Machine Learning-Based Early Cardiovascular Disease Prediction: A Comparative Analysis of Supervised Learning Algorithms Using a Pakistani Clinical Dataset

Awais Khursheed, Soban Ahmed, Sibghat Ullah, Zaryab Khattak, Bilal Ur Rehman*
Department of Electrical Engineering, University of Engineering and Technology, Peshawar, Pakistan.
*Correspondence: bur@uetpeshawar.edu.pk

Citation | Khursheed. A, Ahmed. S, Ullah. S, Khattak. Z, Rehman. B. U, "Machine Learning-Based Early Cardiovascular Disease Prediction: A Comparative Analysis of Supervised Learning Algorithms Using a Pakistani Clinical Dataset", IJIST, Vol. 8 Issue. 5 pp 2098-2112, August 2026.
DOI | https://doi.org/10.33411/IJIST/1963
Received | July 20, 2026 **Revised |** Aug 17, 2026 **Accepted |** Aug 18, 2026 **Published |** Aug 21, 2026.

Cardiovascular Diseases (CVDs) continue to be one of the leading causes of deaths in the world, claiming some 17.9 million lives every year. This burden is higher in Pakistan because of "Asian Indian Phenotype" which makes them vulnerable to early coronary artery disease. The commonly used traditional risk prediction models, including the Framingham Risk Score, have been developed in Western populations and are poorly predictive in South Asian populations. This study aims to fill this important gap by designing, implementing and comparative evaluation of six supervised machine learning algorithms for early detection of cardiovascular disease using a locally collected clinical dataset of 411 patient records with 13 independent clinical attributes. The models tested are Logistic Regression, K Nearest Neighbor, Support Vector Machine, Random Forest, Gradient Boosting and XGBoost. A rigorous gender-based mean imputation and Z-score normalization was done and split in 80/20 ratio. Empirical results show that the Random Forest classifier has Area under the Curve (AUC) of 0.9842, accuracy of 95.2%, precision of 96.0% and recall of 96.0%. The model was then exported and used to create a browser-based, predictive application that could be embedded in an interactive dashboard for real-time cardiovascular risk without the need for a server. These results confirm the effectiveness of ensemble learning approaches for medical diagnostics and highlight the potential of implementing ML-based screening tools in the limited resource healthcare environment in Pakistan.

Keywords: Cardiovascular Disease, Machine Learning, Random Forest, Support Vector Machine, XGBoost, Early Diagnosis, Pakistan, Asian Indian Phenotype, Interactive Dashboard.

Introduction:

Cardiovascular diseases (CVDs) are a major health problem worldwide, accounting for an estimated 32% of deaths worldwide [1]. According to the World Health Organization, ischemic heart disease and stroke are the two main causes of death worldwide, with most deaths from CVDs in low and middle-income countries (LMICs), where healthcare infrastructure is lacking [2]. The total number of deaths due to CVDs in Pakistan is nearly 30% of all deaths, and this has been steadily increasing over the last 20 years [3]. The genetic predisposition, termed the 'Asian Indian Phenotype'[4], is complicated by several modifiable risk factors: hypertension, diabetes mellitus, dyslipidemia, tobacco use and sedentary lifestyle, which contribute to this burden in an escalating manner.

The problem faced by Pakistan is complex in nature. Tertiary cardiac care centers are available in big cities like Karachi, Peshawar, Lahore and Islamabad, but in remote rural and semi-urban areas, majority of the people don't have access to even basic cardiac diagnostic centers. There are a substantial proportion of patients who have ST-Elevation Myocardial Infarction (STEMI) outside the desired treatment window, because of the delay in recognition, referral and transport [5]. This highlights the need for tools that can help to accurately stratify risk at the primary healthcare level and early in the course of the illness.

Machine learning (ML) algorithms have become a significant advancement in medical diagnostics with their potential of uncovering hidden patterns in multi-dimensional clinical data. Various ML techniques such as Random Forest, XGBoost, Support Vector Machines, and deep learning architectures have been shown to achieve high predictive accuracy in the detection of heart disease in recent studies, playing a crucial role in precisely predicting the disease based on historical data, thereby enhancing the efficiency of diagnostic processes [6][7].

Despite the existence of vast literature, a critical review of it reveals that most studies that apply ML to CVD prediction use publicly available datasets that are based on western populations (e.g the UCI Cleveland dataset or the Kaggle Cardiovascular dataset). Models developed on this type of data will be inherently biased to the demographic and clinical characteristics of these populations. These models have been poorly calibrated for use in South Asian patients, who have very different risk profiles: metabolic syndrome, abdominal obesity, and insulin resistance occur at lower body mass indices. This study aims to fill this gap directly as it is based on a dataset that comes from a local clinical setting in Pakistan.

Related Work:

Over the past decade, researchers have intensively studied the use of machine learning for predicting cardiovascular diseases. The development of predictive models has been investigated through numerous studies using different algorithms and feature engineering methods to make the models more efficient. A few studies emphasized the need for data quality and the significance of certain features to increase the accuracy of the model [8].

Seven ML models, such as Logistic Regression, CNN, SVM, Gradient Boosting, KNN, XGBoost, and Random Forest, were used on two publicly available datasets in a comprehensive study. They used Scikit-Learn and Keras libraries for their evaluation and produced good results in various metrics. They found that ensemble techniques, specifically XGBoost, showed excellent accuracy and a well-balanced F1 score on the Cardiovascular Heart Disease Dataset. Another study proposed a hybrid machine learning technique for effective heart disease prediction, highlighting the importance of using multiple algorithmic approaches. Other work concentrated on the prediction of heart disease using commonly used ML algorithms to establish baseline performance. A follow-up study [9] evaluating prediction algorithms yielded encouraging results on structured clinical data. To handle temporal pattern recognition in cardiac data, a hybrid CNN and LSTM network [10] was introduced for heart disease prediction. An Optimal Feature Selection based cardiovascular disease detection framework using ML was proposed [11] with competitive accuracy over benchmark datasets.

The role of feature engineering techniques in enhancing the prediction of heart diseases using machine learning was also shown.

Further research points to the need for locally trained models, as models trained on local hospital data [12] are better than those based only on European cohort data [13]. Machine-learning systems have also been automated for the prediction of heart failure across multiple modalities of data [14] and foundational reviews [15][16] show the general utility of AI in healthcare applications. To improve diagnostic accuracy, advanced optimization methods have been developed, for example, genetic algorithm with support vector machine [17]. Deep learning methods have been used to directly obtain cardiac dynamics from a raw ECG signal [18] and sequential data by using a Long Short-Term Memory (LSTM) network [19].

Recent studies have further emphasized the importance of combining machine learning prediction models with explainable artificial intelligence approaches to improve transparency and clinical acceptance [20]. Ensemble learning-based approaches have demonstrated improved robustness for cardiovascular risk prediction by combining multiple classifiers and reducing model variance [21].

Recent advancements have further strengthened the role of machine learning and artificial intelligence in cardiovascular disease prediction. A comprehensive study by Ogunpola *et al.* investigated machine learning-based predictive models for cardiovascular disease detection and emphasized the importance of robust preprocessing strategies, feature selection, and model evaluation for improving clinical prediction performance [22]. Similarly, Bouqentar *et al.* developed an early heart disease prediction framework using feature engineering and comparative machine learning approaches, demonstrating that optimized feature representation can significantly enhance predictive accuracy in clinical datasets [23]. More recently, Kaushik *et al.* proposed a neural network-based heart disease prediction model that utilized advanced machine learning techniques for improving diagnostic accuracy and highlighted the potential of AI-driven systems for supporting early cardiovascular risk assessment [24]. These recent studies demonstrate the continuous evolution of ML-based cardiovascular prediction systems; however, most existing approaches still rely on publicly available datasets, emphasizing the need for locally representative clinical datasets such as the Pakistani cohort used in this study.

However, a major limitation identified across the existing literature is that many cardiovascular disease prediction models have been developed using publicly available datasets collected from Western populations, which may limit their generalizability to South Asian communities due to differences in genetic background, lifestyle factors, metabolic characteristics, and disease patterns. Models trained on population-specific clinical data have the potential to achieve better calibration and practical relevance for local healthcare settings. Furthermore, most previous studies primarily report model performance through offline evaluation, with limited emphasis on practical deployment for real-time clinical decision support. Therefore, the development of locally validated machine learning models integrated with accessible prediction platforms represents an important direction for improving cardiovascular risk assessment in underrepresented populations.

Research Gap and Scientific Novelty:

The research gap and novelty are as follows:

Existing cardiovascular prediction studies mainly rely on publicly available Western datasets, limiting their applicability to South Asian populations.

This study utilizes a locally collected Pakistani clinical dataset consisting of 411 patient records to develop population-specific prediction models.

Six supervised machine learning algorithms representing different learning paradigms are systematically compared under the same experimental framework.

The study identifies the optimal model considering clinically important evaluation metrics, particularly recall and false-negative rate.

The developed Random Forest model is integrated into a browser-based prediction dashboard, enabling practical deployment in resource-limited healthcare environments.

Research Contributions:

The major findings of this study are:

To develop a machine learning-based cardiovascular disease prediction framework using a locally collected Pakistani clinical dataset containing 411 patient records and 13 clinical attributes.

To preprocess clinical data using gender-based mean imputation and Z-score normalization techniques to improve model reliability.

To implement and compare six supervised machine learning classifiers, including Logistic Regression, KNN, SVM, Random Forest, Gradient Boosting, and XGBoost.

To evaluate the predictive performance of each model using accuracy, precision, recall, F1-score, and AUC-ROC metrics.

To identify the most effective classifier based on predictive performance and false-negative reduction for healthcare screening applications.

To develop a browser-based deployment framework for real-time cardiovascular risk prediction.

Table 1 gives a brief overview of related studies to give a concise cross-study comparison of the datasets, models and reported outcomes.

Table 1. Related Work Summary

Study	Dataset/Context	Methods	Best Reported Result	Limitation
[1]	Public CVD datasets	LR, CNN, SVM, GB, KNN, XGB, RF	High accuracy with XGBoost	No Pakistani local data
[6]	Heart disease benchmark data	Hybrid ML framework	Improved prediction over single models	General population focus
[9]	Cardiac clinical signals	Hybrid CNN-LSTM	Strong Deep Model Performance	Higher computational complexity
[10]	Benchmark cardiovascular data	ML + feature selection	Competitive Accuracy	Limited local calibration

Research Methodology:

As illustrated in Figure. 1, the methodology follows a structured pipeline from data collection to model deployment.

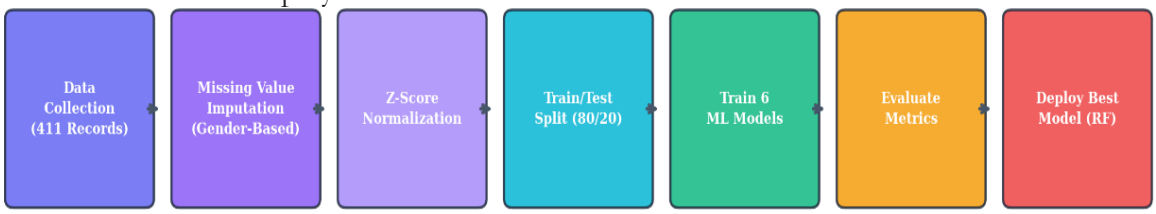

Figure 1. Research Methodology Workflow

Overview of the Proposed Machine Learning Framework:

The proposed framework consists of several sequential stages, beginning with clinical data acquisition and ending with model deployment for cardiovascular disease prediction. Each stage of the workflow is described below.

Step 1: Clinical Data Collection and Dataset Preparation:

The first stage involves collecting clinical records from patients in a Pakistani healthcare setting. The dataset contains 411 patient records with 13 independent clinical attributes, including demographic characteristics, physiological measurements, exercise-related parameters, and cardiac indicators [25]. Each patient record is associated with a binary target variable representing the presence or absence of cardiovascular disease. The collected dataset serves as the foundation for developing population-specific prediction models.

Step 2: Data Inspection and Preprocessing:

Since clinical datasets may contain missing values, inconsistent entries, and variables with different numerical ranges, preprocessing is performed before model development. Missing values are identified and handled using appropriate imputation techniques. In this study, gender-based mean imputation was applied for missing cholesterol values to preserve possible differences in lipid profiles between male and female patients. Continuous variables are then normalized using Z-score normalization to ensure that all numerical features contribute proportionally during model training.

Step 3: Feature Selection and Data Transformation:

After preprocessing, the clinical features are transformed into a suitable format for machine learning algorithms. The prepared feature matrix contains the independent variables, while the target vector represents cardiovascular disease status. The processed dataset is divided into training and testing subsets using an 80/20 stratified split to maintain the original class distribution and ensure reliable model evaluation.

Step 4: Machine Learning Model Development:

The training dataset is used to develop and optimize six supervised machine learning models representing different learning approaches. These include Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Random Forest, Gradient Boosting, and XGBoost. Each algorithm is trained using the same preprocessed dataset to ensure a fair comparative evaluation.

Step 5: Model Performance Evaluation:

The trained models are evaluated using multiple performance indicators, including accuracy, precision, recall, F1-score, and AUC-ROC. These metrics provide a comprehensive assessment of classification performance. Particular emphasis is placed on recall because minimizing false-negative predictions is critical in cardiovascular disease screening applications.

Step 6: Comparative Analysis and Best Model Selection:

The performance of all developed models is compared to identify the most suitable algorithm for cardiovascular risk prediction. The model achieving the best balance between predictive accuracy, sensitivity, and overall discrimination capability is selected as the final prediction model.

Step 7: Model Interpretation and Deployment:

The selected model is further analyzed to identify important clinical predictors contributing to cardiovascular disease classification. The final trained model is then integrated into a browser-based application to demonstrate its potential for real-time prediction and practical deployment in healthcare environments.

This stepwise workflow ensures that the proposed framework follows a systematic process from raw clinical data collection to an operational machine learning-based cardiovascular prediction system.

Dataset Description:

This research uses the dataset of 411 patients from a local clinical environment in Pakistan [25]. It has 13 independent clinical features and 1 binary target variable denoting the presence (1) or absence (0) of cardiovascular disease. The dataset is important in the context of healthcare and the field of machine learning and can be used for tasks related to prediction

and classification of cardiovascular diseases in the Pakistani population. Table 2 provides detailed descriptions of all the attributes.

Table 2. Dataset Feature Description

Feature	Description	Type	Range
age	Patient age in years	Continuous	20–80
gender	Binary (1 = Male, 0 = Female)	Categorical	0, 1
chestpain	Chest pain type (0=TA, 1=AA, 2=NAP, 3=ASY)	Nominal	0–3
resting_BP	Resting blood pressure (mm Hg)	Continuous	90–200
serum_cholestrol	Serum cholesterol (mg/dL)	Continuous	100–600
fastingbloodsugar	Fasting blood sugar > 120 mg/dL	Binary	0, 1
resting_electro	Resting ECG results (0=Normal, 1=ST-T, 2=LVH)	Nominal	0–2
maxheartrate	Maximum heart rate achieved (bpm)	Continuous	70–200
exercise_angina	Exercise-induced angina (0=No, 1=Yes)	Binary	0, 1
oldpeak	ST depression induced by exercise	Continuous	0–6.2
slope	Slope of peak exercise ST segment	Nominal	0–2
no_major_vessels	Number of major vessels colored by fluoroscopy	Discrete	0–3
target	Presence of heart disease (0=Healthy, 1=Disease)	Binary	0, 1

This dataset includes 411 patient records, of which 59.9% (246 patients) have CVD, and 165 patients are healthy. Most of the data is of males (78%). The age range is 20 to 80 years, and the mean age is 49.8 years

Data Preprocessing:

Real data from a clinical application is always noisy, incomplete and features may be on different scales [25]. To present the data for machine learning, a thorough data preparation pipeline was followed, adhering to best practices in medical informatics.

Data Cleaning and Imputation:

First, on checking the data, it was observed that the serum_cholestrol column had missing values (21 records, 5.1% of data) [25]. We used a gender based mean imputation approach, in which males and females were imputed separately from each other, instead of using a naive mean imputation. This method maintains the gender-specific distribution of cholesterol levels, which is of great importance in a clinical context because there are significant differences between genders in lipid profiles. Other missing values were found for oldpeak (7 records), resting_BP (5 records), maxheartrate (3 records) and less than 2 records of other features.

Feature Normalization:

The distance between data points (such as using Euclidean distance) is the basis of ML algorithms like K-Nearest Neighbors (KNN) and Support Vector Machines (SVM). When features are on different numerical scales, the larger-scale features will have a greater influence on the distance calculation. To fix this, all continuous features were normalized using Z-score normalization (StandardScaler).

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

Where:

z is the calculated standard score, representing the normalized value.

x is the original, raw numerical data point for the clinical feature.

μ is the mathematical mean of the feature dataset.

σ is the standard deviation of the feature dataset.

It is a transformation to make all features equally useful in the decision boundary of the model: normalizing the features to have a mean and standard deviation of 0 and 1, respectively.

Train-Test Split:

The preprocessed dataset was split into training and testing sets in a stratified manner with 80% of the data in the training set and 20% in the testing set, with 328 data samples in the training set and 83 data samples in the testing set. To avoid class imbalance artifacts during evaluation, stratified sampling was used to ensure that the original class distribution was maintained in both sets.

Machine Learning Models:

We chose 6 algorithms that sample a wide range of learning approaches, from linear to instance-based, geometric and ensemble approaches. Each algorithm contributes distinct characteristics to unveil predictive revelations in the analysis of cardiovascular diseases.

Logistic Regression (LR): A model that uses the logistic (sigmoid) function as a statistical model for a binary dependent variable. It is a robust baseline, and can be interpreted with learned coefficients or odds ratios. Configuration: max_iter=1000, random_state=42.

K Nearest Neighbor (KNN): Non-Parametric and instance-based algorithm which classifies a new patient according to the most frequent class among its closest neighbors in the feature space. It is sensitive to the k that is chosen and the distance metric used. k=5 and Euclidean distance.

Support Vector Machine (SVM): A supervised learning model that seeks the best hyperplane separating the two classes with the largest margin. We used the Radial Basis Function (RBF) kernel that maps data to a higher dimension space and allows to capture non-linear relationships. Configuration: kernel=rbf, p.

Random Forest (RF): An ensemble learning approach that works by building many decision trees while training. It returns the mode of the classes of individual trees using bagging to reduce overfitting. Using default values: n_estimators=100, max_depth=10 and random_state =42.

Gradient Boosting (GB): An ensemble technique that builds models sequentially, where each new tree improves the accuracy of already trained trees. This boosting method iteratively minimizes a loss function. Parameters: n_estimators = 100, max_depth = 5, random_state = 42.

XGBoost is an optimized distributed gradient boosting library that implements parallel tree boosting with L1 and L2 regularisation for better generalisation (XGBoost or XGB). It has become the leading algorithm in machine learning competitions and medical applications. Model parameters: n_estimators=100, max_depth=5, eval_metric=logloss.

Performance Metrics:

To get a comprehensive assessment, we used the following metrics from the confusion matrix:

Accuracy: Overall correct predictions relative to total instances.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (2)$$

Precision: Positive prediction accuracy; high precision reduces false positive diagnoses.

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

Recall (Sensitivity): The model's ability to capture all positive instances. In medical screening, this is the most critical metric, as a false negative (missing a diseased patient) carries severe clinical consequences.

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

F1-Score: Harmonic mean balancing precision and recall, essential for imbalanced datasets.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

AUC-ROC: Measures the model's discriminative ability across all classification thresholds.

$$AUC = \int TPR\, d(FPR) \quad (6)$$

Where definitions for the metric terms are:

TP (True Positives): The number of diseased patients correctly identified.

TN (True Negatives): The number of healthy patients correctly identified.

FP (False Positives): The number of healthy patients incorrectly identified as diseased.

FN (False Negatives): The number of diseased patients incorrectly identified as healthy.

TPR (True Positive Rate): Mathematically equivalent to Recall (TP/TP+FN).

FPR (False Positive Rate): The proportion of actual negatives incorrectly categorized as positive (FP/FP+TN).

$\int$ represents the mathematical integral evaluating the total area bound under the curve across all possible decision thresholds.

Results and Performance Evaluation:

The following section presents the results of the study. 80% of the dataset (328 records) was used for training of all the models and 20% (83 records) was used as a test set for validation. Comprehensive testing reveals the algorithm's strengths and weaknesses for predicting cardiovascular diseases.

Feature Correlation Analysis:

To get an idea of the underlying structure of the data, a correlation matrix was calculated. Figure. 2 shows the structure of the correlation for the features which shows the maximum correlation of the target class.

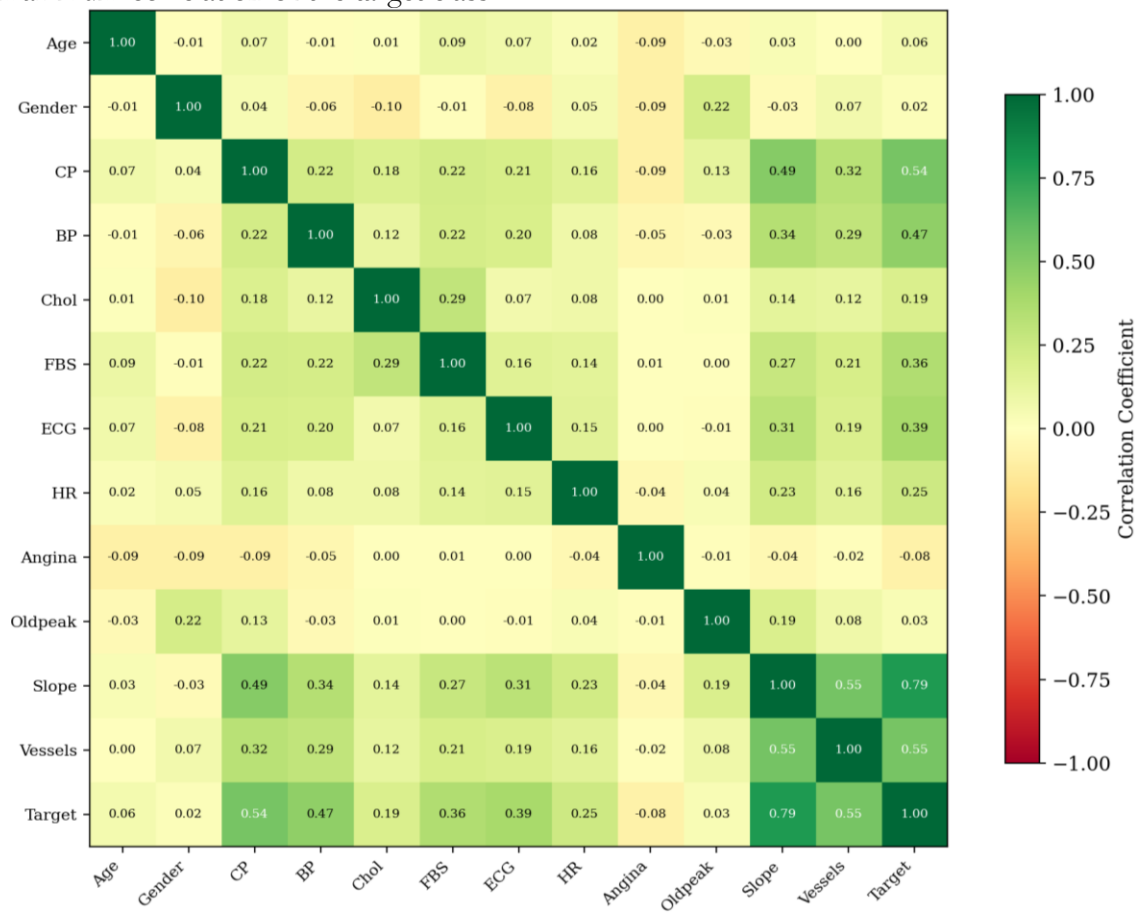

Figure 2. Feature Correlation Heatmap

It was observed that Chest Pain type (cp) and Max Heart Rate (thalach) were significantly correlated with the Target variable, which is in line with medical knowledge about angina and heart rate variability being cardinal signs of cardiac dysfunction. Both Exercise-Induced Angina (exang) and ST Depression (oldpeak) had good correlation with disease presence suggesting that these exercise stress-test parameters are important for discriminating between healthy and diseased patients.

Comparative Model Performance:

Table 3 gives an overview of the comparative performance of the six algorithms on the test set. These models have different characteristics, ranging from linear models to ensemble models. Based on the empirical evaluation, we evaluated the accuracy, precision, recall, F1-measure, and AUC 1 of all models presented in this paper, and found that the ensemble models generally show better and more balanced predictive power in the above metrics, as shown in Figure. 3.

Table 3. Comparative Performance Analysis of ML Models

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	95.2%	96.0%	96.0%	96.0%	0.9788
KNN	91.6%	97.8%	88.0%	92.6%	0.9727
SVM	95.2%	96.0%	96.0%	96.0%	0.9782
Random Forest	**95.2%**	**96.0%**	**96.0%**	**96.0%**	**0.9842**
Gradient Boosting	92.8%	95.8%	92.0%	93.9%	0.9761
XGBoost	95.2%	96.0%	96.0%	96.0%	0.9818

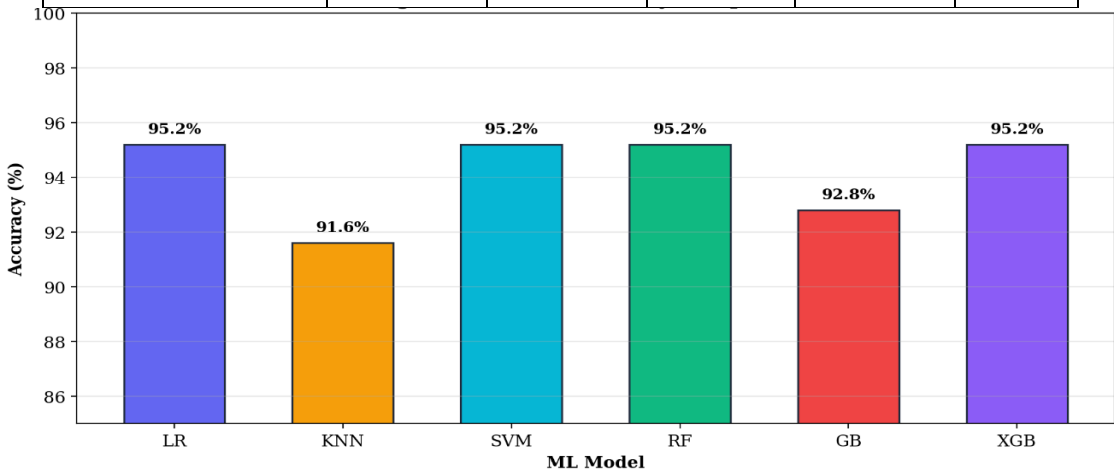

Figure 3. Model Accuracy Comparison

Analysis of Results:

Random Forest was the best model with an AUC of 0.9842 and accuracy, precision and recall of 95.2%, 96.0% and 96.0% respectively. The model achieved an accuracy of 79 out of 83 test patients, 2 false negative(s) and 2 false positive(s). Random Forest has a favorable balance between bias and variance reduction due to the bagging which combines the prediction of 100 different decision trees.

Both Logistic Regression and XGBoost models had an accuracy of 95.2%, precision of 96.0%, and recall of 96.0%. What's remarkable, however, is the performance of Logistic Regression, which indicates that a significant part of this decision boundary is linearly separable, consistent with the structured nature of clinical data. The AUC of XGBoost is 0.9818 which proves that this optimized version of Gradient Boosting is effective.

SVM with RBF Kernel recorded accuracy of 95.2% and recall of 96.0% and the AUC of 0.9782. The RBF kernel was shown to be able to capture the non-linearities in the feature space, and kernel-based methods were still competitive for structured medical data.

The worst performance was KNN with an accuracy of 91.6% and recall of 88.0%. This confirms the previously mentioned issue with the instance-based learning in higher dimensional medical data, where the "curse of dimensionality" might make the distance measures less significant. The lower performance of KNN is also due to the sensitivity to feature scaling and the value of k.

Confusion Matrix Analysis:

In Table 4 and Figure 4, the confusion matrix of the Random Forest model on the test set of 83 samples is shown. Based on this matrix, the metrics of vital precision, recall and F1 are calculated.

Table 4. Confusion Matrix of the Random Forest model

	Predicted Healthy (0)	**Predicted Disease (1)**
Actual Healthy (0)	31 (TN)	2 (FP)
Actual Disease (1)	2 (FN)	48 (TP)

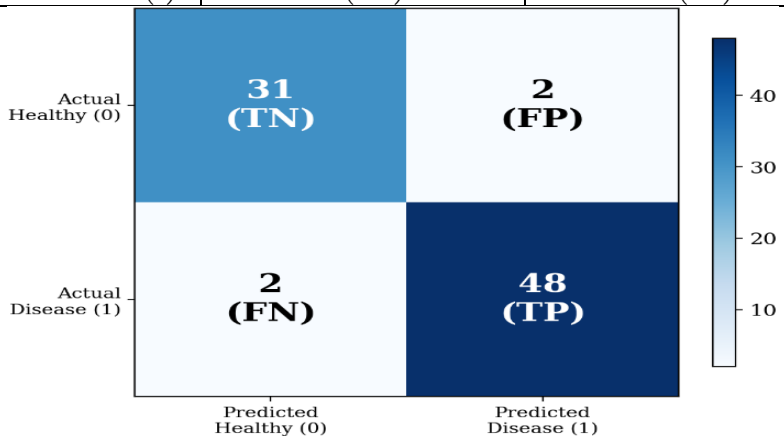

Figure 4. Random Forest Confusion Matrix Heatmap

ROC Curve Analysis:

ROCs were plotted for each of the six models to evaluate their performance in discrimination at different cut-off points. The best discriminative profile across thresholds is obtained by Random Forest, as shown in the ROC comparison in Figure 5. The Area Under the Curve (AUC) range was 0.9727 for KNN and 0.9842 for the Random Forest. The AUC values for all the models were significantly higher than 0.5 (random classification), indicating their predictive value. Random Forest's AUC of 0.9842 indicates near-perfect discrimination between healthy and diseased patients at all thresholds.

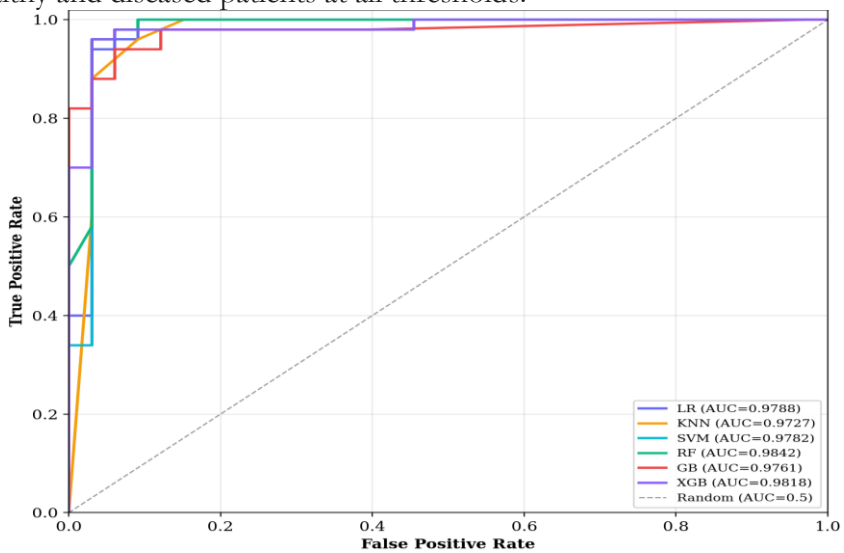

Figure 5. ROC Curves for All Six Models

Feature Importance Analysis:

Random Forest model has built-in feature importance that is based on the Gini impurity criterion. Table 5 shows the top five most predictive features.

Table 5. Random Forest Feature Importance

Feature	Importance Score
Slope	0.3746
Resting BP	0.1359
18Chest Pain	0.1243
Major Vessels	0.1178
Cholesterol	0.0565
Max Heart Rate	0.0436
Resting ECG	0.0435
Oldpeak	0.0331
Fasting BS	0.0300
Age	0.0263
Gender	0.0098
Exercise Angina	0.0046

Among the various features, the feature with the highest importance score was ST Slope (0.3746), which highlights the significance of the exercise stress response in cardiovascular assessments. The relative importance of each feature for the final decisions made by Random Forest is shown in Figure 6. This was followed by Resting Blood Pressure (0.1359) and Chest Pain Type (0.1243), both clinically known to be risk factors for CVD. Interestingly, demographic characteristics such as Age and Gender were not as important, indicating that the clinical characteristics have more predictive power than the demographic attributes in this cohort.

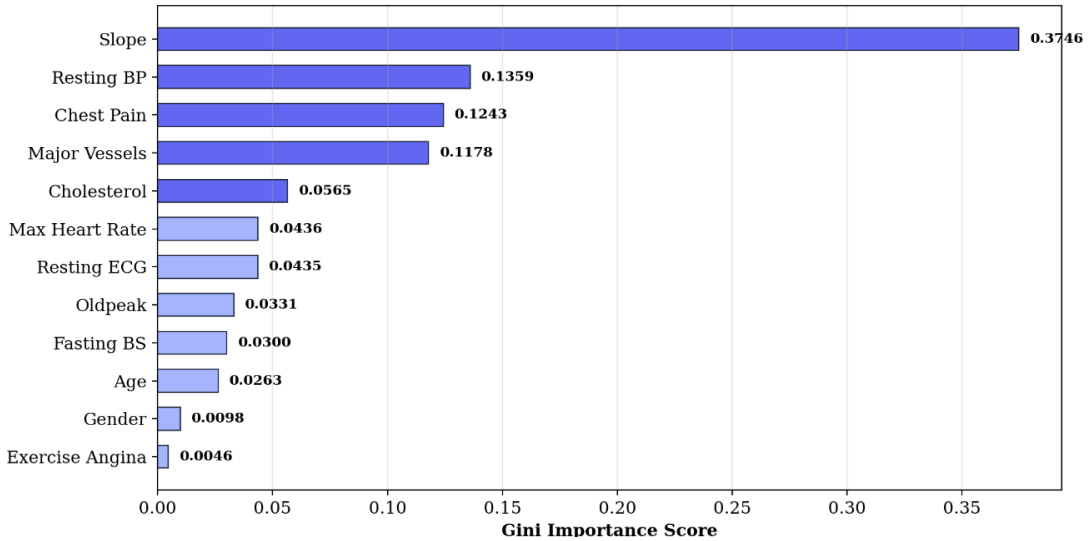

Figure 6. Random Forest Feature Importance

Discussion:

The findings of this research demonstrate the effectiveness of machine learning approaches for cardiovascular disease prediction, particularly when models are developed using population-specific clinical data. The comparative evaluation of six supervised learning algorithms provides a comprehensive assessment of different learning paradigms, including linear, instance-based, kernel-based, and ensemble approaches. Unlike models developed on generalized datasets, the proposed framework utilizes a locally collected Pakistani clinical

dataset, allowing the prediction models to capture region-specific cardiovascular characteristics and improve their practical relevance for local healthcare applications.

Among all evaluated models, Random Forest achieved the most balanced and reliable performance, obtaining an AUC of 0.9842 with an accuracy of 95.2%. This superior performance can be attributed to the ensemble nature of Random Forest, where multiple decision trees collectively improve prediction stability by reducing variance and minimizing overfitting. The model successfully identified 48 diseased patients while producing only two false-negative predictions in the test dataset, which is particularly important in cardiovascular screening applications where missed disease cases may lead to serious clinical consequences. The high recall value indicates that the model is capable of effectively identifying patients requiring further medical evaluation.

Although Random Forest demonstrated the highest overall performance, Logistic Regression achieved comparable accuracy, precision, and recall values. This observation indicates that a considerable portion of the relationship between clinical variables and cardiovascular outcomes follows a relatively structured pattern. The strong performance of Logistic Regression highlights its continued importance in medical prediction systems due to its simplicity, low computational requirements, and interpretability. In resource-limited healthcare settings, where transparency and rapid decision-making are critical, interpretable models may provide significant advantages despite the availability of more complex algorithms.

The performance differences among algorithms also provide important insights into the characteristics of cardiovascular clinical data. The lower accuracy of KNN compared with other approaches indicates that distance-based learning methods may be less suitable for medical datasets containing complex interactions among heterogeneous clinical variables. Patient characteristics such as blood pressure, cholesterol level, chest pain type, exercise response, and cardiac indicators do not always exhibit simple similarity patterns; therefore, algorithms capable of learning nonlinear relationships, such as ensemble models, are more effective for this type of prediction problem.

The comparative analysis also demonstrates the trade-off between predictive performance and model complexity. Advanced ensemble models such as Random Forest and XGBoost provide improved predictive capability by capturing complex interactions between features; however, their decision-making processes are less transparent compared with traditional statistical models. Conversely, simpler approaches such as Logistic Regression provide easier interpretation but may have limitations when dealing with nonlinear relationships. Therefore, the selection of a prediction model should consider not only accuracy but also computational requirements, interpretability, and the intended clinical application.

A major strength of this study is the development and evaluation of models using clinical data collected from Pakistani patients. Cardiovascular prediction models trained on external populations may not always accurately represent local disease patterns due to differences in genetic background, lifestyle factors, metabolic characteristics, and healthcare conditions. The use of locally representative data improves model calibration and increases the possibility of successful implementation within the Pakistani healthcare environment. However, the relatively limited dataset size remains a limitation, and future work should focus on expanding the dataset through multicenter collaboration to improve model generalization and reliability.

Implications for Pakistan's Healthcare System:

The high predictive performance of the developed models demonstrates their potential as supportive tools for cardiovascular risk screening in Pakistan. The proposed system could assist healthcare providers at primary care facilities, including Basic Health Units (BHUs) and rural healthcare centers, by enabling early identification of individuals at elevated

cardiovascular risk. Such an approach may improve patient prioritization and reduce delays in referral to specialized cardiac facilities.

From a practical implementation perspective, the Random Forest model offers advantages due to its low computational complexity and efficient prediction capability. The successful conversion of the trained model into a browser-based application demonstrates its feasibility for deployment on commonly available computing devices without requiring expensive infrastructure. This characteristic makes the proposed approach suitable for remote and resource-constrained healthcare environments.

Furthermore, the localized nature of the developed model provides an important advantage over generalized prediction systems. By learning patterns directly from Pakistani clinical data, the model has the potential to better capture cardiovascular risk characteristics specific to the local population. Nevertheless, before clinical adoption, further validation with larger and more diverse patient groups is required. Future improvements may include integration of additional biomarkers, lifestyle-related parameters, longitudinal patient information, and explainable artificial intelligence techniques to enhance model transparency and clinical acceptance.

Overall, this study highlights that machine learning-based cardiovascular prediction systems can provide accurate, efficient, and scalable solutions for early disease detection. The results emphasize the importance of combining advanced computational methods with locally relevant clinical data to develop reliable artificial intelligence tools for healthcare applications.

Computational Efficiency Consideration:

Although the present study primarily evaluates machine learning models based on predictive performance, computational efficiency represents an important factor for real-world deployment of AI-based healthcare systems. Training time, inference latency, memory requirements, and hardware dependency were not systematically evaluated in this work. However, the proposed framework uses lightweight supervised learning algorithms trained on structured clinical data, making it suitable for implementation in resource-constrained environments. Future studies will incorporate detailed computational complexity analysis to compare model efficiency and evaluate scalability for real-time cardiovascular risk prediction applications.

Conclusion:

This study developed, analyzed, and evaluated machine learning models specifically for the early detection of cardiovascular diseases in the context of Pakistan's demographics and healthcare system. A rigorous methodology involving preprocessing of data from 411 patients was applied, along with a comparative evaluation of six supervised learning algorithms, which led to the identification of Random Forest as the best diagnostic model. Furthermore, Random Forest demonstrated superior performance over single classifiers such as KNN and SVM, achieving 95.2% accuracy with an AUC of 0.9842. Its bagging mechanism provides an optimal balance of bias and variance reduction. All models achieved 96.0% recall, indicating that the models can meet high safety standards in medical screening with minimal false negatives. For future work, the incorporation of additional biomarkers and lifestyle factors is recommended.

References:

[1] A. Ogunpola, F. Saeed, S. Basurra, A. M. Albarrak, and S. N. Qasem, "Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases," *Diagnostics 2024, Vol. 14, Page 144*, vol. 14, no. 2, p. 144, Jan. 2024, doi: 10.3390/DIAGNOSTICS14020144.

[2] WHO, "Global Health Risks," 2009, [Online]. Available: http://www.who.int/healthinfo/global_burden_disease/GlobalHealthRisks_report_full.pdf

[3] A. K. Siddiqi, E. Mubashir, A. A. A. Cheema, M. Noshab, and M. Naeem, "Pakistan's

rising cardiovascular disease burden associated with modifiable risk factors: an urgent appeal for evidence-driven action," *Ann. Med. Surg.*, vol. 88, no. 1, p. 59, Jan. 2025, doi: 10.1097/MS9.0000000000004565.

[4] Z. Samad and B. Hanif, "Cardiovascular Diseases in Pakistan: Imagining a Postpandemic, Postconflict Future," *Circulation*, vol. 147, no. 17, pp. 1261–1263, Apr. 2023, doi: 10.1161/CIRCULATIONAHA.122.059122.

[5] M. S. Awan *et al.*, "Factors Leading to Late Presentation of ST-Segment Elevation Myocardial Infarction for Thrombolytic Therapy," *J. Saidu Med. Coll.*, vol. 14, no. 3, pp. 249–253, Jul. 2024, doi: 10.52206/JSMC.2024.14.3.854.

[6] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[7] "(PDF) Heart Disease Prediction using Machine Learning Algorithms." Accessed: Mar. 09, 2022. [Online]. Available: https://www.researchgate.net/publication/348408218_Heart_Disease_Prediction_using_Machine_Learning_Algorithms

[8] A. M. Qadri, A. Raza, K. Munir, and M. S. Almutairi, "Effective Feature Engineering Technique for Heart Disease Prediction With Machine Learning," *IEEE Access*, vol. 11, pp. 56214–56224, 2023, doi: 10.1109/ACCESS.2023.3281484.

[9] N. Rajesh, M. T, S. Hafeez, and H. Krishna, "Prediction of Heart Disease Using Machine Learning Algorithms," *Int. J. Eng. Technol.*, vol. 7, no. 2.32, pp. 363–366, May 2018, doi: 10.14419/IJET.V7I2.32.15714.

[10] V. K. Sudha and D. Kumar, "Hybrid CNN and LSTM Network For Heart Disease Prediction," *SN Comput. Sci. 2023 42*, vol. 4, no. 2, pp. 172-, Jan. 2023, doi: 10.1007/S42979-022-01598-9.

[11] "Machine Learning-Based Cardiovascular Disease Detection Using Optimal Feature Selection", [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10416957

[12] W. Alsabhan and A. Alfadhly, "Effectiveness of machine learning models in diagnosis of heart disease: a comparative study," *Sci. Reports 2025 151*, vol. 15, no. 1, pp. 24568-, Jul. 2025, doi: 10.1038/s41598-025-09423-y.

[13] T. H. Rim *et al.*, "Deep-learning-based cardiovascular risk stratification using coronary artery calcium scores predicted from retinal photographs," *Lancet Digit. Heal.*, vol. 3, no. 5, pp. e306–e316, May 2021, doi: 10.1016/S2589-7500(21)00043-1.

[14] A. Javeed, S. U. Khan, L. Ali, S. Ali, Y. Imrana, and A. Rahman, "Machine Learning-Based Automated Diagnostic Systems Developed for Heart Failure Prediction Using Different Types of Data Modalities: A Systematic Review and Future Directions," *Comput. Math. Methods Med.*, vol. 2022, p. 9288452, 2022, doi: 10.1155/2022/9288452.

[15] "What is Machine Learning? | IBM." Accessed: Jul. 25, 2026. [Online]. Available: https://www.ibm.com/think/topics/machine-learning

[16] R. Bhardwaj, A. R. Nambiar, and D. Dutta, "A Study of Machine Learning in Healthcare," *Proc. - Int. Comput. Softw. Appl. Conf.*, vol. 2, pp. 236–241, Sep. 2017, doi: 10.1109/COMPSAC.2017.164.

[17] M. Abdar, W. Książek, U. R. Acharya, R. S. Tan, V. Makarenkov, and P. Pławiak, "A new machine learning technique for an accurate diagnosis of coronary artery disease," *Comput. Methods Programs Biomed.*, vol. 179, Oct. 2019, doi: 10.1016/j.cmpb.2019.104992.

[18] M. Deng, C. Wang, M. Tang, and T. Zheng, "Extracting cardiac dynamics within ECG signal for human identification and cardiovascular diseases classification," *Neural Networks*, vol. 100, pp. 70–83, Apr. 2018, doi: 10.1016/j.neunet.2018.01.009.

[19] M. Liu and Y. Kim, "Classification of Heart Diseases Based on ECG Signals Using

Long Short-Term Memory," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, vol. 2018-July, pp. 2707–2710, Oct. 2018, doi: 10.1109/EMBC.2018.8512761.

[20] N. Pratheek, D. H. Balaji Yashas, Y. R. Bhanu Prakash, A. M. Bharadwaj, A. Kodapalli, and T. Rao, "Cardiovascular Disease Prediction with Machine Learning Algorithms and Interpretation using Explainable AI methods: LIME & SHAP," *2024 3rd Int. Conf. Adv. Technol. ICONAT 2024*, 2024, doi: 10.1109/ICONAT61936.2024.10774972.

[21] N. Nagasoudhamani, A. Revathi, D. A. K. Rao, G. Dianakamal, T. R. Tulasi, and S. Rajasekhar Reddy, "Machine Learning Ensemble Model for Heart Disease Prediction," *Proc. Int. Conf. Intell. Comput. Control Syst. ICICCS 2025*, pp. 1403–1408, 2025, doi: 10.1109/ICICCS65191.2025.10985246.

[22] A. Ogunpola, F. Saeed, S. Basurra, A. M. Albarrak, and S. N. Qasem, "Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases," *Diagnostics 2024, Vol. 14, Page 144*, vol. 14, no. 2, p. 144, Jan. 2024, doi: 10.3390/DIAGNOSTICS14020144.

[23] M. A. Bouqentar *et al.*, "Early heart disease prediction using feature engineering and machine learning algorithms," *Heliyon*, vol. 10, no. 19, p. e38731, Oct. 2024, doi: 10.1016/J.HELIYON.2024.E38731.

[24] P. Kaushik, Z. M. Khan, M. M. Khan, S. K. Saranya, S. Shamsi, and G. Parasa, "Heart Disease Prediction: Leveraging Machine Learning and Clinical Data," *2025 IEEE Int. Conf. Interdiscip. Approaches Technol. Manag. Soc. Innov. IATMSI 2025*, 2025, doi: 10.1109/IATMSI64286.2025.10984659.

[25] A. Sahu, H. Gm, M. K. Gourisaria, S. S. Rautaray, and M. Pandey, "Cardiovascular risk assessment using data mining inferencing and feature engineering techniques," *Int. J. Inf. Technol.*, vol. 13, no. 5, pp. 2011–2023, Oct. 2021, doi: 10.1007/S41870-021-00650-W/METRICS.