Vol. 6, No. 2 | July – December 2023

Efficient Machine Learning Techniques to Classifying Cardiovascular Disease and Improve Prediction Analysis

Reehan Ali Shah [1,*], Samavia Riaz [1], Dileep Kumar [2], Muhammad Bux Alvi [1], Syeda Rida Bibi[3], Zulfiqar Zardari [4]

Abstract:

Cardiovascular Disease (CVD) account for a large portion of the global health burden and are one of the main causes of decease worldwide. In the classification and forecasting of CVDs, Machine Learning (ML) techniques have demonstrated encouraging outcomes. In this research report, a comparative analysis of classification and prediction models for CVD is presented, including both linear and ensemble ML approaches. The paper compares ensemble models like Catboost, Histogram Gradient Boosting Machine (HGBM), and Extra Trees against linear models like Gaussian Nave Bayes, SVM, and KNN. The objective is to identify the most effective CVD prediction model by assessing its performance through accuracy, precision, sensitivity and F1 score as key evaluation metrics. Moreover, results show that ensemble models outperform linear models using advanced techniques such as boosting and histogram-based algorithms. The results underscore the critical role ensemble models play in accurately diagnosing and predicting cardiovascular disease and provide important new information to researchers and healthcare providers. Using these models has the potential to significantly improve patient outcomes and health management by enabling early detection and intervention.

Keywords: *Cardio Vascular Diseases (CVD), Ensemble Models, classification, Machine learning (ML), Linear Models.*

1. Introduction

Worldwide, heart disease stands as the primary cause of mortality, with a staggering 17.9 million reported deaths in 2019 [1], accounting for nearly 15% of all natural deaths. Diseases of the heart and blood vessels encompass heart attack, stroke, and other vascular disorders. Lifestyle-related factors, such as smoking, obesity, high cholesterol, and hypertension, contribute to an increased risk of heart disease. Nevertheless, it is crucial to take into account additional non-lifestyle risk variables such as age, family history, and elevated fibrinogen levels, alongside lifestyle risk factors. Furthermore, heart disease could appear even in the absence of any of the aforementioned risk factors or evident symptoms. Consequently, heart disease ranked among the most widespread conditions globally, significantly contributing to the

[1] Department of Computer Systems Engineering, Faculty of Engineering, The Islamia University of Bahawalpur, Pakistan.

[2] Department of Electronics Engineering, Faculty of Engineering, The Islamia University of Bahawalpur, Pakistan.

[3] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Italy.

[4] Department of Information & communication techniques, Begum Nusrat Bhutto Women University, Sukkar, Pakistan.

mortality rate and presenting itself as one of the most formidable ailments to manage.

The electrocardiogram (ECG) is a widely utilized and non-invasive diagnostic technique for cardiovascular disease. It provides a visual representation of the heart's electrical activity. Despite its quick and easy execution, an electrocardiogram (ECG) has various limitations as a diagnostic tool for anticipating the onset of future cardiovascular illness. The manual prediction of the likelihood of developing heart disease is challenging due to several risk factors [2].

The vast majority of fatalities occur in low- and middle-income nations [3]. Early detection and reliable CVD prognosis are essential for timely and efficient CVD prevention and treatment. Consequently, the healthcare industry must establish and enhance strategies to mitigate the socioeconomic impact of chronic diseases. Within the healthcare sector, there exists a considerable volume of data pertaining to heart disease, which can be scrutinized to facilitate well-informed decision-making. Machine learning (ML) methods have shown great potential in classifying and predicting cardiovascular disease based on multiple clinical and demographic features. Machine learning relies on its ability to handle large datasets, has high processing speed, and make early-stage predictions [4].

While machine learning has the potential to increase the accuracy of heart attack prediction, further research is needed to optimize the diagnosis and simplify the algorithms used in ML. Various machine learning methodologies, such as classification, prediction, and pattern recognition, have the capability to predict cardiovascular disorders [5].

There is an abundance of research data and hospital patient records available. The objective of this research is to create a hybrid dataset that can assist in the design and development of an optimal CVD risk prediction model. The Cleveland Heart Disease dataset, obtained from the University of California, Irvine (UCI) dataset repository, is widely used for predicting heart disease (HD). Nevertheless, these datasets are limited by a maximum of 303 instances that have missing values in their features, which greatly affects the precision of prediction models. To address this limitation and improve the model's performance, we expanded the scope of our research by integrating a locally collected dataset from the Ch. Pervaiz Elahi Institute of Cardiology Multan. This local dataset, consisting of 199 instances and having identical features as the UCI dataset, was combined with the UCI dataset to create a comprehensive hybrid dataset.

Specifically, this research objective is to evaluate and compare the performance of different Linear and Ensemble Classification Machine Learning algorithms utilizing this hybrid dataset. Our findings will contribute to the growing body of literature on ML approaches for CVD diagnosis and prediction, offering valuable insights into model performance.

2. Literature Review

Cardiovascular disease continues to be a major global cause of death, making a significant impact to mortality rates in many populations. Early diagnosis and preventive steps are essential for saving lives and reducing the strain on healthcare systems. Machine learning (ML) has become a promising tool in the diagnosis and prediction of cardiovascular disease (CVD), providing new opportunities in comparative classification and prediction. This literature review examines the latest developments in machine learning (ML) for CVD prediction, emphasizing the significance of accurate

classification as a crucial tool for decision-making in medical science. Our research aims to conduct a thorough analysis of different methodologies employed in previous studies, including their combined application, in order to determine the most effective approaches for CVD prediction.

In this study, sequential feature selection strategy was adopted to identify crucial features linked to mortality events in patients undergoing treatment for heart disease. Multiple Machine learning techniques, including LDA, KNN, RF, SVM, DT, and GBC were utilized. The outcomes of the SFS algorithm were validated using validation metrics such as the confusion matrix, receiver operating characteristic curve, precision, recall rate, and F1-score. The experimental results demonstrate that the sequential feature selection strategy attains an accuracy rate of 86.67% for the random forest classifier [6].

This study focused to optimize the prediction of cardiac disease using machine learning algorithms. This was achieved by employing a limited number of features and running a few tests. The researchers utilized 14 essential features from the Cleveland dataset and performed a series of performance tests on four Machine Learning Algorithms (MLAs). Their findings demonstrated that the K-Nearest Neighbor (KNN) algorithm attained the highest level of accuracy in predicting heart disease [7].

The hybrid dimensionality reduction method, CHI-PCA, was proposed [8], which combines Chi-square and principal component analysis techniques for the prediction of heart disease. This approach was thoroughly evaluated using three separate datasets obtained from the UCI Machine Learning Repository: the Hungarian, Cleveland, and Hungarian-Cleveland datasets. The evaluation of the suggested method includes the use of five classifiers: random forests, gradient-boosted tree, decision tree, multilayer perceptron, and logistic regression. Using CHI-PCA in combination with random forests (RF) resulted in remarkable accuracies. Specifically, the Cleveland dataset achieved a recording rate of 98.7%, the Hungarian dataset achieved 99.0%, and the Cleveland-Hungarian (CH) dataset achieved 99.4%. The results highlight the effectiveness of combining CHI-PCA with RF in obtaining exceptionally accurate predictions in several datasets related to heart disease.

A recent investigation examined the utilization of several feature selection methods and data mining strategies for the purpose of predicting heart disease. The study employed the Correlation-based Feature Selection (CFS) approach in conjunction with the Naive Bayes classifier, resulting in an impressive classification accuracy of 84.81%. Furthermore, the utilization of the CHI square feature selection technique in combination with the RBF network resulted in a commendable accuracy rate of 81.1%. The results highlighted the substantial improvements that may be achieved by incorporating feature selection approaches alongside the original classification algorithms. This emphasizes the importance of these methodologies in enhancing predictive models specifically for heart disease [9].

A study [10] suggested a model that combines ensemble methods (boosting and bagging) with feature extraction techniques (LDA and PCA) to predict cardiac disease. The study performed a comparative analysis between ensemble approaches (bagging and boosting) and five classifiers (SVM, KNN, RF, NB, and DT) utilizing specific features from the Cleveland heart disease dataset. The results showed that the bagging ensemble learning method, combined with DT and PCA

SN	Feature Name	Description
1	age	Age (years)
2	sex	Sex (Male=1, Female = 0)
3	cp	Chest pain (4 types)
4	trestbps	Blood Pressure at rest (in mm Hg)
5	chol	Cholesterol concentration in mg/dl
6	fbs	Patient's Fasting Blood Sugar (> 120 mg/dl).
7	restecg	ECG at rest.
8	thalach	The patient's heart rate (maximum).
9	exang	Exertion Induced Angina (1 = Present; 0 = Absent)
10	oldpeak	Depression from exercise ("ST" is ECG plot location)
11	slope	The slope of the "ST" segment during maximal exertion.
12	ca	Major vessels (4 values).
13	thal	Thalassemia (1-3).
14	Target (Class)	Heart disease (0=Absent, 1=Present).

feature extraction, achieved the highest level of performance compared to the other methods that were assessed.

3. MATERIALS AND METHODOLOGY

3.1. Data Collection

In this research article, two datasets employed one as a public dataset and second is a local dataset. Both datasets contain the same features and attributes. The public dataset was collected from UCI machine learning repository (Dua and Karra Taniskidou, 2017). There are four databases: Cleveland, Hungary, Switzerland, and the VA Long Beach database. The Cleveland database was chosen for this study because machine learning researchers frequently use it and find its information to be most comprehensive. There are 303 instances in the dataset. Only a subset of the 76 attributes in the Cleveland dataset are covered by the dataset that is available in the repository. The data source of the Cleveland dataset is Cleveland Clinic Foundation [11], whereas the local dataset was collected from Ch. Pervaiz Elahi Institute of Cardiology Multan. Table 1 provides a summary of the thirteen features common to heart failure patients. In addition, the fourteenth column, referred to as the target column, signifies whether or not the patients have heart disease.

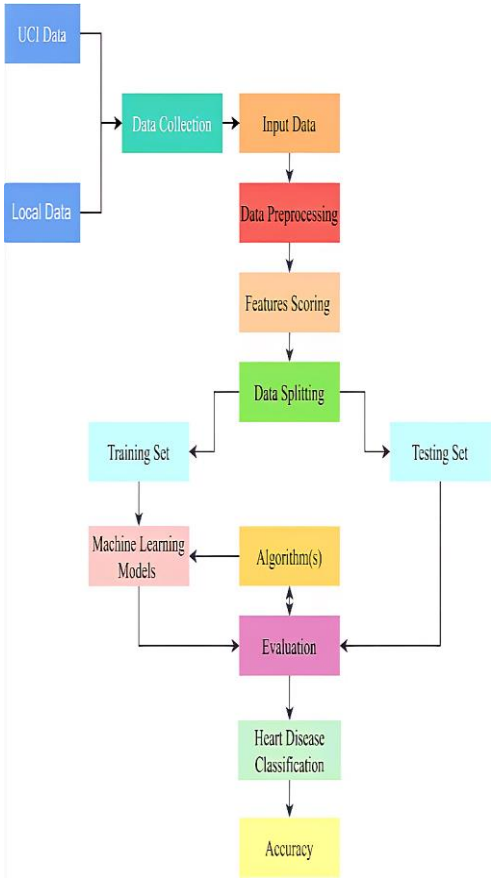
The dataset consists of 502 patients, of which 47.21% had heart disease compared to 52.79% did not have it, among patients with heart disease, 177 (35.26%) of the patients are men, while 60 (11.95%) are women. Table 1 exemplifies the essentials of each attribute.

Table 1: Features of CV Datasets

In this study, the results are visually presented in the graphic represented below. This section provides an insight into the methodology employed during our research process. When working with a dataset, our approach encompasses several essential phases: feature engineering, model creation, and performance assessment. Figure 1 offers a

comprehensive flowchart that illustrates the sequential progression of our research, allowing readers to grasp the systematic journey from data preparation to the final assessment of model performance.

Figure 1. General architecture for CVD prediction System.



Data has been trained and tested using several machine learning models. Cross-validation was subsequently used to carefully evaluate the models' performance in order to validate the results. The best accurate classification models were chosen when the validation process was finished, maximizing the prediction power for the diagnosis of heart disease.

4. Data Pre-processing

The phase of pre-processing of data incorporates multiple steps to ensure that the data are clean and available for analysis. Data pre-processing is a method employed to transform raw data into a refined dataset. In other words, when data is obtained from many sources, it is acquired in its original, unprocessed form, which is not feasible for analysis. To optimize the performance of the machine learning model and obtain better results, the data must be formatted correctly.

5. Cleaning

Initially, we examine the dataset to determine whether it contains any missing values or not. There are several methods for handling missing values, like ignoring them entirely, substituting them with a numerical value, replacing them with the most frequently occurring value for that attribute, or substituting them with the mean value for that attribute. This paper addresses the issue of missing values by replacing them with the mean value of that attribute.

The local dataset does not contain any missing values, however, in the Cleveland dataset, there are 4 missing values in the "ca" feature and 2 missing values in the "thal" attribute.

6. Outlier Detection and Handling

Interquartile range (IQR) was used to identify and eliminate outliers.

To find outliers, the dataset was split into three quartiles, Q3, Q2, and Q1, where Q1 and Q3 represent the lower and higher data limits,

respectively. The interquartile range (IQR) was computed as,

$$IQR = Q3 - Q1 \quad (1)$$

Using the following equations, the values of the lower boundary, B_l, and the upper boundary, B_u, were then calculated:

$$B_l = Q1 - 1.5 * IQR \quad (2)$$

$$B_u = Q3 + 1.5 * 1QR \quad (3)$$

Here, data points falling below the lower boundary (B_l) or exceeding the upper boundary (B_u) are classified as outliers [12].

7. Feature Engineering

Some of the categorical values mentioned have only a few unique values. Categorical encoding is employed to prevent Machine Learning algorithms to not overfit to unique values. Transforming these values into binary values enables Machine Learning algorithms to process the data in a less biased manner without losing all of the information.

8. Scale Data

Data scaling is crucial to prevent Machine Learning algorithms from overfitting to irrelevant features. The Min Max Scaler function is utilized to scale the values of each feature, ensuring that they fall within the range of 0 to 1, based on the minimum and maximum values. This preserves the information from potential loss and enables the Machine Learning algorithms to effectively train the data.

9. Class Imbalance Handling (SMOTE)

In addressing the inherent class imbalance within the dataset, Synthetic Minority Over-sampling Technique (SMOTE) to rectify the intrinsic class imbalance present in the dataset. SMOTE was essential in maintaining a balanced distribution of classes, with 47.21% of the 502 patients having a heart disease diagnosis and the remainder 52.79% not. It was made sure that both classes were more accurately represented in the dataset by creating synthetic instances for the minority class, which is individuals with heart disease. This helped to reduce the possibility of biases in predictive modeling.

10. Splitting Data

The dataset was split into an 80% training set and a 20% testing set, allowing the model to be trained on one portion while reserving the other to be used for assessing the model's ability to generalize to new data.

10.1. Supervised Machine Learning Models

This section focuses on the analysis of ML methods for the purpose of their implementation in the aforementioned dataset. The article analyses the implementation of multiple ML models in the finding of heart illness. The models can be categorized into 2 distinct classes: linear models and ensemble models (boosting and bagging). Labelled data is used in computational learning to train the model and generate predictions. Ensemble models, on the other hand, combine multiple models to improve performance and accuracy. The models can be further split into two categories: bagging and boosting. Bagging entails training models independently on distinct subsets of data and combining their predictions. On the other hand, boosting is a sequential strategy where models are trained iteratively, assigning greater importance to previously misclassified data.

The training data is utilized to train various linear and ensemble models A sensitivity analysis is conducted by iterating the algorithms with different hyperparameters

using GridSearchCV to optimize each model [13]. The best model is the one that exhibits the maximum accuracy while avoiding overfitting, as determined by evaluating both the training and testing data. This decision is made by examining the outcomes of both the training and testing data. The evaluation of these models is conducted using k-fold Cross-Validation with k-fold = 5, employing GridSearchCV to iterate on various hyperparameters of the algorithms. The accurateness, positive predictive value, sensitivity and F1 count for each procedure used in this article were observed and recorded. The next section provides a brief description of these algorithms.

11. K-Nearest Neighbor

The K-Nearest Neighbor (KNN) algorithm is a non-parametric method employed in supervised classification tasks. The method depends on the proximity of K neighboring instances, which is calculated using the Euclidean distance metric [14].

$$d(x, x') = \sqrt{(x_1 - x'_1)^2 + \cdots + (x_n - x'_n)^2} \quad (4)$$

Finally, the input x is allocated to the highest probability class.

$$P(y = j|X = x) = \frac{1}{K} \Sigma_{iEA} I\left(y^{(i)} = j\right) \quad (5)$$

12. Support Vector Machine

Support Vector Machines (SVMs) have exhibited outstanding performance in disease prediction due to their robustness. The method uses hyperplanes to differentiate disease classes while also maximizing the distance between them. Mathematically, SVM can be expressed as follows:

$$If\ Y_i = +1; wxi + b \geq 1 \quad (6)$$

$$If\ Y_i = -1; wxi + b\ \leq 1 \quad (7)$$

$$Fall\ i;\ Y_i(wi + b) \geq 1 \quad (8)$$

In the equation, "x" is a vector point and "w" is a weight and vector. Equation (6) must always create values greater than zero and Equation (7) must always produce values less than zero for efficient data separation. SVM chooses the hyperplane that optimizes data point distance [15].

13. Gaussian Naïve Bayes

Gaussian Naive Bayes, a Bayesian algorithm, assumes feature independence. Gaussian Naive Bayes is especially beneficial when the features are assumed to follow a normal distribution, making it a helpful tool for a variety of classification tasks. For continuous feature values in each class, this method relies on the assumption of a Gaussian (normal) distribution. To create accurate predictions, it requires estimating the mean and variance. Mathematically, it can be expressed as follows:

$$P(X|Y = c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{\frac{-(x-\mu_c)^2}{2\sigma_c^2}} \quad (9)$$

In the aforementioned equations, σ and μ represent the variance and mean of the continuous variable X, respectively, which are calculated for a certain class c of Y [16].

14. Catboost

CatBoost, an advanced Boosting algorithm, provides exceptional efficiency and robust generalization capabilities. It exhibits exceptional proficiency in managing categorical data, minimizing overfitting, and achieving model universality. This method utilizes Prediction Values Changes (PVC) or

Loss Function Change (LFC) to rank the features of the model. The basic methodology in CatBoost-based models is to utilize PVC, which measures the impact of varying feature values on predictions. PVC and LFC combine to rank features in the CatBoost machine learning model. LFC is often used to rank a model inside a group.

$$F = \{f_1, f_2, f_3, \dots\dots, f_n\} \quad (10)$$

$$P_i = \beta_i F_j \quad (11)$$

In Eq (10), the symbol "F" represents the input feature set, "β" denotes numeric factors, and "P" represents the prediction.

In Eq. (11), "P" is the prediction with substituted numeric factor, "β_i" is the numeric factor, and "F_j" is the selected feature.

$$P_{i+1} = \beta_{i+1}\, F_j \quad (12)$$

$$P_{i=0} \neq P_i \neq P_{i+1} \quad (13)$$

In Eq. (12), "P_{i+1}" signifies prophecy with a modified numerical attribute and "β_{i+1}" is the adjusted numeric factor.

Eq. (13) indicates that a modification in the numerical attribute affecting the prophecy value implies the importance of that specific feature [17].

15. Histogram Gradient Boosting Machine

Histogram Gradient Boosting Machine (HGBM) is a modified version of gradient boosting that offers improved prediction accuracy, efficient data processing, robustness, and feature ranking. It efficiently manages extensive healthcare datasets, reduces loss functions, and creates decision tree ensembles by gradient descent. With its exceptional performance, this tool is the most effective for accurately identifying individuals at risk of developing heart disease, improving model interpretability, and developing interventions that align with our study goals [18].

16. Extra Trees

The Extra Tree classifier, sometimes referred to as the Extremely Randomized Trees classifier, enhances prediction accuracy by reducing model variance through the use of ensemble learning. The system operates based on the principles of decision trees and random forests, incorporating concepts such as entropy and information gain. It combines many models and averages their predictions.

$$E = \sum_{i=1}^{c} -pilog_2 pi \quad (14)$$

To help with the computation of information gain or impurity reduction, in this case, "E" stands for entropy, "c" for the number of classes in the dataset, and "pi" for the number of rows associated with each class in the dataset.

$$\gamma(S, A) = E(S) - \sum_{v\varepsilon(A)} \frac{S_v}{S} E(S_v) \quad (15)$$

γ denotes Mutual Information, E for entropy, S represents condition's prospect, A is attribute, v denotes individual values in the feature, and S_v is the probability of a specific value occurring. It's three times faster than the Random Tree Classifier, making it a potent tool for heart disease prediction [19].

1. Performance Evaluation Metrics

Performance metrics such as accuracy, positive precision, sensitivity, F1 score, ROC and Precision-Recall curve used to assess the classification performance of machine

learning models for cardiovascular diseases: **Accurateness**: Accuracy is determined by the percentage of correctly identified cases [20,21,22]. while **Positive predictive value** is the percentage of correct positive predictions [23,24]. **Sensitivity** (recall) is the percentage of positive predictions that are accurate [23]. The **F1 score** is the consonant mean for precision and sensitivity. These measures are frequently used in studies of medical diagnosis to evaluate the performance of machine learning algorithms [23,24].

Each method's accuracy, precision, sensitivity, and F1 score were determined using a confusion matrix. Following are the formulas used to calculate each parameter [20,22].

$$Accuracy = \frac{(TP+TN)}{TP+TN+FP+FN} \quad (16)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (17)$$

$$Precision = \frac{TP}{TP+FP} \quad (18)$$

$$Recall = \frac{TP}{TP+FN} \quad (19)$$

$$F1\ Score = \frac{2*Precision*Recall}{Precision+Recall} \quad (20)$$

Receiver Operation Characteristic (ROC):

This graph serves as a visual representation of the classifier's performance. It displays both the accurately classified instances and the inaccurately classified instances. The AUC represents the level or extent of separability. It quantifies the model's ability to differentiate between different classes. The Area Under the Receiver Operating Characteristic (ROC) curve represents the balance between the true positive rate (TPR) and false positive rate (FPR) of a classification model. A higher AUC indicates a greater level of accuracy in distinguishing between patients with and without the disease [25].

$$TPR = \frac{[\text{Number of True Positives}]}{[\text{Number of True Positives + False Negatives}]} \quad (21)$$

$$\text{FPR} = \frac{[\text{Number of False Positives}]}{[\text{Number of False Positives + True Negatives}]} \quad (22)$$

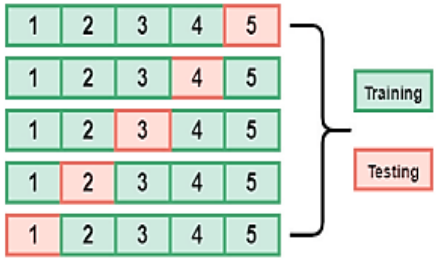
Precision Recall curve

The precision-recall curve is a graphical representation employed to assess the effectiveness of a classifier. This demonstrates the trade-off between precision and recall at varying classification thresholds.

In this study, the model was trained and evaluated using k-fold cross validation to ensure robust and reliable performance assessment. Using this technique, the data set is divided into several categories. K represents the classification factor, also known as the fold. “Cross Validation” is a method for simultaneously evaluating a ML model. The process of k-fold cross validation demands separating the data array into k distinct groups, followed by training the model using (k-1) of these sets, and having the other sets participate in testing and evaluating the skilled model. The strategy involves training the typical k times and evaluating it using a different fold each time. This indicates that in K-fold resampling authentication, a model is trained and tested on each fold. A 5-fold cross-validation method is shown in Figure 2. The diagram demonstrates the five folds within the dataset, four of which take part in model drill and remaining evaluates the training for each iteration. 5-fold cross-validation is utilized in

our investigation. This step is utilized to avoid over-optimization in predictive models.

Figure 2. K-Fold Cross Validation graphical representation



17. Results and Discussion

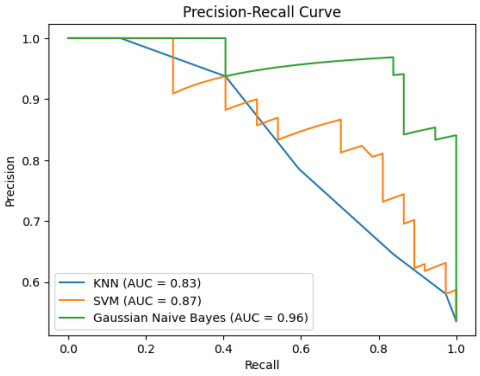
17.1. Results of Machine Learning Models

In this comparative analysis, Python 3.0 was used as the programming language to build the analytical model within the Jupyter (Anaconda) Notebook environment. A hybrid dataset on cardiac disease was evaluated for this research. Various classification techniques were applied, and outliers were identified and removed. With 5-fold cross-validation, these classification techniques were used. To find the best performing method for forecasting the occurrence of CVD, the cross-validation performance metrics were examined. Figure 1 illustrates the entire procedure.

The outputs and accuracies generated by the classifiers are assessed and presented in the following results. All of the models performed well after fine tuning of their hyperparameters, however, the model with the highest accuracy is regarded as the best one. Table 2 presents the performance characteristics of the supervised linear classification algorithms GNB, SVM, and KNN. These metrics include precision, accuracy, sensitivity, and F1 score.

In this analysis, Gaussian Naive Bayes achieved the highest accuracy rate of 85.5%. This was attributed to the utilization of feature independence and adherence to Gaussian distribution assumptions, excelling when these conditions align within the dataset. Following closely, the SVM model attained an accuracy of 75.3% by effectively detecting distinct boundaries between classes in high-dimensional domains. However, when confronted with less separable data, the accuracy of the model may decrease marginally in comparison to models that are better equipped to handle such complexities. The K-nearest neighbors (KNN) technique, which measures instance similarity, has the lowest accuracy of 69.5%. Higher dimensionality or noise in the dataset impacted its performance in these instances.

Table 2. Classification results of Supervised Linear Classification Algorithms.



Model Name	Accuracy	Precision	Sensitivity	F1 score
Gaussian Naive Bayes	0.855	0.864	0.864	0.864
S V M	0.753	0.727	0.864	0.790
K N N	0.695	0.785	0.594	0.676

Figure 3 shows the ROC curve, which includes true and false positive rates. It displays area under ROC (AUROC). The Gaussian Naive Bayes algorithm had the greatest AUC value of 0.96, demonstrating its superior class discrimination. Following that, the SVM achieved an AUC of 0.84. K-nearest neighbors (KNN) had the lowest AUC value of 0.77, showing lower discrimination than the other two models. Gaussian Naive Bayes performs exceptionally well in class differentiation, making it the most proficient model in this particular circumstance.

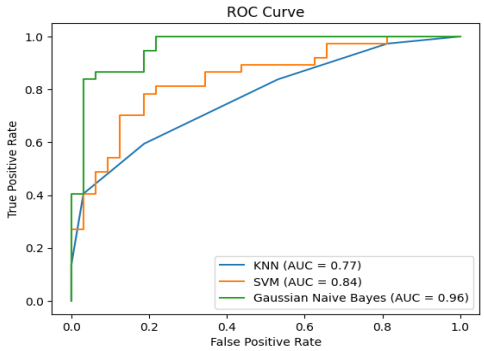


Figure 3. ROC Curve for Supervised linear Learning Algorithms

Figure 4 depicts the Precision-Recall curve for Gaussian Naive Bayes, SVM, and KNN at different recall levels is shown in Figure 4. The Gaussian Naive Bayes algorithm had the highest precision (0.96), reliably finding positive occurrences with few false positives. K-Nearest Neighbors (KNN) has 0.83 precision, while SVM had 0.87. The models' trade-off between true positives and false positives across thresholds is shown by this ROC curve. Gaussian Naive Bayes stands out for precision-centric tasks based on its superior precision in this analysis.

Figure 4. Precision-Recall Curve for Supervised linear Learning Algorithms

Table 3 presents the performance metrics of the Ensemble classification algorithms Catboost, HGBM, and Extra Trees. Accuracy, precision, sensitivity, and the F1 score are among these metrics. CatBoost exhibits the highest accuracy of 91.6%, surpassing Extra Trees at 91.1% and HGBM (Histogram Gradient Boosting Machine) at 90.9%. The improved accuracy of CatBoost can be attributed to its effectiveness in handling categorical features and its robustness against overfitting. Extra Trees, recognized for its ensemble learning and variance reduction approaches, closely followed with impressive accuracy. HGBM's iterative boosting strategy trailed but still exhibited strong predictive capability. CatBoost is the most accurate model in this analysis since it handles diverse data types.

Table 3. Classification results of Ensemble Classification Algorithms.

Model Name	Accuracy	Precision	Sensitivity	F1-score
Catboost	0.916	0.75	1.0	0.857
Extra Trees	0.911	0.75	1.0	0.857
HGBM	0.909	0.75	1.0	0.857

Figure 5 shows the true and false alarm probabilities of ROC curve. AUC scores were calculated using ROC curve analysis of Catboost, HGBM, and Extra Trees. Both Catboost and HGBM performed well with discriminative abilities of 0.92. Extra Trees closely trailed after with a slightly lower AUC of 0.90, indicating a somewhat diminished ability to distinguish between classes compared to Catboost and HGBM.

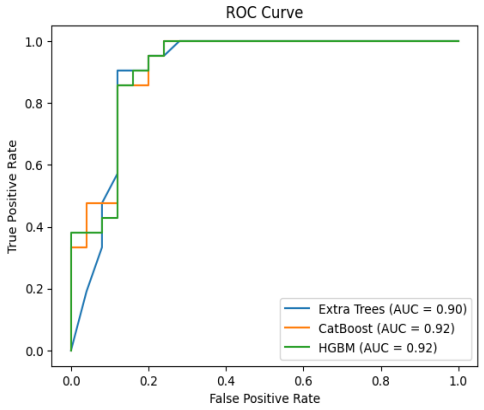


Figure 5. ROC Curve for Ensemble Classification Algorithms.

Figure 6 shows the AUPRC for several classification techniques. AUPRC values determine the area. It offers an illustration of the AUPRC values. With the maximum Area Under the Curve (AUC) score of 0.89, Catboost showed outstanding precision at different recall levels. HGBM's AUC of 0.88 indicated strong positive instance detection accuracy. However, Extra Trees had a lower AUC (Area Under the Curve) of 0.83, indicating lower precision. The results indicate that Catboost and HGBM recall positive cases more precisely than Extra Trees.

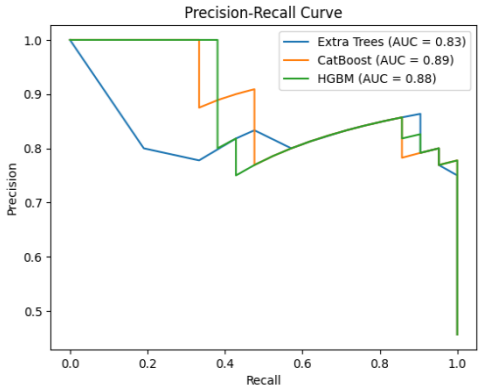


Figure 6. Precision-Recall Curve for Ensemble Classification Algorithms.

Table 4: Comparison of Different Classification Technique Results

Reference No.	Classification Technique	Tool	Accuracy
[26]	NB	WEKA	85.03
[27]	NB	Python 2.7	82
[28]	SVM	Not mentioned	61.86
[29]	SVM	Not mentioned	79.12
[29]	KNN	Not mentioned	70.33
[29]	Catboost	Not mentioned	81.32

[30]	Extra Trees	Python	86.7
[31]	KNN	Not mentioned	74%
[32]	NB	Python	70.26%

The accuracy of Naive Bayes (NB) in our research was 85.5%. These results are consistent with the research conducted by [26] as shown in Table 4, who found that Naive Bayes achieved the maximum accuracy of 85.03% when utilizing the UCI heart disease dataset. Their investigation entailed utilizing various models, and the experiment was carried out using the WEKA tool. The suggestion was to employ a genetic algorithm in MATLAB to decrease the number of feature dimensions prior to inputting the dataset into WEKA for prospective future research. The methodology and emphasis on model comparison in their research fit with our own investigation, therefore contributing to the alignment between our research endeavors.

Our research, unlike the study conducted by [27], utilized feature selection techniques such as PCA and CHI Square in Python 2.7. In comparison to their accuracy of 82%, our research achieved a higher Naive Bayes (NB) accuracy of 85.5%. The disparities in accuracy between their results and ours emphasize the potential discrepancies in dataset composition, feature selection techniques, or model optimization that may affect the performance of models in various research.

Contrasting the study conducted by [28], where they examined different types of SVM algorithms using the UCI dataset for diagnosing heart illness, our SVM model had a superior accuracy of 75.3% compared to their most successful approach, BTSVM, which had an overall accuracy of 61.86%. Their work comprehensively examined many versions of Support Vector Machines (SVM) such as Binary Tree SVM (BTSVM), One-Against-One (OAO), One-Against-All (OAA), Directed Acyclic Graph (DDAG), and Error Correcting Output Codes (ECOC). Prior to that, they performed dataset preprocessing using a min-max scaler. The discrepancies in accuracy among the research may arise from variances in dataset qualities, algorithm setups, or preprocessing approaches employed in the individual investigations.

The study conducted by [29] focused on feature selection in their work on cardiovascular illness prediction, with the objective of improving the dataset by specifically selecting relevant attributes for their models. Compared to our study, we did not use explicit feature selection, but instead chose to include a wider range of attributes from the dataset. The difference in strategy is likely to have had an impact on the performance of the predictive models. In our investigation, Catboost achieved an accuracy rate of 91.6%, SVM reached 75.3%, and KNN attained 69.5%. Nevertheless, Guarneros-Nolasco et al.'s research revealed that CatBoost achieved an accuracy rate of 81.32%, while the SVC model showcased a performance of 79.12%, and KNN presented an accuracy level of 70.33%. Although the studies used different approaches, they both emphasized the significant influence of age, heart rate and blood pressure on predicting cardiovascular disease. This confirms the crucial relevance of these factors in timely diagnosis and preventive measures.

In a similar study conducted by [30], they achieved an accuracy of 86.7% using Extra Trees. However, in our investigation, we acquired a greater accuracy of 91.1% using the same model. The main difference between our methodologies resides in the composition of the datasets used. We employed a hybrid dataset, combining the UCI dataset with our own local dataset. In contrast, their study relied exclusively on the UCI dataset. The variation in the exploitation of the dataset is likely to have led to the disparities in accuracy seen between the two investigations.

In contrast to the study by [31], which reports higher values of 74% accuracy and an AUC of 81%, our KNN model achieves an accuracy of 69.5% and an AUC of 77%. Differences in the datasets used and the data preprocessing techniques employed may be the cause of the disparity. They may have achieved greater model performance by focusing on extensive preprocessing, predicting heart disease using a range of

patient variables, and potentially optimizing hyperparameters.

Unlike the study conducted by [32], which produced an accuracy of 70.26% for their NB model, our research yielded a substantially better accuracy of 85.5%. One significant distinction is in our methodology: whereas they used feature selection approaches and concentrated on BMI as a primary predictor, we used all available data for prediction. The difference in model performance between the two studies was probably caused by these different approaches.

18. Conclusion

In conclusion, the research we conducted used machine learning algorithms to predict heart illness. The study compared linear models like K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Gaussian Naive Bayes to ensemble models like Catboost, HGBM and Extra Trees. The results unequivocally proved that ensemble models outperformed linear models in terms of accuracy and reliability. The ensemble models incorporated advanced techniques consisting of boosting and histogram-based algorithms that enhanced their exceptional predictive performance. Identifying the best machine learning (ML) methods was the study's goal among a collection of algorithms that are well-known and straightforward to implement, it has been observed that these methods demonstrate satisfactory performance, at least for the given dataset. ML methods are still being used at a very early stage, but there is evidence that they could be a very useful adjunct to patient care.

REFERENCES

[1] R. Jagannathan, S. A. Patel, M. K. Ali, and K. M. V. Narayan, "Global Updates on Cardiovascular Disease Mortality Trends and Attribution of Traditional Risk Factors," *Curr Diab Rep*, vol. 19, no. 7, p. 44, Jul. 2019, doi: 10.1007/s11892-019-1161-2.

[2] K. Dissanayake and M. G. Md Johar, "Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms," *Applied Computational Intelligence and Soft Computing*, vol. 2021, pp. 1–17, Nov. 2021, doi: 10.1155/2021/5581806.

[3] V. N. Agbor, N. A. B. Ntusi, and J. J. Noubiap, "An overview of heart failure in low- and middle-income countries," *Cardiovasc Diagn Ther*, vol. 10, no. 2, pp. 244–251, Apr. 2020, doi: 10.21037/cdt.2019.08.03.

[4] C. R. Olsen, R. J. Mentz, K. J. Anstrom, D. Page, and P. A. Patel, "Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure," *Am Heart J*, vol. 229, pp. 1–17, Nov. 2020, doi: 10.1016/j.ahj.2020.07.009.

[5] M. Pires, G. Marques, N. M. Garcia, and V. Ponciano, "Machine learning for the evaluation of the presence of heart disease," *Procedia Comput Sci*, vol. 177, pp. 432–437, 2020, doi: 10.1016/j.procs.2020.10.058.

[6] R. Aggrawal and S. Pal, "Sequential Feature Selection and Machine Learning Algorithm-Based Patient's Death Events Prediction and Diagnosis in Heart Disease," *SN Comput Sci*, vol. 1, no. 6, p. 344, Nov. 2020, doi: 10.1007/s42979-020-00370-1.

[7] D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," *SN Comput Sci*, vol. 1, no. 6, p. 345, Nov. 2020, doi: 10.1007/s42979-020-00365-y.

[8] A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," *Inform Med Unlocked*, vol. 19, p. 100330, 2020, doi: 10.1016/j.imu.2020.100330.

[9] C. Gazeloglu, "Prediction of heart disease by classifying with feature selection and machine learning methods", Progress in Nutrition 2020; Vol. 22, N. 2: 660-670, DOI: 10.23751/pn.v22i2.9830

[10] X.-Y. Gao, A. Amin Ali, H. Shaban Hassan, and E. M. Anwar, "Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method," *Complexity*, vol. 2021, pp. 1–10, Feb. 2021, doi: 10.1155/2021/6663455.

[11] https://archive.ics.uci.edu/ml/datasets/heart+Disease

[12] A. D. Shieh and Y. S. Hung, "Detecting Outlier Samples in Microarray Data," *Stat Appl Genet Mol Biol*, vol. 8, no. 1, pp. 1–24, Jan. 2009, doi: 10.2202/1544-6115.1426.

[13] Reetu Singh, E. Rajesh, "Prediction of Heart Disease by Clustering and Classification Techniques", International Journal of Computer Sciences and Engineering, Vol.-7, Issue-5, May 2019, E-ISSN: 2347-2693, pp. 861-866.

[14] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?," *PLoS One*, vol. 12, no. 4, p. e0174944, Apr. 2017, doi: 10.1371/journal.pone.0174944.

[15] Qing Yang and Fang-Min Li, "Support vector machine for customized email filtering based on improving latent semantic indexing," in *2005 International Conference on Machine Learning and Cybernetics*, IEEE, 2005, pp. 3787-3791 Vol. 6. doi: 10.1109/ICMLC.2005.1527599.

[16] M. Fatima and M. Pasha, "Survey of Machine Learning Algorithms for Disease Diagnostic," *Journal of Intelligent Learning Systems and Applications*, vol. 09, no. 01, pp. 1–16, 2017, doi: 10.4236/jilsa.2017.91001.

[17] P. S. Kumar, A. K. K, S. Mohapatra, B. Naik, J. Nayak, and M. Mishra, "CatBoost Ensemble Approach for Diabetes Risk Prediction at Early

Stages," in *2021 1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology(ODICON)*, IEEE, Jan. 2021, pp. 1–6. doi: 10.1109/ODICON50556.2021.9428943.

[18] Z. Feng, C. Xu, and D. Tao, "Historical Gradient Boosting Machine," pp. 68–54. doi: 10.29007/2sdc.

[19] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach Learn*, vol. 63, no. 1, pp. 3–42, Apr. 2006, doi: 10.1007/s10994-006-6226-1.

[20] H. M. LE, T. D. TRAN, and L. VAN TRAN, "AUTOMATIC HEART DISEASE PREDICTION USING FEATURE SELECTION AND DATA MINING TECHNIQUE," *Journal of Computer Science and Cybernetics*, vol. 34, no. 1, pp. 33–48, Aug. 2018, doi: 10.15625/1813-9663/34/1/12665.

[21] M. Tarawneh and O. Embarak, "Hybrid Approach for Heart Disease Prediction Using Data Mining Techniques," 2019, pp. 447–454. doi: 10.1007/978-3-030-12839-5_41.

[22] C. R, "Heart Disease Prediction System Using Supervised Learning Classifier," *Bonfring International Journal of Software Engineering and Soft Computing*, vol. 3, no. 1, pp. 01–07, Mar. 2013, doi: 10.9756/BIJSESC.4336.

[23] S. Asaduzzaman, Md. R. Ahmed, H. Rehana, S. Chakraborty, Md. S. Islam, and T. Bhuiyan, "Machine learning to reveal an astute risk predictive framework for Gynecologic Cancer and its impact on women psychology: Bangladeshi perspective," *BMC Bioinformatics*, vol. 22, no. 1, p. 213, Dec. 2021, doi: 10.1186/s12859-021-04131-6.

[24] T. Akter *et al.*, "Machine Learning-Based Models for Early Stage Detection of Autism Spectrum Disorders," *IEEE Access*, vol. 7, pp. 166509–166527, 2019, doi: 10.1109/ACCESS.2019.2952609.

[25] S. Pouriyeh et al., "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease," in Proceedings of IEEE Symposium on Computers and Communications (ISCC). Heraklion, Greece: IEEE, July 2017, pp. 204–207.

[26] B. Venkatalakshmi and M. Shivsankar, "Heart disease diagnosis using predictive data mining," International Journal of Innovative Research in Science, Engineering and Technology, vol. 3, no. 3, pp. 1873–1877, 2014.

[27] F. Tasnim and S. U. Habiba, "A Comparative Study on Heart Disease Prediction Using Data Mining Techniques and Feature Selection," in 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), IEEE, Jan. 2021, pp. 338–341. doi: 10.1109/ICREST51555.2021.9331158.

[28] W. Wiharto, H. Kusnanto, and H. Herianto, "Performance analysis of multiclass support vector machine classification for diagnosis of coronary heart diseases," International Journal on Computational Science & Applications, vol. 5, no. 5, pp. 27–37, 2015.

[29] L. R. Guarneros-Nolasco, N. A. Cruz-Ramos, G. Alor-Hernández, L. Rodríguez-Mazahua, and J. L. Sánchez-Cervantes, "Identifying the Main Risk Factors for Cardiovascular Diseases Prediction Using Machine Learning Algorithms," Mathematics, vol. 9, no. 20, p. 2537, Oct. 2021, doi: 10.3390/math9202537.

[30] S. Arunachalam, "Cardiovascular Disease Prediction Model using Machine Learning Algorithms," Int J Res Appl Sci Eng Technol, vol. 8, no. 6, pp. 1006–1019, Jun. 2020, doi: 10.22214/ijraset.2020.6164.

[31] H. Zhang, "Heart disease prediction using machine learning models," in *International Conference on Modern Medicine and Global Health (ICMMGH 2023)*, S. Sukumaran, Ed., SPIE, Sep. 2023, p. 41. doi: 10.1117/12.2692214.

[32] A. Nikam, S. Bhandari, A. Mhaske, and S. Mantri, "Cardiovascular Disease Prediction Using Machine Learning Models," in *2020 IEEE Pune Section International Conference (PuneCon)*, IEEE, Dec. 2020, pp. 22–27. doi: 10.1109/PuneCon50868.2020.9362367.