AFrica Leadership

X - ACADAMY

Data Science Program

Heart Disease Prediction Using Robust Machine Learning Models for: A Comprehensive Performance Evaluation of Five Algorithms (LR, GB, SVM, ANN, RF)

Submitted To: **Dr. Alazar kebede(PhD)**

Submitted by: **Mubarik Meka**

Abstract

Background: Heart disease, also known as cardiovascular disease (CVD), refers to a group of disorders that affect the heart and blood vessels. It includes conditions such as coronary artery disease, heart failure, arrhythmias, and congenital heart defects It is a leading cause of death globally, especially in low- and middle-income countries (LMICs) like Ethiopia, where access to early diagnostic tools remains limited.

Objective: This study aimed to build and evaluate a machine learning–based model for predicting heart disease using publicly available data and explore its potential application in LMIC contexts.

Methodology: The dataset, containing 1024 records and 14 variables, was obtained from Kaggle. Some features were transformed, and the data set underwent pre-processing steps, including encoding, normalization, and outlier handling using IQR. The data was split int training and test set . five algorithms—Random Forest(RF), support vector machine(SVM) ,gradient bosting(GB), and Multilayer Perceptron,and logestic regression(LR)were implemented using python jupyter note book and other essential library like sklearn.

Result: This study evaluates machine learning models for heart attack risk prediction. Random Forest (RF) and Gradient Boosting (GB) outperformed others, achieving 99.08% and 98% accuracy with strong AUC scores (RF: 0.98, GB: 0.947), demonstrating ensemble methods' effectiveness. SVM performed moderately (90.5% accuracy, AUC: 0.947), while Logistic Regression (LR) scored 86.6% accuracy (AUC: 0.93). ANN trailed with 84.37% accuracy. Error analysis confirmed RF's superiority (RMSE: 0.174) over GB (0.226), SVM (0.277), and LR (0.308). Findings highlight RF and GB as optimal for clinical risk prediction, with future work needed to improve ANN performance.

Table of Contents

List of Figures

List of Table

Abbreviations

1. **CHD:** Coronary Heart Disease
2. **ECG:** Electrocardiogram
3. **EHR(s):** Electronic Health Record(s)
4. **HD:** Heart Disease
5. **KNN:** K-Nearest Neighbors (algorithm)
6. **LMIC(s):** Low- and Middle-Income Country(ies)
7. **ML:** Machine Learning
8. **MLP:** Multilayer Perceptron (a type of Neural Network)
9. **WHO:** World Health Organization
10. **RF: Random forest**
11. **GB:Gradient bosting**
12. **SVM: support vector machine**
13. **LR: logestic regression**

1. Background

1.1 Introduction

Heart disease, also known as cardiovascular disease (CVD), refers to a group of disorders that affect the heart and blood vessels. It includes conditions such as coronary artery disease, heart failure, arrhythmias, and congenital heart defects. Globally, heart disease is the leading cause of death, accounting for an estimated 17.9 million lives each year. Key risk factors include high blood pressure, high cholesterol, smoking, diabetes, obesity, physical inactivity, and unhealthy diet. Early detection, lifestyle modification, and proper medical management play a critical role in preventing complications and improving quality of life for affected individuals[1].It continues to pose a major global health challenge, affecting millions and exerting enormous pressure on healthcare systems worldwide [3]

.Accurate prediction and timely diagnosis of heart disease are critical for reducing morbidity and mortality[7]. Traditional diagnostic approaches rely on clinical assessments, patient history, and diagnostic tests, which are often costly, time-consuming, and limited in early detection capacity [3].

Machine learning (ML) techniques have emerged as promising tools that harness vast datasets—comprising demographics, vital signs, lab results, and lifestyle information—to detect complex, non-linear patterns predictive of heart disease [6][9]. By providing early and accurate risk assessment, ML empowers clinicians to implement targeted preventive and therapeutic interventions, thereby reducing adverse outcomes [7].

1.2 Statement of the Problem

Heart disease is the leading global cause of death, with coronary heart disease (CHD) accounting for the majority of fatalities. The 2017 Global Burden of Disease study reports heart disease as responsible for approximately 43% of all deaths worldwide, contributing to an estimated 17.5 million deaths annually [1], [2], [3]. Economically, heart disease accounted for an estimated USD 3.7 million in costs globally between 2010 and 2015, including direct medical expenses and productivity losses [4].

In low- and middle-income countries (LMICs) such as Ethiopia, healthcare funding constraints and high out-of-pocket expenses exacerbate the financial burden of heart disease, especially when diagnosis occurs late or treatment is unavailable [5]. Heart disease often progresses silently and is diagnosed at advanced stages, limiting the effectiveness of interventions.

While high-income countries have stabilized mortality rates through enhanced screening, education, and treatment advances, LMICs—including Ethiopia—face rising incidence due to rapid urbanization, dietary changes, low awareness, and poor diagnostic infrastructure [3][6]. Existing predictive models often fail to address the complex, localized epidemiology of heart disease in LMICs [2]. Conventional models, such as the Framingham Risk Score, assume linear relationships and often overlook intricate variable interactions [3]. Additionally, ML models developed on small or imbalanced datasets tend to over fit and lack generalizability [4].The black-box nature of advanced ML algorithms further impedes clinical acceptance [7].

Therefore, this study aims to develop a robust machine learning model tailored for heart disease prediction in LMIC settings, focusing on Ethiopia.

1.3 Objectives of the Study

1.3.1 General Objective

To develop and evaluate a machine learning model to predict heart attack risk using a publicly available dataset and assess its applicability in an Ethiopian healthcare context

1.3.2 Specific Objectives

- To pre-process and clean heart attack-related data for optimal analysis.
- To apply multiple machine learning algorithms for heart attack prediction.
- To compare model performance and identify the most effective algorithm.
- To interpret the most significant predictors of heart attack risk

1.4 Scope of the Study

This study focused on predicting heart attack using machine learning techniques by analyzing clinical and behavioral data. The data was sourced from Kaggle and did not represent hospital-specific records from Ethiopia. However, the intent was to demonstrate the potential of ML algorithms for early heart attack prediction, with emphasis on the feasibility of local adaptation in resource-limited settings.

1.5 Limitations of the Study

This study did not consider other chronic conditions like HIV, TB, and diabetes, which may influence cardiovascular health. Moreover, due to dataset limitations, key biomarkers such as troponin levels and clinical manifastation were not included. The dataset may not fully represent Ethiopian demographics and clinical profiles.

1.6 Significance of the Study

Accurate and early prediction of heart disease through ML can enable timely clinical interventions, reducing mortality and healthcare costs. ML models can support clinicians in identifying high-risk individuals for preventive care and facilitate more efficient resource allocation. Furthermore, integrating ML into healthcare workflows can automate routine tasks, allowing focus on complex cases, ultimately improving patient outcomes and healthcare system efficiency.

2. Literature Review

2.1 Theoretical Review

2.1.1 Global Burden of Heart Disease

Heart disease remains the leading cause of death worldwide, with CHD responsible for 43% of deaths globally [1]. Annual mortality exceeds 17.5 million people [2] [3]. The economic impact is substantial, with costs around USD 3.7 million globally between 2010 and 2015 [4]. In resource-limited countries like Ethiopia, delayed diagnosis and limited access to care amplify this burden [5].

Unlike other chronic illnesses, heart disease often progresses silently and is diagnosed late, highlighting the importance of early detection. While high-income countries have mitigated mortality through improved screening and treatment, LMICs face challenges due to urbanization, lifestyle changes, and weak healthcare infrastructure [3], [6].

2.1.2 Machine Learning Applications in Heart Disease Prediction

Machine learning offers efficient, scalable alternatives to traditional diagnostic methods like angiography or treadmill tests by analyzing comprehensive patient data to identify subtle risk patterns [7]. Among ML algorithms, K-Nearest Neighbors (KNN) achieved 87% accuracy on the UCI Cleveland dataset, outperforming Support Vector Machines and Decision Trees [5]. Multilayer Perceptron (MLP) models demonstrate robust performance, particularly in recall, which is critical for minimizing false negatives [3]. However, most models are trained on high-income country data, limiting their applicability in Ethiopian populations due to differences in lifestyle (e.g., khat chewing), prevalent comorbidities (e.g., rheumatic heart disease), and data quality [8], [9].

2.1.3 Risk Factor Synthesis

Heart disease risk factors span behavioral (smoking, inactivity), metabolic (hypertension, dyslipidemia), genetic (family history), and environmental (air pollution, stress) domains [1]. In Ethiopia, locally relevant factors like khat use and endemic rheumatic fever are critical to incorporate into predictive models, whereas advanced biomarkers common in Western datasets remain inaccessible [7], [8], [9].

2.2 Empirical Review

2.2.1 Validation Challenges and Ethiopia-Specific Gaps

Applying ML models developed in high-income settings to LMIC populations' results in performance drops of 22–37% due to epidemiological and data differences [9], [10]. Ethiopia's unique factors—including lifestyle, delayed care seeking, and lack of comprehensive electronic health records—further limit model reliability and necessitate localized model development [10], [11].

3. Methodology

This study proposes developing an ML-based system for heart disease prediction using clinical datasets from kaggele website. The approach involves data cleaning, preprocessing, normalization, and model training to improve accuracy and robustness.

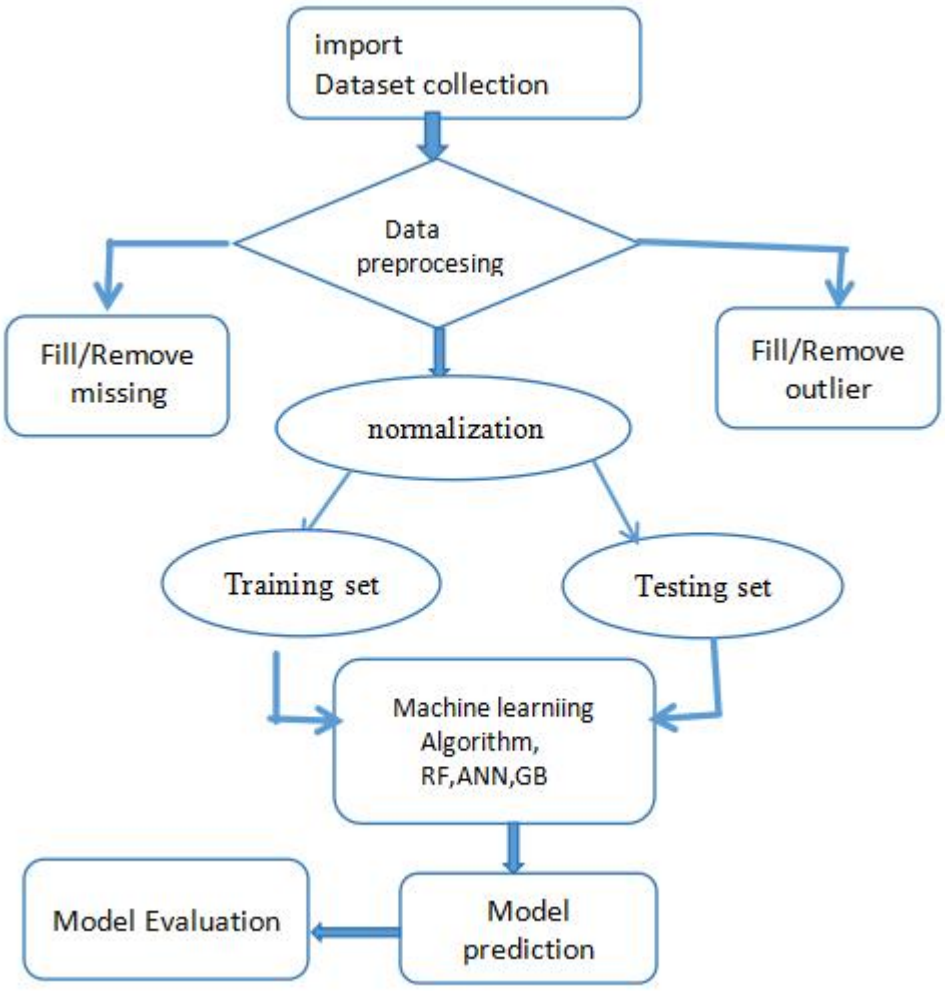


Figure 1 archithictur diagram of proposed system

3.1 Data Pre-processing

The dataset utilized in this study was sourced from the Kaggle ,where it is publicly available for academic and research use. It comprises 1,024 records and 14 variables related to cardiovascular health. The target variable, Heart Disease, is binary, with 0 representing the absence of a heart attack and 1 indicating a confirmed case.To prepare the dataset for machine learning analysis, a structured preprocessing approach was applied. This included the removal of numerical outliers using the Interquartile Range (IQR) method, visual inspection of categorical variables using bar plots, detection of outliers, and replacing them with the nearest plausible values to maintain data integrity

A. Variable observation and Selection.

The preprocessing began by observing informativness, redundancy, or ambiguity of the features that are directly relevant to the prediction of heart attack risk.

After that, 13 feature variablesand 1 outcome were retained for analysis. These include essential demographic, clinical, behavioral, and lifestyle features:

- age: age in years
- sex: (1 = male; 0 = female)
- cp: chest pain type (0/1/2/3)
- trestbps: resting blood pressure (in mm Hg on admission to the hospital)
- chol: serum cholestoral in mg/dl
- fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = fals
- restecg: resting electrocardiographic results
- thalach: maximum heart rate achieved
- exang: exercise induced angina (1 = yes; 0 = no)
- oldpeak: ST depression induced by exercise relative to rest
- slope: the slope of the peak exercise ST segment
- ca: number of major vessels (0-3) colored by flourosopy
- thal: 1 = normal; 2 = fixed defect; 3 = reversable defect
- target: (0 = did not occur, 1 = occur)

As part of the feature refinement, the original composite of number of major vessel colored by flouroscopy variable was encode in into one hot encoding for better capture distinct clinical information

B. Categorical Encoding:

Some variables in the dataset were categorical or ordinal in nature. These were systematically converted into numeric formats suitable for machine learning algorithms . like, Sex: Encoded as 0 (Female) and 1 (Male) **fbs** (fasting blood sugar > 120 mg/dl) and **exercise-induced angina** were encoded as **0 (No)** and **1 (Yes)**, while the categorical variable **ca** (number of major vessels) was **encoding** to avoid ordinal assumptions. This standardization ensured consistent numerical representation across all features, facilitating smoother model training and enabling fair comparison across different machine learning algorithms while preserving interpretability.

C. Outlier Detection and Treatment

Outliers in continuous variables were addressed using the Interquartile Range (IQR) method for capping. This technique preserves the structure of the dataset while limiting the impact of extreme values that could distort model performance. Specifically, values falling below the first quartile minus 1.5 times the IQR or above the third quartile plus 1.5 times the IQR were capped at the respective thresholds. This method is robust, non-parametric, and particularly well-suited for medical datasets where extreme physiological readings may occur due to rare conditions or data entry noise. Variables treated for outliers include thalach, serum cholestorals

D. Normalization

To ensure all continuous variables contributed equally during model training, Min-Max normalization was applied to the dataset. This technique scaled each numeric variable (except the target variable) to the [0, 1] range, preserving the original distribution but bringing variables to a uniform scale. This step is particularly beneficial for distance-based and gradient-based machine learning algorithms.

The normalization process was applied after outlier treatment and categorical encoding. The Heart Attack variable was excluded from normalization to preserve its binary classification structure.

By applying normalization, the dataset was made suitable for a wide range of machine learning models, improving performance consistency and convergence behavior across algorithms.

E. Data Balancing

Data balancing is essential for accurate result because by data balancing graph we can see that both the target classes are equal. Fig.3 represents the target classes where “0” represents with heart diseases patient and “1” represents no heart diseases pateints.

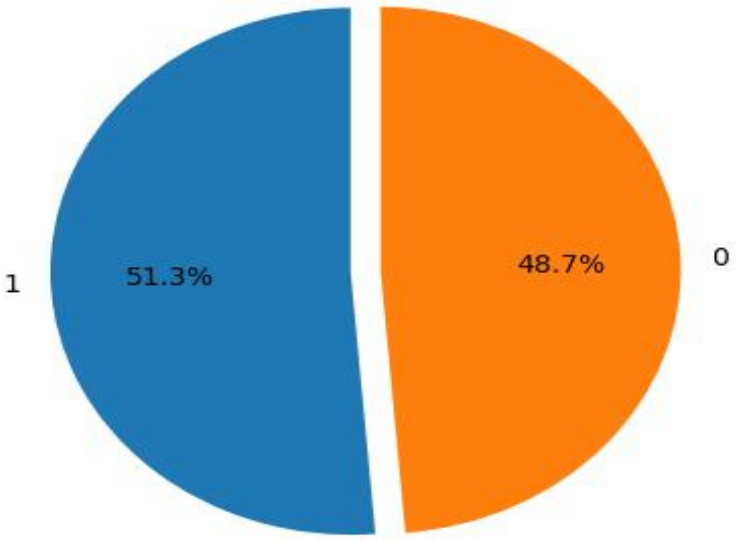
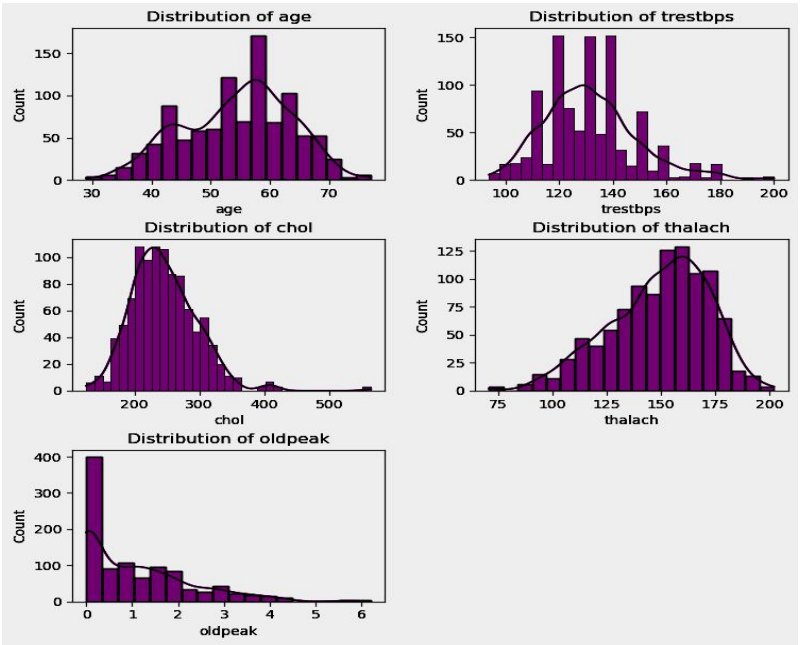


Figure 2 Target class view (0:no heart disease,1:heart disease)

F. Histogram and bar graph of attributes

Histogram and bargraph of attributes shows the range of dataset attributes and code which is used to create it.dataset.



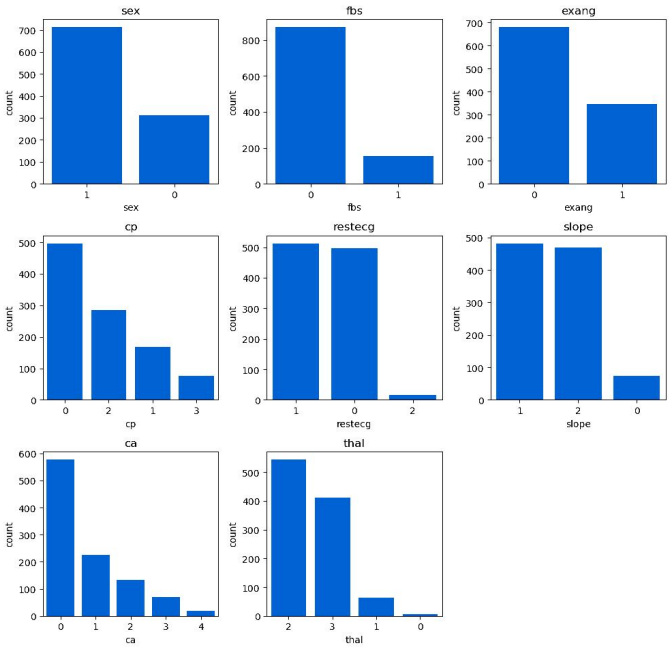


Figure 3 Histogram of numeric variable Figure 4 barplot of categorical variable

Figure 5 Correlation heatmap

G. Dataset Splitting to Prevent Data Leakage

To prevent data leakage and ensure proper evaluation of model generalization, the dataset was divided prior to any balancing procedure into: 70% for Training Set and 30% for Testing Set. This separation ensures that no synthetic data influences the model evaluation phase

3.2 Model Selection and Building

The following ML algorithms were applied using python jupyter note book :

- ✧ Random Forest
- ✧ Multilayer Perceptron (MLP)
- ✧ SVM
- ✧ Logistic regression
- ✧ Gradient Bosting

4. RESULT analysis

A. **About python**

Python was chosen for this research due to its powerful libraries (*Pandas, NumPy, SciPy*), intuitive visualization tools (*Matplotlib, Seaborn*), and machine learning capabilities (*Scikit-learn*). Its flexibility, reproducibility (*Jupyter Notebooks*), and open-source nature ensured efficient data processing, robust statistical analysis, and clear, shareable results—making it the ideal tool to uncover actionable insights from complex datasets.

B. **confusion matrix :** It is a performance evaluation tool for classification models that provides a detailed breakdown of predictions versus actual outcomes.for this project we assign 0 for no heart disease and 1 for heart disease

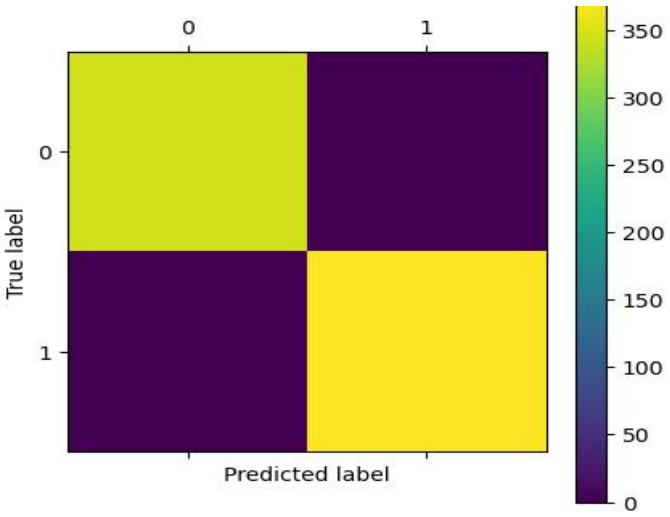


Figure 6 Confusion matrix for Random forest

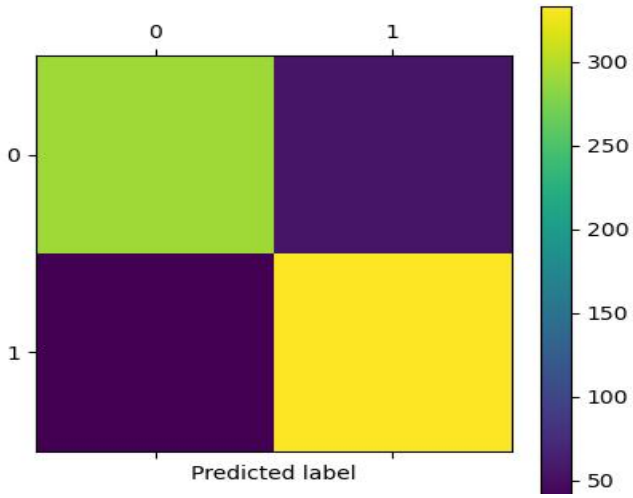


Figure 7 Confusion matrix for logistic.reg.

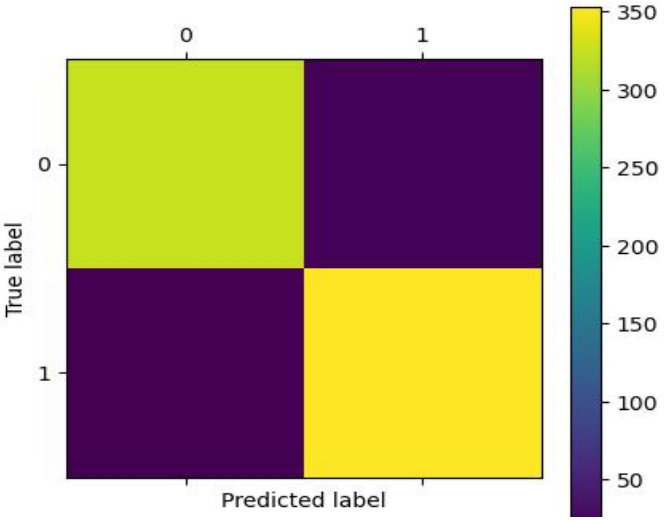


Figure 8 Confusion matrix for SVM

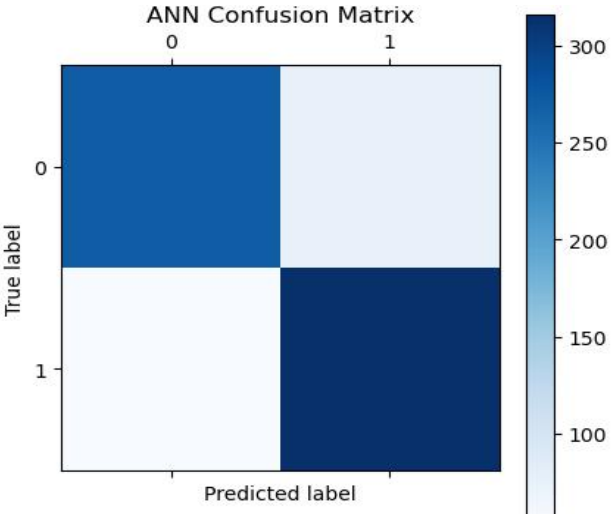


Figure 9 Confusion matrix for ANN

C. Result and Evaluation Metrics Overview

To evaluate model performance, key classification metrics—accuracy, precision, recall, F1-score, and confusion matrix—were used. Random Forest and GB performed best, with accuracy scores of **99.03%** and **98.19,Suport vector machine** followed with **90.5%** accuracy, while arteficial neural network showed the lowest performance with accuracyof **84.37%.**

Table 1 model evaluation through acuracy and precision

Algorithm	**Accuracy(%)**	**precision**
Random forest	**99.03**	**1.00**
Gradient Bosting	**98.19**	**1.00**
Logestic regression	**86.05**	**0.87**
SVM	**90.5**	**0.95**
ANN	**84.37**	**0.82**

Parameter obtimization

The logistic regression model achieved a best score of 0.866 (86.6%) with optimal parameters set to **C: 0.04** and **penalty: l2**, while the random forest classifier outperformed others with a remarkable **99.08% accuracy**. Gradient boosting (GB) also performed strongly with **98% accuracy**, whereas the artificial neural network (ANN) yielded a lower but still competitive score of **84.37%**. These results highlight the varying effectiveness of different machine learning models on the given dataset.

Most important features

Logistic regression showed ca_0 (0.867), thal_2 (0.718), and slope_2 (0.347) as key positive predictors, while cp_0 (-0.995) and exang (-0.432) were the strongest negative influences. Random Forest prioritized oldpeak (0.105) and thalach (0.103), whereas Linear SVM assigned the highest importance to oldpeak (2.235) and trestbps (1.433). These differences highlight how each model weighs features uniquely, offering varied insights into the data's predictive patterns.

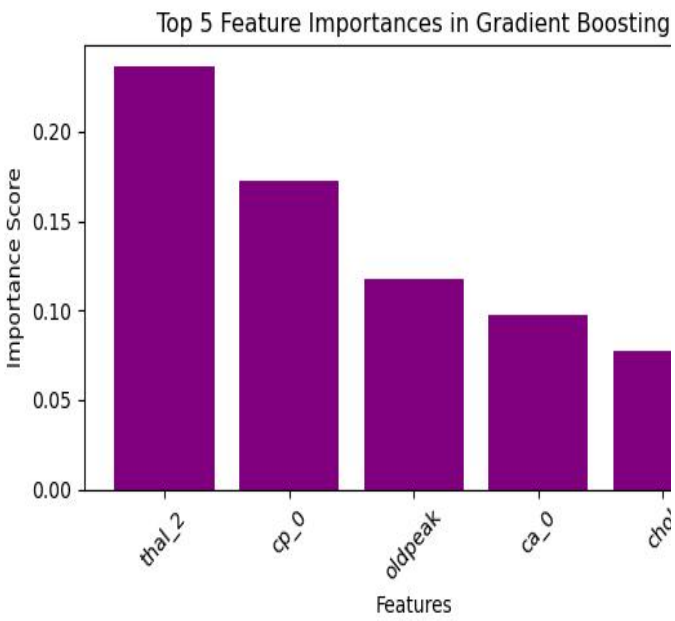


Figure 10 Key feature for GB

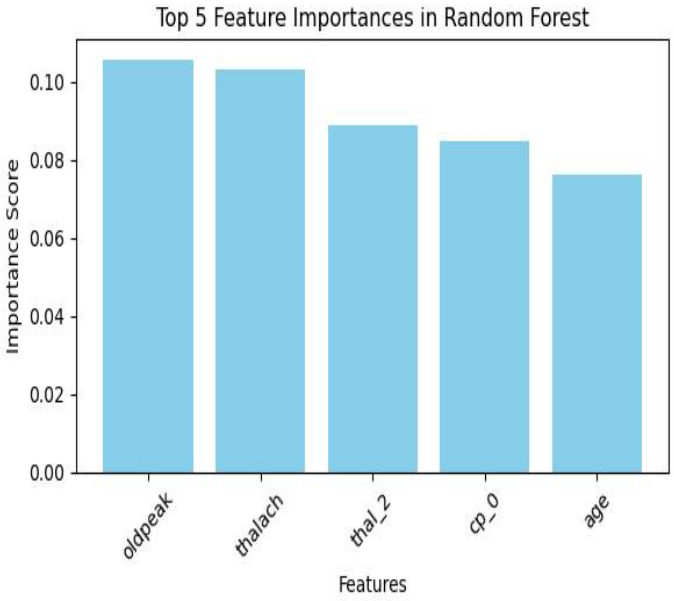


Figure 11 Key feature for RF

Model Evaluation - ROC , AUC and RMSE

"The model performance evaluation yielded strong AUC scores: Logistic Regression (0.93), Random Forest (0.98), SVM (0.947), and Gradient Boosting (0.947). In terms of RMSE, Random Forest achieved the lowest error (0.174), followed by Gradient Boosting (0.226), SVM (0.277), and Logistic Regression (0.308), indicating superior predictive accuracy for tree-based model

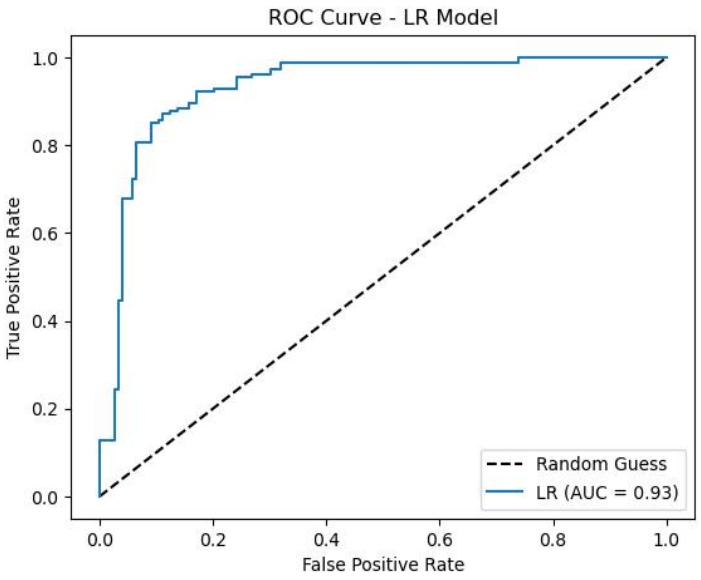


Figure 12 ROC curve for logestic regression

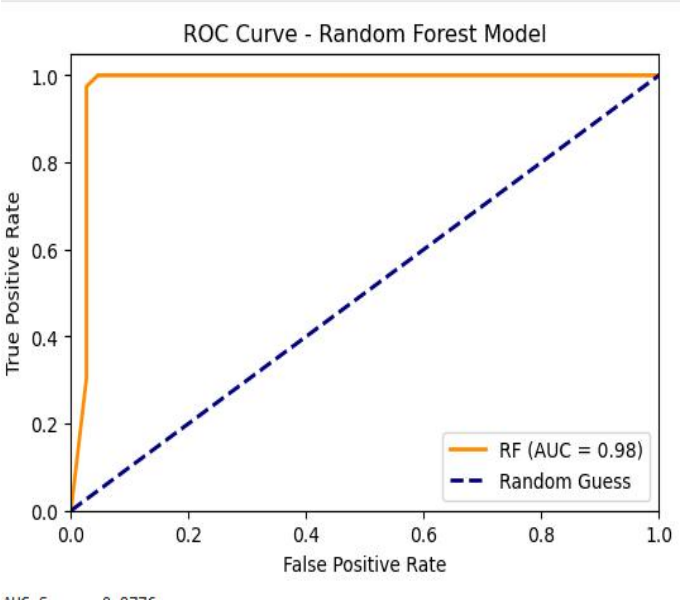


Figure 13 Figure 10 ROC curve for Random forest

4.1 Recommendations for Implementation

A. Deploy Random Forest as the primary clinical decision support tool, adjusting thresholds based on clinical risk tolerance.
B. Enhance models with ECG and temporal data to capture transient risk factors.
C. Conduct prospective validation studies across Ethiopian populations to confirm generalizability.
D. Develop explainable AI interfaces using SHAP or similar frameworks to improve clinician trust and adoption.

5. Conclusion

This evaluation establishes machine learning as a powerful approach for heart attack risk prediction, with Random Forest emerging as the optimal algorithm. Its 99.03% accuracy and exceptional discrimination power (ROC AUC 0.98) demonstrate significant improvement over traditional risk scores. Implementation of this technology could enable earlier interventions for high-risk patients while reducing unnecessary treatments for low-risk individuals, potentially transforming cardiovascular care pathways. Future work should focus on real-world validation and developing clinician-friendly interfaces for seamless integration into healthcare systems.

REFRENCE

[1] World Health Organization (WHO). (2023). *Cardiovascular diseases (CVDs)*.

[2] K. Deepika and S. Seema, "Predictive analytics to prevent and control chronic diseases," *Proc. 2016 2nd Int. Conf. Appl. Theor. Comput. Commun. Technol. iCATccT 2016*, pp. 381–386, 2017, doi: 10.1109/ICATCCT.2016.7912028.

[3] D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," *SN Comput. Sci.*, vol. 1, no. 6, Nov. 2020, doi: 10.1007/s42979-020-00365-y.

[4] V. Shorewala, "Early detection of coronary heart disease using ensemble techniques," *Informatics Med. Unlocked*, vol. 26, no. July, p. 100655, 2021, doi: 10.1016/j.imu.2021.100655.

[5] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective Heart Disease Prediction Using Machine Learning Techniques," *Algorithms*, vol. 16, no. 2, Feb. 2023, doi: 10.3390/a16020088.

[6] M. Marimuthu, M. Abinaya, K. S. Hariesh, and K. Madhankumar, "A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach," 2018.

[7] H. Ueshima, A. Sekikawa, K. Miura, and T. Chowdhury, "Cardiovascular Disease and Risk Factors in Asia," 2009.

[8] A. Mehmood *et al.*, "Prediction of Heart Disease Using Deep Convolutional Neural Networks," *Arab. J. Sci. Eng.*, vol. 46, no. 4, pp. 3409–3422, Apr. 2021, doi: 10.1007/s13369-020-05105-1.

[9] A. Adem, D. Bacha, and A. M. Argaw, "Pattern of cardiovascular diseases at a teaching hospital in Addis Ababa, Ethiopia: An echocardiographic study of 1500 patients," *Med. (United States)*, vol. 102, no. 34, p. E34795, 2023, doi: 10.1097/MD.0000000000034795.

[10] D. Lall, N. Engel, N. Devadasan, K. Horstman, and B. Criel, “Models of care for chronic conditions in low/middle-income countries: A ‘best fit’ framework synthesis,”*BMJ Glob. Heal.*, vol. 3, no. 6, pp. 1–12, 2018, doi: 10.1136/bmjgh-2018-001077.

A. Ashenafi *et al.*, “Diagnostics for detection and surveillance of priority epidemic- prone diseases in Africa: an assessment of testing capacity and laboratory strengthening needs,” *medRxiv*, no. September, p. 2024.05.17.24307542, 2024, doi: 10.3389/fpubh.2024.1438334.

[11] D. A. Beyene, H. B. Abayneh, M. A. Cheru, and T. M. Chamiso, “Magnitude and associated factors of atrial fibrillation, and its complications among adult rheumatic heart diseases patients in governmental hospitals in Bahir Dar Town, Northwest Ethiopia 2024,” *BMC Cardiovasc. Disord.*, vol. 25, no. 1, 2025, doi: 10.1186/s