most relevant attributes.

3.3.7. Class Balancing

Class imbalance can significantly bias machine learning models, especially when one class (e.g., healthy patients) is overrepresented in the dataset. To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) was employed to artificially generate synthetic samples for the minority class, thereby improving the model's fairness, stability, and generalization ability.

It is important to note that SMOTE was only applied to the Cleveland and Framingham datasets, both of which exhibited a significant imbalance between positive (heart disease) and negative (non-disease) cases. The main Heart Disease Dataset (Kaggle), on the other hand, is almost balanced, therefore, there is no need for any resampling process.

To prevent data leakage and ensure valid model evaluation, the SMOTE algorithm was strictly applied before splitting the full data into training and testing subsets.

This approach ensures that all evaluation metrics including Accuracy, Precision, Recall, F1-score, ROC-AUC, and Matthews Correlation Coefficient (MCC) faithfully reflect the true generalization performance of the models without contamination from synthetic data.

3.3.8. Data Splitting

The preprocessed dataset was divided into two subsets: a training set (typically 80% of the data) and a testing set (20%). Alternatively, k-fold cross-validation was used to ensure stability and robustness of the results.

3.4. Classification Models

For the task of cardiovascular disease detection, five supervised machine-learning algorithms were selected due to their popularity in the literature, their complementary learning mechanisms, and their proven effectiveness in the medical domain [14]. Each model was trained on the preprocessed datasets and fine-tuned using a grid search strategy to optimize hyperparameters.

3.4.1. Random Forest (RF)

A Random Forest is an ensemble of multiple decision trees $T_1, T_2, \cdots, T_k$, each trained on a bootstrap sample of the data. Majority voting, in classification, obtains the final prediction:

$$y = maj\left\{T_1(x), T_2(x), \cdots, T_k(x)\right\} \tag{1}$$

This method reduces variance and increases robustness compared to a single decision tree [15].

3.4.2. Decision Tree (DT)

A decision tree recursively partitions the feature space by minimizing an impurity measure. The Gini impurity is commonly used [16]:

$$Gini(D) = 1 - \sum_{i=1}^{n} p_i^2 \tag{2}$$

where p_i is the proportion of class *i* instances in node *D*.

3.4.3. Support Vector Machine (SVM)

SVM seeks a separating hyperplane $w^{\mathrm{T}}x+b=0$ that maximizes the margin between two classes:

$$\min_{w,b}\frac{1}{2}\|w\|^2 \quad \text{subject to} \quad y_i\left(w^{\mathrm{T}}x_i+b=0\right)\geq 1 \tag{3}$$

where $\left(x_i,y_i\right)$ are the training examples. A kernel function $K\left(x_i,x_i\right)$ can be used to map data into higher-dimensional space [17].

3.4.4. K-Nearest Neighbor (KNN)

KNN predicts the label of an instance *x* based on the *k* closest neighbors using a distance function, typically Euclidean distance [18]:

$$d\left(p,q\right)=\sqrt{\sum_{j=1}^{n}\left(p_j-q_j\right)^2} \tag{4}$$

3.4.5. Artificial Neuronal Network (ANN)

A multilayer perceptron consists of several layers of neurons. Each neuron computes:

$$a_j^{(l)}=f\left(\sum_{i=1}^{n}w_{ji}^{(l)}a_i^{(l-1)}+b_j^{(l)}\right) \tag{5}$$

where f is an activation function (ReLU, sigmoid, tanh). Training involves minimizing a cost function, such as cross-entropy, via backpropagation and gradient descent optimization algorithm [19].

3.5. Confusion Matrix and Cross Validation

The confusion matrix is a standard tool for evaluating the performance of classification models by summarizing predictions into four categories: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). It provides a detailed view of classification errors, enabling deeper insights beyond global accuracy. The matrix is expressed as [20]:

$$CM=\begin{bmatrix}\mathrm{TP} & \mathrm{FP}\\ \mathrm{FN} & \mathrm{TN}\end{bmatrix} \tag{6}$$

From this matrix, performance measures such as Precision, Recall, Specificity, and F1-score are derived.

To ensure robustness and avoid overfitting, *k*-fold cross-validation was employed. In this approach, the dataset is divided into *k* equal subsets (folds). At each iteration, one-fold is used as the test set while the remaining folds are used for training. This process is repeated *k* times, and the final performance is averaged over all folds. A common choice is *k = 10*, which provides a reliable trade-off between bias and variance. Cross-validation enhances the generalization capability of the model, ensuring that results are not dependent on a single data split [21].

3.6. Performance Evaluation

To rigorously assess the performance of the classification models, multiple evaluation metrics were employed. These metrics provide complementary perspectives on model effectiveness, particularly important in medical contexts where both sensitivity and specificity are crucial. The effectiveness of the classification methods employed in this study was evaluated using Table 2, which presents the performance metrics, their corresponding formulas, and the evaluation conditions.

Table 2. Table of performance measurements, formulas, and evaluation conditions.

Performance Metrics	Formula	Description
Accuracy	$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$	Accuracy measures the proportion of correctly classified instances over the total number of instances.
Precision	$\text{Precision} = \frac{TP}{TP + FP}$	Precision evaluates the proportion of true positives among all instances predicted as positive.
Recall (Sensitivity)	$\text{Rappel} = \frac{TP}{TP + FN}$	Recall, also known as sensitivity, measures the proportion of actual positives correctly identified.
F1-Score	$\text{F1-score} = \frac{2 * \text{Precision} * \text{Rappel}}{\text{Precision} + \text{Rappel}}$	The F1-score is the harmonic mean of precision and recall, providing a balanced evaluation metric.
Matthews Correlation Coefficient (MCC)	$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TP + FN)(TN + FN)}}$	MCC is a correlation coefficient between observed and predicted classifications, particularly effective for imbalanced datasets.
ROC-AUC		The Receiver Operating Characteristic-Area Under Curve (ROC-AUC) evaluates a classifier's ability to distinguish between classes across different thresholds. A higher AUC indicates better discriminative performance.

3.7. Hyperparameter Optimization and Grid Search Strategy

To ensure optimal model performance and fair comparison, a Grid Search Cross-Validation (GridSearchCV) approach was employed for hyperparameter tuning. Each classification algorithm was systematically optimized based on its most influential hyperparameters identified in prior literature on medical prediction tasks.

The main hyperparameters tuned for each model are summarized as follows:

- Random Forest (RF):

Number of decision trees (n_estimators), maximum tree depth (max_depth), minimum samples required to split a node (min_samples_split), minimum samples required at a leaf node (min_samples_leaf).

- Decision Tree (DT):

Maximum depth (max_depth), minimum samples for node splitting (min_samples_split), and minimum samples per leaf (min_samples_leaf).

- Support Vector Machine (SVM):

Kernel type (kernel), regularization parameter (C), kernel coefficient (gamma).

- K-Nearest Neighbors (KNN):

Number of neighbors (n_neighbors), distance metric (metric), weighting function (weights).

- Artificial Neural Network (ANN-MLP):

Number of hidden layers and neurons (hidden_layer_sizes), activation function (activation), optimization solver (solver), learning rate (learning_rate), maximum number of iterations (max_iter), and regularization term (alpha).

All the models were optimized using 5-fold cross-validation, ensuring robust performance evaluation while avoiding overfitting. The best-performing configurations were selected based on the F1-score and ROC-AUC metrics to balance precision and recall in clinical prediction.

4. Experimental Results and Discussion

This section presents the experimental results obtained from the evaluation of five machine learning models (Random Forest, Decision Tree, Support Vector Machine, Artificial Neural Network, and K-Nearest Neighbors) on the Heart Disease Dataset. The results are summarized in terms of numerical performance metrics, confusion matrices, ROC curves, and comparative metrics visualization, followed by a detailed discussion.

4.1. Performance Metrics

This section presents the experimental results obtained from the evaluation of five machine learning models—Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), Artificial Neural Network (ANN), and K-Nearest Neighbors (KNN)—on the Heart Disease Dataset.

The results are summarized through numerical performance metrics, confusion matrices, ROC curves, and comparative performance visualizations, followed by a detailed discussion of each model's behavior and comparative evaluation across multiple datasets.

4.1.1. Experimental Results on the Heart Disease Dataset

The primary evaluation was conducted on the main Heart Disease Dataset comprising 1025 patient records and 14 key clinical attributes. The performance of each classifier was measured using six evaluation metrics: Accuracy, Precision, Recall, F1-Score, Area Under the ROC Curve (AUC), and the Matthews Correlation Coefficient (MCC). The computational time for model training and prediction was also recorded to assess efficiency see **Table 3**.

4.1.2. Comparative Evaluation Using External Datasets

To validate model generalizability, the different models were further evaluated using two benchmark datasets—the Cleveland Heart Disease Dataset and the Framingham Heart Study Dataset. These datasets differ in sample size, population characteristics, and variable distributions, offering insight into model robustness

under varied conditions. Table 4 reports the performance evaluation results of the classification models on the Cleveland Heart Disease Dataset, whereas Table 5 provides the corresponding results for the Framingham Heart Study Dataset.

Table 3. Performance comparison of different machine learning models on the heart disease dataset.

No.	Models	Accuracy	Precision	Recall	F1-Score	AUC (ROC)	MCC	Computation Time (s)
1	RF	99.26	99.30	99.30	99.28	99.96	98.56	2.95
2	DT	98.78	98.64	99.05	98.82	98.77	99.62	0.21
3	SVM	91.58	90.99	93.15	91.97	97.42	83.33	1.44
4	ANN (MLP)	98.78	98.64	99.06	98.82	99.51	97.61	22.36
5	KNN	84.51	84.40	86.78	85.32	94.88	69.12	0.41

Table 4. Experimental results on the cleveland heart disease dataset.

No.	Models	Accuracy	Precision	Recall	F1-Score	AUC (ROC)	MCC	Computation Time (s)
1	RF	90.34	90.20	90.20	90.66	90.96	90.06	5.41
2	DT	82.97	84.98	85.55	84.82	89.76	83.92	0.42
3	SVM	86.07	85.99	78.25	77.36	87.42	81.43	2.27
4	ANN (MLP)	83.78	85.74	88.10	84.72	89.61	82.78	34.21
5	KNN	80.51	78.50	87.66	82.22	85.69	80.26	0.48

Table 5. Experimental results on the framingham heart study dataset.

No.	Models	Accuracy	Precision	Recall	F1-Score	AUC (ROC)	MCC	Computation Time (s)
1	RF	86.34	89.78	89.77	88.98	89.96	89.66	9.38.43
2	DT	81.60	82.95	83.44	82.22	86.65	80.82	0.48
3	SVM	84.07	75.89	72.33	73.16	82.69	82.97	16.43
4	ANN (MLP)	82.26	83.34	83.13	83.79	87.91	83.18	68.92
5	KNN	74.60	73.20	74.06	71.69	82.49	79.88	1.11

4.1.3. Confusion Matrices

To validate model generalizability, the different models were further evaluated using two benchmark datasets—the Cleveland Heart Disease Dataset and the Framingham Heart Study Dataset. These datasets differ in sample size, population characteristics, and variable distributions, offering insight into model robustness under varied conditions.

The confusion matrices provide further insights into the classification performance of the models. As shown in Figure 2, the Random Forest achieved the most accurate predictions with nearly perfect classification, while ANN and Decision Tree also demonstrated strong performance. The SVM achieved moderate results,

and KNN showed the weakest performance due to a higher number of false positives.

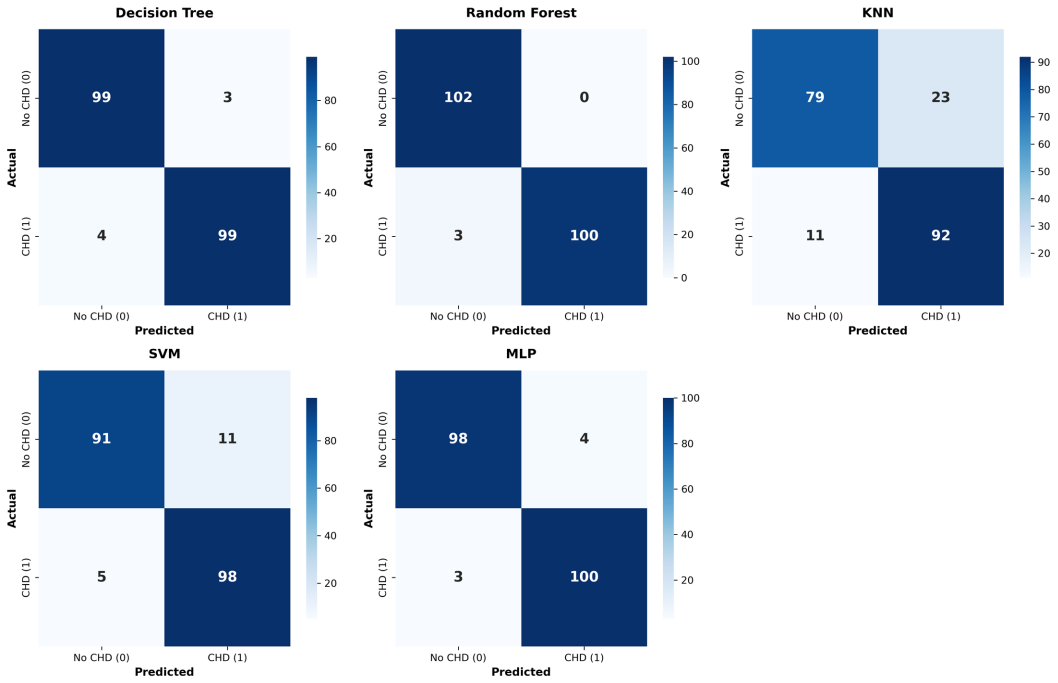

Figure 2. Confusion matrices for the five models.

4.1.4. ROC Curve Analysis

The ROC curve highlights the discriminative power of a model. As shown in **Figure 3**, the Random Forest achieved an AUC of 1.000, indicating perfect separation of classes. The ANN (MLP) followed closely with an AUC of 0.991, while the Decision Tree achieved an AUC of 0.985. The SVM and KNN models showed lower AUC values of 0.963 and 0.949, respectively, reflecting reduced discriminative capabilities. Overall, ensemble methods and neural networks outperformed distance-based algorithms for this classification task.

4.1.5. Comparative Metrics Visualization

Figure 4 presents a comparative bar chart of the performance metrics for the five models. The results confirm the superiority of Random Forest, which consistently achieved the highest scores across all metrics, followed closely by ANN (MLP) and Decision Tree. These models demonstrated balanced performance in terms of both sensitivity (Recall) and specificity (Precision), making them highly suitable for medical decision support. In contrast, SVM exhibited moderate performance with variability across metrics, while KNN showed the weakest results, particularly in terms of MCC, indicating reduced reliability in handling imbalanced or complex data. This visual comparison reinforces the earlier conclusions drawn from the confusion matrices and ROC curves, highlighting ensemble and neural network methods as the most effective approaches for cardiovascular disease prediction.

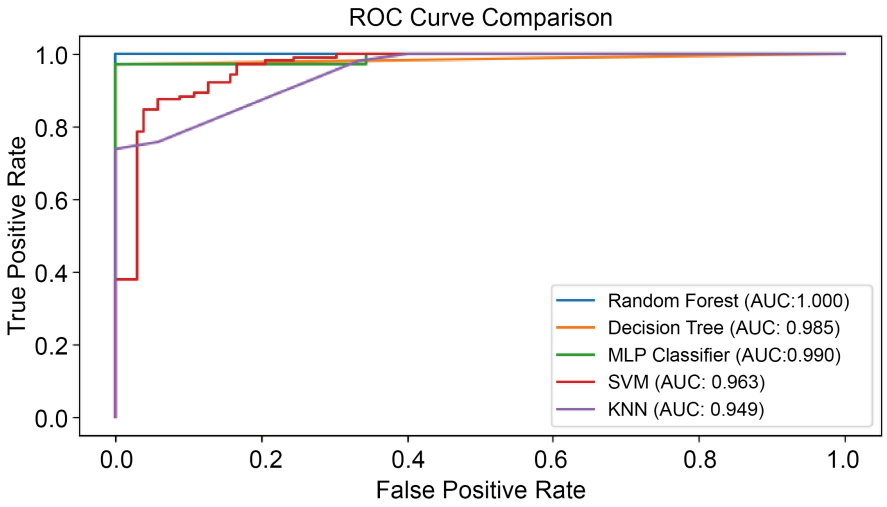

Figure 3. ROC curves comparison of the five models.

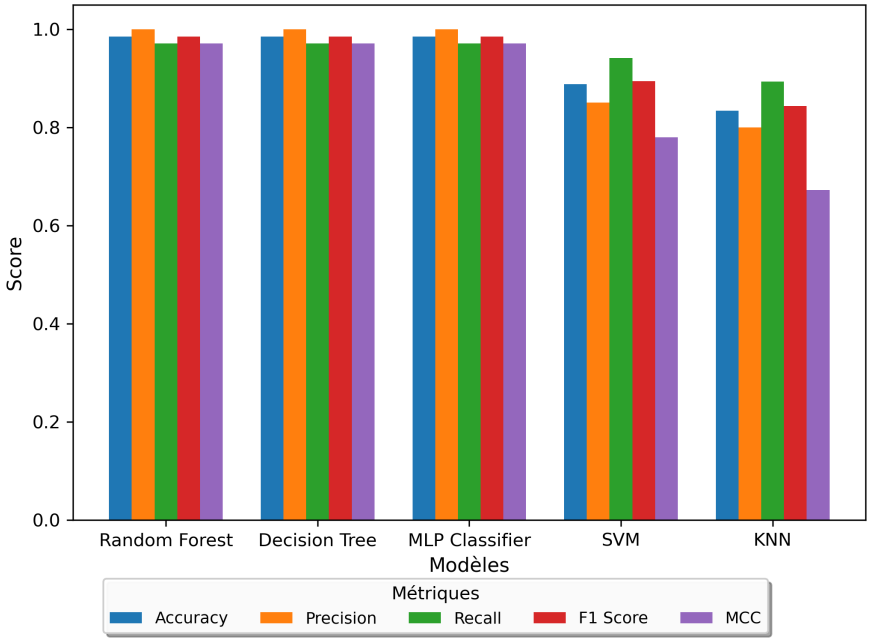

Figure 4. Comparative performance metrics of the five models.

4.2. Discussion

The experimental evaluation demonstrated that the proposed machine learning framework achieved excellent predictive performance on the Heart Disease Dataset. Among the five classifiers tested, Random Forest (RF) emerged as the most effective, reaching an accuracy of 99.26%, an F1-score of 99.28%, and an almost perfect AUC of 1.000. Artificial Neural Network (ANN) and Decision Tree (DT) models also produced highly competitive results, both achieving accuracies close to 98.78%. By contrast, Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) performed less effectively, with accuracies of 91.58% and 84.51%, respectively.

These findings confirm the robustness of ensemble-based methods, particularly Random Forest, in handling high-dimensional clinical data with complex interactions between features. The outstanding performance of RF can be attributed to its ensemble nature, which reduces overfitting and leverages feature subsampling to capture non-linear relationships. Similarly, ANN performed well due to its ability to model non-linear interactions through hidden layers, while DT benefited from its interpretability and simplicity.

On the other hand, the relatively lower performance of SVM and KNN reflects their sensitivity to feature scaling and noise, as well as their weaker generalization ability when dealing with imbalanced or high-dimensional datasets. Moreover, these algorithms tend to be less effective in capturing non-linear and hierarchical relationships that are typical of biomedical data, where interactions between physiological variables such as age, cholesterol, blood pressure, and ECG features often follow complex, non-linear patterns. In contrast, ensemble methods and deep learning models can capture higher-order dependencies through feature aggregation and hierarchical representation learning.

When compared to existing literature, the results align with recent studies emphasizing the superiority of RF and ANN for cardiovascular disease prediction. For instance, Kamble *et al*. [11] reported an RF accuracy of 90% on the UCI dataset, while Talukdar & Singh [8] obtained 81% accuracy using ANN on hospital-collected data. Munmun *et al*. [12] also found RF and SVM to be strong contenders, with RF achieving 93.5% accuracy. Our results surpass these benchmarks, likely due to the larger dataset used (14 features, 1025 samples), rigorous preprocessing steps (normalization, standardization, imputation, and SMOTE balancing), and hyperparameter optimization via grid search.

Nevertheless, the near-perfect performance reported here must be interpreted with caution. The possibility of optimistic bias due to oversampling (SMOTE), cross-validation without strict fold separation during preprocessing, or potential data leakage cannot be excluded. As highlighted by Chong *et al*. [10], many machine learning models for cardiovascular disease prediction demonstrate high in-sample accuracy but struggle with reproducibility on external datasets. To address this limitation, future work should prioritize external validation on independent cohorts such as the Framingham Heart Study and Cleveland Heart Disease Dataset, while employing nested cross-validation to eliminate bias in model selection.

From a clinical perspective, predictive models for heart disease must not only achieve high accuracy but also minimize false negatives, as missing high-risk patients could lead to severe outcomes. In this regard, SVM, despite lower accuracy, demonstrated relatively high recall, highlighting its potential utility when sensitivity is prioritized. Furthermore, explainability remains critical for real-world adoption in clinical settings. While Random Forest provides global feature importance, such transparency alone is insufficient for medical decision-making, where clinicians must understand the reasoning behind each individual predic-

tion. Explainable AI (XAI) techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) enable physicians to interpret how each clinical feature such as blood pressure, cholesterol level, or ECG abnormalities contributes to the final risk classification.

This interpretability allows clinicians to verify the model's logic against established medical knowledge, ensuring that predictive reasoning aligns with known physiological mechanisms rather than statistical artifacts. Beyond interpretability, XAI plays a key role in building clinical trust by making algorithmic decisions transparent, reproducible, and verifiable. It also facilitates error detection, supports regulatory compliance, and promotes responsible deployment of AI systems in healthcare. In fact, for any human–machine interaction in medicine, trust is paramount, and explained individual predictions are essential to achieving and sustaining that trust [22].

Such transparency not only fosters trust and accountability but also facilitates collaborative decision-making between AI systems and healthcare professionals. Integrating interpretable ML pipelines is therefore essential to bridge the gap between algorithmic accuracy and clinical usability, paving the way for safe, transparent, and trustworthy AI-driven decision-support systems in cardiovascular healthcare.

In summary, this study demonstrates that RF and ANN outperform traditional classifiers for heart disease prediction, confirming findings in prior research while achieving higher performance metrics. However, to ensure clinical applicability, future investigations should emphasize reproducibility through external validation, statistical significance testing, model calibration, and explainability. This will facilitate the translation of machine learning models from experimental evaluation into reliable and trustworthy tools, for cardiovascular risk prediction in healthcare practice.

5. Conclusions

Cardiovascular diseases remain the leading cause of mortality worldwide, highlighting the urgent need for accurate and early detection tools. In this study, we conducted a comparative analysis of five supervised machine learning algorithms—Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), Artificial Neural Network (ANN), and K-Nearest Neighbors (KNN)—using the Heart Disease Dataset, complemented by the Cleveland and Framingham datasets for reference.

The experimental findings revealed that ensemble and neural-based approaches deliver the most reliable predictive performance. Random Forest achieved the highest results, with an accuracy of 99.26%, an F1-score of 99.28%, and an almost perfect AUC of 1.000, closely followed by ANN and DT, both of which reached 98.78% accuracy. In contrast, SVM and KNN showed lower predictive capabilities, with accuracies of 91.58% and 84.51%, respectively. The analysis of confusion matrices and ROC curves further confirmed the superiority of ensemble and neu-

ral models in capturing the complex, non-linear relations among cardiovascular risk factors.

When compared to previous works, our results align with the growing consensus in the literature that RF and ANN represent the most effective techniques for clinical decision support in cardiovascular disease prediction. However, the near-perfect accuracy observed in this study must be interpreted with caution. Issues such as potential data leakage, oversampling effects, and lack of external validation could contribute to optimistic estimates.

Future research should therefore explore external validation on independent cohorts, employ robust evaluation strategies such as nested cross-validation, and incorporate explainable AI (XAI) methods to enhance clinical interpretability. Furthermore, attention should be given to model calibration and decision curve analysis to ensure that predictions align with clinical utility and patient safety.

Finally, this study demonstrates that ensemble methods, particularly Random Forest, hold significant promise for the early detection of cardiovascular diseases. Nevertheless, the path toward clinical deployment requires not only technical performance but also transparency, robustness, and reproducibility. Addressing these challenges will be essential to transform machine learning models from experimental tools into trustworthy, real-world solutions that can support clinicians in reducing the global burden of cardiovascular diseases.

Acknowledgements

This work is supported by the Abdou Moumouni University of Niamey (Niger Republic), through its grant for study trips and academic exchange.

Conflicts of Interest

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

References

[1] World Health Organization (2023) Cardiovascular Diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] Chong, B., Jayabaskaran, J., Jauhari, S.M., *et al*. (2024) Global Burden of Cardiovascular Diseases: Projections from 2025 to 2050. *European Journal of Preventive Cardiology*, **32**, 1001-1015.

[3] D'Agostino, R.B., Vasan, R.S., Pencina, M.J., Wolf, P.A., Cobain, M., Massaro, J.M., et al. (2008) General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study. *Circulation*, **117**, 743-753. https://doi.org/10.1161/circulationaha.107.699579

[4] Deo, R.C. (2015) Machine Learning in Medicine. *Circulation*, **132**, 1920-1930. https://doi.org/10.1161/circulationaha.115.001593

[5] Weng, S.F., Reps, J., Kai, J., Garibaldi, J.M., *et al*. (2017) Can Machine-Learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data? *PLOS One*, **12**, e0174944.

[6] Joshi, N. and Dave, T. (2025) Improved Accuracy for Heart Disease Diagnosis Using

Machine Learning Techniques. *Journal of Informatics and Web Engineering*, **4**, 42-52. https://doi.org/10.33093/jiwe.2025.4.1.4

[7] Vembandasamy, K., Sasipriya, R. and Deepa, E. (2015) Heart Diseases Detection Using Naive Bayes Algorithm. *International Journal of Engineering Technology and Scientific Innovation*, **2**, 441-444.

[8] Talukdar, J. and Singh, T.P. (2023) Early Prediction of Cardiovascular Disease Using Artificial Neural Network. *Paladyn, Journal of Behavioral Robotics*, **14**, Article 20220107. https://doi.org/10.1515/pjbr-2022-0107

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

[10] Chong, L., Husain, G., Nasef, D., Vathappallil, P., *et al*. (2025) Machine Learning Strategies for Improved Cardiovascular Disease Detection. *Medical Research Archives*, **13**, 1-16.

[11] Kamble, V.B., Gulabani, B., Narkhede, S. and Godse, S. (2025) Predicting Heart Disease with Machine Learning: Enhancing Accuracy through Algorithmic Approach. *International Journal of Computers and Applications*, **183**, 12-18.

[12] Munmun, Z.S., Akter, S. and Parvez, C.R. (2025) Machine Learning-Based Classification of Coronary Heart Disease: A Comparative Analysis of Logistic Regression, Random Forest, and Support Vector Machine Models. *Open Access Library*, **12**, 1-12. https://doi.org/10.4236/oalib.1113054

[13] Smith, J. (2019) Heart Disease Dataset. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset/data

[14] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I. and Chouvarda, I. (2017) Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*, **15**, 104-116. https://doi.org/10.1016/j.csbj.2016.12.005

[15] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/a:1010933404324

[16] Quinlan, J.R. (1996) Improved Use of Continuous Attributes in C4.5. *Journal of Artificial Intelligence Research*, **4**, 77-90. https://doi.org/10.1613/jair.279

[17] Cortes, C. and Vapnik, V. (1995) Support-Vector Networks. *Machine Learning*, 20, 273-297. https://doi.org/10.1007/bf00994018

[18] Cover, T. and Hart, P. (1967) Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, **13**, 21-27. https://doi.org/10.1109/tit.1967.1053964

[19] LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep Learning. *Nature*, **521**, 436-444. https://doi.org/10.1038/nature14539

[20] Kohavi, R. (1995) A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: *Proceedings of the* 14*th International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, 1137-1145.

[21] Fawcett, T. (2005) An Introduction to ROC Analysis. *Pattern Recognition Letters*, **27**, 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

[22] Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) Why Should I Trust You?. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 1135-1144. https://doi.org/10.1145/2939672.2939778