Enhancing Cardiovascular Stroke Prediction with Feature Selection and Machine Learning Models

Arun Pratap Srivastava[1], Dr Vivekanandam[2], Dr. Mudassir Khan[3]
Lloyd Institute of Engineering & Technology; Lincoln University, Malaysia; King Khalid University, Saudi Arabia
[1]apsvgi@gmail.com, [2]vivekanandam@lincoln.edu.my, [3]mkmiyob@kku.edu.sa

Abstract: Cardiovascular diseases (CVDs) continue to be the major cause of death worldwide, emphasizing the necessity for early, precise, and scalable diagnostic solutions. This research investigates the combination of machine learning (ML) models with feature selection and interaction-based improvements to enhance heart disease prediction. With a patient record dataset of 918 with 12 diagnostic features, we tested various classifiers such as Logistic Regression, Random Forest, Gradient Boosting, SVM, XGBoost, Neural Networks, and Naive Bayes. The performance was tested over accuracy, ROC AUC, and Average Precision (AP) metrics prior to and subsequent to feature selection and incorporation of interaction terms. Results indicate that Logistic Regression performs the best consistently with the highest AUC (0.934) and AP (0.940) after enhancement, followed by Random Forest and SVM. A genetic algorithm feature selection method dramatically enhanced model interpretability and generalization. Analysis of feature importance confirmed the predictive ability of features such as Oldpeak, RiskScore, and Cholesterol. This paper confirms the utility of interpretable ML methods in healthcare diagnostics and suggests a strong, data-driven model for early cardiovascular risk estimation.
Keywords: Cardiovascular diseases; early-stage indicators; healthcar; machine learning; risk prediction

Introduction

Cardiovascular diseases (CVDs) are the most common cause of death worldwide, accounting for millions of fatalities each year and having a considerable economic and social impact. Increased stress levels, sedentary life, and unhealthy diets have increased the rate of cardiac ailments in all age groups. Early detection of CVDs is still a major challenge despite technological advances in diagnostics and therapeutic interventions. This is especially because of the fine or unusual character of early-stage symptoms, which tend to pass unnoticed or become misconstrued in clinics.

In the context of India, one of the emerging issues is that of women's growing prevalence of heart disease, particularly amongst youth. Traditionally not thought to be as susceptible as men, women are now confronting an epidemic of cardiovascular risk factors like diabetes, hypertension, smoking, and obesity. Exacerbated by social norms and failure of patients and healthcare providers to recognize symptoms, most women receive delayed diagnosis and unfavorable outcomes. This accentuates the necessity for gender-sensitive and data-driven diagnostic approaches that are capable of correctly recording early warning signs of CVDs.

Conventional diagnostic approaches tend to be based on subjective interpretation and rule-based findings, both of which are susceptible to variability and error. These shortcomings can be avoided by employing

artificial intelligence (AI) and machine learning (ML) approaches that facilitate the examination of complex, high-dimensional clinical data. Machine learning algorithms have been particularly promising in medicine by enabling automated predictions, detecting subtle patterns, and improving diagnostic accuracy—provided they are augmented with proper feature engineering strategies.

The combination of feature selection methods like permutation importance and genetic algorithms enables one to find the most important predictors from high-dimensional datasets. Combined with interaction-based feature augmentation, the methods can greatly enhance the performance of ML models. This is particularly critical in healthcare where model interpretability, generalizability, and efficiency are as crucial as brute accuracy. In addition, integrating explainability into such models increases clinician trust and enables adoption of responsible AI in practice.

The purpose of this research is to create and test a machine learning-based framework for predicting heart disease based on a curated set of 918 patients. Several classifiers such as Logistic Regression, Random Forest, SVM, and XGBoost are compared with and without interaction terms and feature selection. The performance is evaluated on accuracy, precision, recall, F1-score, and ROC AUC. Our results show that less complex models such as Logistic Regression, with expert-guided feature selection, can surpass more advanced models while being transparent and clinically relevant.

Related work

The recent developments in cardiovascular disease (CVD) prediction models illustrate the promise of machine learning (ML) and deep learning (DL) methods to enhance diagnostic performance and patient outcomes. Ogunpola et al. (2024) aimed at improving the detection of heart disease, specifically myocardial infarction, by comparing seven ML/DL models—KNN, SVM, Logistic Regression, CNN, Gradient Boosting, XGBoost, and Random Forest—on imbalanced datasets. Their research was a tremendous success, with XGBoost delivering 98.5% accuracy. A major strength of this study is its successful management of data imbalance and focus on algorithmic tuning. The main limitation of the study, though, is the absence of external validation, limiting its generalizability across wide-ranging populations.

In an independent endeavor, Drouard et al. (2024) investigated the prediction of CVD risk factors based on multi-omic datasets that included genomics, proteomics, and transcriptomics. Their approach compared six ML classifiers and state-of-the-art methods such as unsupervised/semi-supervised autoencoders and transfer learning. They established that multi-omic models outperformed single-omic models significantly, and model generalization was enhanced using transfer learning. Though these findings are encouraging, the models' complexity and dependence on omic data—something not easily obtainable in routine clinical practice—are a concern for large-scale deployment.

DeGroat et al. (2024) focused on CVD biomarker discovery to predict precision CVD with a hybrid approach that integrated statistical analysis with an ensemble of machine learning (ML) algorithms such as RF, SVM, XGBoost, and KNN. They identified and ranked transcriptomic biomarkers with a respectable 96% accuracy. Targeted diagnostics are enabled with the integration of ML for biomarker prioritization. However, the dependence of the model on data quality and interpretability issues related to ensemble approaches remain challenges for clinical translation.

While most research centers on technical accuracy, Cai et al. (2024) approached differently by critically examining the methodological flaws in existing ML models for predicting CVD. Their review has covered data quality, overfitting of the model, bias, and reproducibility. They suggested a systematic framework of

best practices to improve the reliability and explainability of the model. Though their work is very valuable in informing the development of future studies, the lack of empirical testing restricts its direct real-world application.

For enhancing the early detection of CVD, an ensemble ML model was proposed by Korial et al. (2024), based on voting among Naïve Bayes, Random Forest, Logistic Regression, and KNN, coupled with chi-square feature selection. Predictive accuracy was enhanced to 92.11% and computational needs were reduced by half. Their dataset, however, consisted of just 303 samples, limiting the model's scalability and subjecting it to possibilities of overfitting and lack of generalizability.

Yet another innovative work is presented by Alghamdi et al. (2024), which designed a hybrid ML system combining an arithmetic optimization algorithm as a feature selector and an MLP as a classifier. The pipeline produced an accuracy of 88.89% and had a very robust preprocessing structure. Even though the system performed well with data, it suffered from imbalanced data and was compared only with conventional classifiers, so it was not fully evaluated from a complete set of recent models.

Moreno-Sánchez et al. (2024) provided a structured review of ML and DL techniques used to diagnose and predict CVD from ECG-based approaches. The research centered on data modalities, principles of trustworthy AI, and ethical issues like bias, explainability, and transparency. Their review offered an integrated perspective of the prevailing trends and challenges for ECG-based diagnostics. Nevertheless, its reliance on secondary data and absence of novel experimentation limit its impact on algorithm design and real-world deployment.

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

Reference	Objective	Methodology	Advantage	Limitations
[7] Ogunpola et al., 2024	Improve heart disease detection, especially myocardial infarction	Evaluated 7 ML/DL models (KNN, SVM, LR, CNN, GB, XGBoost, RF) and addressed dataset imbalance	XGBoost achieved 98.5% accuracy; strong performance on imbalanced data	Focused on model tuning, lacks external validation
[8] Drouard et al., 2024	Predict CVD risk factors using multi-omic data	Compared 6 ML classifiers with unsupervised/semi-supervised autoencoders and transfer learning	Multi-omics outperformed single-omics; transfer learning improved generalization	High complexity; relies on omic data not widely available
[9] DeGroat et al., 2024	Identify biomarkers for precision CVD prediction	Used statistical tests + ML ensemble (RF, SVM, XGBoost, KNN);	Achieved 96% accuracy; effective	Data-dependent; interpretability of ensemble

		ranked transcriptomic biomarkers	biomarker discovery	model can be limited
[10] Cai et al., 2024	Highlight pitfalls in ML models for CVD prediction	Reviewed issues in data quality, model design, overfitting, and reproducibility	Provided comprehensive framework of solutions and guidelines	No experimental validation; theoretical perspective
[11] Korial et al., 2024	Improve early CVD detection using ensemble ML with feature selection	Used voting ensemble (NB, RF, LR, KNN) with chi-square feature selection	Ensemble improved accuracy (92.11%) and reduced computation by 50%	Small dataset (303 records); limited generalizability
[12] Alghamdi et al., 2024	Propose accurate ML system for CVD diagnosis	Applied arithmetic optimization algorithm for feature selection + MLP for classification	Achieved 88.89% accuracy; robust preprocessing pipeline	Suffers from data imbalance; performance compared only with traditional models
[13] Moreno-Sánchez et al., 2024	Review ECG-based ML/DL solutions for CVD diagnosis/prognosis	Systematic review focusing on data modalities, DL techniques, and Trustworthy AI aspects	Provides ethical insights, model explainability, bias analysis	Lack of primary experiments; depends on secondary data

Method, Experiments and Results

Dataset: This dataset appears to be related to predicting the presence or absence of heart disease based on various clinical and demographic features shown in table 2.the data set is comprised of 918 patient records, each containing 12 heart disease diagnosis-relevant attributes. It contains both numerical (such as Age, RestingBP, Cholesterol) and categorical variables (such as Sex, ChestPainType, ST_Slope). The target column, HeartDisease, is binary, with 1 representing the presence of heart disease and 0 representing the absence of heart disease. The mean patient age is about 53.5 years with the resting blood pressure mean at 132 mm Hg. Cholesterol ranges quite widely and in some was found to be 0 and this could indicate unrecorded or missing information. The database also captures exercise-induced angina, maximum heart rate, as well as ST depression on exertion. Categorical features such as ChestPainType and ST_Slope yield clinically relevant information regarding symptoms and ECG traces. Interestingly, most patients in this dataset are male (Sex = M), and the most prevalent chest pain type is ASY (asymptomatic). In general, the

dataset is good for classification models that can predict heart disease with a limited class imbalance towards positive diagnoses (55.3%).

Table 2: Dataset description

Column	Type	Description	Example Values / Stats
Age	Numeric	Age of the patient	Mean: 53.5, Min: 28, Max: 77
Sex	Categorical	Biological sex (M/F)	Most common: M (725 out of 918)
ChestPainType	Categorical	Type of chest pain	ASY, NAP, ATA, TA (most common: ASY)
RestingBP	Numeric	Resting blood pressure (mm Hg)	Mean: 132.4, Min: 0, Max: 200
Cholesterol	Numeric	Serum cholesterol (mg/dl)	Mean: 198.8, Min: 0, Max: 603
FastingBS	Binary	Fasting blood sugar > 120 mg/dl (1 = yes)	23.3% have FastingBS = 1
RestingECG	Categorical	ECG results at rest	Normal, ST, LVH (most common: Normal)
MaxHR	Numeric	Maximum heart rate achieved	Mean: 136.8, Range: 60–202
ExerciseAngina	Categorical	Exercise-induced angina (Y/N)	Most common: N (547 out of 918)
Oldpeak	Numeric	ST depression induced by exercise	Mean: 0.89, Range: -2.6 to 6.2
ST_Slope	Categorical	Slope of the peak exercise ST segment	Up, Flat, Down (most common: Flat)
HeartDisease	Binary	Target: presence of heart disease	55.3% positive cases

Pre-processing:

Handling Missing Values: Although the dataset currently has no missing values, it's always good practice to check for and handle any missing data in real-world scenarios.

The graphical inspection of cholesterol levels offers several views on cholesterol variability in the population and its association with heart disease and resting blood pressure. The top-left histogram displays a close-to-normal distribution of cholesterol with a mean of around 244.66 and median of 237.00, reflecting a slight right skew of the data. This is confirmed by the KDE line superimposed on the histogram. The top-right boxplot provides a comparison between cholesterol levels among people with and without heart disease. It indicates that people who have heart disease have higher cholesterol, and using a t-test gives a

statistically significant p-value of 0.0047, which underscores that the variation between the two groups is very unlikely to have occurred by chance.

In the bottom-left violin plot, the same comparison between distributions is again shown in more detail. The story emphasizes the mean cholesterol levels of both groups: approximately 238.77 for those without heart disease and 251.06 for those with heart disease. This plot serves to support the upward trend of cholesterol levels of the heart disease group. Finally, the bottom-right scatter plot investigates the association between cholesterol and resting blood pressure (RestingBP). The correlation is low (0.096), indicating that there is a weak positive trend, but cholesterol and resting blood pressure are not linearly related strongly. The scatter plot is supplemented with regression lines divided by heart disease status, giving a detailed perspective on how the relationship can differ between groups. Generally, the evidence upholds that increased cholesterol is linked to heart disease, though it has limited correlation with resting blood pressure.

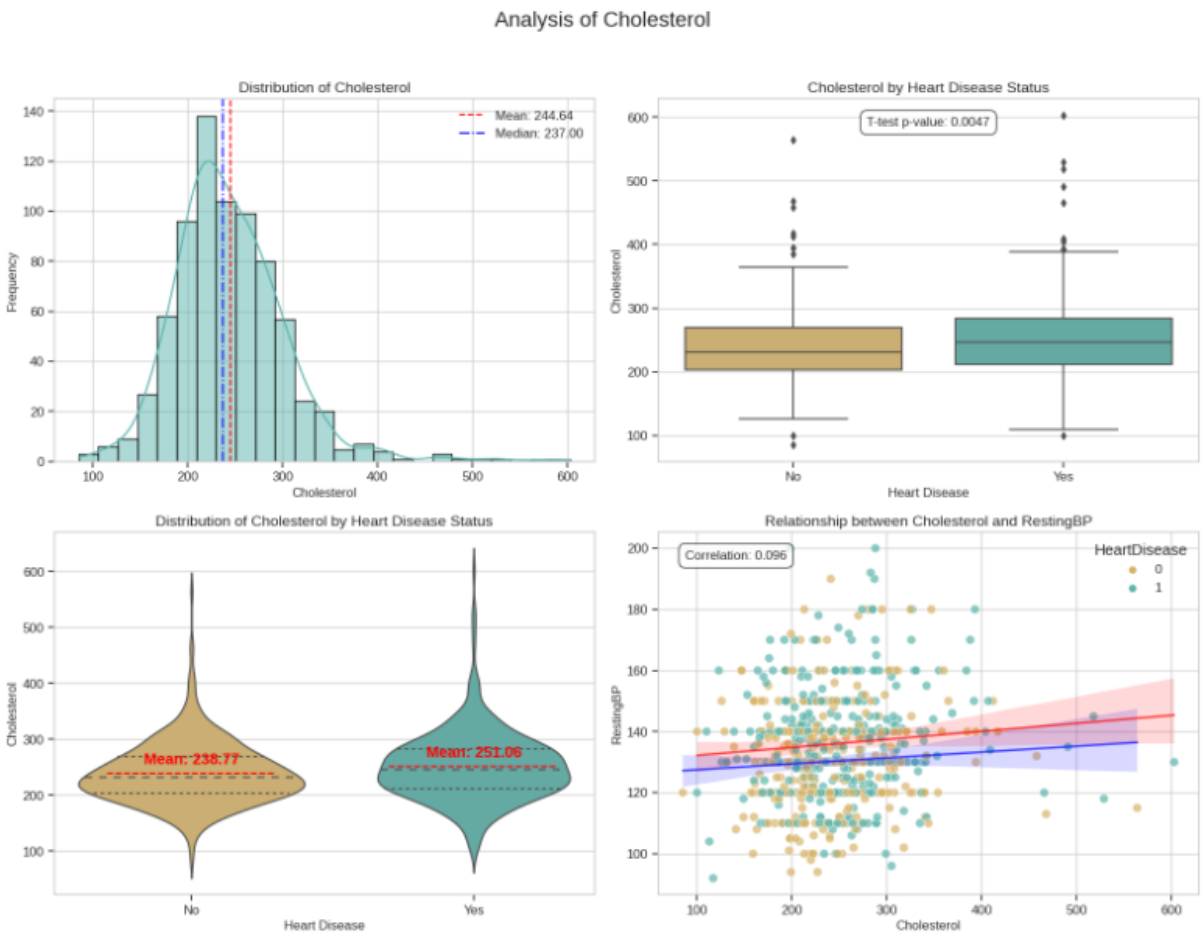

Figure 2: Cholesterol Analysis

The plot of Resting Blood Pressure (RestingBP) provides information on its distribution and correlation with heart disease and cholesterol. The top-left histogram shows the histogram of RestingBP values, having a mean of around 132.54 and a median of 130.00, indicating a relatively symmetric distribution with a mild right skew. The KDE overlay aids in the visualization of the overall pattern. The top-right boxplot offers a comparison of those with and without heart disease. It is easily seen that those with heart disease have the

higher resting blood pressure, backed up by a statistically significant t-test p-value of 0.0003, showing a large difference between both groups.

The bottom-left violin plot gives a closer split breakdown of RestingBP by heart disease status. The average resting blood pressure in a group that doesn't have heart disease is roughly 130.18, compared to about 134.45 in people suffering from heart disease, identifying a significant rise in blood pressure levels among the patient group. Last but not least, the bottom-right scatter plot investigates RestingBP's association with cholesterol, indicating an extremely weak positive relationship (0.089). Even though the regression lines by heart disease status indicate a slight upward trend, the relationship between these two variables is limited. In general, the analysis shows that higher resting blood pressure is strongly related to heart disease, whereas its relationship with cholesterol is weak.

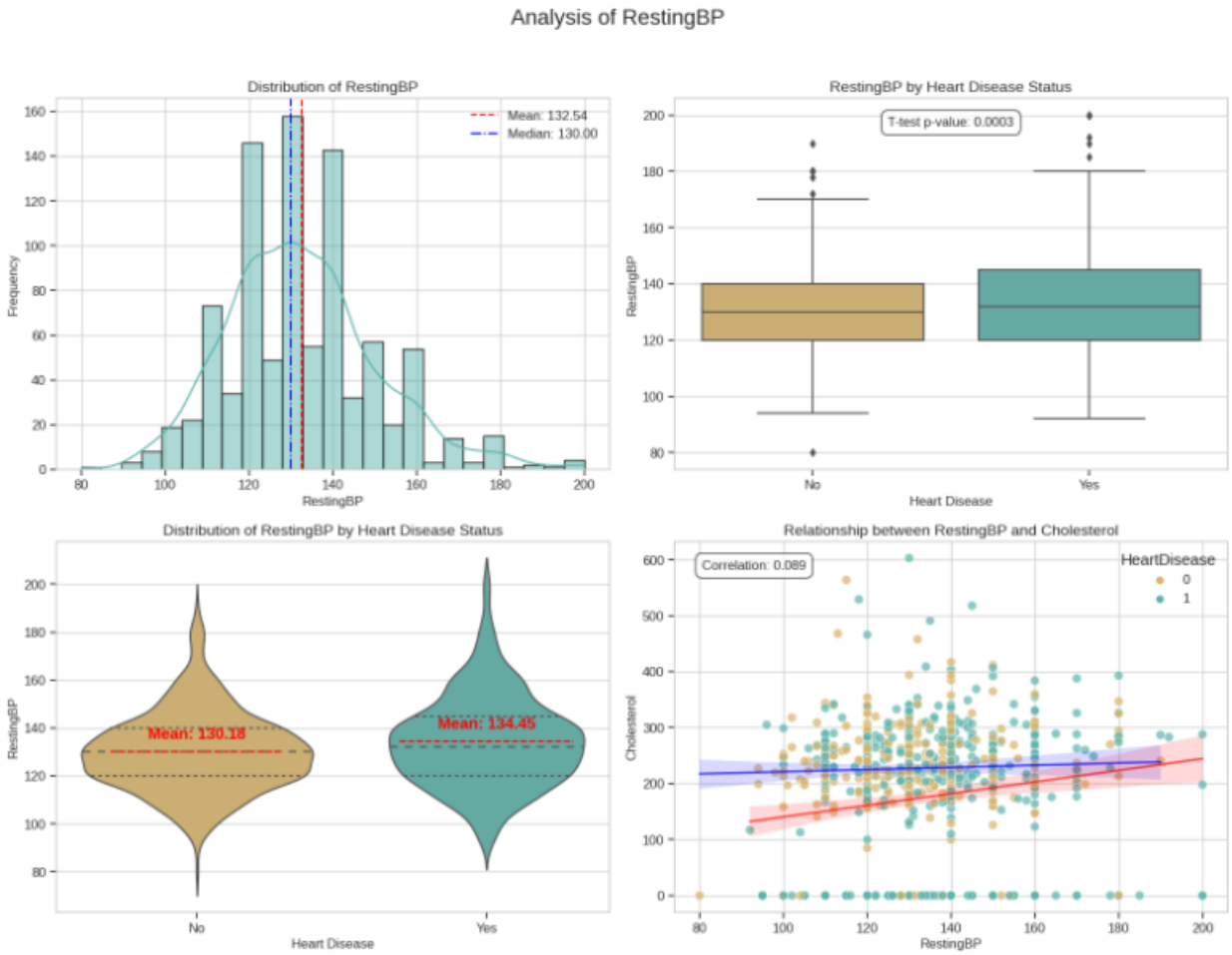

Figure 3: RestingBP Analysis

Different ML Models:

comparison of the K-Nearest Neighbors (KNN), Support Vector Machine with RBF Kernel (SVM_RBF), Decision Tree (DT), Random Forest (RF), and Multilayer Perceptron (MLP) algorithms in terms of their key characteristics, advantages, and limitations describe in table 3 and table 4 provides ML model performance metric.

Table 3: ML model Key characteristics, Advantage, and Limitation

Algorithm	Key Characteristics	Advantages	Limitations
KNN	- Instance-based learning - Non-parametric - Lazy learner	- Simple to implement and understand - No training phase - Effective for small datasets	- Computationally expensive for large datasets - Sensitive to irrelevant features and noise
SVM_RBF	- Kernel-based method - Effective for non-linear data - Margin maximization	- High accuracy for complex datasets - Robust to overfitting in high-dimensional spaces	- Computationally intensive - Requires careful tuning of hyperparameters (e.g., C, gamma)
DT	- Tree-based model - Splits data based on feature values - Interpretable	- Easy to visualize and interpret - Handles both numerical and categorical data	- Prone to overfitting - Sensitive to small changes in data
RF	- Ensemble of decision trees - Bagging technique - Reduces overfitting	- High accuracy and robustness - Handles missing data and outliers well	- Computationally expensive - Less interpretable than single decision trees
Neural Network	- Feedforward neural network - Multiple layers of neurons - Non-linear mapping	- Can model complex, non-linear relationships - Scalable to large datasets	- Requires large amounts of data - Computationally expensive and hard to interpret

Table 4: ML model performance metric

Proposed feature selection model:

Algorithm 1: Proposed feature selection model

1. **Initialize Dataset**
   - Load the dataset (X, y)
   - Split the dataset into training and testing sets

2. **Define Fitness Function**
   - For each individual (feature subset):
     - Select features based on individual (binary encoding: 1 means selected, 0 means not selected)
     - Train a machine learning model (e.g., classifier or regressor) using the selected features
     - Evaluate the model performance (e.g., accuracy, F1-score, AUC)
     - Return the performance score as the fitness value

3. **Initialize Population**
   - Generate an initial population of candidate solutions (individuals), where each individual is a binary list (each bit represents whether a feature is selected or not)

4. **Define Genetic Algorithm Parameters**
   - Set crossover probability (CXProb)
   - Set mutation probability (MutProb)
   - Set the number of generations (GenMax)
   - Set population size (PopSize)

5. **Main Genetic Algorithm Loop** (for generation in 1 to GenMax):
   - **Selection**:
     - Select individuals based on fitness (e.g., using tournament selection or roulette wheel selection)
   - **Crossover**:
     - Randomly pair individuals and perform crossover with probability CXProb to create offspring
   - **Mutation**:
     - Apply mutation to individuals with probability MutProb (flip some bits in their binary representation)
   - **Evaluation**:
     - Evaluate the fitness of all new offspring using the fitness function
   - **Replacement**:
     - Replace less fit individuals in the population with the newly generated offspring

6. **Termination**:
   - Terminate when the maximum number of generations (GenMax) is reached, or when fitness no longer improves
   - Retrieve the best solution (feature subset) from the population

7. **Output**
   - The best feature subset
   - Model performance based on selected features

Result:

1. Accuracy: Measures overall correctness of the model [14].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Measures the proportion of correct predictions.

2. **Precision:** How many of the predicted positives were actually correct.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

3. **Recall:** How many actual positives were correctly predicted.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

4. **F1-Score:** Harmonic mean of precision and recall; balances false positives and false negatives.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

5. **ROC AUC (Receiver Operating Characteristic - Area Under Curve):** No simple formula; it is calculated from the ROC curve which plots:

Measures the ability of a classifier to distinguish between classes.

$$\text{TPR} = \frac{TP}{TP + FN} \quad \text{vs} \quad \text{FPR} = \frac{FP}{FP + TN} \tag{5}$$

6. **Average Precision**

Again, no closed-form formula; it is calculated as the area under the precision-recall curve.Captures both precision and recall across thresholds.

$$\text{AP} = \sum_n (R_n - R_{n-1}) \cdot P_n \tag{6}$$

7. **Cross-Validation F1 Mean:** Average F1 score across k cross-validation folds.

$$\text{CV F1 Mean} = \frac{1}{k} \sum_{i=1}^{k} \text{F1}_i \tag{7}$$

8. **Cross-Validation F1 Std:** Standard deviation of F1 scores from k folds; measures variability.

$$\text{CV F1 Std} = \sqrt{\frac{1}{k} \sum_{i=1}^{k} (\text{F1}_i - \overline{\text{F1}})^2} \tag{8}$$

9. **Training Time:** How long the model took to train on the dataset.

$$\text{Train Time} = \text{End Time} - \text{Start Time} \tag{9}$$

The table shows a complete performance comparison of different machine learning models prior to feature selection, on the basis of several evaluation metrics. Logistic Regression has the highest accuracy (0.8859), recall (0.9118), and F1 score (0.8986), which makes it a good choice for classification problems where sensitivity is important. It also has high ROC AUC (0.8827) and average precision (0.8565), which shows a well-balanced model.

Gradient Boosting follows suit closely, providing industry-leading performance with a well-rounded profile on metrics, specifically leading in precision (0.8922) and recall (0.8922). The Random Forest model also fares competitively, particularly in recall (0.8922) and F1 score (0.8878), but with slightly lower ROC AUC.

Support Vector Machine (SVM) provides consistent performance across the board and is in the lead on cv_f1_mean (0.8746), indicating very good generalization across cross-validation folds. Neural Networks and XGBoost, though still performing well, trail behind in such performance metrics as F1 and ROC AUC, with Neural Network model's longest training time (2.2007 seconds) being of concern to computation efficiency.

Notably, the Naive Bayes model with less complexity yet gets fairly high scores—more so in precision (0.8800) and ROC AUC (0.8582)—and takes the least amount of time to train (0.0022 seconds).

Logistic Regression and Gradient Boosting are the optimal trade-off in terms of performance and efficiency before feature selection, whereas models such as Neural Networks can perhaps improve competitiveness more via feature engineering or tuning.

	accuracy	precision	recall	f1	roc_auc	average_precision	cv_f1_mean	cv_f1_std	train_time
Logistic Regression	0.8859	0.8857	0.9118	0.8986	0.8827	0.8565	0.8631	0.0244	0.1498
Random Forest	0.8750	0.8835	0.8922	0.8878	0.8729	0.8480	0.8684	0.0293	0.2425
Gradient Boosting	0.8804	0.8922	0.8922	0.8922	0.8790	0.8557	0.8635	0.0211	0.1656
SVM	0.8641	0.8738	0.8824	0.8780	0.8619	0.8362	0.8746	0.0228	0.0727
Neural Network	0.8207	0.8416	0.8333	0.8374	0.8191	0.7937	0.8390	0.0182	2.2007
Naive Bayes	0.8587	0.8800	0.8627	0.8713	0.8582	0.8353	0.8514	0.0212	0.0022
XGBoost	0.8424	0.8763	0.8333	0.8543	0.8435	0.8226	0.8498	0.0237	0.0989

Figure 3. ML Model Performance before feature selection.

*The above visualizations present the baseline performance of different classification models in terms of Receiver Oper*ating Characteristic (ROC) Curve and the Precision-Recall (PR) Curve, both of which are crucial tools for measuring model discrimination ability.

In the ROC curve (left), algorithms such as SVM (AUC = 0.936), Logistic Regression (AUC = 0.932), and Random Forest (AUC = 0.932) have very good capability to separate classes, with high True Positive Rates and low False Positive Rates. Gradient Boosting (AUC = 0.918) and XGBoost (AUC = 0.906) also perform well, whereas the Neural Network (AUC = 0.880) performs relatively worse, indicating higher misclassifications than other models.

Precision-Recall Curve (right) is a complement of the ROC curve that emphasizes performance under class imbalance. Logistic Regression (AP = 0.942) and Random Forest (AP = 0.937) top the list with the highest Average Precision (AP) value, showing higher precision at a large number of recall values. Gradient Boosting (AP = 0.904) and SVM (AP = 0.941) continue to perform well, and Neural Network (AP = 0.876) trails.

In general, both curves reflect the strength of Logistic Regression, Random Forest, and SVM prior to feature selection as good baseline models for subsequent tuning and optimization.

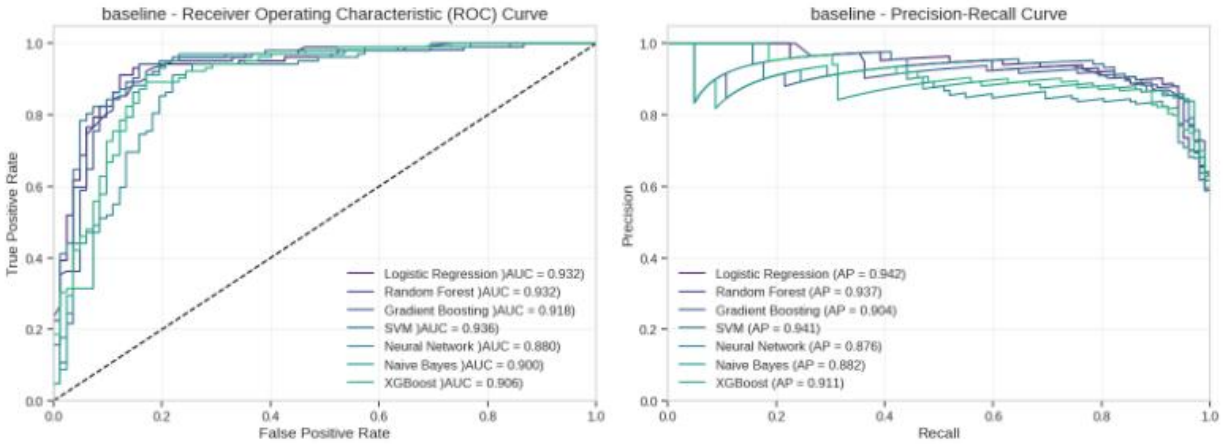

Figure 4. Baseline Model Performance: ROC and Precision-Recall Curves for Classification Models Before Feature Selection

The feature importance Permutation plot above quantifies how every feature influences the performance of the model, in this case, according to the mean decrease in accuracy when values for that feature are randomized. This process can be used to determine the features most essential for accurate prediction.

The most significant feature from the visualization is Oldpeak, with the greatest decline in model accuracy when permuted. This indicates that ST depression during exercise (Oldpeak) is a strong predictor of heart disease. RiskScore comes next, which is an aggregation of different indicators of health and also exhibits good predictive capacity.

Some key features are Sex_M, Cholesterol, and ChestPainType_ASY, which demonstrate the significance of gender, cholesterol, and non-typical chest pain in heart disease risk assessment. Surprisingly, even dichotomous indicators such as FastingBS_1 (fasting blood sugar > 120 mg/dl) and RestingECG_LVH (left ventricular hypertrophy on ECG) are significant contributors.

Least important at the lower end are MaxHR (max heart rate attained), ST_Slope_Up, and RestingECG_ST, having very little influence on accuracy upon shuffling, indicating they are likely to be of lesser utility for model prediction here.

Generally speaking, the chart proves useful for making feature selections, allowing for features to be ranked as important contributors to the performance of the model and arguably enhancing model efficiency through the removal of less important features.

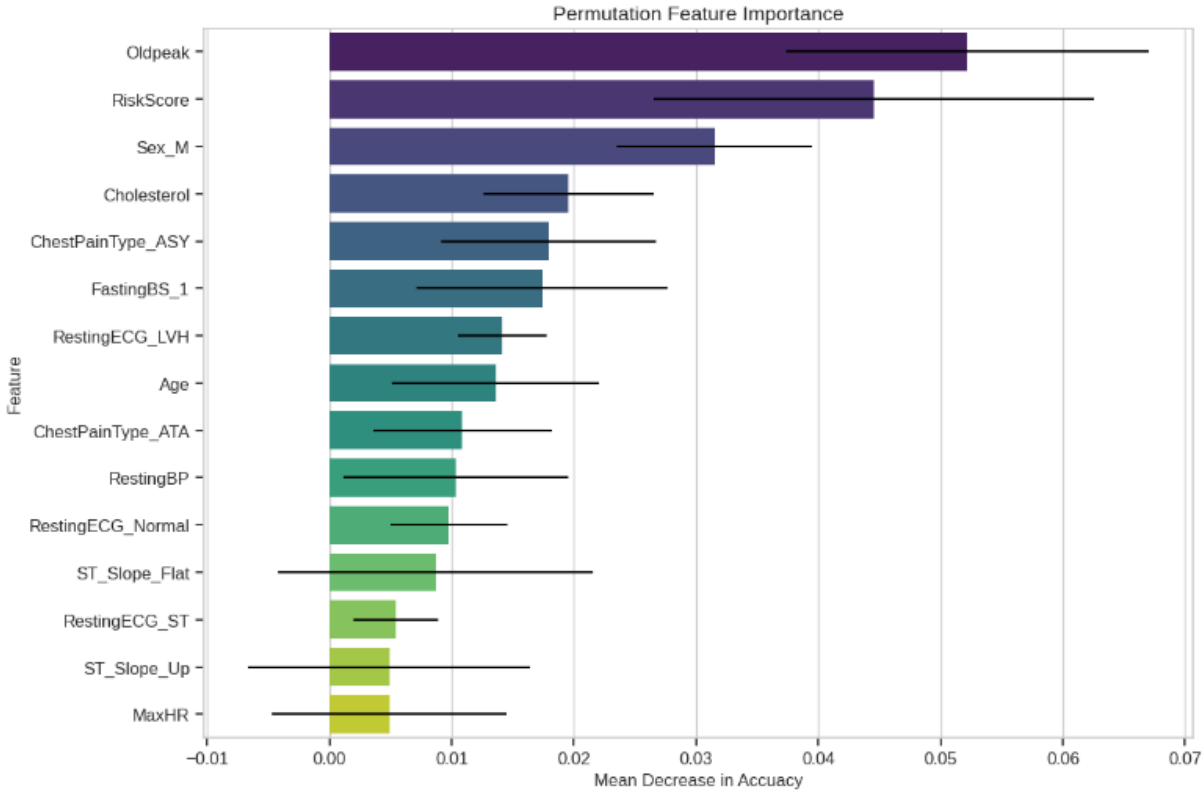

Figure 5. Report based on proposed Feature Selection technique

Following feature selection, various machine learning models were compared with different performance measures. Out of them, Logistic Regression and Random Forest emerged with the highest accuracy (0.8750), recall (0.9020), F1 score (0.8889), ROC AUC (0.8717), and average precision (0.8446), and thus they are good choices for tasks demanding high recall and well-balanced performance. XGBoost reported best precision (0.8878) and optimal cross-validation F1 mean (0.8642), depicting high prediction quality and stability with respect to the validation folds. Interestingly, the lowest cross-validation F1 standard deviation (0.0124) was associated with Neural Network, reflecting best stability, albeit at the expense of longest training time (1.8334 seconds). Compared to that, Naive Bayes provided least training time (0.0019 seconds) but only represented moderate performance over metrics. Overall, XGBoost is a well-balanced trade-off between accuracy, precision, stability, and training time, and Logistic Regression and Random Forest are good options for high-recall tasks. Which model to use depends on the task requirements at hand, such as whether speed, recall, or model stability is the priority.

	accuracy	precision	recall	f1	roc_auc	average_precision	cv_f1_mean	cv_f1_std	train_time
Logistic Regression	0.8750	0.8762	0.9020	0.8889	0.8717	0.8446	0.8733	0.0235	0.2764
Random Forest	0.8750	0.8762	0.9020	0.8889	0.8717	0.8446	0.8727	0.0269	0.2381
Gradient Boosting	0.8587	0.8800	0.8627	0.8713	0.8582	0.8353	0.8597	0.0277	0.1963
SVM	0.8641	0.8738	0.8824	0.8780	0.8619	0.8362	0.8747	0.0162	0.0782
Neural Network	0.8533	0.8788	0.8529	0.8657	0.8533	0.8311	0.8372	0.0124	1.8334
Naive Bayes	0.8587	0.8800	0.8627	0.8713	0.8582	0.8353	0.8502	0.0312	0.0019
XGBoost	0.8587	0.8878	0.8529	0.8700	0.8594	0.8387	0.8642	0.0215	0.0577

Figure 6. ML Model Performance after feature selection.

The performance of the model incorporating interaction features was tested with the Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves. ROC curve analysis showed that Logistic Regression performed the best with the highest AUC (0.934), followed by Random Forest (0.932), SVM (0.931), Gradient Boosting (0.925), and XGBoost (0.916), while Naive Bayes (0.913) and Neural Network (0.893) came in second last. Likewise, in PR curve evaluation—especially insightful for imbalanced sets—Logistic Regression topped again with an AP of 0.940, followed by SVM (0.929), Random Forest (0.924), Gradient Boosting (0.920), XGBoost (0.916), Naive Bayes (0.911), and Neural Network (0.873), which ranked at the bottom. In general, Logistic Regression performed better than other models consistently, whereas tree-based models (Random Forest, Gradient Boosting, XGBoost) showed consistent performance, and the poorer performance of the Neural Network could be due to overfitting, lack of tuning, or sensitivity to preprocessing.

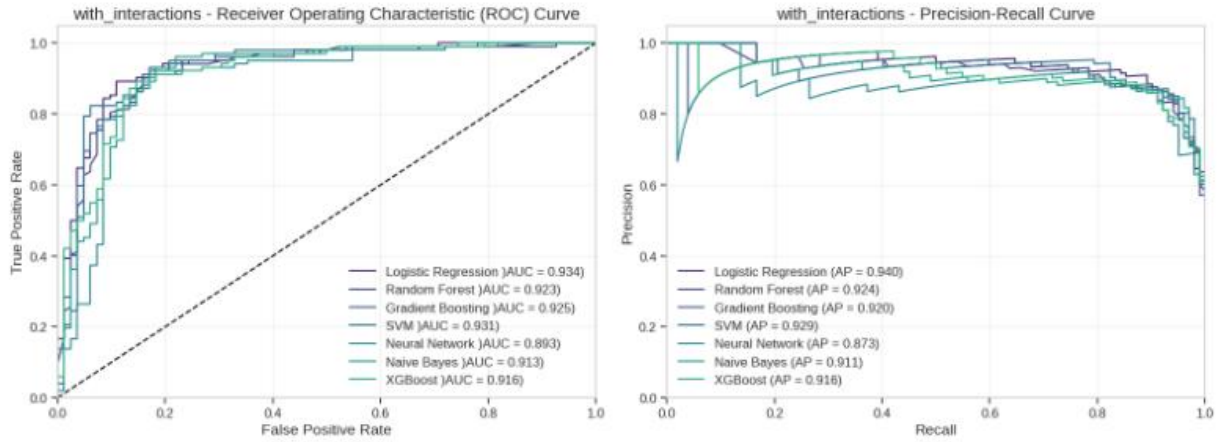

Figure 7. Report based on proposed Feature Selection technique

Discussions: Comparison of the machine learning models revealed remarkable model performance improvements after including interaction terms and feature selection. Logistic Regression and Random Forest performed best across most of the metrics—accuracy (0.875), recall (0.902), F1-score (0.8889), and ROC AUC (0.8717)—indicating their stability and viability in clinical risk prediction where high sensitivity is critical.

XGBoost was the best model for accuracy (0.8878) and cross-validation mean F1-score (0.8642) and thus best suited for applications where high confidence in positive labels is needed. Interestingly, the Neural Network, although with the lowest cross-validation variation (CV F1 Std = 0.0124), had the highest training time, suggesting possible computational inefficiencies.

Adding interaction features further boosted model discrimination power, evidenced by better ROC and PR curves. Logistic Regression achieved optimal performance (AUC = 0.934, AP = 0.940), followed by Random Forest and SVM, confirming their generalization ability. By contrast, the Neural Network still lagged behind (AUC = 0.893, AP = 0.873), perhaps due to overfitting or data representation sensitivity.

Feature importance analysis revealed Oldpeak, RiskScore, Sex_M, and Cholesterol as the predictors with the highest impact, which is consistent with clinical expectations. These results demonstrate the synergy between domain-specific features and algorithmic effectiveness in predictive modeling.

Metric	Best Model	Score
Accuracy	Logistic Regression / Random Forest	0.875
Precision	XGBoost	0.8878
Recall	Logistic Regression / Random Forest	0.902
F1 Score	Logistic Regression / Random Forest	0.8889
ROC AUC	Logistic Regression / Random Forest	0.8717
Average Precision	Logistic Regression / Random Forest	0.8446
CV F1 Mean	XGBoost	0.8642
CV F1 Std	Neural Network	0.0124 (lowest variability)
Train Time	Naive Bayes	0.0019 sec (fastest)

Conclusions: This research validates the significance of feature selection and interaction modeling in boosting machine learning-based cardiovascular disease prediction. Logistic Regression and Random Forest were the most dependable classifiers with a good trade-off between accuracy, recall, and interpretability. XGBoost also proved to have higher precision and stability and was suitable for applications needing high prediction confidence.

The work presents a scalable framework for creating intelligent CVD diagnostic systems by employing permutation-based feature importance and genetic algorithm-based feature optimization. The application of ROC and Precision-Recall curves enabled detailed insight into model performance under class imbalance—ubiquitous in medical data.

The repeated dominance of Logistic Regression over even more complicated models such as Neural Networks reiterates the fact that simple models with expertly crafted features tend to yield the best results in healthcare contexts. Future research can investigate integration with real-time clinical data, model explainability improvements, and deployment in early cardiovascular risk assessment decision support systems.

References

[1] Veroff, D. R., Sullivan, L. A., Shoptaw, E. J., Venator, B., Ochoa-Arvelo, T., Baxter, J. R., ... & Wennberg, D. (2012). Improving self-care for heart failure for seniors: the impact of video and written education and decision aids. Population health management, 15(1), 37-45.

[2] Ahmadli, N., Sarsil, M. A., Mizrak, B., Karauzum, K., Shaker, A., Tulumen, E., ... & Ergen, O. (2024). Voice-Driven

[3] Uddin, K. M. M., Dey, S. K., & Babu, H. M. H. (2024). A Voice assistive mobile application tool to detect cardiovascular disease using machine learning approach. Biomedical Materials & Devices, 2(2), 1246-1257.

[4] Abbas, S., Ojo, S., Al Hejaili, A., Sampedro, G. A., Almadhor, A., Zaidi, M. M., & Kryvinska, N. (2024). Artificial intelligence framework for heart disease classification from audio signals. Scientific Reports, 14(1), 3123.

[5] Idrisoglu, A. (2024). Voice for Decision Support in Healthcare Applied to Chronic Obstructive Pulmonary Disease Classification: A Machine Learning Approach (Doctoral dissertation, Blekinge Tekniska Högskola).
[5] Alsekait, D. M., Shdefat, A. Y., Nabil, A., Nawaz, A., Rana, M. R. R., Ahmed, Z., ... & AbdElminaam, D. S. (2024). Heart-Net: A Multi-Modal Deep Learning Approach for Diagnosing Cardiovascular Diseases. Computers, Materials & Continua, 80(3).
[6] Mayourian, J., El-Bokl, A., Lukyanenko, P., La Cava, W. G., Geva, T., Valente, A. M., ... & Ghelani, S. J. (2024). Electrocardiogram-based deep learning to predict mortality in paediatric and adult congenital heart disease. European Heart Journal, ehae651.
[7] Ogunpola, A., Saeed, F., Basurra, S., Albarrak, A. M., & Qasem, S. N. (2024). Machine learning-based predictive models for detection of cardiovascular diseases. Diagnostics, 14(2), 144.
[8] Drouard, G., Mykkänen, J., Heiskanen, J., Pohjonen, J., Ruohonen, S., Pahkala, K., ... & Kaprio, J. (2024). Exploring machine learning strategies for predicting cardiovascular disease risk factors from multi-omic data. BMC Medical Informatics and Decision Making, 24(1), 116.
[9] DeGroat, W., Abdelhalim, H., Patel, K., Mendhe, D., Zeeshan, S., & Ahmed, Z. (2024). Discovering biomarkers associated and predicting cardiovascular disease with high accuracy using a novel nexus of machine learning techniques for precision medicine. Scientific reports, 14(1), 1.
[10] Cai, Y. Q., Gong, D. X., Tang, L. Y., Cai, Y., Li, H. J., Jing, T. C., ... & Zhang, G. W. (2024). Pitfalls in developing machine learning models for predicting cardiovascular diseases: challenge and solutions. Journal of Medical Internet Research, 26, e47645.
[11] Korial, A. E., Gorial, I. I., & Humaidi, A. J. (2024). An improved ensemble-based cardiovascular disease detection system with chi-square feature selection. Computers, 13(6), 126.
[12] Alghamdi, F. A., Almanaseer, H., Jaradat, G., Jaradat, A., Alsmadi, M. K., Jawarneh, S., ... & Alfagham, H. (2024). Multilayer perceptron neural network with arithmetic optimization algorithm-based feature selection for cardiovascular disease prediction. Machine Learning and Knowledge Extraction, 6(2), 987-1008.
[13] Moreno-Sánchez, P. A., García-Isla, G., Corino, V. D., Vehkaoja, A., Brukamp, K., Van Gils, M., & Mainardi, L. (2024). ECG-based data-driven solutions for diagnosis and prognosis of cardiovascular diseases: A systematic review. Computers in Biology and Medicine, 108235.
[14] Mall, P. K. (2025). Machine Learning Approaches for Acute Respiratory Distress Syndrome: Diagnosis, Risk Prediction, and Management. SGS-Engineering & Sciences, 1(1).