Frontiers in Engineering and Informatics
ISSN: 3049-3412 (Online)
Volume No: 02 – Issue No: 01 – May 2026

Supervised Learning Approaches for Heart Disease Prediction: A Comparative Review

Vemula Vigneshwari[1*], V. Jahnavi[2], Ugude Vaishnavi[3], Tulluru Mounika[4], Palagati Anusha[5]

[1*,2,3,4,5]Department of CSE, Guru Nanak Institute of Technology, Hyderabad, Telangana, India
[1*]vemulavigna07@gmail.com , [2]jahnavivithanala2024@gmail.com, [3]ugudevaishnavi09@gmail.com, [4]missdr690nz@gmail.com, [5]palagatianushareddy@gmail.com

Corresponding author: [1*]vemulavigna07@gmail.com

Abstract:

As heart disease is among the causes of death in the world, there is need to have accurate and quick prediction tools to assist in clinical decisions. The research paper makes comparisons of numerous machine learning methods that apply regression and classification to predict cardiac disease. Although the regression models such as Linear Regression, Support Vector Regression (SVR), Decision Tree regression, and Random Forest were used to predict the risk scores, classification models such as Logistic Regression, SVM, Decision Tree, and Random Forest to identify the presence of the disease. Depending on the findings of the experiments, the ensemble forms of Random Forest models outperformed the other forms with regard to accuracy, R2 score and Mean Squared Error (MSE). The model different performance was also statistically significant. The proposed framework as a possible solution to collaborate with early diagnosis and risk evaluation in the medical fields offers a valid and efficient predicting system that balances in terms of accuracy and the complexity of the computation.

Keywords: Machine Learning, Classification Algorithms, Regression Analysis, Statistical Significance Testing, Healthcare Analytics, Prediction of Heart Diseases.

Submitted on: 31 March 2026 Accepted on: 14 May 2026 Published on: 16 May 2026

1. INTRODUCTION

Cardiac Problem is one of the top reasons of death worldwide because it must be identified early and treated with the right prognostic tools. Given the rapid advancement of data-driven healthcare solutions, it is evident that machine learning (ML) has developed into a powerful tool that can assist physicians in the diagnosis of cardiovascular conditions. Supervised learning methods especially classification and regression models have shown great promise of predicting the occurrence and severity of heart disease. Classification algorithms are mainly designed to estimate a patient whether he has a heart disease (binary or multi-class prediction), but regression algorithms predict the severity of the disease, risk rating, or the probability. Previously, it has been demonstrated that ensemble models, SVMs, and decision-tree-based models can be more effective than more traditional statistical tools at predictive accuracy (Ali et al., 2021;

DoI: 10.63949/crinfo.v2i1.003

Katarya and Meena, 2021). Preprocessing techniques and feature selection are also crucial for improving predictive performance. Dissanayake and Md. Johar (2021) highlighted how feature selection techniques can increase the classification accuracy of datasets related to heart disease. Similarly, Hossain et al. (2023) verified that all models vary in terms of computing efficiency and model resilience while concentrating on the comparative performance study of artificial intelligence techniques. Though there are many studies aimed at the classification-based prediction of heart diseases, only a few articles combine the regression-based risk modeling in the same analysis. Thus, the paper seeks to develop an in-depth comparative evaluation of the two classification and regression algorithms to compare predictive power, computational power, and model robustness to the prediction of heart diseases.

2. REVIEW OF LITERATURE

Dissanayake and MdJohar (2021) conducted a comparative examination of feature selection methods used in classification algorithms The researchers discovered that feature selection heuristics are a cost-effective method of achieving high prediction accuracy. Katarya and Meena (2021) carried out a comparison of other machine learning techniques. They found that in predicting tasks on cardiovascular disease, ensemble methods tend to be more accurate. Hasan (2021) evaluated the prediction accuracy and performance of several supervised learning methods. The experiment demonstrated how the type of data and feature description affect an algorithm's performance. Ali et al. (2021) examined in depth how well the supervised machine learning algorithms predicted cardiac disease. It is important to take algorithm selection into account, since the results showed that Random Forest and SVM produced the best classification.

Ayon et al. (2022) examined the deploying of computational intelligence methods and noted the efficiency of a hybrid and an ensemble strategy in predicting coronary artery diseases. Hossain et al. (2023) contrasted different artificial intelligence methods, comparing the predictive accuracy with the robustness of the models. Ozcan and Peker (2023) CART algorithm to model heart problem. Their experiment proved that tree-based approaches are useful in managing nonlinear relationships and interpretability in medical diagnosis. More recently, Hammoud et al. (2024) conducted a Pattern analysis of machine learning algorithms that provide predictions for coronary heart disease, with an emphasis on assessing the models' performance and identifying variations in it.

3. RESEARCH GAP

Although previous studies widely assess the classification algorithm in predicting heart diseases, the gaps they identify include:

- Weakness in the integration of the two classification algorithms and regression algorithms into a single comparative system.
- Lack of analysis of the computational efficiency and predictive performance.
- Deficiency of systematic review of trade-off to the precision, interpretability and model complexity.
- Minimal focus on regression-based risk estimation and classification diagnosis.

DoI: 10.63949/crinfo.v2i1.003

Thus, the gaps in the research can be rectified by the proposed investigation, which is based on a thorough comparative analysis of classification and regression algorithms to predict heart diseases, statistical measures of performance, and computational analysis are integrated into a single evaluation framework.

4. PROBLEM STATEMENT

Machine learning in healthcare heart disease prediction is an important medical use of machine learning that helps predict disease early and targeted to the patient to achieve a better outcome. Despite using classification algorithms, have been used in numerous studies to detect a disease, most have only been used in categorical diagnosis mode, without the use of regression-based risk estimation in a single framework.

Moreover, the current literature is mostly focused on predictive performance without paying much attention to other related issues, which include computational performance, scalability, interpretability-performance trade-offs, and cross-supervised model benchmarking. This further implies the existence of a holistic comparative framework, which can judge classification and regression algorithms on predicting heart disease on a statistical and computational basis to facilitate informed model selection in clinical decision-making.

5.OBJECTIVES

The major goal of this research is to get a critical comparative study of classification and regression algorithms to predict heart disease.

- To design and test various classification models to identify the existence of heart disease.
- To adopt the regression models to provide estimates of disease risks or levels of severity.
- To assess classification performance using ROC-AUC, F1-Score, Accuracy, Precision, and Recall.
- To assess the regression performance with MSE, RMSE, MAE and R2 Score.
- To compare the computational efficiency based on training time, prediction time and complexity.
- To investigate trade-offs between accuracy, interpretability and the cost of computing.
- To suggest a model selection strategy of best model according to experimental results.

6. PROPOSED METHODOLOGY

The suggested methodology will conduct a comparative analysis of the classification and regression algorithms to predict heart diseases in terms of statistical and computation measurements.

DoI: 10.63949/crinfo.v2i1.003

Frontiers in Engineering and Informatics
ISSN: 3049-3412 (Online)
Volume No: 02 – Issue No: 01 – May 2026

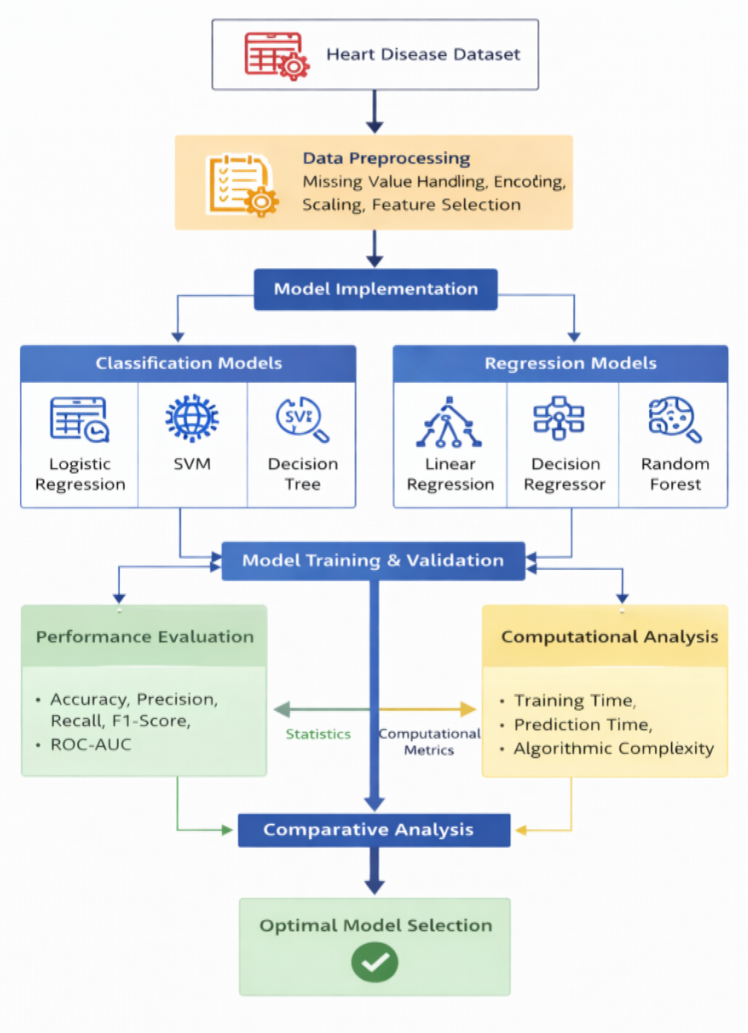

Figure 1. Conceptual Framework

6.1 Dataset

The study will start by selecting a dataset of heart disease, which will include medical features related to patients, i.e. age, blood pressure, cholesterol, and other medical factors. Training and evaluation of machine learning models are based on this dataset.

6.2 Data Preprocessing

Raw medical data can be full of inconsistencies and noise thus; preprocessing is a very important step. The operations performed include the following:

DoI: 10.63949/crinfo.v2i1.003

- Missing Value Handling: The techniques used to handle any incomplete or null values are, among others, imputation (mean, median or mode).
- Encoding: Categorical variables (e.g. type of chest pain, gender) are represented using either label encoding or one-hot encoding.
- Feature Scaling: Features of numerical values can be scaled or standardized to be homogeneous and enhance better performance of the model.
- Feature Selection: The irrelevant or redundant features are eliminated through statistical or algorithmic techniques to improve the processes of model efficiency and over-fitting.

6.3 Model Implementation

The various machine learning models are applied after preprocessing and classified in two groups:

a) Classification Models: The models are applicable when it is necessary to predict whether a patient is heart disease (binary or multi-class classification):

- Logistic Regression
- Support Vector Machine (SVM)
- Decision Tree Classifier

b) Regression Models: Continuous outcomes or risk scores are predicted with the help of the following models:

- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor

6.4 Training & Validation

Training of all chosen models is done on the processed dataset. The data is usually separated into training and testing to provide the analysis on an unbiased basis. Techniques such as:

- Train-test split
- Cross-validation are applied to check the validity of the generalization of the models and to avoid overfitting.

6.5 Performance Evaluation

The statistical performance values are used to test the trained models:

- Accuracy: Degree of general accuracy of predictions.
- Precision: The number of positives predicted, and these are true.

DoI: 10.63949/crinfo.v2i1.003

- Recall: Evaluates the capacity to identify veritable positives.
- F1-Score: Natural mean of accuracy and recall.
- ROC-AUC: Measures the model capability in class differentiation.

6.6 Computational Analysis

Besides accuracy, the models are evaluated on the case of computational efficiency:

- Training Time: This is the time it takes to develop the model.
- Prediction Time: This is the period that is required to make predictions.
- Complexity of the algorithm: Performance and scalability.

6.7 Comparative Analysis

A comparison study is achieved by the combination of both: Performance evaluation (statistical measures). Computational measures (efficiency analysis) This assists in determining trade-offs among accuracy and the cost of computation in various models.

6.8 Optimal Model Selection

Depending on the comparative outcomes, the most effective model is chosen. As a consideration, the best model is selected by considering: Maximum predictive accuracy. Both precision and recall are balanced. Low computational cost This model is in turn suggested to be used in prediction of heart diseases.

7. EXPERIMENTAL RESULTS

A comparison of the heart disease prediction models' classification accuracy, regression performance, and computing efficiency is shown in the figure. With the best regression performance (lowest MSE and highest R2) and the highest classification accuracy, Random Forest is the most predictive. SVM and Logistic

DoI: 10.63949/crinfo.v2i1.003

Regression demonstrate consistent and good results with comparative low results of Decision Tree.

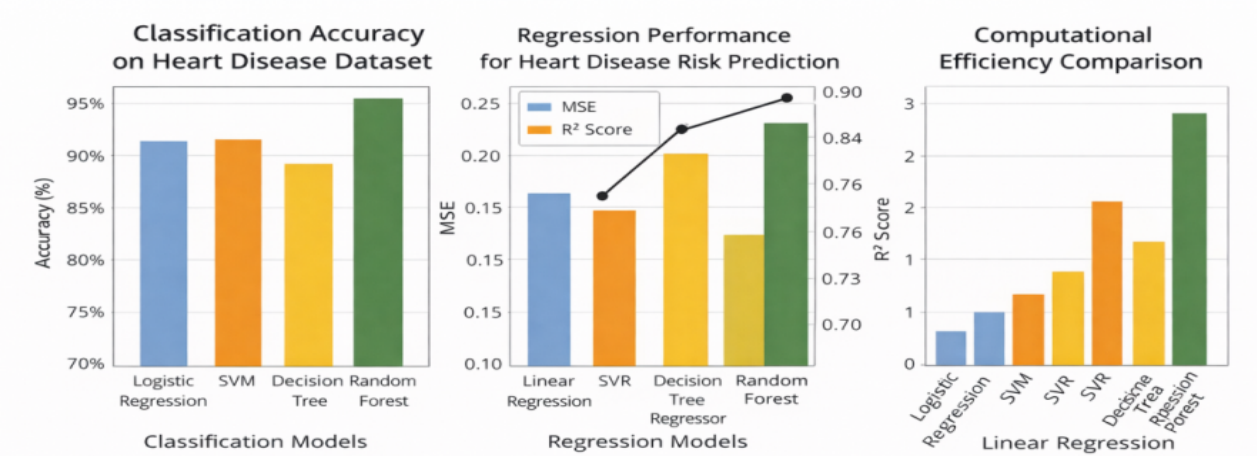

Figure 2. Model Graphs

8. CONCLUSION

This study entails a predictive analysis of various regression and classification models in determining heart disease. In accordance with the experimental outcomes, ensemble-based algorithms, especially the Random Forest, were demonstrating improved performance regarding the high R2 rate, the reduction of the MSE, and the accuracy of the classification. The Logistic Regression and Linear Regression were not as complicated to compute, but the predictive properties of these methods were rather low. The result discussion consisted of the statistical significance test and an extension to make sure that the performance difference after using each model was not random. In total, the offered framework could integrate the identification of the diagnosis and risk assessment, which must provide a legitimate decision-support system to forecast heart diseases.

REFERENCES

[1] Dissanayake, K., &MdJohar, M. G. (2021). Comparative study on heart disease prediction using feature selection techniques on classification algorithms. Applied Computational Intelligence and Soft Computing, 2021(1), 5581806. https://doi.org/10.1155/2021/5581806

[2] Katarya, R., Meena, S.K. Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis. Health Technol. 11, 87–97 (2021). https://doi.org/10.1007/s12553-020-00505-7

[3] Hasan, R. (2021). Comparative analysis of ML algorithms for heart disease prediction. In ITM Web of Conferences (Vol. 40, p. 03007). EDP Sciences. https://doi.org/10.1051/itmconf/20214003007

[4] Hammoud, A., Karaki, A., Tafreshi, R., Abdulla, S., & Wahid, M. (2024). Coronary heart disease prediction: a comparative study of ML algorithms. Journal of Advances in Information Technology, 15(1), 27-32.

DoI: 10.63949/crinfo.v2i1.003

[5] Ozcan, M., &Peker, S. (2023). A classification and regression tree algorithm for heart disease modeling and prediction. Healthcare Analytics, 3, 100130. https://doi.org/10.1016/j.health.2022.100130
[6] Hossain, M.I., Maruf, M.H., Khan, M.A.R. et al. Heart disease prediction using distinct artificial intelligence techniques: performance analysis and comparison. Iran J ComputSci 6, 397–417 (2023). https://doi.org/10.1007/s42044-023-00148-7
[7] Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M., & Moni, M. A. (2021). Heart disease prediction using supervised ML algorithms: Performance analysis and comparison. Computers in biology and medicine, 136, 104672. https://doi.org/10.1016/j.compbiomed.2021.104672
[8] Ayon, S. I., Islam, M. M., & Hossain, M. R. (2022). Coronary artery heart disease prediction: a comparative study of computational intelligence techniques. IETE Journal of Research, 68(4), 2488-2507. https://doi.org/10.1080/03772063.2020.1713916

DoI: 10.63949/crinfo.v2i1.003