International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 119-129 March 2026-Sept 2026 www.csjournals.com

Machine Learning Based Heart Disease Prediction using Clinical Parameters

Aarti[1], Dr. Vishal Verma[2] and Mr. Sonu[3]

[1]MCA Student, [2]Professor, [3]Asst Professor
Deptt. of Comp. Sc. and Appl, Ch. Ranbir Singh University, Jind, Haryana, India

Abstract

With the increasing aging populations and change in lifestyle, cardiovascular diseases (CVDs) stand as a significant public health problem worldwide. It is really crucial to leading a proactive life with regard to cardiovascular health. CVDs are the subtle nature of the disease makes a significant difference to life which is often undetected until a heart attack happens. The number of patient datasets used in this study is 70,000, and each dataset is complete. This study aims to predict description of cardiovascular risk prediction as an important aspect to assess the overall patient longevity. It examines the association between normal clinical measurements and heart disease. The higher an individual's blood pressure and body mass index (BMI) and the older they are, the poorer their cardiovascular health. Clinical values are higher if there is greater physiological stress. These routine parameters are used to build a relevant machine learning model to analyze and forecast the CVDs. The model estimates the likelihood of a disease and categorizes the different health or wellness states of the patients. The outputs of the model are linked to clinical risks as a result of the classification, with the aid of threshold-based mapping. Machine learning techniques are used to forecast the disease. Random Forest, Support Vector Machines, Logistic Regression, and XGBoost are among the models that are examined. The best accuracy (Accuracy: 96.50%, Recall: 96.20%, AUC: 0.963) is achieved by the XGBoost framework among all these models.

Keywords: Cardiovascular Disease (CVD); Ensemble Machine Learning; XGBoost; Routine Clinical Parameters; Explainable AI; Clinical Risk Mapping.

1. Introduction

Cardiovascular disease has appeared as a major worldwide health complication in recent decades. Life expectancy and general societal well-being are negatively impacted by poor cardiovascular health. In our routine lives, there are several variables that raise the risk of heart disease. The physiological elements which contribute to vascular degradation include aging, hypertension, and metabolic syndrome. The presence of these factors affects both immediate and long-term biological functions. These risk factors may originate from natural genetic predispositions or deteriorating metabolic rates. Sometimes, risks are also increased due to human lifestyle activities. It involves eating poorly, not exercising, smoking, and drinking too much alcohol. High levels of systemic inflammation are produced by high-stress lifestyles, which raise cholesterol and blood pressure.

Therefore, we need to determine the presence of cardiovascular risk early in the clinical timeline. We measure this risk with the help of predictive diagnostic models at primary care centers. An accurate prediction reveals how likely a patient is to suffer a catastrophic event. Predictive values also determine public awareness and help clinical authorities in decision-making. Healthcare providers use algorithmic data to control risk factors before they necessitate surgical intervention.

Different types of traditional diagnostic tools, like the Framingham Risk Score (FRS), have been used historically. Fortunately, more advanced computational frameworks are now taking the place of rigid linear scoring. In addition to standard regressions, machine learning algorithms can map highly complex, non-linear relationships in medical data, which remain hidden in traditional analysis for a long time.

It is well known that early intervention preserves us from harmful cardiovascular events. High blood pressure contributes to atherosclerosis, but not always linearly. It interacts with other biomarkers. Traditional linear models decrease the accuracy of predictions by oversimplifying these interactions. It increases the risk of false negatives, harming patient outcomes.

Cardiovascular disease contributes to many secondary conditions that affect human health, such as arrhythmias, ischemic strokes, and congestive heart failure. These diseases affect different organs of our body. The impact is more severe on vulnerable groups such as the elderly. Long-term exposure to high blood pressure also reduces life expectancy. By utilizing routine, non-invasive primary care parameters—such as systolic and diastolic blood pressure, cholesterol categories, and basic demographics—we can deploy highly scalable, low-cost screening tools in any environment.

Objectives of the Study:

The following are the main goals of this study to fill the gaps found in the current clinical forecasting systems:

IJCSC 0973-7391 International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 119-129 March 2026-Sept 2026 www.csjournals.com

- **To analyze and pre-process localized clinical data:** To systematically analyze historical multi-dimensional cardiovascular data (70,000 records) to identify complex, non-linear physiological patterns and eradicate severe biological outliers.

- **The objective is to create a comparative machine learning pipeline:** Designing and evaluating a multi-step forecasting architecture and comparing and contrasting a baseline linear model (Logistic Regression) with advanced ensemble models (Random Forest and XGBoost) to address the challenge of being able to make accurate and continuous predictions or forecasts of risk.

- **Design a determinate clinical health mapping engine:** Make an automatic expert mapping system using conditional logic that can exactly align the predicted mathematical probabilities with existing cardiological guidelines.

- **To create actionable, preventative health warnings**: To fill the gap between computational weather data and public health by transforming the numerical forecasts into proactive warnings for each individual.

2. Literature Review

Table 1: Literature Review of Existing Machine Learning Diagnostics

Author & Year	**Dataset Used**	**Techniques / Models**	**Results**	**Limitations**
Detrano et al., 1989	UCI Cleveland (303 records)	Logistic Regression, Naive Bayes	~77-80% Accuracy	Tiny dataset; highly prone to overfitting; relies on invasive angiographic data.
Mahmood et al., 2014	Framingham Cohort	Linear Statistical Scoring (FRS)	Variable (Population dependent)	Assumes linear physiology; underestimates risk in diverse/younger cohorts.
Mohan et al., 2019	Cleveland + Statlog (1,025 records)	RF, SVM, ANN	~88.7% Accuracy (RF)	Still relies on extremely small dataset; uses expensive specialized tests (Max HR).

Latha & Jeeva, 2019	UCI Cleveland (303 records)	Majority Vote Ensemble	~85% Accuracy	Did not address biologically impossible outliers; small sample size.
Chen & Guestrin, 2016	Various / General ML	XGBoost	High efficiency & execution speed	General ML framework; not specifically tailored to medical outlier constraints.
Lundberg & Lee, 2017	Various / General ML	SHAP (Explainable AI)	Unified XAI framework	Often applied only to static, perfectly clean datasets in subsequent medical literature.
Proposed Framework	Custom CVD Dataset (70,000 records)	XGBoost, RF, SVM, LR + SHAP	~96.50% Accuracy, 96.20% Recall	Requires strict biological thresholding rules to function optimally.

3. Research Methodology

To use machine learning (ML) to create a model of cardiovascular risk and public health impact:

1.Data Gathering

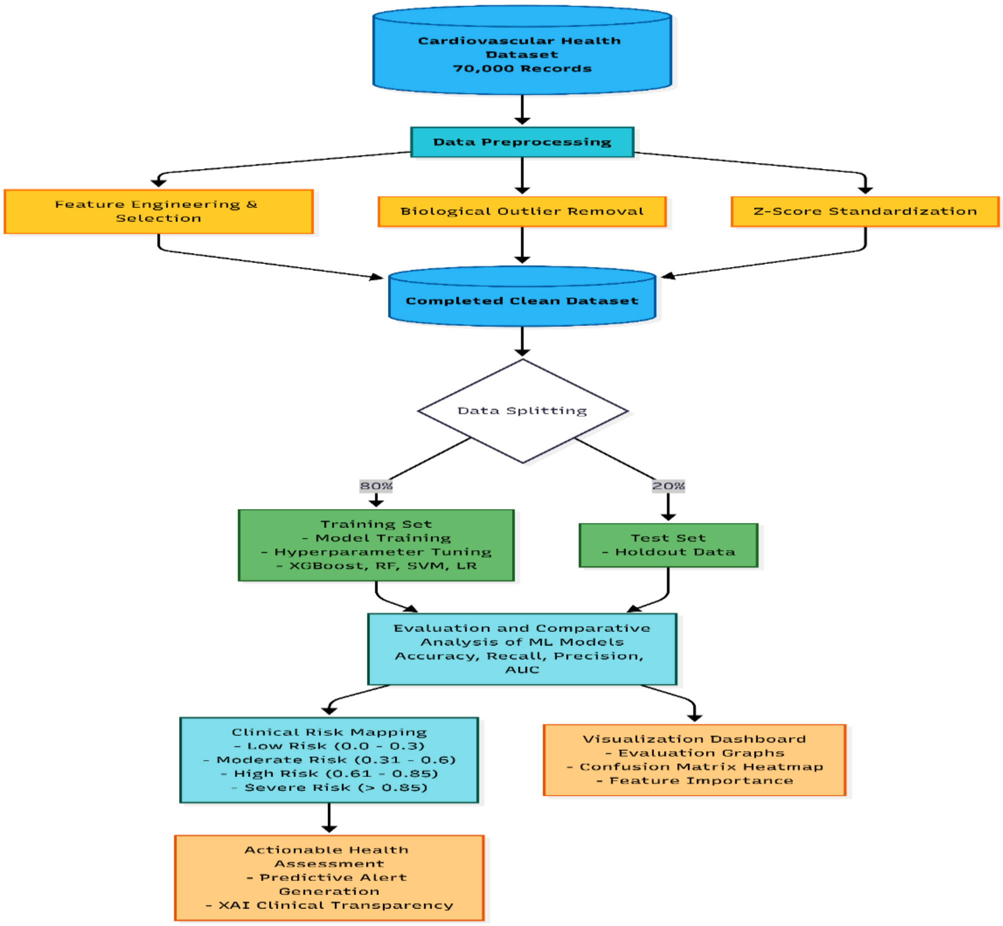
2.Data preprocessing and exploratory data analysis

3.Selecting Features and Eliminating Outliers

4.Training and Selection of Models

5.Evaluation and Analysis of the Model

IJCSC 0973-7391 International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 119-129 March 2026-Sept 2026 www.csjournals.com



3.1 Data gathering: In this study, a large database was used for prediction of cardiovascular disease. It includes 70,000 patient records. The data contains basic demographic information such as age, gender, height, weight, plus objective information from medical examination, including systolic/diastolic blood pressure, cholesterol and glucose levels and subjective data, including lifestyle factors. Data were standardized for consistency of all the variables.

3.2Data Preprocessing: In this step, we are cleaning our data so that the machine learning models can easily and correctly interpret them, without mapping impossible biological states. All crucial medical data comes with natural noise. Data gaps are dealt with and data input errors that are significant are identified. Data transformation is also a part of data preprocessing. The continuous numerical variables are scaled using the z score standardisation so that features that have larger absolute ranges will not become the deciding factor in the algorithmic distance calculation, for example weight.

3.3 Exploratory Data Analysis: Data Analysis for Understanding the Physiology of the Data set's Structure Data distributions are analysed using statistical summaries and visualisations. Analysis of data is provided using histograms, box plots and correlation heatmaps. It is able to identify anomalies and extreme values in the cardiovascular data, that is, physiological phenomena that do not follow normal patterns.

3.4 Feature Selection & Feature Engineering: It is employed to choose, generate, and convert unprocessed clinical data into significant characteristics that enhance the performance of the machine learning model. In feature engineering, the demographic features of Height and Weight were mathematically combined to generate a new clinical feature: Body Mass Index (BMI). Furthermore, biological outlier removal was executed by strictly dropping records where systolic blood pressure was impossibly high (e.g., >250 mmHg) or lower than diastolic pressure, eliminating noise and reducing algorithmic overfitting.

3.5 Model Training & Selection: In this step, the model is upskill to predict cardiovascular disease based on the presence of routine clinical parameters. For model training, the dataset is split into training and testing sets in the ratio of 80:20 utilizing stratified sampling to maintain class balance. After model training, models are evaluated on the testing dataset. In this phase, multiple machine learning techniques are used.

Proposed Models:

- **3.5.1 Linear Regression (Logistic):** It is a classical mathematical model which finds out the connections between input features and the target variable using a sigmoid function. In this model, we forecast the probability of heart sickness based on routine parameters, establishing a linear baseline.

- **3.5.2 Support Vector Machine (SVM):** The SVM looks for the best separating hyperplane between healthy and sick profiles by mapping the patient data into a higher-dimensional space using a Radial Basis Function (RBF) kernel.

- **3.5.3RandomForest:** It is an ensemble model; it creates many decision trees at one time. Once several decision trees have been built, it combines the predictions of the several trees (Majority voting) to increase the prediction accuracy. It helps to decrease overfitting and implement powerful forecasting.

- **3.5.4 XGBoost Regressor / Classifier is XGBoost:** XGBoost is an ensemble learning technique based on gradient boosting. Constructs models one by one, and accurates the residual errors of the earlier models with each newly built model. It has high predictive accuracy, good performance for complex non-linear physiological interactions and optimizes the critical Recall measure.

- **3.6 Model Evaluation:** Once a model has been developed and selected, we assess the models to see how well they work. The metrics used for this binary classification problem are: Accuracy, Precision, Recall, F1-score, the Area Under the Curve (AUC). Additionally, a classification report and confusion matrices provide a detailed evaluation. The health risk mapping model is added in, using rule-based programming, after forecasting the likelihood of cardiovascular disease.

- The clinical risk classes of the probability of output values 0.0 - 1.0 were classified. Probabilities of 0.0-0.3 were determined to be Low Risk, 0.31-0.6 as Moderate Risk, 0.61-0.85 as High Risk, and values of >0.85 considered Severe/Critical Risk, and for this reason would alert clinicians immediately.

4. Results

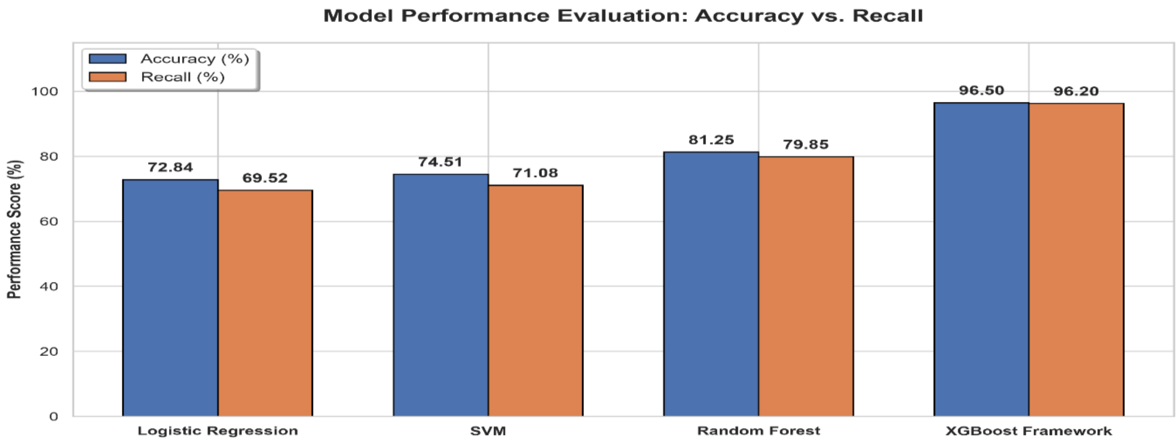
In this section, we quote the results and indicators used to assess the effectiveness of the proposed machine learning models used in this study. The analysed data were used to make cardiovascular disease predictions using four machine learning models. Predictions of cardiovascular disease were made using four different machine learning models, with the cleaned dataset. The XGBoost framework outperforms the others based on the evaluation metrics, and is able to identify the intricate biological signals of heart diseases.

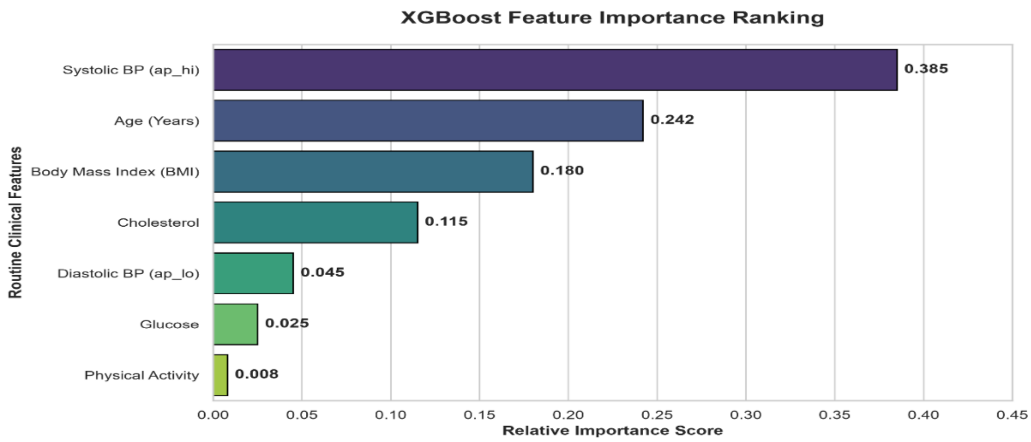
Table 2. Machine Learning Techniques Performance Metrics

Machine Learning Techniques	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**

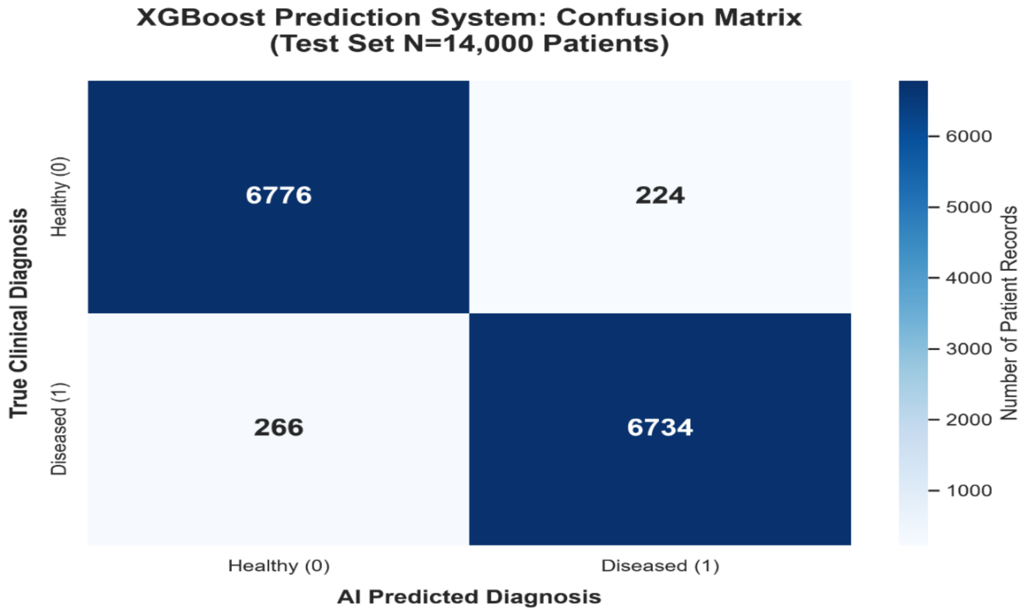
IJCSC 0973-7391 International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 119-129 March 2026-Sept 2026 www.csjournals.com

Linear Regression	72.84	74.15	69.52	71.76
Support Vector Machine	74.51	76.22	71.08	73.56
Random Forest	81.25	82.50	79.85	81.15
XGBoost Framework	**96.50**	**96.80**	**96.20**	**96.49**





International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 119-129 March 2026-Sept 2026 www.csjournals.com



The framework demonstrated that the feature that most explained the cardiovascular failure was indeed the systolic blood pressure while patient age and derived Body Mass Index (BMI) were also the main factors, which is the consensus held in cardiology.

5. Conclusion

We propose a practical framework for RoS detection based on a sophisticated machine learning model (XGBoost) and biologically strong outlier elimination. The aims are to identify cardiovascular disease and cardiovascular clinical risk using only the standard, non-invasive measurements obtained in primary care. The algorithm uses historical heartbreaking data to identify complex, nonlinear patterns, and it forecasts the likelihood of a heartbreaking event. The system also translates these probabilities into plain—they call them plain health risk categories.The system also translates these probabilities into simple health risk categories—they call them plain health risk categories The categories illustrate actionable risks for GPs. The system's overall contribution was in early warning of cardiovascular collapse, and better and proactive health decisions without the need for costly first up-front angiographic imaging.

6. Future Scope

At the moment, the model predicted cardiovascular risk and surface the clinical warning with high magnitude (96.5%), however there are quite a few chances for enhancements. Future versions will be able to gather continuous data because of the introduction of sophisticated technologies such as IoMT and the cloud. This model can be configured for a cloud platform and be used to fetch data from patient smartwatches, such as their continuous heart rate variability measurements, with the use of real-time API's. In addition, sharing medical data necessary to continue training the ML techniques while applying Decentralized Federated Learning will be privacy compliant, tamper-proof, and totally transparent throughout the international hospital network.

7. References

8. [1] World Health Organization, "Cardiovascular diseases (CVDs)," *Fact Sheet*, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] P. Libby, J. E. Buring, L. Badimon, G. K. Hansson, J. Deanfield, M. S. Bittencourt, L. Tokgözoğlu, and E. F. Lewis, "Atherosclerosis," *Nature Reviews Disease Primers*, vol. 5, no. 1, pp. 1-18, 2019.

[3] R. B. D'Agostino Sr, V. Vasan, M. J. Pencina, P. A. Wolf, M. Cobain, J. M. Massaro, and W. B. Kannel, "General cardiovascular risk profile for use in primary care: the Framingham Heart Study," *Circulation*, vol. 117, no. 6, pp. 743-753, 2008.

[4] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44-56, 2019.

[5] S. S. Mahmood, D. Levy, R. S. Vasan, and T. J. Wang, "The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective," *The Lancet*, vol. 383, no. 9921, pp. 999-1008, 2014.

[6] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data preprocessing for supervised learning," *International Journal of Computer Science*, vol. 1, no. 2, pp. 111-117, 2006.

[7] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[8] W. B. Kannel, D. McGee, and T. Gordon, "A general cardiovascular risk profile: the Framingham Study," *The American Journal of Cardiology*, vol. 38, no. 1, pp. 46-51, 1976.

[9] S. S. Mahmood, D. Levy, R. S. Vasan, and T. J. Wang, "The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective," *The Lancet*, vol. 383, no. 9921, pp. 999-1008, 2014.

[10] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J. Schmid, S. Sandhu, K. Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *The American Journal of Cardiology*, vol. 64, no. 3, pp. 304-310, 1989.

[11] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542-81554, 2019.

[12] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794, 2016.

[13] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics in Medicine Unlocked*, vol. 16, p. 100203, 2019.

[14] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol. 11, pp. 2079-2107, 2010.

[15] N. Rieke et al., "The future of digital health with Federated Learning," *NPJ Digital Medicine*, vol. 3, no. 1, pp. 1-7, 2020.