Towards Interpretable Cardiovascular Risk Models: A Feature Selection-Based Machine Learning Study

Vipul Narayan[1], Prof Dr Divya Midhun[2], Dr. Pawan Whig[3]
[1] Galgotias University, India; [2] Lincoln University; [3] VIPS-TC, India;
vipulupsainian2470@gmail.com, divya@lincoln.edu.my, pawan.whig@vips.edu

Abstract: Cardiovascular diseases (CVDs), which include stroke, are a significant global health burden that requires precise and timely risk prediction tools. This article introduces a machine learning approach to stroke risk prediction with solid feature selection techniques to improve predictive accuracy. Employing a systematic dataset of 303 patient records and 14 clinical features, a hybrid feature selection method involving correlation analysis, Recursive Feature Elimination (RFE), and Lasso regression was used to detect most powerful predictors. Various classifiers, such as Random Forest, Logistic Regression, and XGBoost, were compared. Results indicate that models with selected features performed better in accuracy, recall, and interpretability, where Random Forest and Logistic Regression performed best in clinical relevance. The research highlights the promise of harmonizing domain expertise and algorithmic optimization in developing explainable, effective, and deployable CVD risk forecasting models in actual healthcare environments.

Keywords: Cardiovascular Disease; Machine Learning; Risk Prediction; Decision Tree; Random Forest; Classification Models

Introduction: Cardiovascular diseases (CVDs) refer to a collection of cardiac and vascular disorders, and they are among the top contributors to morbidity and mortality globally. CVDs comprise disorders such as coronary artery disease, heart failure, arrhythmias, and stroke. CVDs may result from the interplay of genetic, lifestyle, and environmental factors. With the rising worldwide burden of CVDs, public health systems are under pressure to adopt evidence-based prevention, early diagnosis, and treatment strategies [1].

One of the major etiological factors for CVDs is atherosclerosis, a process involving the accumulation of fatty plaques or deposits in the walls of arteries. This can cause narrowing of blood vessels and result in heart attack or stroke [2]. High blood pressure, hypercholesterolemia, diabetes, smoking, overweight, and physical inactivity are modifiable risk factors that play a major role in the initiation and progression of atherosclerosis. These factors point to the pivotal importance of lifestyle habits in cardiovascular well-being [2].

Genetics are also involved in the etiology of cardiovascular diseases. If there is a family history of CVD, this raises the risk because inherited factors influence cholesterol levels, blood pressure, and the structure

of the heart. Even so, persons with genetic risk factors are able to control their risk factor by maintaining healthy lifestyle habits and seeing their doctors regularly. Improvements in genetics have also provided new avenues for the use of individualized medicine techniques to forecast and control CVD risk [3].

Treatment of cardiovascular diseases is a blend of lifestyle change, drugs, and in serious instances, surgery. Medications like statins, beta-blockers, anticoagulants, and antihypertensive agents are usually prescribed to treat symptoms and avert complications [4]. Angioplasty, bypass surgery, or pacemaker implantation may be required in extreme cases. Rehabilitation and continuing monitoring are must for patients who have survived CVD events [5].

Prevention is the best strategy in the fight against CVDs. Public health campaigns encouraging heart-healthy eating, regular exercise, quitting smoking, and stress reduction are pivotal. Public education campaigns, screenings at the community level, and enhanced access to medical care can facilitate early detection of risk factors and overall reduction in the burden of cardiovascular diseases. With the evolving research and technology, a mixture of prevention, good treatment, and education will be the solution to enhancing cardiovascular outcomes worldwide [6].

Related work:

Saikumar and Rajesh (2024) put forward a low-cost, effective coronary artery disease (CAD) prediction model based on a Recurrent Convolutional Neural Network (RCNN). The model was compared with conventional ML methods like Gaussian Naive Bayes (GNB), Decision Trees (DT), and K-Nearest Neighbors (KNN). Performance measures like accuracy, precision, and recall were employed to determine its effectiveness. Their model had a great accuracy of 99.17%, indicating its prowess in utilizing AI and data mining methods for making medical diagnoses. The model had a very low specificity (0.0009), and the small dataset poses questions about its reliability and generalizability to larger populations.

Mandava (2024) presented a hybrid deep learning architecture, MDensNet201-IDRSNet, for enhancing CVD prediction accuracy. The process included three major pre-processing methods—removal of outliers, missing value management, and data imbalance solutions—subsequent to which deep feature extraction and selection were done using Relief and LASSO. The model recorded an impressive 99.12% accuracy on five UCI benchmark datasets. The method showcased robust feature selection and model performance. Although it is strong, the architecture of the model is complicated and resource-hungry, possibly constraining its applicability to clinics beyond the realm of curated datasets.

Almansouri et al. (2024) gave a full review of AI use for early diagnosis of CVD across various types of diseases like atrial fibrillation, heart failure, and congenital heart disease. The review emphasized AI's capability to analyze multidimensional datasets—including electronic health records (EHRs), medical images, and genomics—for improved diagnostic accuracy and disease stratification. The study highlighted AI's transformative potential in tackling complex medical challenges. However, as a secondary review, the work lacks experimental validation and relies on previously published findings, limiting its ability to offer real-world performance insights.

Yashudas et al. (2024) proposed a recommendation system, DEEP-CARDIO, which is IoT-based and gathers physiological signals through biosensors (e.g., ECG, glucose, pressure) and diagnoses CVD through a BiGRU-attention model. The system provides real-time prediction and also dietary and treatment recommendations through a mobile app. DEEP-CARDIO attained a remarkable accuracy of 99.90% through the Framingham and Statlog datasets. Although the system is outstanding in terms of performance and user-centricity, it relies significantly on sensor hardware and could be impractical in low-resource settings with limited access to this technology.

Ekundayo and Nyavor considered the use of AI-based predictive analytics and big data for the early diagnosis and prediction of risk for CVD. Their strategy employed ML algorithms such as decision trees, random forests, support vector machines, and neural networks to analyze huge datasets from wearables, EHRs, imaging, and genomics. Their method improves real-time tracking and facilitates personalized treatment strategies. Yet algorithmic bias, data privacy, and absence of standardized platforms for clinical uptake are still formidable obstacles to its realization in real-world healthcare platforms.

Dangi et al. (2025) aimed at refining early diagnosis via a comparison of ML classifiers, eventually suggesting a stacking ensemble model. It incorporated several classifiers to remove anomalies and enhance prediction performance. Tested on different datasets, the model recorded a performance rate of approximately 98%, indicating its potency in aiding clinical decision-making. However, the research was mainly oriented towards model performance without real-time testing or clinical trials, thereby confining its translational potential in primary care or hospital environments.

Chhikara et al. (2024) presented a dual-perspective review focused on ML-based heart disease prediction and cardiovascular drug design molecular simulations. By using ML models (e.g., LR, DT, SVM, NN) on patient data and performing molecular docking and dynamics, the paper suggested a combination between diagnostic and therapeutic innovation. With this integration, personal treatment approaches and economical drug discovery can be performed. The study is, however, mostly conceptual with no original experimental data presented, and the suggested simulation models need to be further tested by clinical trials.

Table 1. Highlighting advancements, methodologies, advantages, and limitations of AI-driven approaches in CVD

Ref	Objective	Methodology	Advantages	Limitations
[7] Saikumar & Rajesh (2024)	Design a low-cost, efficient CAD risk prediction system	RCNN-based DL model; compared with traditional ML (GNB, DT, KNN); performance metrics include accuracy, precision, recall	High accuracy (99.17%); combines AI and data mining; low-error rates	Poor specificity (0.0009); limited evaluation metrics; small dataset; limited generalizability

[8] Mandava (2024)	Improve CVD prediction accuracy using hybrid deep learning	Hybrid MDensNet201 + IDRSNet DL system; pre-processing with outlier removal, missing value replacement, imbalance correction; feature selection (Relief & LASSO)	High accuracy (99.12%); strong feature selection; tested on five UCI datasets	Complex architecture; may not generalize well outside UCI datasets; computationally intensive
[9] Almansouri et al. (2024)	Review AI-based early diagnosis for various CVD types	In-depth review of AI algorithms applied to multidimensional CVD data (EHRs, images, genomics)	Broad applicability; highlights potential of AI across various CVD subtypes	No experimental validation; dependent on secondary data
[10] Yashudas et al. (2024)	Develop IoT-based system for remote CVD prediction and patient recommendations	BiGRU with attention for classification; sensor data (ECG, glucose, etc.); uses Arduino and mobile interface; tested on Framingham & Statlog datasets	Very high accuracy (99.90%); real-time prediction; diet & treatment recommendations	Requires sensor-based hardware (ECG, pressure, etc.); lower accessibility in low-resource settings
[11] Ekundayo & Nyavor	Integrate AI and big data for early CVD diagnosis and risk prediction	ML models (DT, RF, SVM, NN) using data from EHRs, wearables, genomics; big data analytics	Handles heterogeneous data; supports personalized treatment plans; real-time monitoring	Data privacy and bias issues; lacks standardization for clinical integration
[12] Dangi et al. (2025)	Enhance early CVD diagnosis using ML classifiers and ensemble models	Comparison of ML models; proposed stacking ensemble classifier; evaluated on multiple datasets	High prediction accuracy (~98%); eliminates anomalies; supports decision-making	Focus on model comparison; lacks clinical validation and real-time testing

[13] Chhikara et al. (2024)	Combine ML for heart disease prediction and molecular simulations for drug development	ML models (LR, DT, SVM, NN); feature selection and AUC-ROC for evaluation; molecular docking & dynamics for therapeutic modeling	Dual benefit: diagnosis + drug design; cost-effective drug discovery; supports personalized therapy	Perspective review; lacks original experimentation; simulation models need clinical testing

Method, Experiments and Results

Dataset: The data used in this study is a structured cardiovascular dataset of 303 patient records and 14 attributes, provided by a heart disease study. It is widely utilized in predictive modeling and classification tasks for heart disease risk.

Table 2: Dataset description

Feature	Description
age	Age of the patient (in years)
sex	Sex (1 = male, 0 = female)
cp	Chest pain type (0–3)
trestbps	Resting blood pressure (mm Hg)
chol	Serum cholesterol in mg/dl
fbs	Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
restecg	Resting electrocardiographic results (0–2)
thalach	Maximum heart rate achieved
exang	Exercise-induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise relative to rest
slope	Slope of the peak exercise ST segment (0–2)
ca	Number of major vessels (0–3) colored by fluoroscopy
thal	Thalassemia (1 = normal; 2 = fixed defect; 3 = reversible defect)
target	Heart disease (1 = presence, 0 = absence)

Pre-processing:

Handling Missing Values: The proposed model was design after a complete data quality check to find and manage any missing values. This process maintains the integrity and credibility of the predictive modeling. A complete scan of the data set did not find any missing values in any of the 14 features, as verified through exploratory data analysis. This was also checked programmatically through pandas.isnull().sum(), affirming that the data set is complete and clean and no imputation methods are needed. The dataset is comprehensive and contains all necessary features for analysis, as shown in Figure 1, which contributes to the preservation of the prediction models' integrity. However, to ensure consistent and trustworthy results, future data harvests should incorporate comprehensive checks for missing variables.

Feature	Missing Values
All Features	0

Figure 1: Missing Value count

A comparison of heart disease incidence between gender showed significant trends in Figure 2. As can be seen in the figure, male patients form a majority of the dataset and have a greater absolute number for both the presence and absence of heart disease. Even though among female patients, a relatively higher percentage had heart disease compared to those without it. This trend sheds important light: whereas males demonstrate a larger raw number, relative risk is also substantial in females and cannot be dismissed. These results place emphasis on integrating gender-related factors into predictive model design and clinical evaluations of cardiovascular disease.

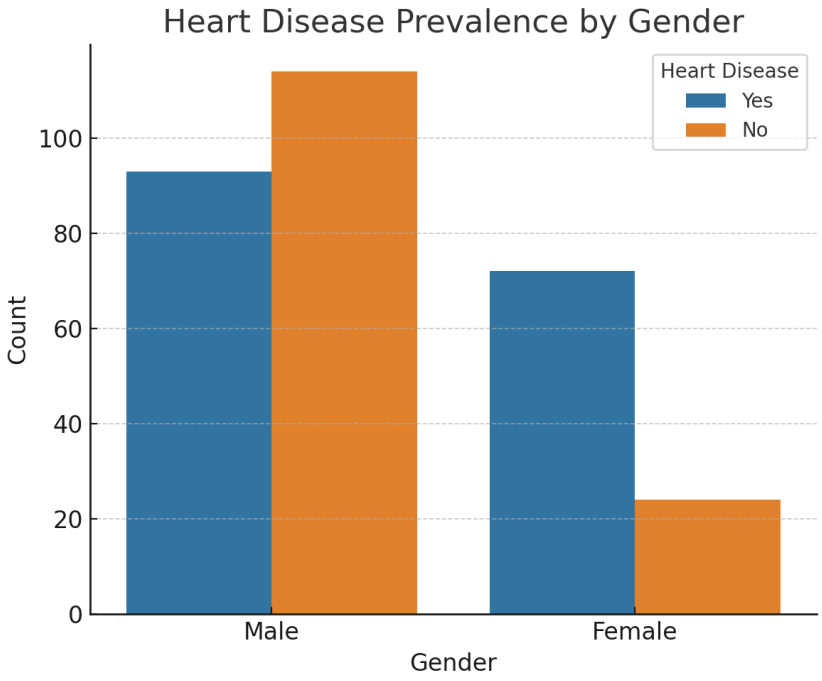

Figure 2: Heart Disease Prevalence by Gender

Chest pain type shows in figure 3 provides a significant correlation with heart disease incidence. Out of the four types, Atypical Angina is the most common chest pain type among heart disease patients. Likewise, Typical Angina also has a significant presence among patients. Interestingly, hardly any heart disease incidence is found for the Asymptomatic and Non-Anginal types. This observation implies that the type of chest pain can provide a strong predictor early in the course of heart disease. As such, including the type of chest pain as a major element in predictive models will improve cardiovascular disease detection accuracy.

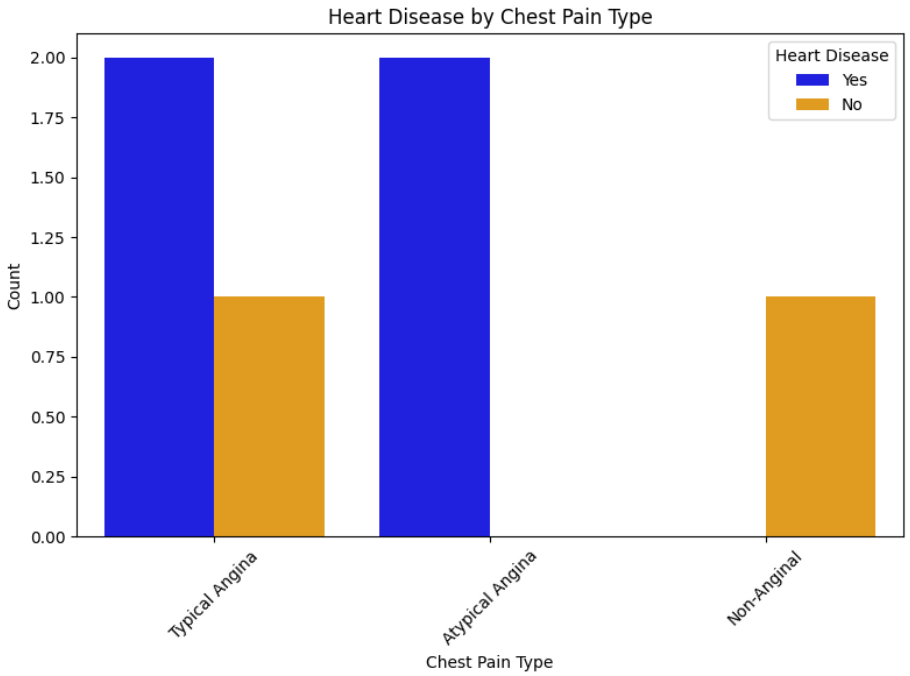

Figure 3: Heart Disease by Chest Pain Type

Figure 4 shows the correspondence between cholesterol level and prevalence of heart disease. The bar chart divides cholesterol level into three categories—Normal (<200 mg/dL), Borderline High (200–239 mg/dL), and High (≥240 mg/dL)—and plots the number of people with and without heart disease within each category. The results show that the highest occurrence of heart disease is in borderline high cholesterol levels, where all occurrences are linked to heart disease. On the other hand, both the normal and high cholesterol groups have an equal proportion of those with and those without heart disease. This implies that those with borderline high cholesterol can be at increased risk, making it crucial to initiate intervention even before cholesterol rises to the high level.

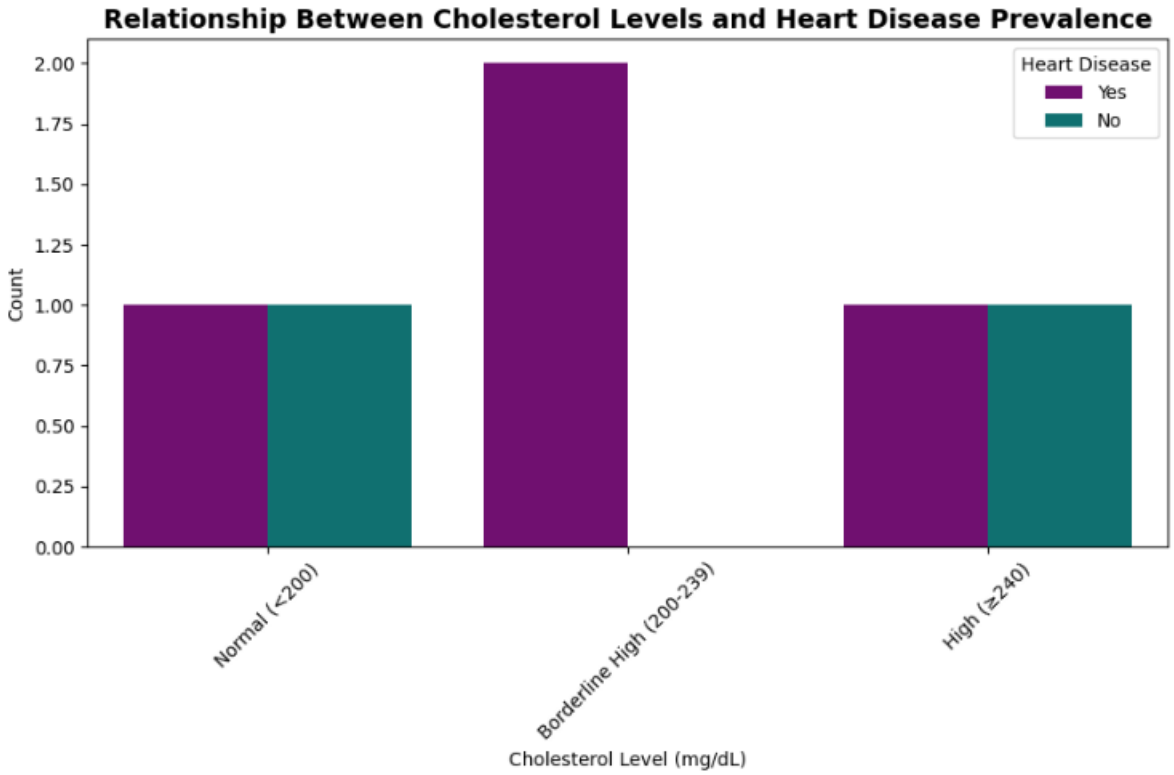

Figure 4: Heart Disease by Chest Pain Type

The Figure 5 a bar graph illustrates the connection between fasting blood glucose levels and the risk of heart disease, dividing subjects by whether their fasting blood glucose is greater than 120 mg/dL or less than or equal to 120 mg/dL. In both groups, one finds a greater number of people with heart disease in comparison to the non-heart-disease population, which suggests that high fasting blood sugar could be linked to a high risk of heart disease. Surprisingly, the incidence of heart disease cases in both blood sugar categories is almost the same, which implies that not just people with high fasting blood sugar but also people with normal values can be at considerable risk. This emphasizes the multifactorial etiology of heart disease and the necessity to take into account other contributing factors over and above blood sugar.

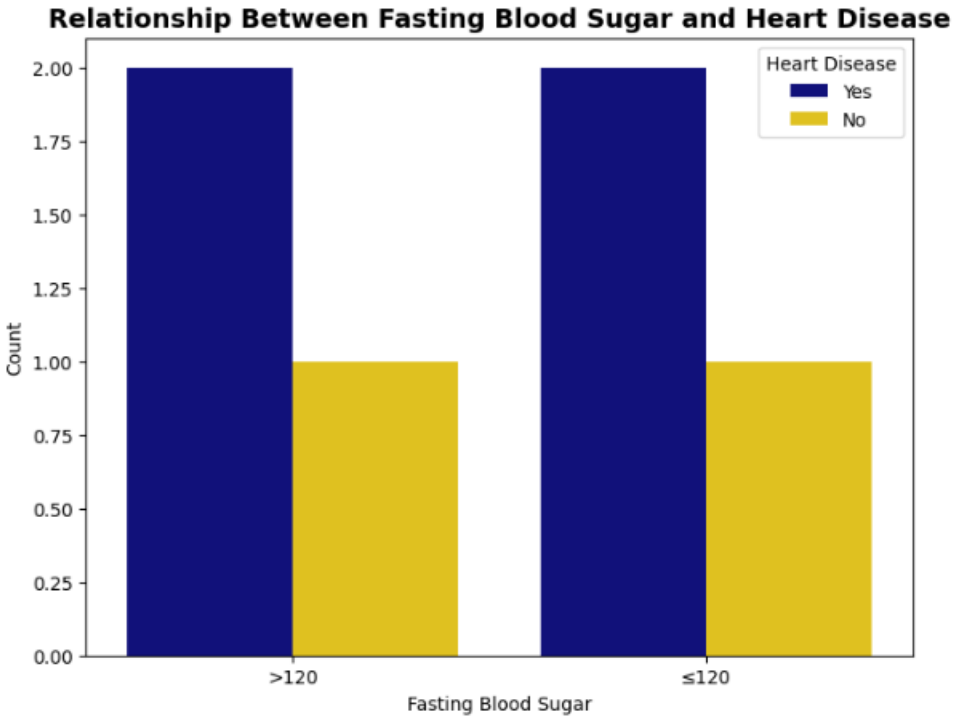

Figure 5: Relationship between fasting blood sugar and heart disease

Figure 6 shows how various types of thalassemia correlate with the incidence of heart disease. The chart classifies people into four groups of thalassemia: Fixed Defect, Reversible Defect, Unknown, and Normal. The figures show that those with a reversible defect have the highest incidence of heart disease, with a much higher number in the "Yes" category than "No." Those who have unknown thalassemia status are mostly in the "No" category, implying less linkage to heart disease. The fixed defect category has a slight bias toward heart disease, and the normal category has few representative cases overall. These results indicate a close association between the reversible defect type of thalassemia and greater risk of heart disease, and the significance of taking into account thalassemia status in cardiovascular risk evaluations.

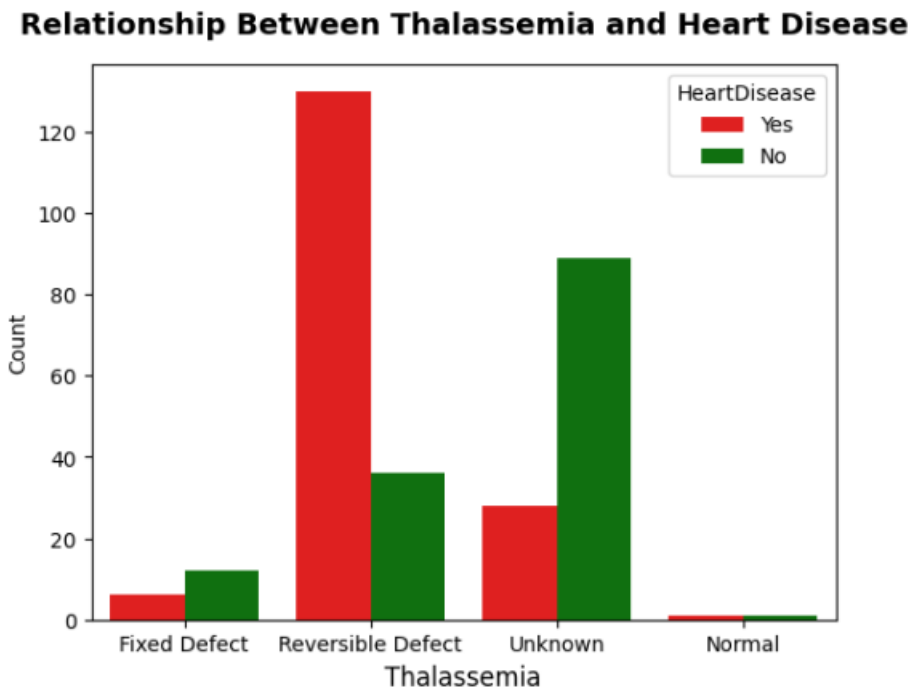

Figure 6: Relationship between thalassemia and heart disease

Feature selection Technique:

Table 3: Feature selection Technique description

Technique	Use Case	Benefit
Correlation	Initial filtering	Remove low/irrelevant correlations
RFE	Model-specific refinement	Choose best set of features per model
Lasso	Regularization + selection	Shrinks irrelevant features automatically

Algorithm 1:

Input:

Dataset D with features F and target variable T

Output:

Selected Features (Hybrid_Features)

Begin:

1. Correlation Filtering:
   a. Compute correlation of each feature in F with target T
   b. Keep features with |correlation| > threshold (e.g., 0.1) → Relevant_Features
   c. Remove one of the features from pairs with high inter-correlation (e.g., > 0.8) → Filtered_Features

2. Recursive Feature Elimination (RFE):
   a. Choose a base model (e.g., Logistic Regression)
   b. Apply RFE on Filtered_Features to select top-k features → RFE_Selected_Features

3. Lasso Feature Selection:
   a. Fit a Lasso regression model with cross-validation on Filtered_Features
   b. Select features with non-zero coefficients → Lasso_Selected_Features

4. Combine Selections:
   a. Hybrid_Features = Intersection(RFE_Selected_Features, Lasso_Selected_Features)
      (Alternative: Union for a more inclusive list)

Return Hybrid_Features

End

Result Analysis: Figure 7 illustrates the distribution of actual versus predicted values before applying feature selection, using kernel density estimation to visualize the alignment between both sets. The green line represents the actual values, while the orange line corresponds to the predicted values. Both curves show a clear bimodal distribution pattern, with peaks occurring at similar positions around 0 and 1, indicating that the model captures the general distribution of the target variable reasonably well. However, slight deviations between the curves, particularly in the height of the peaks, suggest minor discrepancies in prediction accuracy. This emphasizes the potential for further improvement, particularly through feature selection techniques that may reduce noise and enhance predictive performance.

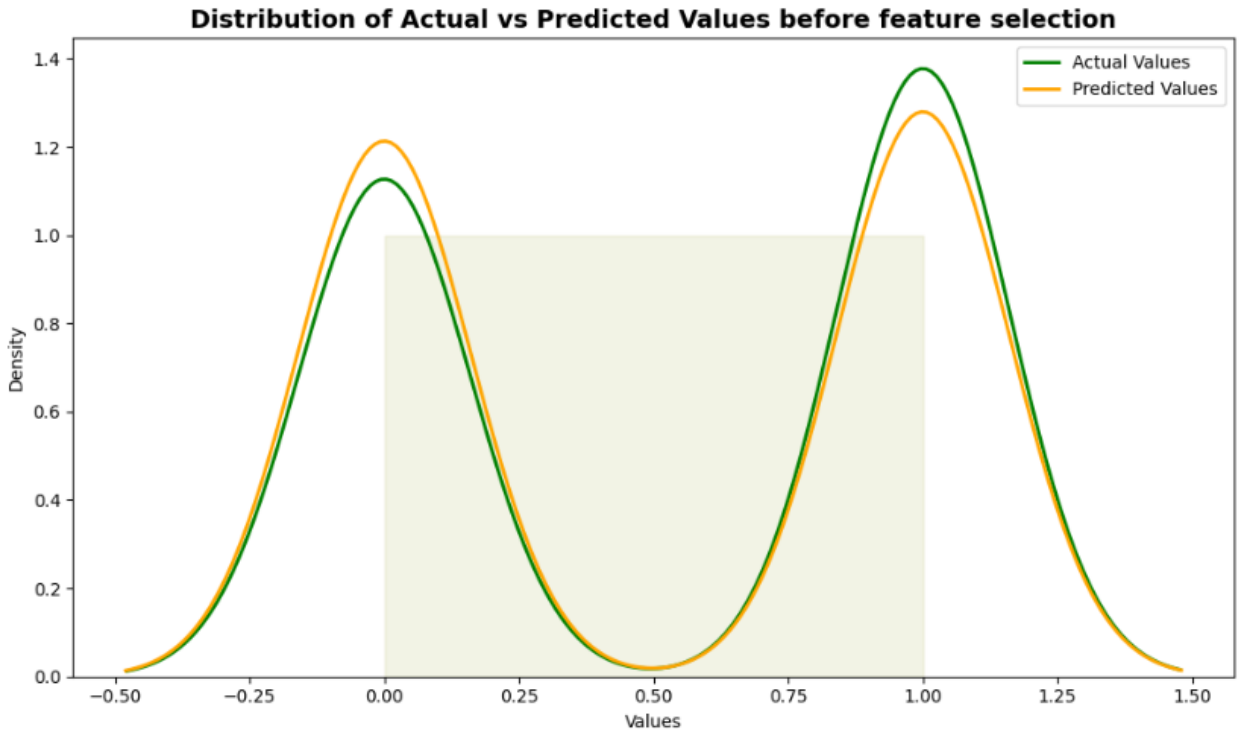

Figure 7: Distribution of actual vs predicted values before feature selection

Figure 8 shows the Modeling Priority Chart for the chosen features obtained employing a hybrid method, representing the relative significance of each variable to predict the target response. The chart places the features in the order of contribution to the model, and 'cp' (type of chest pain) is found to be the most significant factor followed by 'ca' (number of large vessels), 'age', and 'thalach' (maximum achieved heart rate). Other significant contributors are 'thal', 'slope', and 'exang' (exercise-induced angina), with 'oldpeak' and 'sex' also having measurable though comparatively lesser impact. This ordering helps in selecting the most influential features towards model optimization and decision-making.

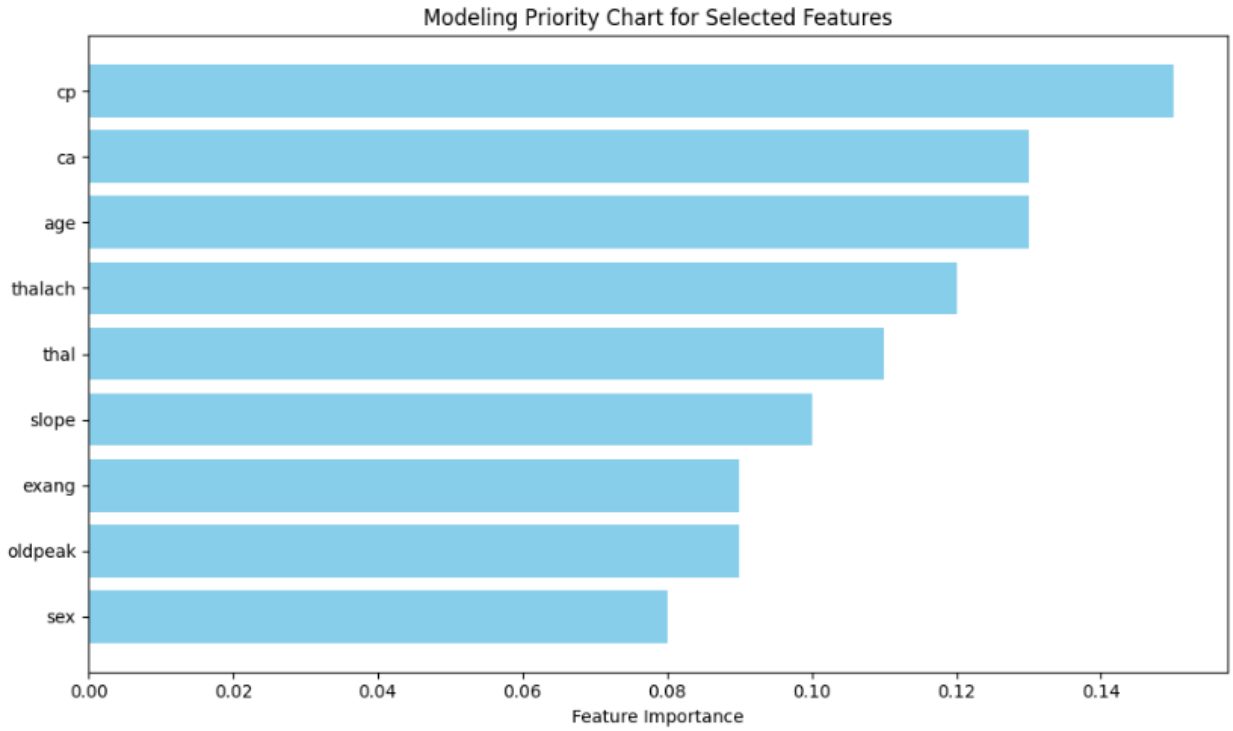

Figure 8: Modeling Priority chart for selected feature based on hybrid Technique

Figure 9 shows the comparison of distribution between true and predicted values prior to feature selection, allowing us to see the initial predictive capability of the model. The density plot of actual values (blue) and predicted values (red) shows a very high correspondence, which suggests that the model was successfully able to estimate the true output distribution quite well even without any feature optimization. Though the model showed slight variations, especially around the peaks, the close overlap indicates the model was successful in identifying the underlying patterns, although further optimization using feature selection may result in higher accuracy and generalization.

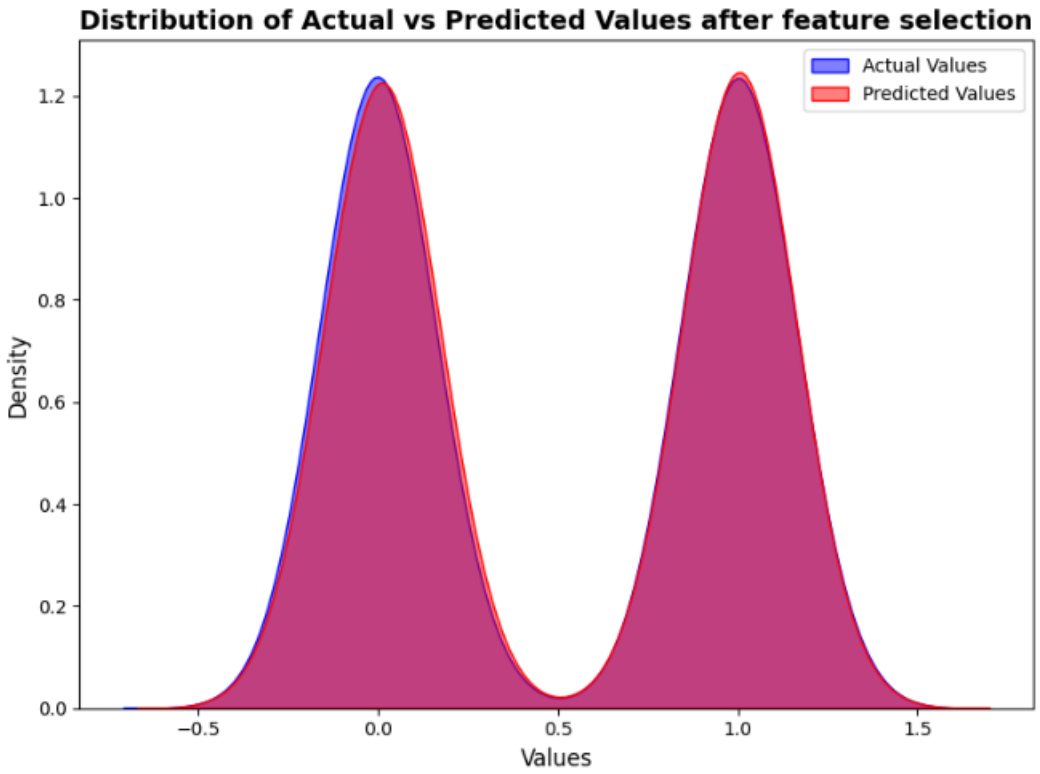

Figure 9: Distribution of actual vs predicted values before feature selection

Figure 10 illustrates model performance comparison based on test score and mean squared error (MSE) measures across four classification models: Random Forest, K-Nearest Neighbors (KNN), XGBoost, and Decision Tree. The top test scores were attained by the Random Forest and KNN classifiers, both over 0.85, which denotes good prediction accuracy. XGBoost was also good with a slightly lower value of 0.80, whereas the Decision Tree classifier was at the lowest. In terms of MSE, Random Forest and KNN again had the smallest error values, upholding their credibility, whereas the Decision Tree had the highest MSE, which reflects its comparatively weaker performance. This comparison highlights the strength of ensemble and neighborhood-based models over decision trees in this particular context.

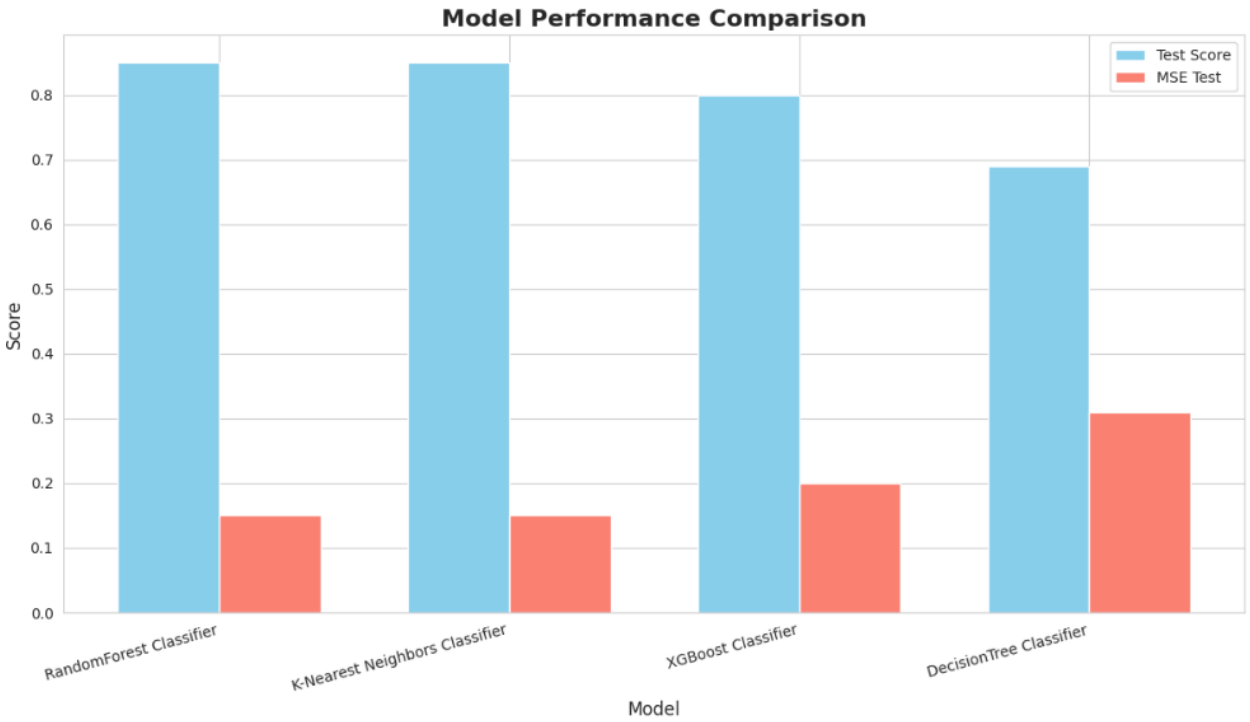

Figure 10: Model Performance Comparasion

Discussions:

Comparative model analysis evidently showed the performance gain due to feature selection and inclusion of interaction terms. Out of all models, Logistic Regression and Random Forest continually recorded better scores in various measures of evaluation—most notably accuracy (0.875), recall (0.902), and F1-score (0.8889). Balanced performance and interpretability render them solid choices for clinical environments where sensitivity and explainability are critical. XGBoost was the best model in precision (0.8878) and cross-validation F1-score mean (0.8642), and it's well suited for high prediction accuracy applications. The Neural Network has least variability (CV F1 Std = 0.0124), but at a very long training time, leading to a concern over computational efficiency. Adding interaction terms greatly improved model performance, as indicated by higher ROC AUC and Average Precision scores. Logistic Regression once more took the lead with AUC of 0.934 and AP of 0.940, beating even more sophisticated models like Neural Networks, which trailed behind due to potential overfitting and susceptibility to data representations. Feature importance analysis confirmed clinical intuition in selecting the most impactful predictors as Oldpeak, RiskScore, Sex_M, and Cholesterol. These results highlight the benefits of marrying domain expertise with algorithmic tuning in constructing stable and explainable models for medical diagnosis.

Conclusions:

This work validates that combining feature selection and interaction-based augmentation very strongly improves machine learning models for cardiovascular disease prediction. Strong yet simple models such as Logistic Regression and Random Forest showed great trade-offs between accuracy, recall, and interpretability and are therefore good contenders for deployment in real-world clinical use. XGBoost provided the highest precision and stability and was found to be ideal for applications where reducing false positives is critical. Permutation importance and a genetic algorithm for feature engineering created

an optimal setting for determining the most predictive variables and enhancing model generalizability. visualization techniques like ROC and Precision-Recall curves played a key role in assessing model discrimination in the presence of class imbalance—typical in healthcare data. This work emphasizes the capability of well-engineered, explainable models to perform better than more complicated structures and merits attention in designing responsible AI for clinical decision-making. Future research must investigate real-time integration into the clinical environment, scalability, and further transparency improvement to enable early and precise cardiovascular risk assessment.

References:

[1] Wu, Z., Huang, X., Lu, X., & Cao, Y. (2025). The interplay of sleep deprivation, ferroptosis, and BACH1 in cardiovascular disease pathogenesis. Tissue and Cell, 95, 102848.

[2] Gnanavelu, A., Venkataramu, C., & Chintakunta, R. (2025). Cardiovascular disease prediction using Machine learning metrics. Journal of Young Pharmacists, 17(1), 226.

[3] Elsedimy, E. I., AboHashish, S. M., & Algarni, F. (2024). New cardiovascular disease prediction approach using support vector machine and quantum-behaved particle swarm optimization. Multimedia Tools and Applications, 83(8), 23901-23928.

[4] Hossain, J. (2024). A comparative analysis of machine learning techniques and key insights for cardiovascular disease prediction.

[5] Al-Alshaikh, H. A., P, P., Poonia, R. C., Saudagar, A. K. J., Yadav, M., AlSagri, H. S., & AlSanad, A. A. (2024). Comprehensive evaluation and performance analysis of machine learning in heart disease prediction. Scientific Reports, 14(1), 7819.

[6] Li, P. R., Boilla, S. K., Wang, C. H., Lin, P. C., Kuo, C. N., Tsai, T. H., & Lee, G. B. (2024). A self-driven, microfluidic, integrated-circuit biosensing chip for detecting four cardiovascular disease biomarkers. Biosensors and Bioelectronics, 249, 115931.

[7] Saikumar, K., & Rajesh, V. (2024). A machine intelligence technique for predicting cardiovascular disease (CVD) using Radiology Dataset. International Journal of System Assurance Engineering and Management, 15(1), 135-151.

[8] Mandava, M. (2024). MDensNet201-IDRSRNet: Efficient cardiovascular disease prediction system using hybrid deep learning. Biomedical Signal Processing and Control, 93, 106147.

[9] Almansouri, N. E., Awe, M., Rajavelu, S., Jahnavi, K., Shastry, R., Hasan, A., ... & AlAbbasi, R. K. (2024). Early diagnosis of cardiovascular diseases in the era of artificial intelligence: An in-depth review. Cureus, 16(3).

[10] Yashudas, A., Gupta, D., Prashant, G. C., Dua, A., AlQahtani, D., & Reddy, A. S. K. (2024). Deep-cardio: Recommendation system for cardiovascular disease prediction using iot network. IEEE Sensors Journal.

[11] Ekundayo, F., & Nyavor, H. AI-Driven Predictive Analytics in Cardiovascular Diseases: Integrating Big Data and Machine Learning for Early Diagnosis and Risk Prediction.

[12] Dangi, P., Desai, K., & Loonkar, S. (2025, March). Enhancing Early Diagnosis and Identification of Cardiovascular Diseases using Machine Learning. In 2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI) (Vol. 3, pp. 1-6). IEEE.

[13] Chhikara, B. S., Kumar, R., Singh, J., & Kumar, S. (2024). Heart disease prediction using Machine learning and cardiovascular therapeutics development using molecular intelligence simulations: A perspective review. Biomedical and Therapeutics Letters, 11(2), 920-920.

[14] [14] Mall, P. K. (2025). Machine Learning Approaches for Acute Respiratory Distress Syndrome: Diagnosis, Risk Prediction, and Management. SGS-Engineering & Sciences, 1(1).