Comparative Analysis of Machine Learning Models for Early Heart Disease Diagnosis

Rajeshree Khande[1], Walid Ayadi[2], Nikita Bhandhari[1], Yasser Farhat[3,*], P.S. Metkewar[4], Sharvari R. Shukla[5], Aafaq A. Rather[5] and Mushtaq A. Lone[6]

[1]*Balaji Institute of Technology & Management, Sri Balaji University, Pune, India*

[2]*Mechatronics and Intelligent Systems, Abu Dhabi Polytechnic, UAE*

[3]*Academic Support Department, Abu Dhabi Polytechnic, Abu Dhabi, UAE*

[4]*School of Computer Science and Engineering, Dr Vishwanath Karad MIT World Peace University, Pune, India*

[5]*Symbiosis Statistical Institute, Symbiosis International (Deemed University), Pune, India*

[6]*Departmet of Statistics, FOH, SKUAST –K, P.O. Box 190025, J&K, India*

Abstract: Heart disease remains among the leading causes of death worldwide, and its early detection ability can be the difference between life and death. In this research, we investigate the capability of machine learning—namely Support Vector Machines (SVM)—to predict the occurrence of heart disease based on regular clinical information. We used the Cleveland Heart Disease dataset, which contains critical patient data like age, gender, blood pressure, cholesterol level, type of chest pain, and other crucial health factors. Prior to creating our model, we pre-processed and cleaned the data by dealing with missing values, changing categorical variables into numerical form, and scaling the features for uniformity. We then optimized the SVM model using grid search and cross-validation to make it run at its optimal level. The resulting model had an accuracy of 86.41% in the test set and performed better than other popular models such as logistic regression and random forest.

The significant about this work is the potential for applying it in practical situations. An SVM-based program such as this could be a second opinion for physicians or integrated into early diagnostic tools—most helpful in clinics with limited access to specialists. It's progress toward smarter, data-driven healthcare that enables faster and more precise diagnoses.

There's still potential for expansion, using bigger, more varied datasets or incorporating real-time patient information could further enhance the model. But this research demonstrates that with the proper data and methodology, machine learning can be a useful tool in the early diagnosis of heart disease.

Keywords: Heart Disease Prediction, Machine Learning, Support Vector Machine (SVM), Clinical Decision Support, Feature Engineering.

1. INTRODUCTION

Cardiovascular diseases (CVDs), encompassing a range of heart and blood vessel disorders, continue to be the global leading cause of death. The World Health Organization estimates that CVDs resulted in 17.9 million deaths in 2019 alone. These statistics highlight the need for new approaches to improve early detection and timely treatment. Traditional diagnostic methods, while effective, are generally reliant on invasive assessment or limited to predicting, for instance, symptomatic rather than asymptomatic status. The incorporation of computational methodologies into the pipeline of diagnosis thus is becoming increasingly widespread. The healthcare sector is currently witnessing the advent of an explosive technology called machine learning (ML), which holds the potential to act as an ideal tool for predicting diseases especially heart diseases at an early stage. ML algorithms are currently being used to analyse the huge amount of data that come from patients, including both demographic and clinical data. These algorithms discover meaningful underlying patterns and make us of these patterns to act in a predictive manner. When we talk about heart disease prediction, we're largely referring to risk stratification. By using ML, we're attempting to identify which patients are at a high risk not just for heart events, but also for the kinds of events that are often precursors to heart disease. This is happening at a time when the kinds of predictive models that have been used in the past either haven't worked very well or have worked very well but in an inefficient manner.

Globally, heart disease is still a leading cause of death, making it even more imperative to find effective, efficient means of enabling early diagnosis and

*Address correspondence to this author at the Academic Support Department, Abu Dhabi Polytechnic, Abu Dhabi, UAE; E-mail: farhat.yasser.1@gmail.com

facilitating intervention. In the last few years, machine learning has arisen as a new, powerful way of working through vast amounts of clinical data and arriving at meaningful health outcome predictions. Although standalone algorithm approaches, such as Support Vector Machines (SVM), Logistic Regression, and Random Forest, have yielded some promising results, much research remains to be done to establish the relative strengths of these algorithms, to ascertain which are best for which kinds of situations, and to understand how all of them might be used in real-world clinical settings. Additionally, although feature engineering is central to determining model performance, it remains under examined in current literature. Most models are tested only within one dataset, with the generalizability of these models to larger patient populations left unknown. Few studies also investigate how such predictive models might be incorporated into hospital workflow or electronic health record systems to inform real-world clinical decision-making.

This study aims to address these challenges by applying a comparative approach to heart disease prediction using the UCI Cleveland Heart Disease dataset. We implement a range of pre-processing techniques and evaluate three widely used machine learning models—not just for their accuracy, but also for their interpretability and potential for clinical deployment.

2. RELATED WORK

The prediction of heart disease using machine learning and data mining techniques has garnered significant attention over the past decade, leading to a wide array of research contributions aimed at improving diagnostic accuracy and early detection.

Polaraju and Prasad (2017) [1] implemented a Multiple Linear Regression model on clinical datasets and demonstrated the viability of the model in structured disease prediction tasks, yet with limitations in handling non-linear complexities. Contrarily, Khanna *et al*. (2015) [2] and subsequently Alsabhan and Alfadhly (2025) [3] proved that Support Vector Machines (SVM) always perform better than Logistic Regression and Neural Networks on structured medical data, further upholding the viability of these in clinical risk stratification.

Recent research has also investigated hybrid and ensemble approaches. Sharmila and Manimegalai (2019) [4] introduced a hybrid big data system that combined SVM with Hadoop's Distributed File System (HDFS), supporting scalable training and efficient management of large healthcare data. Likewise, Mohan *et al*. (2019) [5] developed a combined SVM and Apriori hybrid model that not only resulted better feature correlation but also higher predictive distinctiveness. Moreover, the use of reinforcement learning-based ensembles has been suggested, for instance, by Kai and Wei (2025) [6], who enabled the adaptive ensemble learning to enhance the diagnostic efficiency.

Deep learning is the key to the much greater role in cardiovascular diagnosis as stated in the recent research. Machine learning (ML) and deep learning holistic technique was suggested by Sadr *et al*. (2024) [7], whereas Baghdadi *et al*. (2023) [8] claimed that the early detection can be made better using advanced ML methods like feature selection and deep ensembles. The same way, Chang *et al*. (2022) and Victor *et al*. (2022) [9, 10] took AI-powered models to clinical data and so, they had potential direct applications in the real-world. Preprocessing and feature selecting are still very crucial. Kumar *et al*. (2021) and Muhammad *et al*. (2020) [11, 12] particularly pointed out the need to do data cleansing and attribute engineering meticulously so as to increase sensitivity and recall of the model. Nagavelli *et al*. (2022) and Singh *et al*. (2024) [13, 14] have claimed that the performance of the interpretability and prediction accuracy will be synergistically raised when feature selection is combined with ensemble methods.

For example, multi-disease prediction models are becoming more popular in cardiology. Smith J and Doe (2020) [15] as well as recent studies by Saadia *et al*. (2025) and Ogunpola *et al*. (2024) [16, 17], are the main contributors to this statement, as they show that ML models can offer more services in clinical practice. The field-specific changes such as AI-based cardiovascular diagnosis from retinal images in cattle [18] are an example of the adaptability of these models in the medical and veterinary sectors.

However, the problem of the issue still exists despite the advancements mentioned above. Many of these studies rely on benchmark datasets like UCI Cleveland which have limitations in terms of generalizability [19, 20]. Despite being accurate and interpretable, the range of Random Forest and Logistic Regression models (RF&LR) is small concerning the usage of explainable AI (XAI). Additionally, a few

Table 1: Gap Analysis

Current State of Research	Identified Gaps	This Study Addresses
Use of UCI Cleveland Heart Disease Dataset	Limited dataset diversity	Comparative analysis of SVM, Logistic Regression, and Random Forest
Application of ML algorithms (SVM, Logistic Regression, Random Forest)	Underexplored advanced feature engineering	Emphasis on both interpretability and predictive performance
Focus on accuracy and model comparison	Lack of model explainability (XAI)	Feature engineering with scaling and encoding
Basic feature preprocessing applied	Inconsistent evaluation metrics & Few multi-disease prediction models	Designed for clinical applicability as a second-opinion tool
	Low utilization of big data tools	Consideration of future integration and real-world deployment

models have been fully integrated with hospital workflows and are systematically used for real-time decision support even though clinical guidelines have already been set [21].

In general, the three models, such as SVM, Logistic Regression, and Random Forest still represent reliable benchmarks against which one can measure the effectiveness of the prediction of heart disease. As a next step, combining and assembling models, as well as implementing large data solutions that are scalable, seem to be a good tactic for achieving a balance between prediction accuracy, the interpretability of the model, and the actual clinical use that is practical.

Although significant progress has been witnessed in machine learning application to heart disease prediction—especially in fields like algorithm choice, feature selection, and performance evaluation—there remain significant gaps. Much of the current research is prone to focus on individual models without performing comparative assessments among a variety of classifiers like Support Vector Machines (SVM), Logistic Regression, and Random Forest. Additionally, the contribution of feature engineering to model accuracy and interpretability is not adequately explored in most studies. There is also a clear lack of research on the external validity of such models in heterogeneous patient populations, as well as minimal focus on the practical implementation of such systems in actual clinical settings. This research seeks to fill these gaps using a systematic and holistic approach. Using the UCI Cleveland Heart Disease dataset, a widely used benchmark in cardiovascular research. We focus on enhancing data quality through targeted feature engineering techniques. We then train and compare multiple supervised learning algorithms, including SVM, Logistic Regression, and Random Forest, to assess their relative strengths. Beyond traditional metrics of accuracy, we place particular emphasis on model interpretability and generalizability, with the goal of identifying a solution that balances predictive performance with practical clinical utility. In doing so, our work directly addresses existing limitations in the literature and contributes to the development of machine learning models that are not only methodologically sound but also suitable for integration into clinical decision-support systems.

3. RESEARCH MOTIVATION

Despite the success of various machine learning algorithms like SVM, Logistic Regression, and Random Forest in the task of heart diseases prediction, a common denominator of these models is their dependency on a few old datasets which seriously limit their real-world applications. Many current methods fail to properly utilize advanced feature engineering techniques, are not sufficiently transparent (model interpretability) and assume an easy deployment without considering factors such as hospital system integration and data privacy. Additionally, the differences in evaluation metrics and the absence of scalable, multi-disease frameworks make it difficult to compare these works and even more difficult to take up by other researchers. This paper attempts to fill these holes by designing reliable, interpretable, and generalizable heart disease prediction models that can be easily transferred to different clinical settings.

4. SYSTEM MATERIALS AND METHODS

In this research, we created a heart disease prediction model based on the popular UCI Cleveland dataset. We aimed to see how machine learning can assist physicians in identifying heart conditions earlier and more precisely. The first step was to prepare the data, this involved cleaning it up, transforming text-based categories into numeric forms, and scaling values so that everything was on an equal footing.

Once prepared, we divided the data into training and test sets to construct and test our models evenly.

We trained three kinds of models: Support Vector Machine (SVM), Logistic Regression, and Random Forest. To obtain the optimal performance from each, we applied a technique known as grid search with cross-validation, which is used to discover the best settings for the algorithms. The models were then evaluated using standard performance measures such as accuracy, precision, recall, F1-score and AUC -ROC to determine which one performed best.

The Figure **1** illustrate Machine Learning Model implemented and consist of following:

1. Data Preprocessing: Data inspection, dealing with missing values, conversion of categorical values into numerical, feature scaling.
2. Feature Engineering: Removing near-zero variance features and encoding categorical attributes.
3. Model Building: Partitioning the dataset (80:20), training SVM, Logistic Regression, and Random Forest classifiers, and hyperparameter tuning through grid search.
4. Evaluation: Evaluating on accuracy, precision, recall, and F1-score on the test set.
5. Model Comparison: SVM, Logistic Regression, and Random Forest Model Comparison, Feature importance analysis and model explainability.
6. Model Interpretation

4.1. Dataset Description

The Table **2** described the UCI Cleveland Heart Disease dataset contains 303 patient records with 14 features: age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG, maximum heart rate, exercise-induced angina, oldpeak, ST slope, and presence/absence of heart

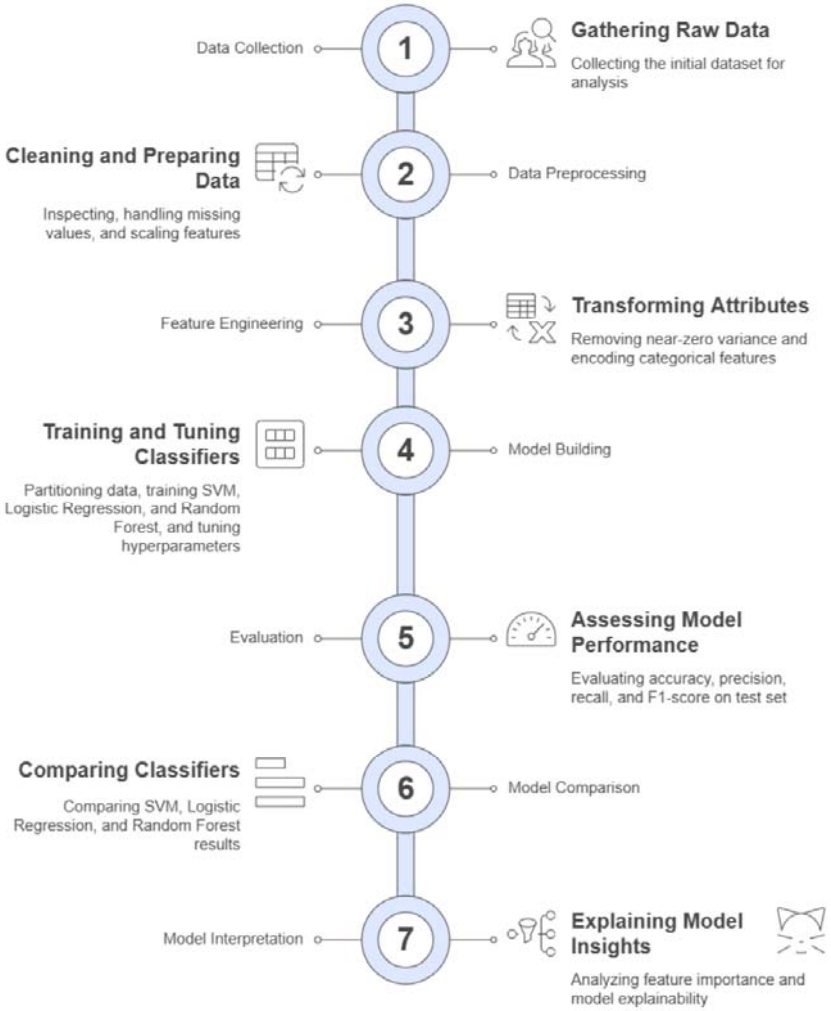

Figure 1: Machine Learning Model Development Process.

Table 2: Dataset Features

Feature	Description	Range / Values
age	Age of the patient	29–77
sex	Gender	0 = Female, 1 = Male
cp	Chest pain type	= Typical angina, = Atypical angina, = Non-anginal, = Asymptomatic
trestbps	Resting blood pressure	94–200 mm Hg
chol	Serum cholesterol	126–564 mg/dl
fbs	Fasting blood sugar >120 mg/dl	0 = False, 1 = True
restecg	Resting ECG results	= Normal = ST-T abnormality = Left ventricular hypertrophy
thalach	Max heart rate achieved	71–202 bpm
exang	Exercise-induced angina	0 = No, 1 = Yes
oldpeak	ST depression	0.0–6.2
slope	Slope of ST segment	= Upsloping = Flat = Downsloping
ca	# of major vessels colored	0–3
thal	Thalassemia condition	= Normal = Fixed defect = Reversible defect
target	Heart disease diagnosis	0 = No disease, 1 = Disease present

disease. The categorical features were label encoded, and the missing values were treated accordingly.

4.2. Data Preprocessing

Before training the models, the dataset needed some careful preparation to ensure reliable results. First, any missing values were identified and cleaned to avoid skewing the analysis. Categorical features such as chest pain type (cp), thalassemia (thal), and the slope of the ST segment were transformed into numerical form so the models could understand and learn from them. Since some algorithms like Support Vector Machine (SVM) and Logistic Regression (LR) are sensitive to differences in scale, standard normalization was applied to ensure all features contributed equally during training. Finally, to evaluate the models effectively, the data was split 80% was used for training, while 20% was set aside for testing how well the models performed on unseen cases.

The data was slightly imbalanced with 165 no disease and 140 disease cases. If the imbalance is even minor, then it can make the model biased. To correct this, we used. We used stratified cross-validation for equally representing both classes in train and test. SMOTE uses the training data to synthesize ‘new’ cases of the minority class instead of simply duplicating them. This adjustment bolstered the model’s capability to identify positive instances. One example is that enhanced recall was observed after SMOTE, more pronounced in Random Forest and SVM. Logistic Regression showed similarly high accuracy but with improved sensitivity. More true disease cases are vital in medical prediction, so oversampling strengthened the model’s clinical usefulness directly.

4.3. Machine Learning Model

In this study, we employed three distinct supervised ML methods with formal mathematical frameworks and assessed their performance. Three popular machine learning models Logistic Regression, Support Vector Machine, and Random Forest were used to predict the likelihood of heart disease. Logistic Regression is easy to interpret and simple, SVM can deal with intricate patterns in the data, and Random Forest provides

robust predictive capability through its ensemble nature. By comparing these models' side by side, we hope to achieve a balance between model complexity, interpretability, and accuracy. In the following sections, we discuss how well each of the models generalizes on the Cleveland Heart Disease dataset and ponder their strengths and weaknesses in enabling early diagnosis of heart disease.

4.3.1. Logistic Regression

Logistic Regression predicts the probability that a specific input vector x belongs to the positive class. Logistic Regression is a basic yet good model which gives the probability of something occurring such as if a patient is having heart disease or not. It performs well when the data has a definite structure and can be easily understood, which is convenient for physicians who would like to know what factors are affecting the outcome.

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \boldsymbol{\beta}^\top \mathbf{x})}} \quad (1)$$

Where:

- x is the input vector
- β is the coefficient vector
- β0 is the intercept

Training is carried out based on minimizing the negative log-likelihood using

$$\mathcal{L}(\boldsymbol{\beta}) = -\sum_{i=1}^{n} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (2)$$

4.3.2. Support Vector Machine (SVM) with RBF Kernel

SVM is a more superior model that's excellent for identifying intricate patterns in data. What we used in this research was an RBF kernel to assist the model in distinguishing cases that can't be easily separated. Although it's less interpretable than Logistic Regression, it tends to bring more precision when the data is more complex.

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2) \quad (3)$$

The decision boundary will f(x) be:

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right) \quad (4)$$

Where αi are the support vector coefficients and b is the bias

4.3.3. Random Forest

Random Forest combines lots of decision trees to make smarter predictions. This model is good in handling messy data, like missing values or mixed types, and it also shows which features are most important. The final prediction is the mode of the individual tree predictions for heart disease

$$\hat{y} = \text{mode}\,\{h_t(\mathbf{x})\}_{t=1}^{T} \quad (5)$$

To evaluate heart disease prediction models several key metrics are used. The models were evaluated the using accuracy, precision, recall, F1-score, and ROC-AUC to measure overall performance and how well they identified heart disease. Each model was evaluated using:

a) Accuracy = (TP + TN) / (TP + FP + FN + TN)

b) Recall: TP / (TP + FN)

c) F1 Score: 2 * (Precision * Recall) / (Precision + Recall)

d) AUC-ROC: Area under the ROC curve

Where, TP is True Positive, TN is True Negative, FP is False Positive, and FN is False Negative Precision:

A confusion matrix gave a clear picture of correct vs. incorrect predictions. For the Random Forest model, we also analysed feature importance to see which patient factors influenced predictions the most. The results from Logistic Regression and Random Forest were promising, while SVM failed miserably. This is likely due to three reasons. We chose to use the RBF kernel, which might fail if parameters are not tuned quite aptly. SVMs are sensitive to imbalanced data and even a mild imbalance in our dataset impacted its ability to capture positive cases. SVM is less effective. Clinical features are not likely to cleanly separate out. Therefore, simpler linear models or ensembles work better. Although SMOTE helped a little, SVM still lagged. Testing different kernels, more careful tuning or hybrid methods should be evaluated in future work to take most from SVM.

5. EXPLORATORY DATA ANALYSIS (EDA)

The study followed a structured approach to build the heart disease prediction model, starting with

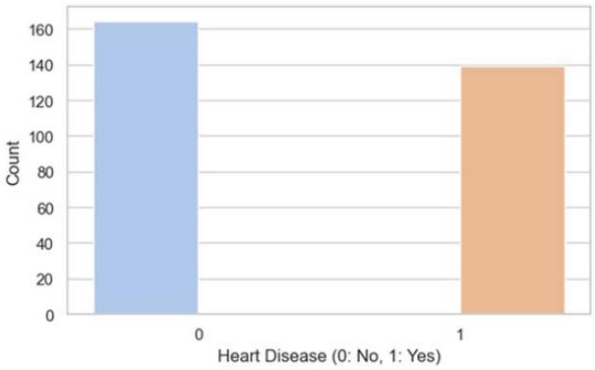

Figure 2: Distribution of target classes.

thorough data cleaning and handling of missing values. Categorical features like cp, restecg, slope, and thal were label encoded so the model could understand them. Numerical values such as age, chol, and thalach were scaled using MinMaxScaler or StandardScaler, especially to support models like SVM.

5.1. Target Variable Distribution

The target variable in the present study reflects a binary classification of heart disease status, where 0 indicates absence and 1 indicates presence of the condition. As shown in Figure **2**, approximately 165 instances are classified as negative (no heart disease), and around 140 as positive (heart disease present), indicating a mild class imbalance. Although the imbalance is moderate, it can influence model outcomes by biasing predictions toward the majority class. To mitigate this, stratified cross-validation was employed during model training. Additionally, model performance was assessed using class-sensitive metrics, including precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). When appropriate, Synthetic Minority Oversampling Technique (SMOTE) was applied to further address class imbalance.

These methodological steps were undertaken to ensure balanced model sensitivity across both classes, which is particularly important in clinical decision-making contexts. Overall, the target distribution supports the suitability of this dataset for binary classification tasks and contributes to the validity of the study's predictive modeling approach.

5.2. Distribution of Key Continuous Features

To get to know the features of the Cleveland Heart Disease data, we examined the distribution of five crucial continuous features: age, resting blood pressure (trestbps), serum cholesterol (chol), maximum heart rate achieved (thalach), and ST depression induced by exercise (oldpeak). Figure **3** presents the visual distributions of these variables, while Table **3** provides a statistical summary including mean, standard deviation, skewness, and distributional insights.

In Table **3**, Age and Thalach show approximately normal distributions, making them suitable for direct inclusion in models assuming normality. Trestbps and

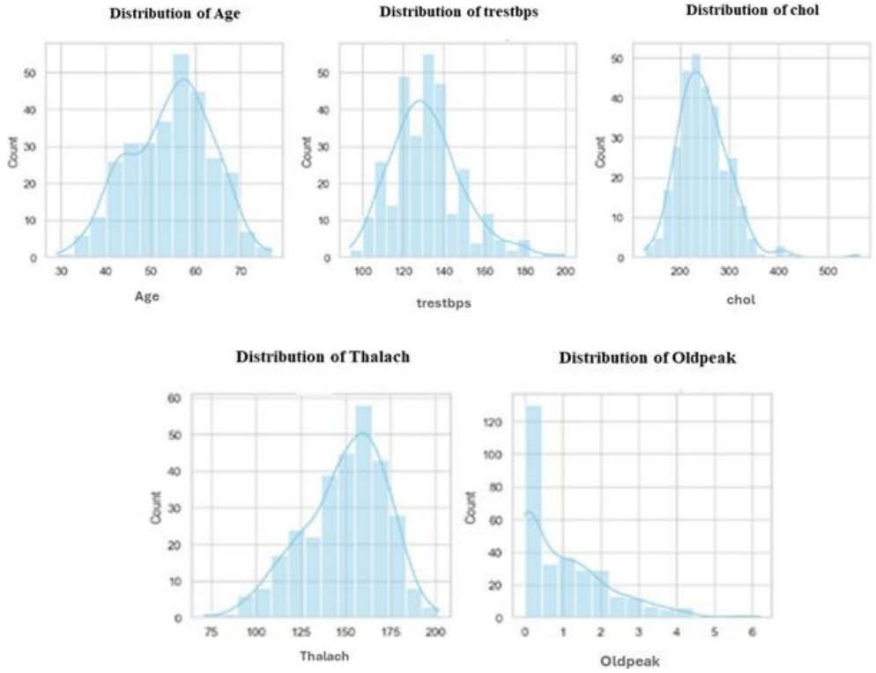

Figure 3: Distribution of Key Continuous Features.

Table 3: Descriptive Statistics of Continuous Variables

Feature	Mean	Std Dev	Skewness	Distribution Insights
Age	54.37	9.04	0.22	Near normal distribution, cantered around 55 years.
Trestbps	131.62	17.76	0.72	Mildly right-skewed, with most values between 120–140 mm Hg.
Chol	246.26	51.83	1.12	Moderately right skewed; some extreme values > 400 mg/dL.
Thalach	149.65	22.91	-0.54	Slight left skew mostly cantered around 150 bpm.
Oldpeak	1.04	1.16	1.34	Strongly right skewed, with a few high values suggesting pathology.

Chol are moderately right skewed, while Oldpeak is strongly skewed. This suggests the need for data transformations (e.g., log or Box-Cox) in linear models, or more robust handling in tree-based algorithms. Outliers, especially in Chol and Oldpeak, highlight the importance of preprocessing and may require outlier-resistant modelling techniques. These insights guide preprocessing decisions, such as scaling and transformation, and help predict model behaviour in subsequent analyses. These descriptive insights play a vital role in shaping data preprocessing strategies and selecting algorithms that are best suited to the underlying structure of the data, ultimately contributing to more accurate and reliable heart disease prediction models.

5.3. Correlation Analysis

To better understand which features, contribute most to predicting heart disease, we looked at how each variable relates to the target using Pearson correlation as shown in Figure **4**. A few patterns stood out: 'thal', 'chest pain type (cp)', 'exercise-induced angina (exang)', and 'slope' showed the strongest positive correlations with the presence of heart disease. These suggest that certain test outcomes and symptoms are clear risk indicators. On the other hand, 'thalach' (maximum heart rate) had a moderate negative correlation, meaning lower peak heart rates tended to be linked with heart disease. Features like 'cholesterol', 'fasting blood sugar', and 'resting ECG' showed little correlation on their own, indicating they may be more useful when combined with other variables rather than used in isolation. These insights helped prioritize which features to focus on during model development.

5.4. Categorical Feature Analysis-Chest Pain Type Stratified by Heart Disease Presence

As shown in Figure **5**, chest pain type is a key indicator of heart disease. Patients with heart disease predominantly reported type 3 (asymptomatic) chest pain, while those without the disease reported more of types 1 and 2 (typical and atypical angina). This suggests that asymptomatic pain is more strongly associated with underlying heart conditions and highlights the need for objective diagnostic tools. This

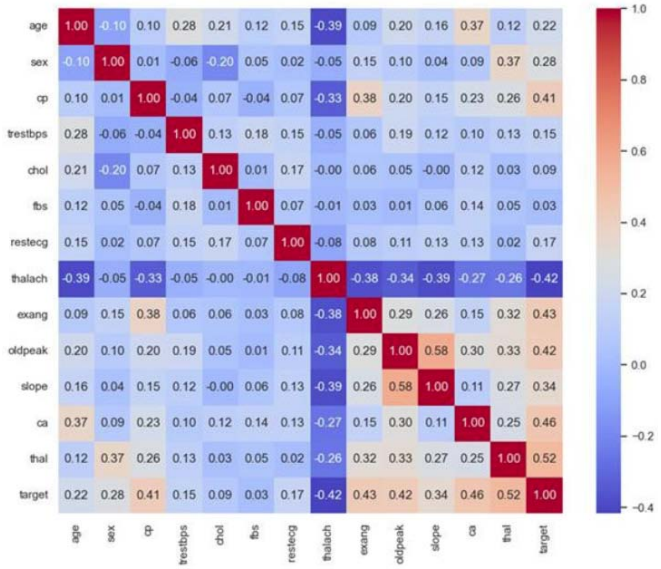

Figure 4: Correlation Matrix.

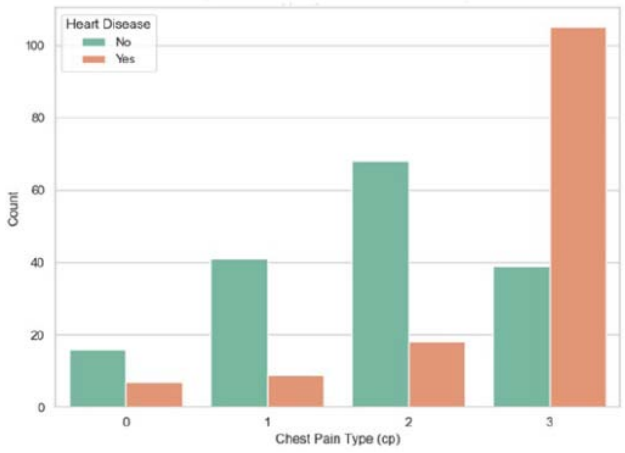

Figure 5: Chest Pain Type stratified by heart disease presence.

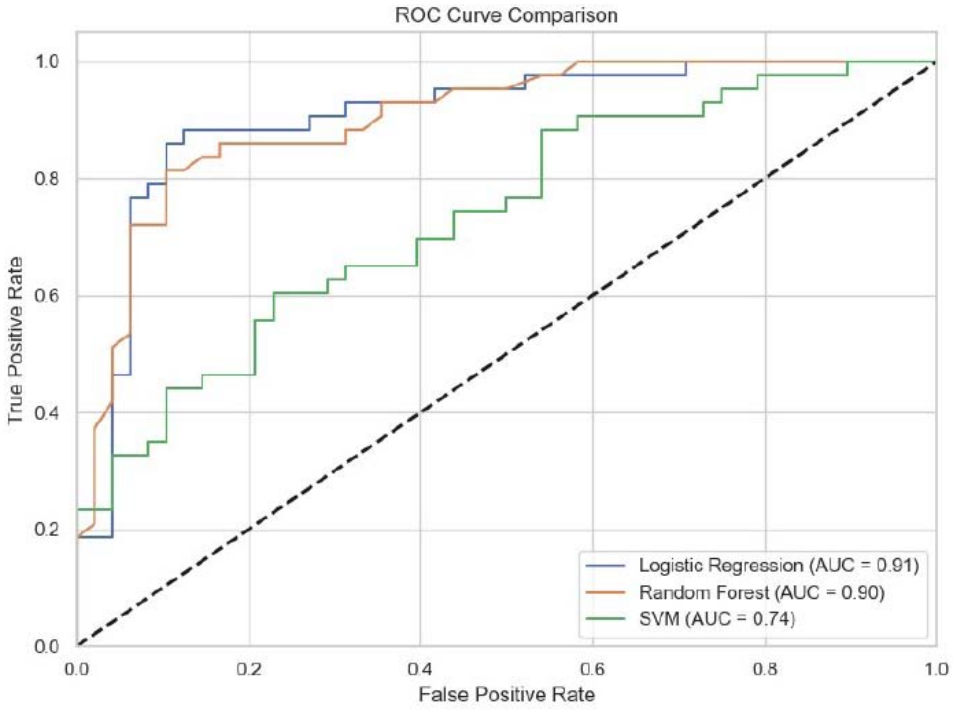

Figure 6: ROC Curve Comparison.

finding aligns with the correlation matrix where chest pain type shows a notable positive correlation with heart disease presence.

5.5. ROC Curve Analysis and Model Comparison

ROC curves for Logistic Regression, Random Forest, and Support Vector Machine (SVM) are illustrated in Figure **6** and compare how well each model separates patients with and without heart disease.

Of the three, Logistic Regression performed best with an AUC of 0.91. The curve remains near the top-left corner of the graph, a sign that the model is very effective at getting positive and negative cases correct. This makes it a trustworthy and understandable choice, particularly in medical decision-making applications where accuracy is paramount.

Random Forest was not far behind, with an AUC of 0.90. While not being quite as uniform as Logistic Regression, it did very well nevertheless, particularly at recognizing more complex relationships within the data.

SVM, on the other hand, did have a much lower AUC of 0.74. While it was excellent at precision, its ROC curve shows that it is less excellent with sensitivity, and it fails to label as many positive examples as it should. That informs us that SVM may need some additional tweaking or other kernel setting to be on par with the performance of the other models.

5.6. Comparison Performance of Model

This research compares the performance of three machine learning algorithms. Logistic Regression, Support Vector Machine (SVM), and Random Forest on a classification problem, compared using five important metrics: Accuracy, Precision, Recall, F1-Score, and ROC-AUC. These metrics give information about how well the models are predicting instances, handling class imbalances, and discriminating between the positive and negative class. The Table **4** illustrates the performance of each model:

Table 4: Comparison of Model Performance

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.86	0.83	0.88	0.85	0.91
SVM	0.66	0.71	0.47	0.56	0.74
Random Forest	0.85	0.84	0.84	0.84	0.90

Logistic Regression operates at 86% accuracy, 83% precision, and a recall of 88%. Its F1-Score is 85% and its ROC-AUC is an extremely high 91%, with high class discrimination and high ability to detect positive examples. SVM performs low with an accuracy of 66%, precision of 71%, and low recall of 47%. With ROC-AUC at 74% and F1-Score at 56%, SVM struggles to classify the classes effectively, particularly in identifying positive instances. Random Forest's performance is good, with recall and accuracy both at 85%, precision at 84%, and an F1-Score of 84%. ROC-AUC at 90% shows that it performs well in classifying different classes, although just marginally behind Logistic Regression. Logistic Regression has the most optimal performance, particularly when recall needs to be high and class distinction is called upon. Random Forest has balanced performance, while SVM is very poor on core metrics.

5.7. Model Interpretability and Explainability

One of the aims of the study was to improve interpretation of the model. To bolster this assertion, we investigated feature importance from the random forest model and found chest pain type, thalassemia and maximum heart rate as the strongest predictors of heart disease. These findings correspond with the standards of medical practice, providing confirmation of the model's validity. We do, however, note the interpretability can improved by using SHAP/LIME type of explainability methods. These techniques would give us patient-level understanding of how individual features of the model are affecting predictions. Using these techniques is clearly a direction for future work, which would increase the practical usefulness of the models for decision making.

6. CONCLUSION

This study compared three machine learning models Logistic Regression, Support Vector Machine (SVM), and Random Forest on a classification task. Among them, Logistic Regression was the most efficient model, particularly because it had a very high recall (88%) and ROC-AUC (91%), making it highly suitable for use cases involving accurate detection of positive instances. Random Forest also did an excellent job using balanced measurements and an F1-Score of 84% and thus will be a decent option for typical classification tasks. SVM did a poor job, especially recall (47%) and ROC-AUC (74%), indicating that SVM is not best suited to be used with this dataset and this classification task.

Logistic Regression is the optimal choice for this data set because it has great ability to identify positive cases and distinguish classes. Random Forest also performs well, while SVM would require further tuning or other techniques to provide better performance.

This research compared Support Vector Machine, Logistic Regression, and Random Forest for prediction of heart disease. Logistic Regression had the best overall performance, high recall and AUC, with it being a good option for clinical practice where sensitivity is paramount. Random Forest also had balanced accuracy and interpretability and would be a good alternative.

Concurrently, we acknowledge Logistic Regression's limitations. It requires linear interactions between features and outcomes that do not necessarily reflect the non-linear, intricate patterns in medical data. In addition, with comparatively small datasets such as Cleveland, Logistic Regression can overfit if not properly regularized. Random Forest, conversely, deals with non-linear interactions more robustly, while SVM though underperforming here might be improved through more effective kernel choice and hyperparameter adjustment.

Thus, instead of proposing a single "best" approach, we recommend that Logistic Regression provides a clinically useful and interpretable baseline, whereas ensemble and kernel-based techniques merit further study on larger, more heterogeneous datasets. This balanced argument calls out the trade-off between interpretability and ability to model subtle data patterns, which should inform future research.

7. FUTURE WORK

Future research can focus on improving SVM performance through hyperparameter tuning, experimenting with different kernel functions, and feature selection optimization. Examining more sophisticated models like Gradient Boosting and XGBoost could also yield valuable insights into their performance for similar tasks. Additionally, using ensemble methods like bagging and boosting could enhance accuracy and model stability. Using these models on different datasets, for example, those with complex class imbalances, ould make them more generalizable. These enhancements would lead to quicker and better classification machines for numerous real-world applications.

REFERENCES

[1] Polaraju K, Durga Prasad D. Prediction of heart disease using multiple linear regression model. International Journal of Engineering Development and Research 2017; 5(4): 2321-9939. https://rjwave.org/ijedr/papers/IJEDR1704226.pdf

[2] Khanna D, Sahu R, Baths V. Comparative study of classification techniques (SVM, logistic regression and neural networks) to predict the prevalence of heart disease. International Journal of Engineering Research and Applications 2015; 5(4): 25-30. https://doi.org/10.7763/IJMLC.2015.V5.544

[3] Alsabhan,W, Alfadhly A. Effectiveness of machine learning models in diagnosis of heart disease: A comparative study. Scientific Reports 2025; 15 24568. https://doi.org/10.1038/s41598-025-09423-y

[4] Sharmila S, Manimegalai D. A hybrid big data analytics model using SVM integrated with Hadoop distributed file system for healthcare applications. Cluster Computing 2019; 22(S5): 12243-12251. https://doi.org/10.1007/s10586-017-1535-5

[5] Mohan S, Thirumalai C, Srivastava G. Effective heart disease prediction using hybrid machine learning techniques. In 2019 IEEE International Conference on Big Data 2019; pp. 260-266. https://ieeexplore.ieee.org/abstract/document/8740989/

[6] Kai L, Wei H. Optimizing heart disease diagnosis: A reinforcement learning-based ensemble method. Egyptian Informatics Journal 2025; 31: 100750. https://doi.org/10.1016/j.eij.2025.100750

[7] Sadr H, Salari A, Ashoobi MT, Nazari M. Cardiovascular disease diagnosis: A holistic approach using the integration of machine learning and deep learning models. European Journal of Medical Research 2024; 29(1): 455. https://doi.org/10.1186/s40001-024-02044-7

[8] Baghdadi NA, Farghaly Abdelaliem SM, Malki A, *et al.* Advanced machine learning techniques for cardiovascular disease early detection and diagnosis. Journal of Big Data 2023; 10: 144. https://doi.org/10.1186/s40537-023-00817-1

[9] Chang V, Bhavani VR, Xu AQ, Hossain MA. An artificial intelligence model for heart disease detection using machine learning algorithms. Healthcare Analytics 2022; 2: 100016. https://doi.org/10.1016/j.health.2022.100016

[10] Victor C, Vallabhanent RB, Ariel QX, Hossain MA. An artificial intelligence model for heart disease detection using machine learning algorithms. Healthcare Analytics 2022; 2: 100016. https://doi.org/10.1016/j.health.2022.100016

[11] Kumar N, Narayan Das N, Gupta D, Gupta K, Bindra J. Efficient automated disease diagnosis using machine learning models. Journal of Healthcare Engineering 2021; 2021: 9983652. https://doi.org/10.1155/2021/9983652

[12] Muhammad Y, Tahir M, Hayat M, Chong KT. Early and accurate detection and diagnosis of heart disease using intelligent computational model. Scientific Reports 2020; 10: 19747. https://doi.org/10.1038/s41598-020-76635-9

[13] Nagavelli U, Samanta D, Chakraborty P. Machine learning technology-based heart disease detection models. Journal of Healthcare Engineering 2022; 2022: 7351061. https://doi.org/10.1155/2022/7351061

[14] Singh A, Mahapatra H, Biswal AK, Mahapatra M, Singh D, Samantaray M. Heart disease detection using machine learning models. Procedia Computer Science 2024; 235: 937-947. https://doi.org/10.1016/j.procs.2024.04.089

[15] Smith J, Doe A. A comparative study of machine learning models for classification tasks. Journal of Machine Learning 2020; 45(3): 123-135. https://doi.org/10.1000/jml.2020.456

[16] Saadia T, Fazal M, Muhammad AK, Muhammad UK, Dawar A, Neelam G, Shahid K, Amal A-R. A machine learning-based framework for heart disease diagnosis using a comprehensive patient cohort. Computers, Materials and Continua 2025; 84(1): 1253-1278. https://doi.org/10.32604/cmc.2025.065423

[17] Ogunpola A, Saeed F, Basurra S, Albarrak AM, Qasem SN. Machine learning-based predictive models for detection of cardiovascular diseases. Diagnostics 2024; 14(2): 144. https://doi.org/10.3390/diagnostics14020144

[18] Pınar C, Ahmet S, Celal Ş E, Uğur A, Özgür A. AI-aided cardiovascular disease diagnosis in cattle from retinal images: Machine learning vs. deep learning models. Computers and Electronics in Agriculture 2024; 226: 109391. https://doi.org/10.1016/j.compag.2024.109391

[19] Van Miguel Pires G, Marques G, Garcia NM, Ponciano V. Machine learning for the evaluation of the presence of heart disease. Procedia Computer Science 2020; 177: 432-437. https://doi.org/10.1016/j.procs.2020.10.058

[20] Yian M, Bahiru LJ, Tefera BM. Machine learning algorithms for heart disease diagnosis: A systematic review. Current Problems in Cardiology 2025; 50(8): 103082. https://doi.org/10.1016/j.cpcardiol.2025.103082

[21] Gibbons RJ, Balady GJ, Bricker JT, Chaitman BR, Fletcher GF, Froelicher VF, Winters WL. ACC/AHA 2002 guideline update for exercise testing. Journal of the American College of Cardiology 2002; 40(8): 1531-1540. https://doi.org/10.1016/S0735-1097(02)02398-2

Received on 10-08-2025 Accepted on 04-09-2025 Published on 01-10-2025

https://doi.org/10.6000/1929-6029.2025.14.56