Open Access International Journal of Science and Engineering

ISSN: 2456-3293

Volume 9 Issue 1s (2026) | Pages: 254-257

Heart Disease Prediction using Machine Learning

Sudarshana A.[1], Shubhankar Singh[2], Mr. Tanishq[3]

[1,2,3]Dept. of Computer Science Business Systems, College of Engineering, Bharati Vidyapeeth University, Pune, India

Peer Review Information

***Type**: Article*
***Received**: 04 April 2026*
***Revised**: 30 April 2026*
***Accepted**: 05 June 2026*
***Published**: 26 June 2026*

Abstract

Early detection and timely medical intervention significantly improve patient outcomes. However, traditional diagnostic methods can be time-consuming and rely heavily on expert interpretation. With advancements in Artificial Intelligence and the increasing availability of medical datasets, machine learning offers a reliable and efficient approach to heart-disease prediction. This project, Heart Disease Prediction Using Machine Learning, presents an end-to-end predictive system developed using machine learning algorithms. The objective is to analyse clinical parameters such as age, sex, chest pain type, cholesterol, resting blood pressure, maximum heart rate, exercise-induced angina, ST depression (oldpeak), and thalassemia to predict the likelihood of heart disease. The methodology includes data preprocessing, Exploratory Data Analysis (EDA), baseline model comparison, correlation study, hyperparameter tuning, and evaluation using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Among the models evaluated—Logistic Regression, K-Nearest Neighbours, and Random Forest—the optimized Logistic Regression model exhibited the best-balanced performance, achieving strong accuracy and interpretability. The system demonstrates the potential of machine learning in enhancing early detection, supporting clinical decision-making, and improving healthcare delivery.

Keywords: Heart Disease; Machine Learning; Logistic Regression; Data Analysis; Predictive Modelling; Healthcare Artificial Intelligence.

How to Cite This Article

A., S., Singh, S., & Tanishq. (2026). Heart disease prediction using machine learning. *Open Access International Journal of Science and Engineering*, 9(1s), 254–257.

Introduction

Cardiovascular diseases (CVDs) – including coronary artery disease and related conditions –remain the leading cause of death worldwide, killing on the order of 20 million people per year[1]. The burden is especially severe in low- and middle-income regions, where limited resources and late presentation mean many patients are diagnosed only after a major event. Compounding this problem, heart disease often progresses silently. "Often, there are no symptoms of the underlying disease," so that a heart attack or stroke may be a patient's first warning sign[2]. This asymptomatic nature, along with the varied causes of heart disease (genetic factors, lifestyle, comorbidities) and its subtle early manifestations, makes timely diagnosis inherently difficult. In practice, routine screenings and risk questionnaires frequently miss high-risk individuals, underscoring why better predictive tools are so urgently needed.

Traditional diagnostic workflows rely on a combination of clinical evaluation and specific tests such as exercise ECG, echocardiography, or invasive angiography. While these techniques can be effective in detecting established disease, they have well-known limitations. For example, angiography – the gold standard for coronary artery disease – is costly and carries procedural risks, and even noninvasive tests are operator-dependent and may fail to detect early-stage pathology [2]. Interpretation of these tests often rests on physician judgment, introducing variability in outcomes. Moreover, large-scale screening using these methods is impractical: equipment shortages, high costs, and the subjective nature of reading scans or ECGs mean that many at-risk patients remain undiagnosed until symptoms appear. In short, conventional pathways struggle to deliver the early, reliable detection that experts deem essential [3][2].

Machine learning (ML) has been proposed as a way to help bridge this gap. By ingesting large, heterogeneous clinical datasets, ML algorithms can model complex nonlinear relationships among risk factors and uncover subtle patterns that elude traditional statistics [4]. Studies have shown that advanced ML classifiers – particularly ensemble methods like XGBoost or LightGBM – frequently outperform classic risk scores or single-model approaches for heart disease prediction [4]. In principle, an ML-driven tool could automatically flag high-risk patients (for example, based on electronic health record data or routine exams) and thus enable earlier intervention. It is important, however, to view this potential realistically. Its true contribution is to synthesize many inputs and highlight risk in a data-driven way, not merely to automate what physicians already do.

Despite these promising aspects, implementing ML-based diagnostics in real clinical environments faces significant hurdles. First, many published models rely on limited or homogeneous datasets and lack robust external validation – raising questions about their reliability for new patients [2]. Second, the "black box" nature of complex models can undermine trust: if a high-accuracy algorithm cannot explain why it makes a prediction, clinicians may be hesitant to use it[2][5]. In practice, integrating ML tools requires solving engineering and workflow issues as well. Patient data are often fragmented, and effective deployment would demand seamless connections to electronic health records, real-time processing, and clear interpretability for users.

Regulatory and logistical barriers are nontrivial as well: rigorous privacy safeguards (e.g. HIPAA/GDPR compliance) and formal clinical evaluations (FDA or equivalent approval) are typically required before an ML tool can be trusted in care [6].

Literature Review

Early work on heart disease prediction typically employed simple, transparent models such as logistic regression and decision trees. These are easy to interpret – for example, linear weights or IF–THEN rules map directly to risk factors – which is attractive for clinical use. Indeed, one study showed a CART decision-tree achieving about 87% accuracy while producing human-readable rules for clinicians [1]. In practice, however, these models often oversimplify complex relationships. Linear models like logistic regression assume additive effects of predictors, which can miss nonlinear interactions among clinical features. Similarly, a single decision tree may be unstable and prone to overfitting small changes in the data unless carefully pruned [1].

More advanced machine-learning techniques generally yield higher accuracy but at the cost of complexity. Ensemble methods like Random Forests and Gradient Boosting machines combine many weak learners to capture intricate patterns in the data. These models routinely outperform single classifiers on structured heart-disease datasets [2]. For example, recent reviews note that ensemble models "dominate prediction tasks on structured data, achieving high accuracy" [2]. However, such models become black boxes; a forest of trees or a boosted ensemble has no simple interpretation. Tools like feature-importance scores, SHAP or LIME plots are therefore used to interpret them, but these are "an attempt to retrofit transparency" onto opaque models [2][3]. Support Vector Machines (SVMs) are another popular choice: they can learn nonlinear boundaries via kernels and often do well on modestly sized datasets. SVMs can achieve strong classification performance, but selecting kernels and tuning parameters can be difficult, and they offer little intrinsic insight into feature contributions. K-Nearest Neighbors provides a simple similarity-based baseline, but it scales poorly with data size and also lacks a global model to interpret.

Across all these approaches, several practical challenges recur. A common issue is class imbalance: heart-disease datasets typically have fewer positive cases than negatives, so naïve learners can be biased toward the majority class. In fact, systematic reviews note that many

models "deal with the effects of imbalanced data" as a key problem [2]. Researchers often apply resampling or SMOTE techniques to correct this, but imbalance still limits real-world performance. Another problem is featuring quality: clinical attributes can be noisy or correlated. If irrelevant or redundant variables are included, even powerful models can suffer degraded performance. Careful data preprocessing and feature selection can mitigate this. Indeed, studies find that selecting the right subset of features can substantially boost accuracy. For instance, one work reported that applying L1-based feature selection before training a logistic model raised its accuracy from the usual ~80% up to about 89% [5]. This underscores that much of the performance gain in recent literature comes not just from new algorithms but from better data preparation.

Finally, there is a clear accuracy–interpretability trade-off. Simple models (linear regression, single trees) are transparent but may leave clinically relevant patterns unmodeled. On the other hand, ensembles and deep nets often achieve the highest accuracy on standard benchmarks [2], but their complexity necessitates additional explanation techniques [3][2]. In short, the existing literature suggests that practical promise lies in balancing these concerns: the most useful systems combine advanced learners for high sensitivity/specificity with explicit interpretability layers to gain clinician trust.

Methodology

This chapter explains the practical development of the Heart Disease Prediction system. The implementation includes data preprocessing, exploratory data analysis, model training, evaluation, and visualization using Python and standard machine-learning libraries.

The dataset contains 1200 records and 10 clinical features, including age, chest pain type, resting blood pressure, cholesterol, maximum heart rate, exercise-induced angina, ST depression, number of major vessels, and thalassemia.

The target variable shows whether a patient has heart disease (1) or not (0).

To prepare the dataset: Missing values were removed, categorical columns (cp, slope, thal, ca) were encoded, continuous features were scaled using StandardScaler and, data was split into 80% training and 20% testing

EDA was performed to understand feature behavior. Key insights: thalach, cp, and oldpeak show strong correlation with heart disease

Patients with higher ST depression values have higher disease probability Exercise-induced angina is more common in positive cases. Correlation heatmaps and distribution graphs supported deeper understanding.

Conclusion

The increasing global burden of heart disease, combined with the limitations of traditional diagnostic approaches, underscores the need for innovative, data-driven solutions. Machine learning, with its ability to analyse complex, multi-dimensional clinical data, holds genuine promise in transforming early diagnosis and risk stratification. However, the adoption of such technologies in real-world healthcare settings is not straightforward. Success hinges not just on achieving high accuracy but also on ensuring that the systems are interpretable, ethically sound, and clinically validated.

This project attempts to address these challenges by building a machine-learning-based heart disease prediction system that is not only technically robust but also designed with practical clinical considerations in mind. By focusing on both predictive performance and model transparency, the work aspires to bridge the gap between academic machine learning research and real-world clinical utility. Ultimately, the goal is not to replace medical expertise, but to augment it—providing clinicians with reliable, explainable decision-support tools that enhance patient care and enable earlier, more personalized interventions in cardiovascular health.

Acknowledgement

The authors are very much thankful to management for providing us with facilities to conduct the experiments in the institution's laboratory and colleagues for their support to prepare this paper. We are also thankful to the Computer Science and Business System department for approving our research proposal and providing us with relevant resources to conduct research activities.

References

1. A classification and regression tree algorithm for heart disease modeling and prediction. *ScienceDirect*. Available at: https://www.sciencedirect.com/science/article/pii/S2772442522000703
2. A systematic review of machine learning in heart disease prediction. *PubMed Central*. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC12614364/
3. Heart disease prediction with a feature-sensitized interpretable framework for the Internet of Medical Things sensors. *PubMed Central*. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC12522401/
4. Neural network-based coronary heart disease risk prediction using feature correlation analysis. *PubMed Central*. Available at:

https://pmc.ncbi.nlm.nih.gov/articles/PMC5606055/

5. Optimal feature selection for heart disease prediction using modified Artificial Bee Colony (M-ABC) and K-nearest neighbors (KNN). *Scientific Reports*. Available at: https://www.nature.com/articles/s41598-024-78021-1
6. Personalized health monitoring using explainable AI. *Scientific Reports*. Available at: https://www.nature.com/articles/s41598-025-15867-z
7. SFC Whitepaper on personalized cardiovascular health monitoring. *Société Française de Cardiologie (SFC)*. Available at: https://www.sfcardio.fr/wp-content/uploads/2025/11/SFC-ACVD-2511.pdf