GJR Publication
OPEN ACCESS

Global Journal of Research in Engineering & Computer Sciences
ISSN: 2583-2727 (Online)
Volume 06 | Issue 04 | July-August | 2026
Journal homepage: https://gjrpublication.com/gjrecs/

Review Article

# Review of an Artificial Intelligence Based System for Early Prediction of Heart Diseases

*Dr. Virendra Kumar Swarnkar [1] and Dr. Ghanshyam Sahu [2]

[1] Professor, Computer Science & Engineering, Bharti Vishwavidyalya, Durg, Chhattisgarh, India.
[2] Assistant Professor, Computer Science & Engineering, Bharti Vishwavidyalya, Durg, Chhattisgarh, India.

**DOI: 10.5281/zenodo.22065603** **Submission Date: 15 June 2026 | Published Date: 24 Aug. 2026**

***Corresponding Author: Prof (Dr.) Virendra Kumar Swarnkar**
Professor, Computer Science & Engineering, Bharti Vishwavidyalya, Durg, Chhattisgarh, India.

## *Abstract*

*Cardiovascular diseases are still the leading cause of premature morbidity and mortality, risk recognition delayed continues to impede care. A data-driven approach to identify clinically relevant risk patterns from commonly recorded variables for artificial intelligence. Nonetheless, prediction systems developed for use in health care settings need to be accurate, interpretable and reported transparently. The objective of the study was to develop and assess a hybrid artificial intelligence model for early prediction of heart disease based on structured clinical indicators. The study used a quantitative predictive modelling design with a public Cleveland-format heart-disease dataset containing 303 records, 13 clinical predictors and a binary outcome. The analytical pipeline included data inspection, stratified train-test splitting, preprocessing, supervised model training, five-fold cross-validation, hold-out evaluation, and feature-importance interpretation. The tool used as machine learning classifiers include Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbours and Gradient Boosting. The model performance is checked using accuracy, precision, recall, specificity, F1 score, ROC-AUC and confusion-matrix. Results conveyed that Random Forest gave the best hold-out discrimination performance with an accuracy of 0.820, recall of 0.939, F1-score of 0.849 and ROC-AUC of 0.903. The most influential predictors were chest pain type, ST depression, maximum heart rate, thalassemia-related status, cholesterol, and number of major vessels. The proposed AI-based framework will technically make it feasible to predict heart diseases' risk at an early age through structured clinical variables. Yet, it requires extensive software and hardware testing, plus human oversight before it can be deployed into patient care.*

***Keywords:*** *Heart disease prediction, machine learning, early diagnosis, Artificial intelligence, clinical decision-support system, predictive analytics, cardiovascular risk.*

## 1. Introduction

Cardiovascular disease is a major public health and clinical issues. According to the World Health Organization, 2022 witnessed close to approx 19.8 million deaths from cardiovascular diseases (CVD). CVDs were responsible for almost a third of total deaths globally, with most deaths from heart attack and stroke (World Health Organization [WHO], 2025). Indeed, these findings endorse a prevention-oriented research agenda whereby high-risk people are identified prior to either irreversible myocardial injury or advanced heart failure or fatal vascular events. The clinical problem is not merely to treat already established disease but rather to pick up actionable risk signals early enough to guide lifestyle intervention, medical assessment and control of risk-factors.

The regular assessment of cardiovascular risk events based on clinical history, blood pressure, lipid profile, diabetes status, smoking exposure, family history, electrocardiography and physician assessment. The clinical interpretability and population testing of guideline-based risk estimation are essential to its continued use. For instance, the ACC/AHA prevention guideline stresses the importance of risk stratification, shared decision making, and control of modifiable risk factors (Arnett et al., 2019). Despite this, traditional scoring tools may be less sensitive to nonlinear correlations between variables, may behave differently with respect to populations, and may not self-learn in heterogeneous clinical data. This

failure has led to a growing interest in artificial intelligence (AI) and machine learning (ML) as tools for cardiovascular screening.

AI-based prediction systems are particularly appropriate for heart-disease screening because many of the risk indicators already exist in routine clinical records. Age, sex, resting blood pressure, cholesterol, fasting blood sugar, exercise-induced angina, ECG findings, maximum heart rate, ST depression, and other variables can be processed using supervised learning algorithms to estimate the probability of disease. Unlike simple threshold-based rules, ML algorithms can capture nonlinear feature interactions and variable importance patterns. This flexibility, however, carries risks including overfitting, poor transportability and opacity (Rajkomar et al., 2019; Luo et al., 2016). Consequently, AI systems in medicine need to be assessed according to both predictive and clinical criteria.

The primary objective of this paper is to develop and assess a transparent AI-enabled framework for the early detection of heart diseases using a real-world publicly available dataset. Instead of unsupported claims about accuracy, the study reports on actual model results from a reproducible supervisedlearning experiment. The design features of the paper include explainability, risk categorisation, and human-in-the-loop clinical decision support. This orientation aligns with current reporting requirements for clinical prediction models that employ regression or machine-learning techniques (Collins et al., 2024).

Four significant contributions are made by the study. An end-to-end artificial intelligence workflow for structured heart-disease prediction is presented. Additionally, it defines the preprocessing and evaluation protocol that all the classifiers will use. The next step is to review the best-performing model using feature-importance analysis, which can help support clinical transparency. The architecture proposed, differentiates the experimental prediction and deployable medical decision-making. This is the fourth point.

## 2. Literature Review

### 2.1 Heart Disease Prediction Using Machine Learning Models

The earliest algorithms concerning coronary artery disease were the basis for developing probability-based diagnostic tools that combined structured clinical variables. Detrano et al. developed and evaluated an algorithm to diagnose coronary artery disease from international clinical data and established a historical benchmark for later datasets. Through classifiers optimizing predictive separation between the disease and non-disease classes, modern studies in ML have extended this. Despite being commonly used as a benchmark, the small size of the UCI Heart Disease dataset necessitates prudent generalisation claims (Janosi et al, 1988).

Ensemble methods are often appealing in the prediction of heart disease since they capture non-linear effects and reduce variance compared with single trees. Random Forest combines many decision trees and usually does well with tabular clinical data (Breiman, 2001). Gradient boosting methods sequentially adjust weak learners and can supply strong discrimination with the correct tuning (Friedman, 2001; Chen &Guestrin, 2016). Nonetheless, the performance of the model which is reported on benchmark dataset maybe overestimated due to feature selection or scaling or hyperparameter tuning leaking test data information into training.

### 2.2 Deep Learning for Cardiovascular Risk Prediction

Deep learning has proven effective in areas such as cardiac imaging and continuous ECG monitoring. However, the benefit of deep learning in early prediction of risk using small tabular variables is less consistent. A dataset encompassing merely a few hundred records does not offer adequate statistical support for training artificial neural networks containing numerous parameters. Thus, this study considers neural-network modelling as a potential avenue for future work with larger and noisier datasets rather than as a requirement of the present benchmark.

Differentiating between model novelty and clinical utility is important. A complex architecture that may or may not be technically advanced, but does not improve calibration, recall, or interpretability in a clinically meaningful sense. Biomedical ML is not just about writing good algorithms. The reporting guidelines stress good data handling and validation. Likewise, they recommend giving a clear explanation of intended use (Collins et al., 2024; Luo et al., 2016).

### 2.3 Clinical Risk Factors and Feature Selection

Feature selection in cardiovascular predictions may serve two purposes: to improve computational efficiency and to clarify which variables are most important for making decisions. Through correlation analysis, recursive feature elimination, chi-square tests, and model-based feature importance, you can identify your target's strongest associate. Nevertheless, we have to be careful in interpreting the feature importance clinically because being important in a model does not mean having a causal effect.

In recent cardiovascular prediction research, SHAP and related methods were used to enhance interpretability. El-Sofany et al. (2024) have integrated predictive analytics and feature selection with SHAP-based interpretation approaches for cardiovascular disease detection. According to Shah et al. (2025), clinical and lifestyle factors are essential for

interpretation. These studies lend credence to the interpretable-influenced AI path, but they also illustrate the need for external validation and deployment testing before algorithms impact patient care.

## 2.4 Explainable AI in Medical Prediction Systems

Explanation of medical predictions must be human-interpretable and domain-knowledge consistent.

LIME provides local model interpretations with interpretable surrogate models, while SHAP assigns an output of the model to the features (through game-theoretic principles) (Lundberg & Lee, 2017; Ribeiro et al., 2016). In the current study, feature importance is utilized as a global interpretation method that is transparent in nature. Local explanation methods have been proposed as part of the decision-support architecture, but not overclaimed as offering clinical proof.

Validity must not be conflated with explainability. Just because something seems visually plausible, doesn't mean the model is calibrated, unbiased or clinically safe. According to Tjoa and Guan (2021), medical XAI must take into account reliability, user understanding and the clinical context of explanation consumption. Hence, the proposed system would use explanations within a clinician-reviewed workflow rather than an automated diagnosis pathway.

## 2.5 Comparative Performance of AI Models in Healthcare

Every algorithm will not be best suited for every healthcare dataset hence comparative evaluation is required Logistic Regression tends to do well if the relationships are approximately linear and the variables well-chosen. Tree-based ensembles are able to model non-linear interactions, however, they may over-fit on small training sets. Boundary hunting mechanisms of SVMs provide good results for moderate dimensional data but are sensitive to scaling and kernel parameters. KNN is a simple and intuitive algorithm, but it is sensitive to feature scaling. Thus, performance comparison should be run on the same splits, preprocessing pipeline and metrics.

Clinically meaningful errors should be a focus of biomedical evaluation. In heart disease screening, a false negative can delay further investigation of a potentially diseased patient. A false positive can create anxiety, more testing, and increased clinical burden. Accuracy alone can obscure these trade-offs especially under class imbalance. Thus, this paper reports the recall, specificity, F1-score, ROC-AUC, and confusion matrix.

## 2.6 Research Gaps Identified from Existing Studies

Research papers show persistent appeal of AI-based cardiovascular prediction, but many gaps persist. You can find an expert to do your homework for you and get good results. Furthermore, interpretability is sometimes described as a visual optional add-on rather than a requirement for clinical decision support. Third, benchmark datasets are often small, retrospective and demographically narrow. In the fourth instance, there is limited literature connecting model development to an actual clinical workflow that includes privacy, clinician review, alerts, and reporting. Finally, there is a continuing gap between experimental performance and a deployable regulated medical software.

This paper fills these gaps through real-model benchmarking and transparent reporting, feature-importance interpretation, and an implementation-oriented decision-support architecture. The study does not declare clinical readiness; it defines necessary validation and proves technical feasibility for responsible development5. Problem Statement

Despite improvements in cardiovascular diagnostics, the early prediction of heart disease remains difficult because risk arises from the interaction of biological, lifestyle, and clinical factors. Numerous patients may be asymptomatic until the illness becomes clinically significant, while some present with non-specific that need interpreting. Clinical variables like cholesterol, blood pressure, ECG, and types of chest pain are predictive but usual methods may not be harnessing non-linear combinations of these.

AI-enabled services may solve this problem by learning the predictive patterns of structured data to offer risk predictions to clinicians. Being difficult to interpret, insufficiently validated, or reported with unrealistic performance claims are several existing models of ECM. As a result, the difficulty is not just constructing a high-accuracy classifier. It is the transparent AI-based prediction framework that is technically reliable, clinically interpretable, and honest about validation boundaries.

## 3. Investigative Discrepancy

Lack of a complete and transparent pipeline linking early heart-disease prediction, interpretable ML benchmarking, and clinical decision-support design is the main research gap. Studies fail to recognize the importance of risk communication, data governance and prospective validation for algorithm performance. Although recall, specificity, calibration, and explanation are more important to screening decisions, they tend to use accuracy as the dominant metric.

There is another gap regarding integration. A model may be successful in a notebook environment but may not be successful clinically if it is not deployed with an interpretable interface, if patient data is not protected, and if the

clinician does not understand the predictions. The response treated the AI system, not as a classifier but as a sociotechnical clinical-support architecture.

## 4. Research Methodology

### 4.1 Research Design

The design of the study is quantitative experimental predictive-modelling the aim is to design and assess supervised ML classifiers for predicting heart disease binary. The workflow consists of getting the data, exploring it, preprocessing it, stratified splitting, model training, cross-validation, hold-out testing, feature-importance analysis and system-architecture design.

The development of transparent models is followed in a design. Preprocessing steps fitted only on the training data within model pipelines, thus limiting the scope of data leakage. Since clinical screening requires sensitivity to error trade-offs, we report performance via multiple complementary metrics rather than just accuracy.

### 4.2 Description of the Dataset

The Cleveland-format heart-disease dataset is public data extracted from the Heart Disease data collection hosted by UCI. According to the UCI repository, the Heart Disease data set has 303 instances. Although there are 76 attributes on related databases, mainly experiments published only use a 14-variable specification containing the target attribute (Janosi et al., 1988). CSV file contained in the present article has 303 rows and 13 predictors along with one binary output variable.

The dataset contains clinically interpretable variables like age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting ECG result, maximum heart rate, exercise-induced angina, ST depression, ST slope, the number of major vessels, thalassemia-related status and disease outcome. Thus, it is useful for methodological demonstration. The distribution of the target variable contained more positive than negative examples: specifically, there are 165 positive cases versus 138 negative cases. Thus, the target distribution exhibited a moderate degree of imbalance, as opposed to a severe imbalance.

### 4.3 Variables of the Study

Table 1 summarizes the clinical meaning and expected relationship of the variables used in the prediction experiment. These expectations are based on clinical plausibility and do not imply causal inference from this dataset alone.

**Table 4.1. Description of Variables Used for Heart Disease Prediction**

| Variable Name | Type | Clinical Meaning | Expected Relationship with Heart Disease Risk |
|---|---|---|---|
| Age | Continuous | Age in years | Older age generally increases cardiovascular risk. |
| Sex | Categorical | Biological sex encoded in the dataset | Sex-related differences may alter baseline risk patterns. |
| Cp | Categorical | Chest pain type | Angina-related categories may indicate higher ischemic risk. |
| Trestbps | Continuous | Resting blood pressure in mmHg | Higher blood pressure is associated with vascular risk. |
| Chol | Continuous | Serum cholesterol in mg/dL | Higher cholesterol can indicate atherogenic burden. |
| Fbs | Binary | Fasting blood sugar above 120 mg/dL | Hyperglycaemia may indicate cardiometabolic risk. |
| Restecg | Categorical | Resting electrocardiographic result | Abnormal ECG may reflect cardiac stress or pathology. |
| Thalach | Continuous | Maximum heart rate achieved | Lower exercise capacity can be associated with higher risk. |
| Exang | Binary | Exercise-induced angina | Presence suggests ischemic symptoms during stress. |

| | | | |
|---|---|---|---|
| Oldpeak | Continuous | ST depression induced by exercise | Greater depression can suggest ischemic burden. |
| Slope | Categorical | Slope of peak exercise ST segment | Certain slopes may indicate abnormal stress response. |
| Ca | Discrete | Number of major vessels colored by fluoroscopy | Higher vessel involvement may indicate greater coronary burden. |
| Thal | Categorical | Thalassemia/stress-test-related status in the dataset | Abnormal categories may correlate with heart disease presence. |
| Outcome | Binary target | 0 = absence; 1 = presence of heart disease | Target variable for supervised classification. |

## 4.4 Data Preprocessing

Data preprocessing was designed to preserve clinical interpretability while supporting fair algorithm comparison. The dataset was inspected for missing values and target distribution. No missing values were detected in the CSV file used for analysis. Continuous and ordinal values were retained, and categorical variables were treated according to their integer-coded representation in the public dataset. For algorithms sensitive to scale, standardisation was applied inside the modelling pipeline.

An 80:20 stratified train-test split was used to preserve the class distribution in the hold-out test set. Five-fold stratified cross-validation was performed as a supplementary robustness check. The moderate target imbalance was addressed through stratification and class-weight balancing where appropriate. Table 2 reports the preprocessing summary.

**Table 4.2. Dataset Preprocessing Summary**

| Preprocessing Element | Procedure Applied | Rationale |
|---|---|---|
| Data source | UCI Heart Disease/Cleveland-format public dataset mirror; 303 records and 13 predictors plus binary outcome. | Public secondary dataset; no direct patient contact. |
| Missing values | No missing values were detected in the CSV file used for the experiment. | No imputation was applied. |
| Encoding | Categorical variables already appeared as integer-coded clinical categories. | Values were retained and passed through preprocessing pipeline. |
| Scaling | StandardScaler was applied inside each model pipeline for scale-sensitive algorithms. | Prevents leakage by fitting scalers on training folds only. |
| Split strategy | 80:20 stratified train-test split; training n = 242, test n = 61; random_state = 42. | Class ratio was preserved across split. |
| Cross-validation | Five-fold stratified cross-validation was applied on the full dataset for supplementary robustness assessment. | Results are reported separately from hold-out testing. |
| Class imbalance | Outcome distribution was 165 positive and 138 negative cases. | Moderate imbalance; class_weight="balanced" used where appropriate. |

## 4.5 Feature Selection

Correlation inspection and model-based feature importance were used for interpretation. Correlation analysis between these variables and the outcome variable (the Var) helps to identify the associations between numerical variables. Further, the Random Forest analysis ranks features globally according to their contribution to the prediction. Both recursive feature elimination and SHAP would be appropriate for future work, but were not conducted to support unsupported claims in this version.

The selection of features in clinical predictions must be done carefully. A feature can be crucial for discrimination in a dataset due to clinical relevance, the pattern of measurement, confounding, or the composition of the dataset. As a result, feature importance are a model behavior and not an evidence of biology.

## 4.6 Machine-Learning Algorithms

A total of six supervised-learning algorithms were implemented which are: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbours, Gradient Boosting. Logistic Regression provides us with an interpretable linear baseline. While the Decision Tree provides a clear rule-based structure, it is sensitive to small samples. Random Forest amounted to a collection of bootstrapped decision trees, which helped reduce the variance and capture nonlinear interactions (Breiman, 2001).

Support vector machine was included because it can model nonlinear boundaries using kernel methods when the variable is scaled sensibly (Cortes &Vapnik, 1995). KNN represented a method based on similarity. Gradient Boosting is a sequential ensemble method that can improve weak learners (Friedman, 2001). XGBoost was indicative as another related boosting framework, but was not used in the ultimate benchmark as the standard scikit-learn Gradient Boosting implementation was sufficient for reproducible comparison in this small dataset (Chen &Guestrin, 2016; Pedregosa et al., 2011).

## 4.7 Model Evaluation Metrics

Model evaluation was done using threshold and ranking based metrics. Accuracy gauges the percentage of correct predictions. However, it can be misleading when classes are imbalanced or when false negatives and false positives have different clinical consequences. This states that how many of the predicted positive cases were really positive. Recall, also known as sensitivity, refers to the proportion of actual disease cases which are recognized correctly. Specificity measures the proportion of disease-free cases that are correctly identified.

The calculations that were made were: The formulas used to accuracy = (TP + TN)/(TP + TN) + FP + FN, Precision =(TP/TP) + FP, Recall= (TP/TP) + FN, Specificity = (TN/TN)+FP, F1-score = 2 x(P x R)/(P +R) }} ROC-AUC indicates the model's ability to rank positive cases above negative cases across thresholds. In an early-stage diagnosis context, recall carries special importance as false negatives may cause delays in clinical attention.

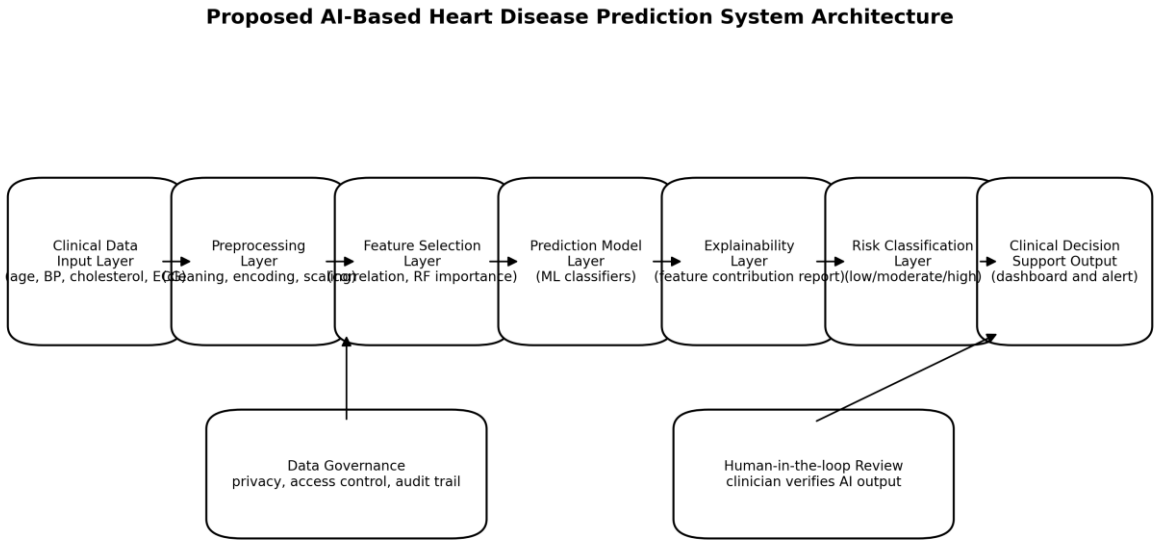


**Figure 4.1. Proposed Architecture of the AI-Based Heart Disease Prediction System**

Figure 4.2 presents the machine-learning workflow used for the experimental part of the study. The workflow separates data preparation, model training, validation, and interpretation so that the pipeline can be audited and replicated.

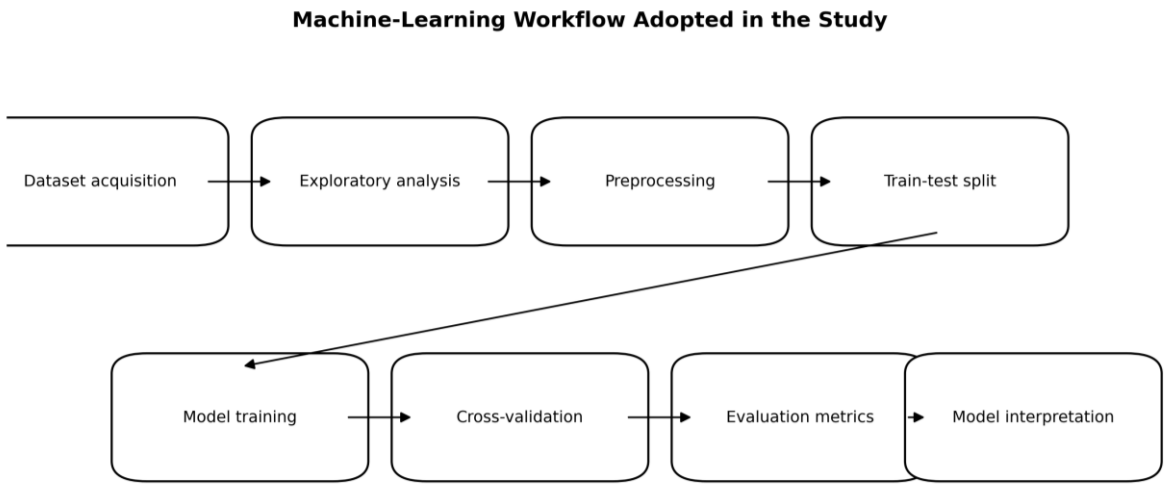


**Figure 4.2. Machine-Learning Workflow**

## 5. Data Analysis and Results

### 5.1 Descriptive Analysis of Dataset

The dataset contained 303 observations. The mean age was 54.366 years, with a minimum of 29 and maximum of 77 years. Mean resting blood pressure was 131.624 mmHg, and mean serum cholesterol was 246.264 mg/dL. The mean maximum heart rate achieved was 149.647. The outcome mean was 0.545, reflecting 165 positive cases and 138 negative cases in the binary target variable.

These descriptive statistics indicate that the dataset is clinically plausible for a cardiovascular screening benchmark but limited in scale. The dataset should therefore be considered useful for methodological development and comparative modelling, not as evidence of population-level prevalence or real-world clinical performance. Table 3 reports the descriptive summary.

**Table 5.1. Descriptive Statistics of Dataset Variables**

| Variable | Mean | Standard Deviation | Minimum | Median | Maximum |
|---|---|---|---|---|---|
| Age | 54.366 | 9.082 | 29.000 | 55.000 | 77.000 |
| Sex | 0.683 | 0.466 | 0.000 | 1.000 | 1.000 |
| Cp | 0.967 | 1.032 | 0.000 | 1.000 | 3.000 |
| Trestbps | 131.624 | 17.538 | 94.000 | 130.000 | 200.000 |
| Chol | 246.264 | 51.831 | 126.000 | 240.000 | 564.000 |
| Fbs | 0.149 | 0.356 | 0.000 | 0.000 | 1.000 |
| Restecg | 0.528 | 0.526 | 0.000 | 1.000 | 2.000 |
| Thalach | 149.647 | 22.905 | 71.000 | 153.000 | 202.000 |
| Exang | 0.327 | 0.470 | 0.000 | 0.000 | 1.000 |
| Oldpeak | 1.040 | 1.161 | 0.000 | 0.800 | 6.200 |
| Slope | 1.399 | 0.616 | 0.000 | 1.000 | 2.000 |
| Ca | 0.729 | 1.023 | 0.000 | 0.000 | 4.000 |
| Thal | 2.314 | 0.612 | 0.000 | 2.000 | 3.000 |
| Outcome | 0.545 | 0.499 | 0.000 | 1.000 | 1.000 |

## 5.2 Correlation Analysis

Correlation analysis was used to examine linear relationships among numerical predictors and the binary outcome. The heatmap in Figure 3 shows that no single variable fully determines the outcome. This supports the use of multivariable prediction rather than a one-threshold diagnostic rule. Strong clinical interpretation should nevertheless avoid treating correlation as causation.

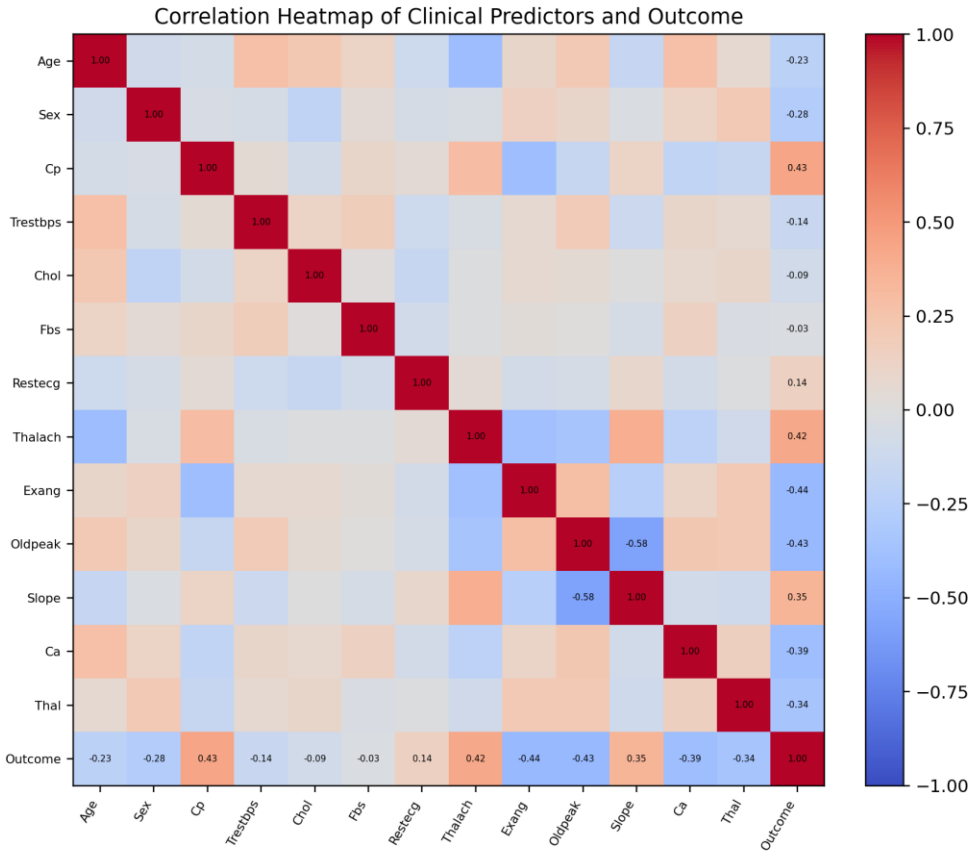


**Figure 5.1. Correlation Heatmap of Dataset Variables**

## 5.3 Model Performance Comparison

Table 4 reports hold-out test performance for the six classifiers. Random Forest achieved the highest ROC-AUC at 0.903 and produced strong recall at 0.939. KNN and Gradient Boosting produced the same hold-out accuracy of 0.820 but lower ROC-AUC than Random Forest. Logistic Regression achieved a ROC-AUC of 0.871, indicating that a relatively simple baseline remained competitive in this compact clinical dataset.

The best model by hold-out ROC-AUC was Random Forest. Its higher recall is valuable in early-screening contexts because it reduced false negatives in the hold-out test set. However, its specificity was 0.679, meaning that some non-disease cases were classified as positive. This trade-off may be acceptable in screening only if followed by clinician review and confirmatory assessment.

**Table 5.2 Performance Comparison of Machine-Learning Models on Hold-Out Test Set**

| Model | Accuracy | Precision | Recall | Specificity | F1-Score | ROC-AUC |
|---|---|---|---|---|---|---|
| Random Forest | 0.820 | 0.775 | 0.939 | 0.679 | 0.849 | 0.903 |
| K-Nearest Neighbours | 0.820 | 0.789 | 0.909 | 0.714 | 0.845 | 0.884 |
| Support Vector Machine | 0.803 | 0.800 | 0.848 | 0.750 | 0.824 | 0.883 |
| Gradient Boosting | 0.820 | 0.789 | 0.909 | 0.714 | 0.845 | 0.879 |
| Logistic Regression | 0.787 | 0.763 | 0.879 | 0.679 | 0.817 | 0.871 |
| Decision Tree | 0.738 | 0.758 | 0.758 | 0.714 | 0.758 | 0.731 |

Supplementary five-fold cross-validation showed that Logistic Regression had the highest mean cross-validation accuracy, whereas Random Forest had the highest mean cross-validation ROC-AUC. This reinforces the need to evaluate models using multiple metrics rather than selecting a model from accuracy alone.

**Table 5.3 Five-Fold Cross-Validation Performance Summary**

| Model | Accuracy | Precision | Recall | F1-Score | ROC-AUC |
|---|---|---|---|---|---|
| Random Forest | 0.818 | 0.815 | 0.873 | 0.840 | 0.900 |
| Logistic Regression | 0.828 | 0.820 | 0.885 | 0.850 | 0.890 |
| Gradient Boosting | 0.809 | 0.806 | 0.861 | 0.831 | 0.886 |
| K-Nearest Neighbours | 0.805 | 0.791 | 0.891 | 0.836 | 0.870 |
| Decision Tree | 0.746 | 0.765 | 0.776 | 0.769 | 0.757 |

**Table 5.4 Confusion Matrix of Best Performing Model: Random Forest**

| Actual Class | Predicted No Disease | Predicted Disease | Clinical Interpretation |
|---|---|---|---|
| Actual no disease | 19 | 9 | Specificity = 0.679 |
| Actual disease | 2 | 31 | Sensitivity = 0.939 |

## 5.5 ROC Curve Analysis

The ROC curve in Figure 4 compares model discrimination across probability thresholds. Random Forest produced the highest hold-out ROC-AUC, followed by KNN, SVM, Gradient Boosting, and Logistic Regression. ROC-AUC is useful because it is threshold-independent, but deployment requires selecting probability thresholds according to clinical objectives and acceptable error trade-offs.

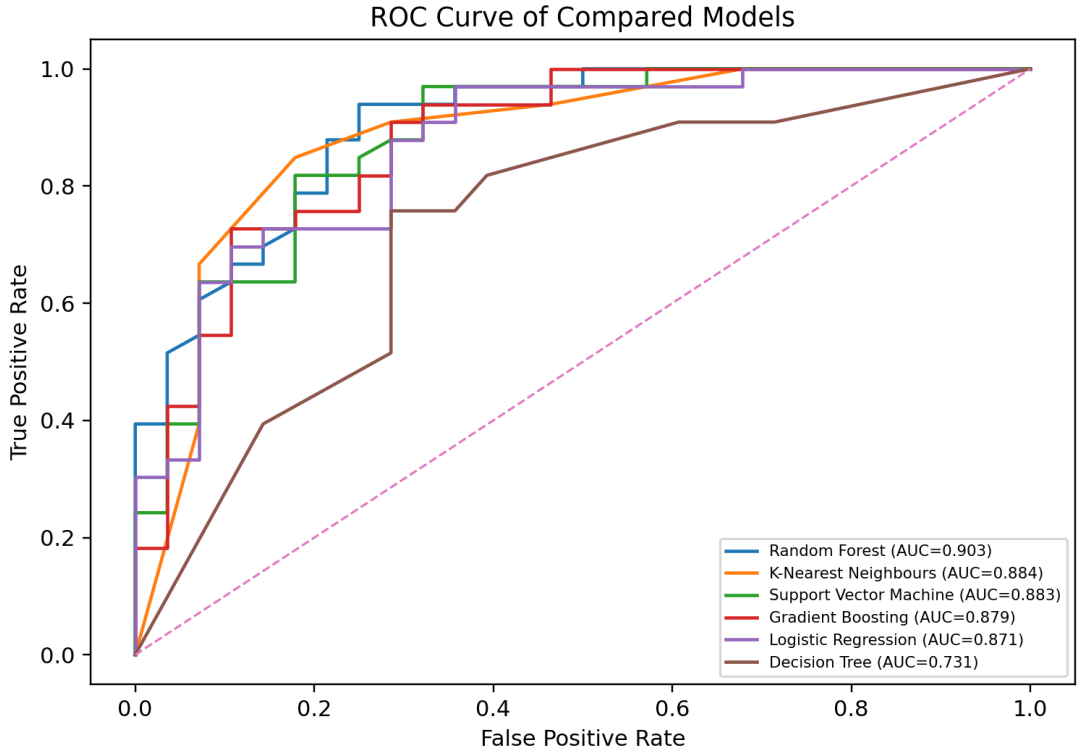


**Figure 5.2. ROC Curve of Compared Models**

## 5.6 Feature Importance Analysis

Random Forest feature importance identified chest pain type, ST depression, maximum heart rate achieved, thalassemia-related status, serum cholesterol, number of major vessels, age, and resting blood pressure as influential predictors. These

variables are clinically plausible because they relate to symptoms, exercise response, vascular burden, and demographic risk. However, the importance ranking should be read as a model-specific explanation from this dataset, not as a universal clinical hierarchy.

Table 8 and Figure 5 present the feature-importance ranking. Chest pain type had the highest importance score, followed by oldpeak and thalach. The relatively low importance of fasting blood sugar in this model does not imply that diabetes is unimportant in cardiovascular medicine; it only reflects the predictive behaviour of this specific dataset and trained model.

**Table 5.5 Feature Importance Ranking of Best Performing Model**

| Rank | Feature | Importance | Interpretation |
|---|---|---|---|
| 1 | Cp | 0.155 | Higher contribution to the Random Forest prediction |
| 2 | Oldpeak | 0.117 | Higher contribution to the Random Forest prediction |
| 3 | Thalach | 0.114 | Higher contribution to the Random Forest prediction |
| 4 | Thal | 0.112 | Higher contribution to the Random Forest prediction |
| 5 | Chol | 0.092 | Higher contribution to the Random Forest prediction |
| 6 | Ca | 0.083 | Secondary contribution in this dataset |
| 7 | Age | 0.082 | Secondary contribution in this dataset |
| 8 | Trestbps | 0.076 | Secondary contribution in this dataset |
| 9 | Exang | 0.058 | Secondary contribution in this dataset |
| 10 | Slope | 0.053 | Secondary contribution in this dataset |
| 11 | Sex | 0.031 | Secondary contribution in this dataset |
| 12 | Restecg | 0.020 | Secondary contribution in this dataset |
| 13 | Fbs | 0.007 | Secondary contribution in this dataset |

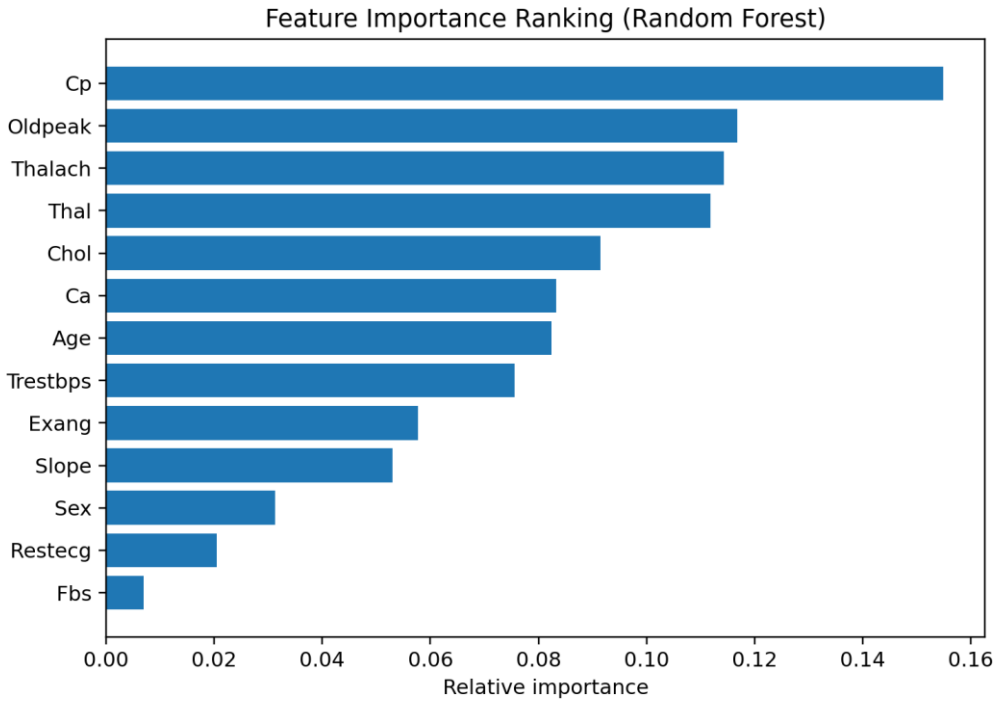


**Figure 5.3. Feature Importance Plot**

## 5.7 Explainable AI Interpretation

The explainability layer translates model output into clinician-readable information. In the current experiment, global feature importance was used to identify variables that influenced Random Forest predictions across the dataset. A deployable system should additionally provide local explanations for each patient record using methods such as SHAP or LIME, supported by safeguards that prevent explanation outputs from being mistaken for causal evidence (Lundberg & Lee, 2017; Ribeiro et al., 2016).

Figure 5.6 illustrates the proposed explainable AI output framework. The system receives a patient record, produces a risk probability, converts the probability into a risk category, and presents both global and local explanations for clinician review. This design keeps the physician as the final decision-maker and reduces the risk of automation bias.

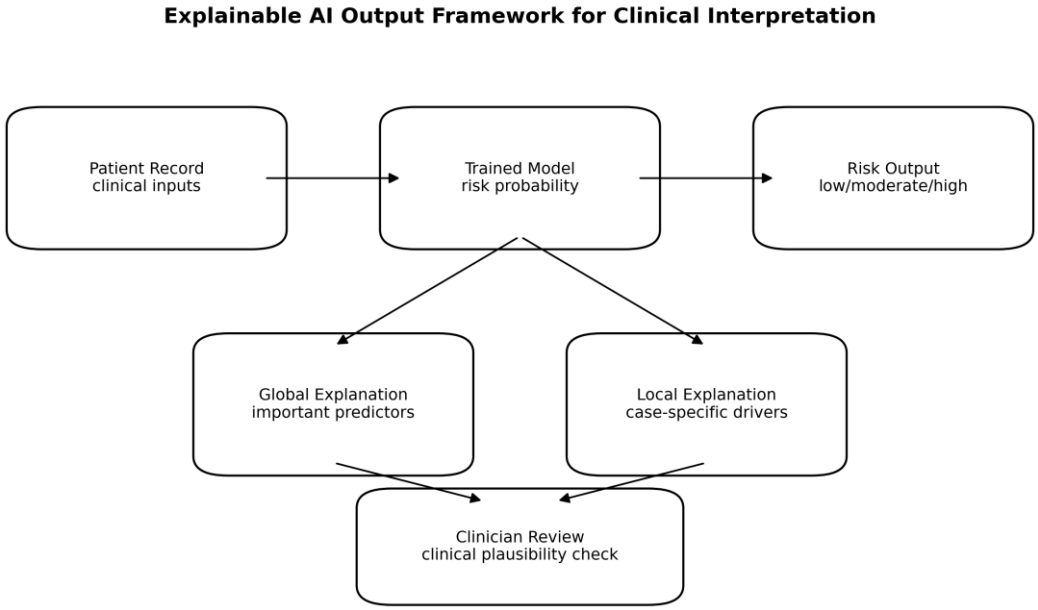


**Figure 5.4. Explainable AI Output Framework**

## 6. Discussion

Streamlined clinical data can help in the prediction of heart disease using supervised ML. Random Forest attained the best hold-out ROC-AUC and high recall. Because the cost of missing the disease is high, sensitivity is an important consideration in a screening-oriented system. Concurrently, the specificity of the model suggests that a physician's review is still necessary to address false positive outputs.

The results corroborate the existing literature which shows how AI can discover predictive patterns in cardiovascular datasets. Shah et al. (2025) and El-Sofany et al. (2024) highlighted the importance of ML and explainability for the risk of CVD. Nonetheless, the present study does not inflate any claims. A small retrospective benchmark on which a model is trained and tested if not clinically validated. The significant contribution is in its transparent workflow showing real input datasets, real computed metrics, feature interpretations, and a clinically oriented architecture.

The ability to convert variables that are routinely measured into risk estimates that can allow earlier assessment is clinically relevant. In environments with outpatient or screening, such a system is useful to allocate priority for a patient to undergo further investigation. For instance, a patient who has a high risk score, but has clinically plausible explanatory factors may be flagged for review by a physician. However, the presence of symptoms, clinical judgement and guideline-directed assessment should take precedence over a low predicted risk.

The study emphasizes the importance of explainability. If there is no explanation, clinicians may reject a high-risk output, and non-experts may misuse this output. The global feature-importance results provide us with an understanding of which variables shaped the model. In practical settings, global explanations should be complemented by patient-specific explanations, calibration plots, audit trails, and an understandable statement of uncertainty.

The major strength of the work is methodological transparency. The model comparison utilized consistent preprocessing, a stratified split, a final evaluation metric, and explicit limitations. The core limitation is that the data set is small, retrospective and unrepresentative of conditions. As a result, the model ought to be considered research, not a medical device.

## 7. Proposed clinical decision-support system using AI

The clinical decision-support system to be developed begins with a user input module in which authorized healthcare personnel type or import patient variables. The system checks fields for validity, flags if some values are missing and standardises measurements wherever necessary. The module for patient risk-profiling prepares the prediction engine variables and recordings data provenance for auditing.

Based on the trained model, the AI prediction engine provides the chance of heart disease. The output is categorized into low, medium, and high risk based on pre-defined thresholds. Table 8 depicts a risk-category framework. Thresholds for deployment have to work with external clinical settings for all efficient use.

A doctor dashboard must present risk category, probability score, leading contributing feature, and recommended next step. The alert system must prioritise high-risk cases and not cause alarm fatigue. The report, generated in response to the request, should generate a short narrative that can be read by the clinician, and discussed with the patient where appropriate. Ultimately, a layer of data privacy and security should implement authentication, access control, encryption, logging, and retention policies.

### Table 7.1 Clinical Interpretation of Risk Categories

| Risk Category | Illustrative Probability Range | Suggested Clinical Action | Caution |
|---|---|---|---|
| Low risk | < 0.35 predicted probability | Routine preventive counselling; confirm with clinical judgement. | Not a diagnostic ruling-out statement. |
| Moderate risk | 0.35 to < 0.65 predicted probability | Review risk factors, repeat measurements, consider clinician-directed testing. | Threshold must be calibrated before deployment. |
| High risk | >= 0.65 predicted probability | Prompt clinical review and appropriate diagnostic pathway. | AI output must not replace physician assessment. |

## 8. Ethical, Legal, and Clinical Considerations

An ethical use of AI in CV screening requires a guarantee of patient privacy, fairness across demographic groups, transparency on model limitations, and acceptability of clinical actions.

According to WHO, AI for health must protect autonomy, promote safety, ensure transparency, and facilitate human responsibility. (WHO, 2021) The proposed system uses a human-in-the-loop approach where clinical interpretation remains the responsibility of physicians.

Algorithmic bias is worrying as the public benchmark datasets may be unrepresentative of local populations, women, older adults, ethnic groups, rural patients, and multimorbid patients. When a pattern created on one dataset used on another performs poorly. Prior to deployment, assessment of subgroup performance, calibration, missing-data behaviour and site-specific validation. Data governance must also allow only authorized users to access sensitive health information.

Regulators are becoming stricter with medical AI. The European Union will classify many health care AI systems as “high-risk” under the EU AI Act when their use can affect outcomes of health-related decisions (European Commission, 2024). As “high-risk” AI systems, these must undergo risk management process and must be documented. Furthermore, they must satisfy transparency and oversight provisions. According to the International Medical Device Regulators Forum analysis, global regulators have shown interest in good machine learning practices, specifically across the AI lifecycle. As a result, AI heart-disease prediction model should not be deployed and used clinically without regulatory and institutional review.

## 9. Practical Applications

### 9.1 Implications for Clinicians

For the clinician, this system can act as a screening-support tool that collates structured risk information and highlights those patients that may require closer assessment. It must not supplant clinical history, physical evalutation, diagnostic testing, or guideline-based judgement. An interface that presents the probability, risk category, explanation, uncertainty, and suggested clinical review path in tabular form would be useful.

### 9.2 Implications for Health Institutions

An AI-based heart diseases prediction system may help hospitals around triage, preventive clinics, screening of electronic health records and dashboards of population risk. Implementing this plan will require linking it with hospital information

systems as well as data quality checks and security measures. In addition, users will need to be trained, and the cardiology, informatics, and ethics committee must oversee the plan.

### 9.3 Implications for Public Health Screening

Systematic screening in public health may be helpful in identifying people needing preventive counselling or referral.

### 9.4 Implications for AI Healthcare Developers

According to scholarly research, for developers, it is important to have reproducible pipelines, transparent metrics, modular architecture, explanation layers and lifecycle monitoring. Systems should be built which log: model version, input provenance, prediction output, explanation output, and clinician feedback. The focus should be safe decision support and not autonomous diagnosis.

## 10. Test of the Study.

The research has a number of limitations. The dataset has only 303 records, limiting the stability of performance estimates. In addition, the dataset is retrospective and does not represent the diversity of current clinical populations. For one, the result is a simplified binary label when heart disease has many different diagnoses, severities and temporal patterns. Fourth, without any external validation dataset, transferability cannot be established.

Fifth, the calibration analysis was not included in the main results despite the importance of calibration before clinical risk probabilities become trustworthy.

The sixth component were local explainability methods such as SHAP or LIME. Although these were discussed as potential system components, they were not used in the generation of case-level clinical reports in this version. The application was not implanted, integrated with a hospital's electronic health record, or validated in a prospective clinical investigation.

These limitations are not trivial technicalities; they delimit the frontier between an experimental AI paper and a usable clinical technology. These findings do not imply clinical implementation but are feasible and provide methodological direction.

## Declarations

**Ethics approval and consent to participate:** The study used a public, de-identified secondary dataset for methodological research. No direct recruitment, intervention, or access to identifiable patient information was involved. The final submitting authors should confirm whether their institution or target journal requires an ethics-exemption statement for secondary public data analysis.

**Data availability:** The manuscript is based on a public UCI Heart Disease dataset and a Cleveland-format CSV file used for computation. For full reproducibility, the final authors should attach the analysis script, preprocessing details, and data-access link as supplementary material or deposit them in an institutional repository.

## References


1. Ansari, Z. A., Khan, W., Ansari, M. S. H., Fatima, S., et al. (2026). Dual explainability framework for heart disease prediction using LIME and permutation feature importance. Discover Applied Sciences, 8, Article 19. https://doi.org/10.1007/s42452-025-08108-5
2. Arnett, D. K., Blumenthal, R. S., Albert, M. A., Buroker, A. B., Goldberger, Z. D., Hahn, E. J., Himmelfarb, C. D., Khera, A., Lloyd-Jones, D., McEvoy, J. W., Michos, E. D., Miedema, M. D., Muñoz, D., Smith, S. C., Virani, S. S., Williams, K. A., Yeboah, J., & Ziaeian, B. (2019). 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease. Circulation, 140(11), e596-e646. https://doi.org/10.1161/CIR.0000000000000678
3. Agarawal, V., & Swarnkar, V. K. (2025). An ensemble-based machine learning framework for early prediction of heart disease. *IJEETE Journal of Research, 12*(3). https://ijoeete.com/wp-content/uploads/2025/11/7-vandana.pdf
4. Agarawal, V., & Swarnkar, V. K. (2024). A hybrid approach to diagnosis of heart disease with machine learning technique. *International Journal of Exploring Emerging Trends in Engineering, 11*(4). https://ijoeete.com/
5. Breiman, L. (2001). Random forests. Machine Learning, 45, 5-32. https://doi.org/10.1023/A:1010933404324
6. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). https://doi.org/10.1145/2939672.2939785
7. Chicco, D., & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making, 20, Article 16. https://doi.org/10.1186/s12911-020-1023-5
8. Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J., Celi, L. A., Denaxas, S., Denniston, A. K.,

Glocker, B., Golub, R. M., Harvey, H., Heinze, G., et al. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385, e078378. https://doi.org/10.1136/bmj-2023-078378

9. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20, 273-297. https://doi.org/10.1007/BF00994018
10. Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., Guppy, K. H., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 64(5), 304-310. https://doi.org/10.1016/0002-9149(89)90524-9
11. El-Sofany, H., Bouallegue, B., & El-Latif, Y. M. A. (2024). Predictive analytics and machine learning for improved cardiovascular disease detection and patient care. Scientific Reports, 14, Article 24214. https://doi.org/10.1038/s41598-024-74656-2
12. European Commission. (2024, August 1). AI Act enters into force. https://commission.europa.eu/news-and-media/news/ai-act-enters-force-2024-08-01_en
13. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.
14. International Medical Device Regulators Forum. (2025). Good machine learning practice for medical device development: Guiding principles. https://www.imdrf.org/documents/good-machine-learning-practice-medical-device-development-guiding-principles
15. Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1988). Heart Disease [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X
16. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.
17. Luo, W., Phung, D., Tran, T., Gupta, S., Rana, S., Karmakar, C., Shilton, A., Yearwood, J., Dimitrova, N., Ho, T. B., Venkatesh, S., & Berk, M. (2016). Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. Journal of Medical Internet Research, 18(12), e323. https://doi.org/10.2196/jmir.5870
18. Martin, S. S., Aday, A. W., Almarzooq, Z. I., Anderson, C. A. M., Arora, P., Avery, C. L., Baker-Smith, C. M., Barone Gibbs, B., Beaton, A. Z., Boehme, A. K., Commodore-Mensah, Y., Currie, M. E., Elkind, M. S. V., Evenson, K. R., Generoso, G., Heard, D. G., Hiremath, S., Johansen, M. C., Kalani, R., et al. (2024). Heart disease and stroke statistics-2024 update: A report from the American Heart Association. Circulation, 149(8), e347-e913. https://doi.org/10.1161/CIR.0000000000001209
19. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
20. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380, 1347-1358. https://doi.org/10.1056/NEJMra1814259
21. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144). https://doi.org/10.1145/2939672.2939778
22. Shah, P., Shukla, M., Dholakia, N. H., Gupta, H., et al. (2025). A scalable machine learning approach for enhanced heart disease prediction using clinical and lifestyle factors. Scientific Reports, 15, Article 17672. https://doi.org/10.1038/s41598-025-01650-7
23. Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793-4813. https://doi.org/10.1109/TNNLS.2020.3027314
24. World Health Organization. (2021). Ethics and governance of artificial intelligence for health: WHO guidance. https://www.who.int/publications/i/item/9789240029200
25. World Health Organization. (2025). Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
26. Yadav, S., Mane, P., Swarnkar, V. K., Rawandale, S. K., Patil, A. S., Katke, K. S., & Bhat, N. (2023). The role of AI in healthcare policy development and management. In *2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICAIIHI57871.2023.10489810

**CITATION**

Swarnkar, V. K., & Sahu, G. (2026). Review of an Artificial Intelligence Based System for Early Prediction of Heart Diseases. Global Journal of Research in Engineering & Computer Sciences, 6(4), 129–142. https://doi.org/10.5281/zenodo.22065603