Article

Enhanced Machine Learning Model for CVD Prediction using Principal Component Analysis (PCA)

(IJGASR) International Journal For Global Academic & Scientific Research
ISSN Number: 2583-3081
Volume 4, Issue No. 2, 22–45
© The Authors 2025
journals.icapsr.com/index.php/ijgasr
DOI: 10.55938/ijgasr.v4i2.202

IJGASR

Shailendra Chaurasia[1] and A.K Sachan[2]

Abstract

The World Health Organization (WHO) report says that each year, cardiovascular diseases are the leading reason for around 17.9 million deaths across the globe. This is a more serious problem in low- and middle-income countries where there are barriers to early check-ups and specific treatments. The quicker and better detection of heart attacks helps reduce the risk of death. Based on previous methods, the study takes the Cleveland Heart Disease dataset from the UCI Machine Learning repository and uses it to design and check best machine learning models that take advantage of standardization, Principal Component Analysis (PCA) and hyperparameter tuning. We used machine learning algorithms such as Support Vector Machine, k-Nearest Neighbors, Logistic Regression and a Multi Layer Perceptron model, all combined under a Voting Classifier. With a 98.33% accuracy, 98.25% F1-score, 96.55% precision, and 100% recall on test data, the enhanced hybrid model (Voting Classifier) leaves all other models far behind in performance. The hybrid model had small gaps between train and test values for metrics, with 1.24% accuracy difference, 1.29% F1-score difference, 3.45% precision difference and -0.92% recall difference. Incorporating Principal Component Analysis (PCA) lowered the number of dimensions used while increasing accuracy, precision and F1 scores for a number of models. The results suggest that the use of Principal Component Analysis (PCA)-combined hybrid models leads to better, more understandable and trustworthy tools for predicting CVD. Strengthening predictive models for CVD risk assessment is now possible, supporting prompt clinical choices and helping patients improve.

Keywords

WHO, Heart Disease, Cardiovascular Disease (CVD, UCI Dataset, Machine Learning, PCA, Feature Selection, SVM (Support Vector Machine), KNN (K-Nearest Neighbors), LR (Logistic Regression) And A MLP (Multi Layer Perceptron), Hybrid Model (Voting Classifier)

Received: 26 May 2025; **Revised:** 03 July 2025; **Accepted:** 08 July 2025; **Published:** 15 July 2025

Introduction

Cardiovascular diseases consist of several heart and blood vessel illnesses, including coronary artery disease, stroke and heart failure, all of which significantly threaten worldwide health. Cardiovascular diseases are the leading cause of deaths across the globe. [15],[20] The World Health Organization (WHO)

[1,2]Department of Computer Science Engineering, LNCT University, Bhopal Madhya Pradesh, India. chaurasia.shailendra@gmail.com, sachank_12@yahoo.com

Corresponding Author:
E-mail: chaurasia.shailendra@gmail.com

© 2025 by Shailendra Chaurasia and A.K Sachan Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license,(http://creativecommons.org/licenses/by/4.0/). This work is licensed under a Creative Commons Attribution 4.0 International License

says that almost 18 million people lost their lives to cardiovascular diseases in 2019 (32% of all deaths worldwide). [1] The World Heart Federation found that each year, an estimated number of deaths caused by heart disease rose from 12.1 in 1990 to 18.6 in 2019, owing to an expanding population with risk factors. A report by the World Heart Federation states that 80% of the deaths from cardiovascular diseases are premature.[25] Heartattacks and strokes could be prevented if risk management is done promptly, yet in reality, it can be difficult to identify heart disease early [2]. ECG and angiography are common methods, but they can be expensive or a bit risky and many areas in need do not possess the equipment for screening many patients [5]. Therefore, we need ways to detect heart problems early that are available and accurate for guiding interventions.

With the help of machine learning, it is possible to detect cardiovascular diseases early by reviewing medical data and finding patterns that could indicate a disease. Lately, various forms of Machine Learning such as logistic regression and deep neural networks have been used to identify people at risk of heart disease. Still, it is difficult to use these approaches as dependable clinical instruments [6]. Some past studies sometimes select all features despite the fact that including all may not be useful and can result in reduced accuracy. Because most datasets have fewer cardiovascular diseases' patients than non-cardiovascular diseases patients, they may not train the model properly. Moreover, well-known models such as k-Nearest Neighbors (KNN) or Random Forest (RF) find it difficult to handle large and complex data.

We have already used some of these methods and studied their performance in previous papers. For instance, in our previous study [13], we applied Principal Component Analysis to reduce the number of features, then used Gradient Boosting, AdaBoost and Multi-Layer Perceptron to achieve almost 92% accuracy on the Cleveland data set. The heart disease risk model on [30] was based on Support Vector Machine (SVM) and Logistic Regression, among other algorithms and reported SVM had the highest accuracy of 96.66%.

The importance of training performance was discussed in both the previous studies [13][30], but this study changes the goal to analyzing and evaluating how the model behaves on new data. This study also introduces a way of combining conventional machine learning with a deep learning model called the Multi-Layer Perceptron (MLP), organized under a Voting Classifier framework. We opt for the term "hybrid model" rather than ensemble to show the blend of Machine Learning and Deep Learning in our approach.

This report boosts this area of study by effectively incorporating a Principal Component Analysis (PCA) technique in the Machine Learning step and by introducing better model training strategies (such as cross-validation and hyperparameter tuning). It's interesting that by refining our model we were able to overcome the drawbacks of complex models: using fewer features and better understanding the results. To be consistent with previous studies, we focus on the Cleveland Heart Disease dataset from UCI Machine Learning Repository [14].We are particularly interested in how Principal Component Analysis (PCA) will affect the results of machine learning models when utilized and how a hybrid model will perform on this data along with PCA. When we test our new model against the old ones, we can notice the improvements in how efficient and consistent it is.

To support the significance of our study, we mention that many leading organizations promote the advancement of cardiovascular disease prediction models. The World Health Organization (WHO) and others strongly recommend prompt diagnosis to reduce the deaths caused by these diseases [2]. However it is difficult to achieve high accuracy and reliability in CVD diagnostics due to both the complexity of the clinical data and the shortcomings of how features are modeled [4]. To meet this challenge, we use sophisticated ways to process data and improve with Machine Learning techniques. It supports global health by helping clinicians with choice-useful support for heart disease patients.

Literature Review

In the past five years, there has been a fast increase in machine learning being used in predicting heart risks. Several approaches, from the original logistic regression algorithm and decision trees to more advanced solutions, have been used in research and these techniques might also include feature selection or optimization.

Artificial Intelligence/Machine Learning in Medicine: Experts in medicine recognize that using Artificial Intelligence and Machine Learning can greatly enhance the management of cardiovascular diseases. [28][29] A study commissioned by the World Health Organization (WHO) revealed that use of computational tools can improve a doctor's ability to make decisions, allowing them to process medical info more efficiently and accurately. Initially, Machine Learning models for heart disease worked by predicting the diagnosis using risk factors found in medical records. Studies show that approaches such as Decision Trees and Naïve Bayes are effective for forecasting heart disease.[3][14] [10] But, many times, these original models included numerous attributes that were not necessary which could lead to overfitting [12].

Ensemble and Hybrid Models: According to the latest studies, organizations are following ensemble learning and hybrid methods because they result in more accurate documents. They increase the accuracy of predicting heart problems by running neural networks and combining them with voting methods.[16][19] A study carried out in 2025 applied stacking and voting algorithms to datasets on heart disease, concluding that these models achieve better results than the individual models.[6] When combining six different algorithms, the soft-voting classifier had ~93–95% accuracy on two datasets, whereas the best single model got ~90% accuracy. [6] Another study by Anas Maach and team [24] demonstrated that a hybrid model built on MultiLayer Perceptron, Random Forest, and Adaboost classifier was a top performer among all the models with an accuracy of 88.12%. The other models tested in this study included XGBoost, Naive Bayes, LogitBoost, and k-Nearest Neighbors. It becomes clear from these findings that ensembles are helpful in strengthening cardiovascular disease prediction, so we examine them in our research as well.

Support Vector Machine (SVM) and Other Classifiers: Among individual algorithms, Support Vector Machine has often outperformed other algorithms on data from medicine. In [30], we noticed that Support Vector Machine reached an accuracy of 96.66% on the Cleveland dataset, surpassing Random Forest, Logistic Regression and Decision Tree. This result is similar to findings by another research, where it was observed that Support Vector Machine gave above 83% accuracy on a cardiovascular disease dataset and this rose to 88.3% accuracy after using a genetic algorithm to select the best features. [11] Logistic Regression and Decision Trees may compete well with other methods if the data is properly cleaned and organized. According to a study, Logistic Regression and k-Nearest Neighbors outperformed Support Vector Machine and Random Forest on heart disease data through multiple cross-validations.[7] However, since their results were much lower than those of more complex models, the authors found that handling data (in their case, imputation and normalization of features) strongly affected the accuracy ranks of the tested models.[7] Other studies also demonstrated that Logistic Regression, Support Vector Classifier, k-Nearest Neighbors classifiers are generally the potential choice for researchers for cardiovascular disease prediction. [21][22][27][29] While conducting our research, we apply Support Vector Machine as a common baseline and carefully process the data to evaluate all models evenly.

Feature Selection and Dimensionality Reduction: Several recent works agree that the performance of a machine learning model often depends on choosing the appropriate features.

A study by El-Sofany and Hosam F. [26] demonstrated the use of three feature selection methods namely, chi-square, ANOVA (Analysis of Variance) and Mutual Information (MI) out of which three different feature sets were obtained. On these three feature sets, different models were trained such as XGBoost, k-Nearest Neighbors, Naive Bayes, Support Vector Machine, etc. Among all the different models, XGBoost was found to have the highest accuracy of 96.57% on the feature set obtained using ANOVA method. [26] Support vector machines work best when random and repeated features are removed. Multiple methods were suggested such as an approach blending Chi-square, Information Gain and forward/backward selection. It helped the researchers select 8 useful features, leading XGBoost to achieve up to 99% accuracy.[4] In a study by Moshood Abiola Hambali [8], after performing Principal Component Analysis, XGBoost performed well with 98% accuracy and 97% precision. They examined heart disease data with Principal Component Analysis (PCA) and decision trees, finding that a Principal Component Analysis (PCA)-based classifier had 98% accuracy, 100% sensitivity and 98% precision. In their study, Principal Component Analysis (PCA) became part of the classifier design and so highlighted the main variations needed for accurate prediction.[8] In 2021, Principal Component Analysis (PCA) was implemented on a number of classifiers; as a result, the accuracy improved to ~91% (for Support Vector Machine classifier model) which is higher by a few percentage points than when all features were used. [9]. With the evidence in mind, we believed that performing Principal Component Analysis (PCA) would help us preserve the important information in the data, reduce the number of variables and enhance both training efficiency and prevention of overfitting.

In addition to Principal Component Analysis (PCA), other ways of handling features have achieved good results. Researchers have applied both Genetic Algorithm and Particle Swarm Optimization to help decide which features are significant or to refine the parameters in the model. Researchers helped shape current hybrid models by improving the accuracy of classifying heart disease by combining Support Vector Machine with Particle Swarm Optimization. [5] According to Muhammad [23], by applying filtering (using methods like chi-square for feature selection) and wrapping (training models on selected features), we can identify the best features for which accuracy is highest. At the same time, LASSO (Least Absolute Shrinkage and Selection Operator) is used to choose vital features by shrinking those that are not important. One study showed that using LASSO with Pearson correlation and a Random Forest classifier gave 99% accuracy when predicting heart disease. Even so, the authors pointed out that training on a small dataset could lead to overfitting. [12] Actually, no machine learning model can accurately predict all data with 100% confidence, as this may not be true for new data. Studies reveal that Principal Component Analysis (PCA) is broadly applicable and improves the model performance by achieving higher accuracy in lesser time. [18] Therefore, we focus our optimization efforts on using Principal Component Analysis (PCA). Although our research's main focus is enhancing model performance, we emphasize that PCA-based dimensionality reduction not only improves the model performance but also improves computational efficiency by removing redundant and noisy features. It explains the most important features in the data by grouping correlated variables into a few principal components that are not correlated with each other.

Methodology

For this study, we applied an optimized machine learning procedure to predict cardiovascular disease using Principal Component Analysis (PCA) for feature reduction and carried out thorough evaluation. We chose the Cleveland Heart Disease dataset from the UCI repository which is a regular benchmark of 297 samples with 13 features and 1 target variable focusing on whether heart disease is noted. Based on

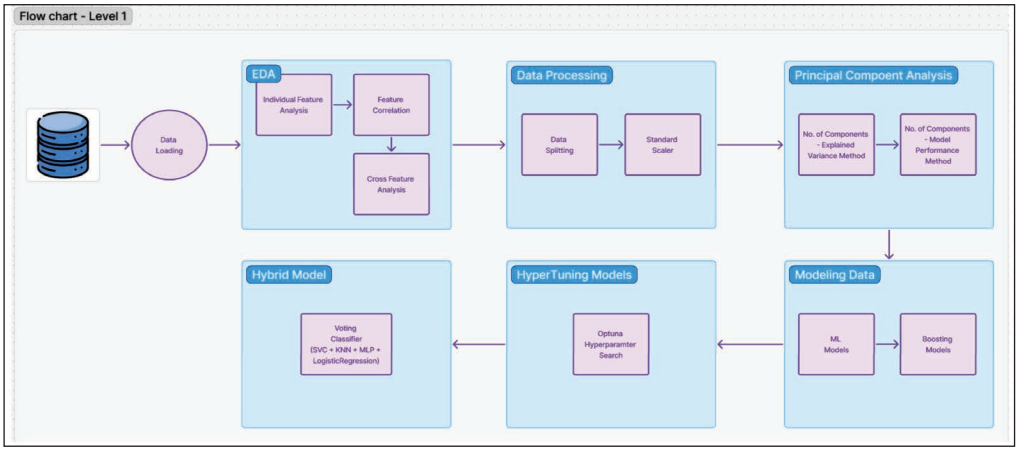

Figure 1. Proposed methodology for CVD prediction.

clinical features, the task involves determining whether a person has heart disease or not. Figure 1 illustrates that our process goes from gathering data to evaluating our models.

Initially, we collected our data (UCI Cleveland dataset [14]) and proceeded with data preprocessing, optimizing features using Principal Component Analysis (PCA), using various Machine Learning algorithms for training and using accuracy, precision, recall, and F1-score to assess your performance.

Data Preprocessing

Data Cleaning & Preparation

We looked for any missing or unusual value in the dataset. The data was already clean. It had no missing values, outliers, and duplicate values.

Feature Scaling:

Since some machine learning algorithms are sensitive to how features are scaled, we standardized all our features before using them. StandardScaler was used to change each of the features so that their means were at zero and variances were at one. Because of this, numbers for cholesterol and resting blood pressure are not far apart on the scales and do not have a major influence on distance computations. Before using Principal Component Analysis (PCA), the data should be standardized since Principal Component Analysis (PCA) looks for the most significant directions of change among the features. Equation (1) demonstrates how the standardization of data should be applied:

$$z_i = (x_i - \mu) / \sigma \qquad \text{Equation (1)}$$

where,

x_i is an original feature value,

μ is the mean of that feature in the training set, and

σ is the standard deviation.

Once the scaling is done, each feature becomes dimensionless and can be roughly seen as the same.

Train-Test Split and Cross-Validation:

Initially, the data was divided into a training set that composed most of the data (80%) and a testing set that left 20%. We used k-fold cross-validation (with k=5) on the training set to select the best hyperparameters and measure how robust the model is. Cross-validation decreases the possibility of overfitting since it checks the model's performance using several folds. Our final evaluation on performance was conducted on the test set (20%) and its elements were not included in the training or validation.

Principal Component Analysis:

After scaling the data, we performed Principal Component Analysis (PCA) to reduce the correlated features and improve the model's results. It helps us take the original features and code them into another orthogonal space, where the largest amount of variance is captured last. StandardScaler was applied to the pipeline before PCA, meaning each feature in the data contributed the same to each of the resulting components.

We tested Principal Component Analysis using the Decision Tree and Random Forest Classifier on a wide range of components, starting with one and ending with the total number of features the data held. Testing and training were done throughout and the results for accuracy and F1-score were plotted for each model combination. Doing this analysis, we found that the accuracy and F1-score were consistent for 9 principal components across both models (Random Forest Classifier and Decision Tree Classifier). So we selected 9 principal components out of 13 which explained most of the variance and made the data easier to read and analyze.

We plotted a feature loadings matrix to analyze how principal components explain most of the variance with respect to the original features. The use of a heatmap on the matrix allowed us to see that cholesterol and resting blood pressure were key features impacting some components.

By analyzing the explained variance ratio (see Figure: Cumulative Variance and Figure2: Variance per Component), it was found that while PC1 held the largest share (22.69%), the variance was generally

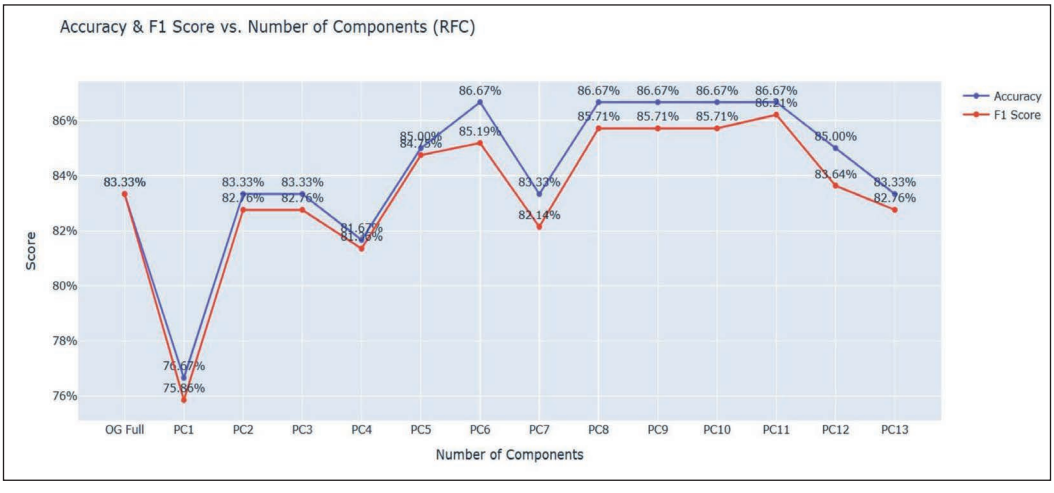

Figure 2. Accuracy and F1 Score vs. Number of Components (RFC.

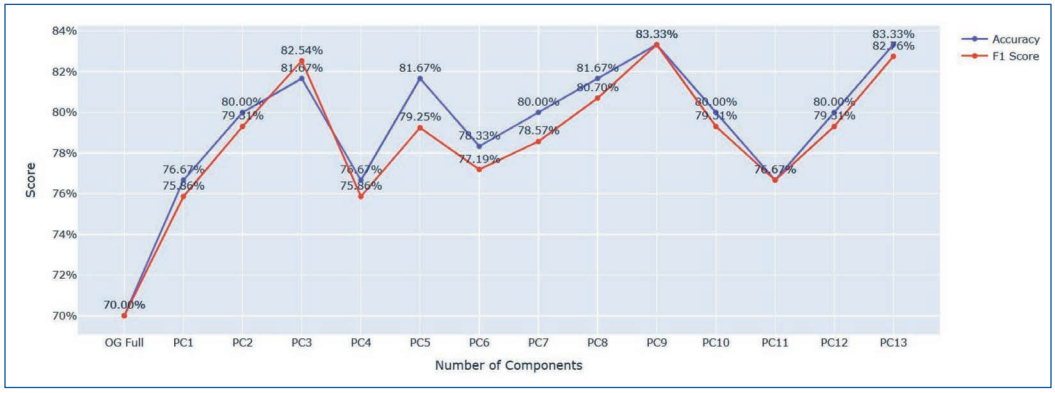

Figure 3. Accuracy & F1 Score vs. Number of Components (DTC).

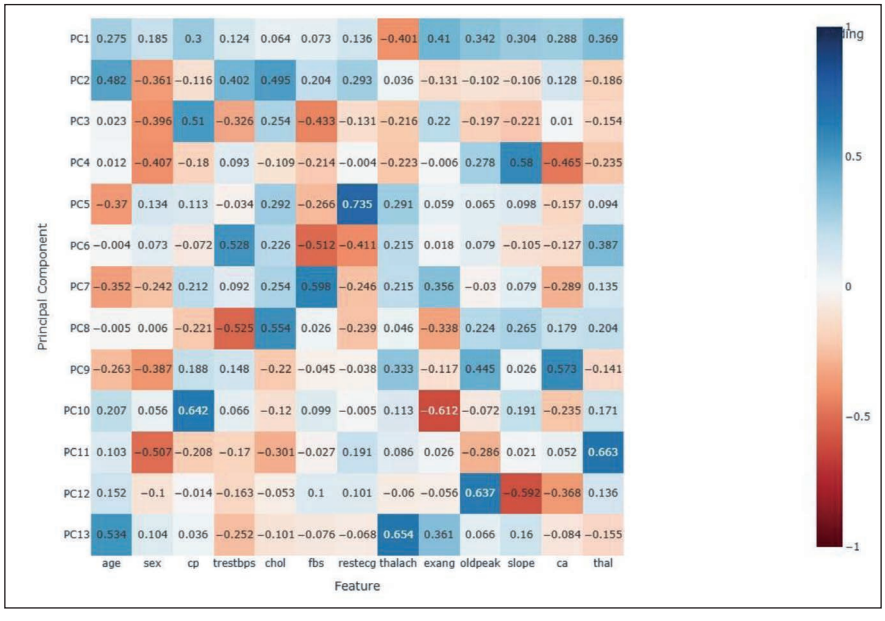

Figure 4. PCA Feature Loadings Heatmap.

distributed across all components. It didn't show a separate elbow point and over 53.9% of the data was covered by the first 4 components.

We trained a number of machine learning models on the transformed data. These models include Logistic Regression, K-Nearest Neighbors, Support Vector Classifier, Decision Tree, Random Forest, Gradient Boosting, AdaBoost, XGBoost, LightGBM and CatBoost. We analyzed model performance using accuracy, precision, recall, and F1-score.

We compared the performance of each model before and after performing Principal Component Analysis (PCA). Improvements in test precision, F1 score, recall and accuracy after PCA were represented

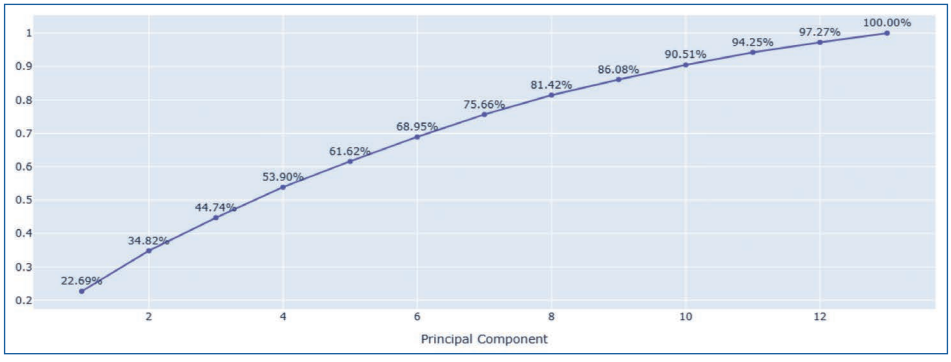

Figure 5. Cumulative Variance.

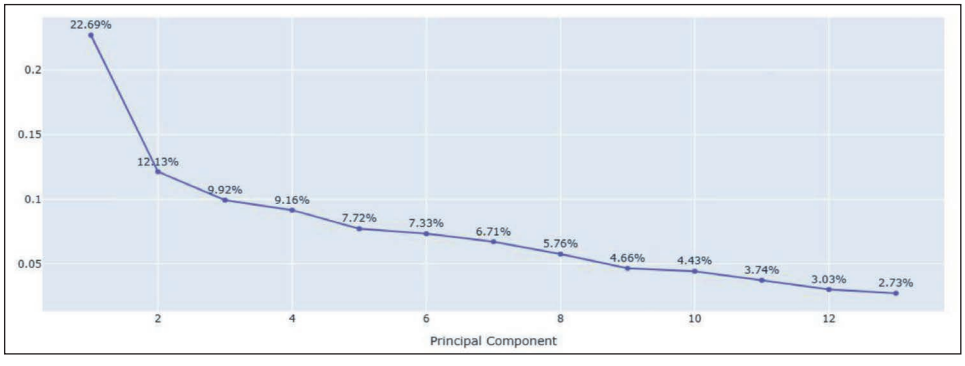

Figure 6. Scree Plot - Variance per Component.

using bar charts. It was found that Principal Component Analysis made it easier for a machine learning model to generalize, as it reduced unwanted noise and took out features that were not useful.

In the end, PCA was used within the training process of the hybrid model, along with Multi Layer Perceptron, Support Vector Classifier, K-Nearest Neighbors and Logistic Regression. The use of PCA-enhanced inputs reduced overfitting and allowed the hybrid model to succeed with an accuracy of 98.33% and no overfitting, while every single model had lower performance.

Machine Learning Models and Training

Based on our studies and literature research, we used several machine learning methods to choose the model that performed best for cardiovascular disease prediction.

Support Vector Machine (SVM): Usually, Support Vector Machine with Radial Basis Function (RBF) kernel performs well in numerous studies and was therefore selected for this work. Support Vector Machine identifies a separating hyperplane in a high number of dimensions by using kernel functions to account for non-linearity. The task was to use a grid search to set the values of the

regularization parameter C and kernel width gamma for the model. After using cross-validation, the model with C≈10 and gamma≈0.1 worked extremely well on the training data.

Ensemble Boosting Algorithms: We wanted to test Gradient Boosting, XGBoost, CatBoost, Light Gradient Boosting Machine (LGBM), and AdaBoost classifiers because these were part of the algorithms used in [13]. Gradient Boosting Classifier (GBC) constructs a series of weak learners and then unites them to ensure the error in predictions is minimized. In each stage, AdaBoost focuses on the instances that are missed by adding more weight to them for the next round. Both techniques have been successful with organized medical information and in some heart disease research, an AdaBoost model achieved more than 90% accuracy. For both experiments, we chose 100 estimators and selected 0.1 as the learning rate, adjusting them both according to results from the validation set.

Multi Layer Perceptron (MLP) Neural Network: We trained a Multi Layer Perceptron and found its best structure using Optuna which is a hyperparameter optimization framework. It has a single hidden layer (like in [13]) to examine instances of non-linearity. So that the data set would not cause overfitting, the network architecture was designed to be simple (a hidden layer with only 50 neurons). We applied ReLU activation and Adam optimizer and set the early stopping feature at 200 epochs. These models are able to find relationships between features that linear models miss, though our previous findings revealed that Multi Layer Perceptron and boosting methods gave similar results when the data was limited (accuracy in the 91%-92% range).

Logistic Regression (LR) and k-Nearest Neighbors (KNN): We decided to include Logistic Regression and K-Nearest Neighbors (instance-based learning) models as baseline algorithms. The main benefit of Logistic Regression is that it creates a straightforward model, whereas k-nearest neighbors model is useful on small datasets because it instinctively learns from the provided samples. In this study, both logistic regression and k-nearest neighbors models were able to reach a 91% accuracy on the test dataset. Although we hoped that our advanced models would outperform the simpler ones, considering the results confirmed the task is challenging and the new approaches are valuable.

Hybrid Model design - Integrating Multi Layer Perceptron with Machine Learning Models: We present a model in this study that uses the best aspects of machine learning and deep learning models together. The hybrid model combines predictions from a Multi Layer Perceptron (MLP) with predictions from Logistic Regression, Support Vector Classifier and k-Nearest Neighbors. The aim of this strategy is to use the outstanding Multi Layer Perceptron (MLP) pattern recognition with the simple and transparent approach of common Machine Learning models to efficiently detect cardiovascular diseases.

The hybrid model integrates the strengths of the Multi Layer Perceptron (MLP) with simpler machine learning models like Logistic Regression, Support Vector Classifier, and k-Nearest Neighbors. This combination uses the Multi Layer Perceptron model's ability to capture complex nonlinear patterns while benefiting from the interpretability and stability of basic models. By combining these approaches with the help of a voting classifier, the hybrid model attains higher accuracy, reduces overfitting, and provides better generalization on unseen data.

Hyperparameter Tuning

We implemented advanced, automated techniques to fine-tune hyperparameters in order to ensure the best performance and generalizability from the model. We selected Optuna as our main method to efficiently search for optimum values for hyperparameters.

Since Optuna is defined by runs, its path for functioning allows simple redefinition of space to search and the construction of the search method to change based on circumstances. Tree-structured Parzen Estimator (TPE), which is a tree-structured method, is used by Optuna as the default method for examining various hyperparameter options. Unlike grid search which checks every option, Optuna predicts where the most promising values might be and searches more near there. As a result, it is an excellent choice for fine-tuning severe models, especially when training time is a constraint.

In our research, Optuna was mainly chosen to find the best combination of architecture and learning parameters for the Multi Layer Perceptron (MLP) model. Parameters such as the number of hidden layer neurons, learning rate, activation function, batch size, optimizer, and early stopping criteria were tuned inside the framework.

Optuna was also used with traditional models such as Support Vector Machine, Logistic Regression and k-Nearest Neighbor to find the best hyperparameters for each model using Accuracy as the goal in 5-fold cross-validation.

Continuously and wisely tuning the models led to each being checked in its optimal configuration, making any comparison fair. We saw that the Multi Layer Perceptron and SVM models performed better overall when using Optuna, giving a big advantage to the final hybrid model.

We combined Optuna with our approach which made sure both deep learning and machine learning gained their highest prediction skills, strengthening the confidence and stability of our system for cardiovascular disease prediction.

Performance Evaluation

We then compared the models using accuracy, precision, recall, F1-score and the area under the curve. The main goal is to report precision and recall for the positive cases (diseases) to see the effectiveness of disease identification by the model. With high recall, most patients with cardiovascular diseases will be identified and with high precision, the majority of people labeled as having the disease truly have it. F1-score measures how well a classifier balances the sensitivity and specificity, using the harmonic mean of precision and recall.

Our objective here is to determine if the new model along with Principal Component Analysis (PCA) and tuned hyperparameters can produce the same or better results as before, using less data. In the section after that, we share the results along with a comparison with previous studies to demonstrate the progress made using this approach.

Results and Discussion

This part provides an in-depth comparison of machine learning algorithms for finding cardiovascular disease (CVD) based on Principal Component Analysis (PCA)-transformed features (9 principal components). The results were assessed using accuracy, F1-score, precision, recall and generalization gap to choose the best model.

Baseline Model Performance

Each chosen model was trained on a Principal Component Analysis (PCA)-transformed dataset that included 9 principal components and their performance was assessed in terms of Accuracy, F1-Score,

Table 1. Performance metrics for all classifiers on 9 Principal Component Analysis (PCA) components.

Model	Train Accuracy	Test Accuracy	Train F1Score	Test F1Score	Train AUCScore	Test AUCScore	Train Precision	Test Precision	Train Reca ll	Test Rec all
Gaussian NB	0.902954	0.916667	0.892019	0.909091	0.900624	0.915179	0.913462	0.925926	0.87156	0.892857
Logistic Regression	0.936709	0.933333	0.930876	0.928571	0.935959	0.933036	0.935185	0.928571	0.926606	0.928571
K Neighbors Classifier	0.924051	0.933333	0.915888	0.925926	0.922198	0.930804	0.933333	0.961538	0.899083	0.892857
SVC	0.957806	0.966667	0.953271	0.964286	0.956171	0.966518	0.971429	0.964286	0.93578	0.964286
Decision Tree Classifier	1	0.816667	1	0.8	1	0.814732	1	0.814815	1	0.785714
Random Forest Classifier	1	0.9	1	0.892857	1	0.899554	1	0.892857	1	0.892857
Gradient Boo sting Classifier	1	0.9	1	0.892857	1	0.899554	1	0.892857	1	0.892857
Ada Boost Cl assifier	0.991561	0.9	0.990826	0.892857	0.991507	0.899554	0.990826	0.892857	0.990826	0.892857
XGB Classifier	1	0.883333	1	0.877193	1	0.883929	1	0.862069	1	0.894737
LGBM Classifier	1	0.916667	1	0.912281	1	0.917411	1	0.896552	1	0.928571
Cat Boost Classifier	1	0.916667	1	0.912281	1	0.917411	1	0.896552	1	0.928571

Precision and Recall. The results for both the training and testing phases are shown in detail in the following table.

Support Vector Classifier scored the highest test accuracy (96.67%) and F1-score (96.43%) and exceeded the performance of Generic Ensemble methods such as Gradient Boosting and Random Forest in terms of generalizability. The support vector classifier exhibited the lowest difference in results for all of the metrics, indicating that it performed reliably and was not prone to overfitting. (Figure 7)

Moreover, the following figure shows that the Support Vector Classifier managed to balance precision and recall very effectively (Figure 8), so it works well for medical uses where errors put people at great risk.

On the training dataset, CatBoost, LGBM and RandomForest obtained perfect scores, although the test results showed F1 and accuracy scores of around 91%. Decision Tree Classifier had the biggest gap in its accuracy (+18.33%), along with the lowest test recall at 78.57%. Due to this, it was not chosen despite its high accuracy on the training set.

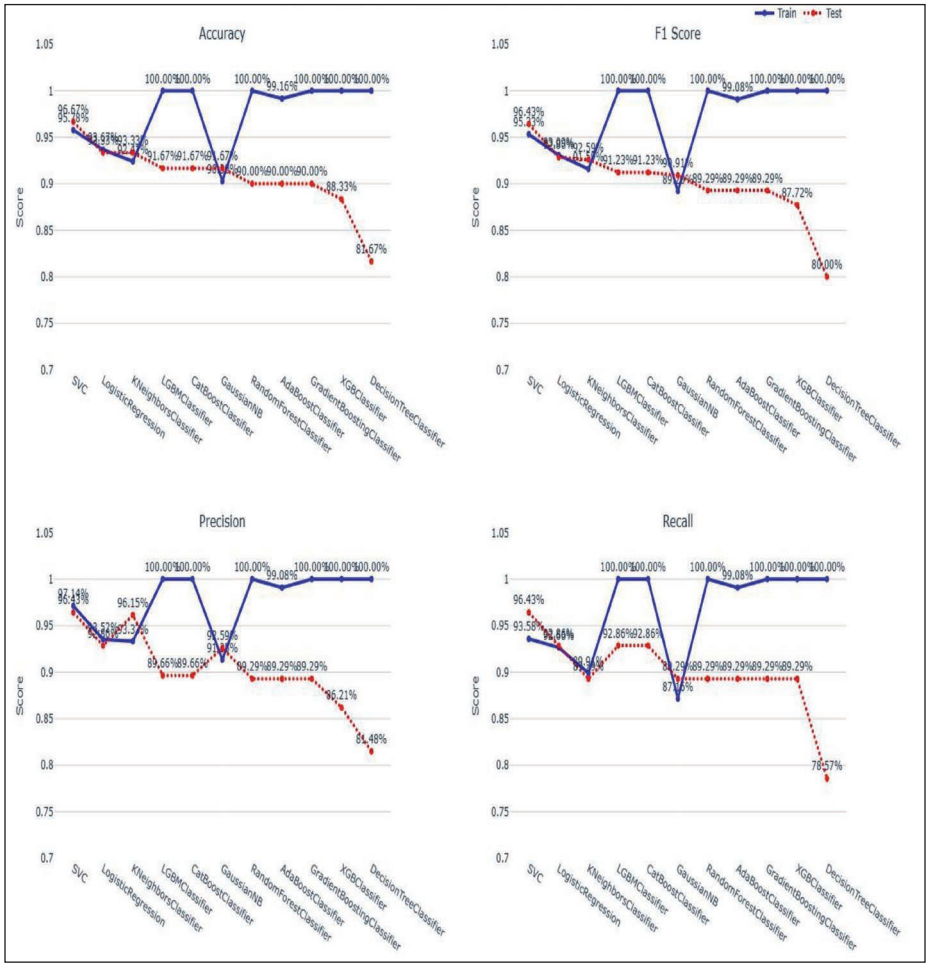

Figure 7. Model Performance Comparison - 9 PCs - Performance Curve.

(Figure 9) uses a Precision-Recall tradeoff plot to provide key examples of performance difference. Among all the tests, SVC had the greatest F1-score of 96.43% while providing similar accuracy and recall of 96.42%. Just after Linear Regression, Logistic Regression and k-Nearest Neighbors placed, with their F1-scores staying highest among the algorithms tested. While both CatBoost and LightGBM were strong on recall, they didn't perform as well on precision. It was apparent that Decision Tree Classifier overfit which leads to a low F1-score of ~0.80.

Overall, the results confirm that Principal Component Analysis improved not only the accuracy but also the steadiness of precision-recall balance among all models and support vector classifier was the clear winner.

Results of Principal Component Analysis

To judge how Principal Component Analysis (PCA) affects the effectiveness of the model, we first used each classifier without PCA and then with PCA. We wanted to find out if Principal Component Analysis improved how well the models fit future data by cutting down noise and extra similarities in the input data.

In terms of precision, Gradient Boosting Classifier improved by 17.857% and came in first, while Decision Tree Classifier and Support Vector Classifier saw 15.86% and 13.67% improvements, respectively, followed by Cat Boost Classifier with 13.185%. Even when testing Logistic Regression and

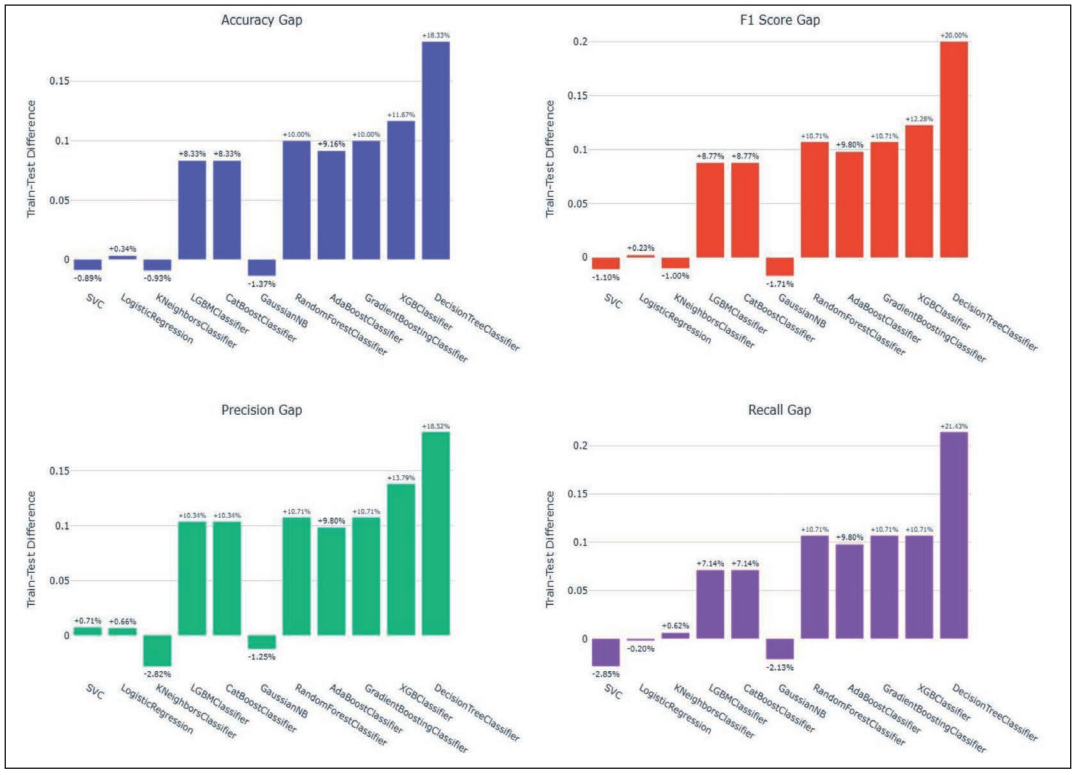

Figure 8. Model Performance Comparison - 9 PCs - Performance Gaps.

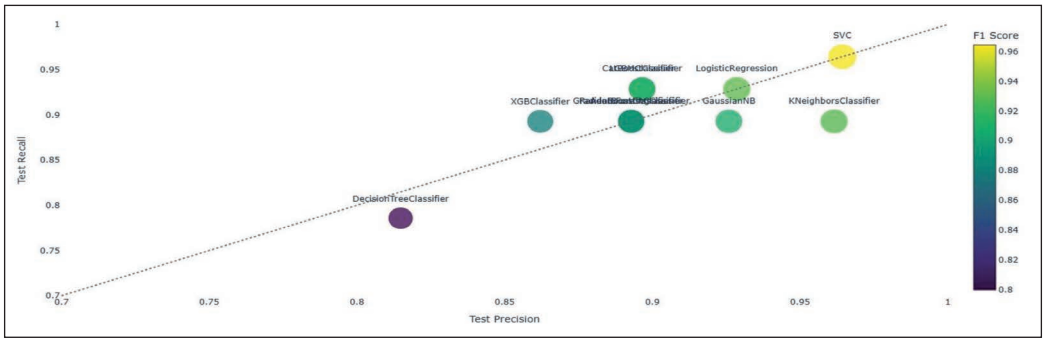

Figure 9. Model Performance Comparison - 9 PCs - Precision-Recall Tradeoff.

Gaussian Naive Bayes, which are usually quite stable, put up precision gains of 7.143% and 9.834% respectively, suggesting that PCA improves both simple and complex models. (Figure 10)

The accuracy of all models was found to improve when using PCA compared to previous results. Here, Support Vector Classifier went up from 85.00% to 96.66%, Decision Tree Classifier went from 70.00% to 81.67% and Random Forest Classifier moved from 83.33% to 90.00%. These improvements mean that PCA probably got rid of unnecessary features, helping to increase the success of the classification. We can see that other classifiers also had significant improvement in accuracy. (Figure 11)

As for F1 Score, Principal Component Analysis (PCA) gave us promising results again. After PCA, the F1 score of Support Vector Classifier rose from 84.211% to 96.429%, with Gradient Boosting Classifier improving from 79.365% to 89.29%. We can see how these numbers change in (Figure 12) which plots grouped bar charts for F1 scores.

Principal Component Analysis showed that the CatBoost Classifier and LightGBM Classifier maintained their strong performance in terms of recall, practically unchanged at about 92.86%, but the recall for Support Vector Classifier rose from 85.71% to 96.42%. (Figure 13)

In conclusion, Principal Component Analysis made a clear and frequently important difference in the performance of each classifier. It improved not only our ability to make correct predictions, but also the match between our training and testing data—mainly for models impacted by redundant features or uneven numbers of classes. As a result, we conclude that integrating PCA with strong classifiers like the Voting Classifier leads to powerful methods for predicting cardiovascular disease.

Results of Hyperparameter Tuning

The Multi Layer Perceptron did the best after tuning, reaching an F1-score of 92.86%, a test accuracy of 93.33% and test precision and recall of 92.86% (Figure 14). It may have a greater difference in train-test accuracy as compared to Support Vector Classifier, but the improvement of +7.14% in precision, recall, and F1-score was still considered acceptable. (Figure 15).

Following Optuna tuning, the MLPClassifier had the highest precision and recall which can be seen by the bright colour and size of its point on the graph. High F1-scores were obtained by Logistic Regression and k-Nearest Neighbors models. On the other hand, after tuning, the Support Vector Classifier achieved lower recall than the untuned model. They prove that Optuna is helpful for boosting model performance and demonstrate its key part in improving generalization for all models.

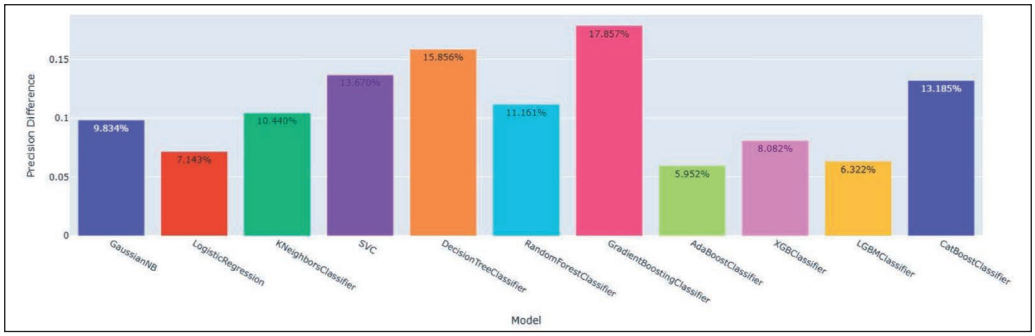

Figure 10. Before & After PCA - Testing Precision Gain After PCA.

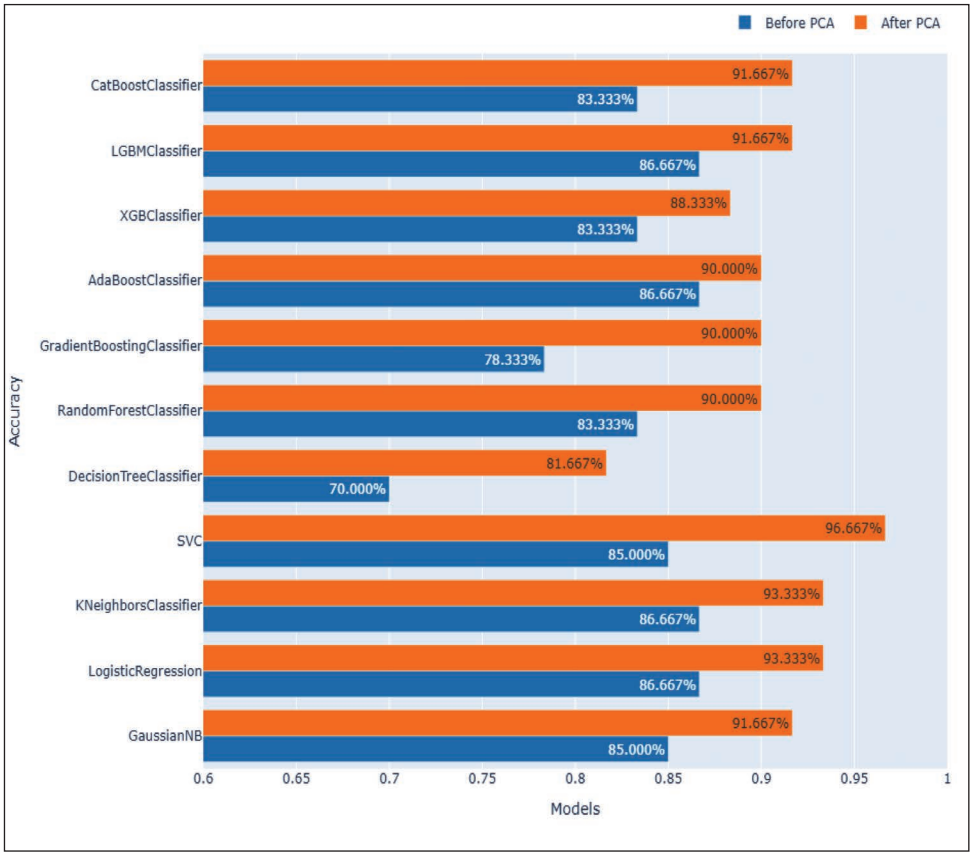

Figure 11. Accuracy Comparison Before vs After PCA.

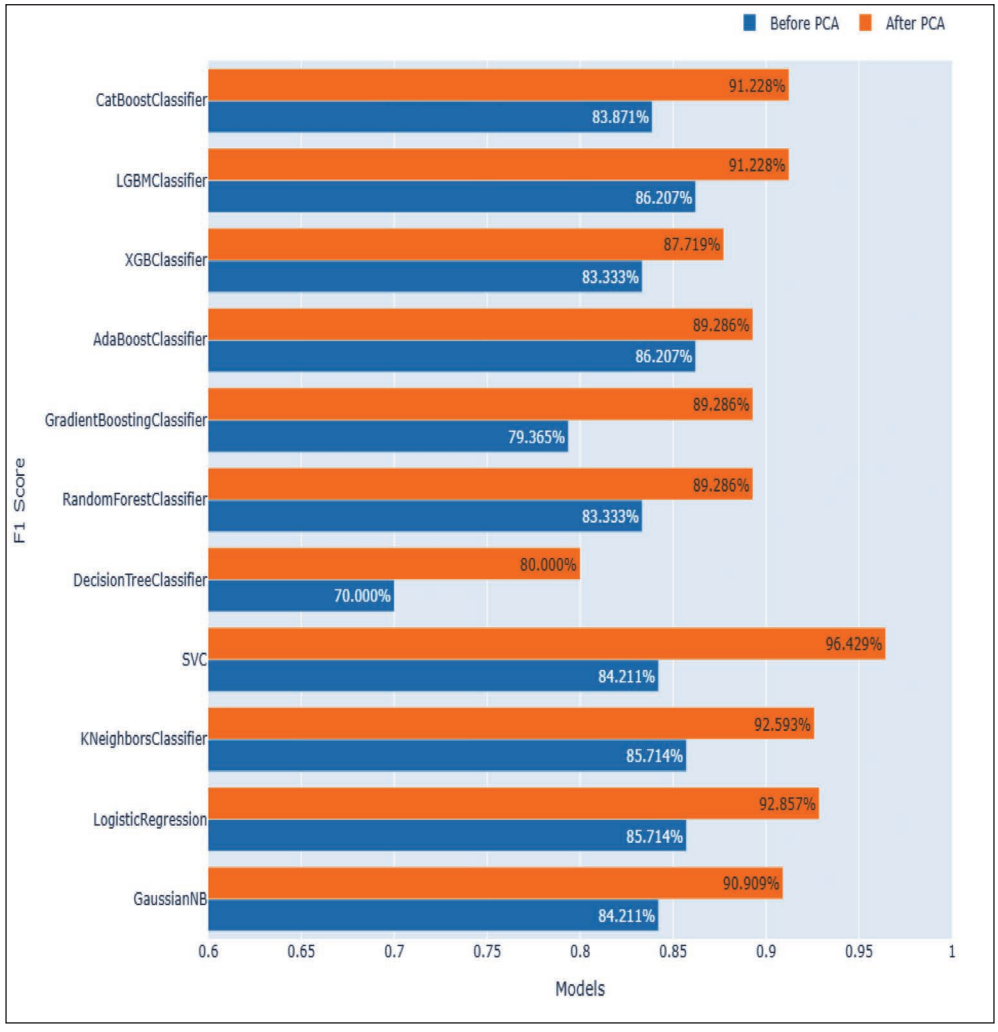

Figure 12. F1 Score Comparison Before vs After PCA.

Final Comparison and Hybrid Model

The next stage is to compare the methods and make a hybrid model. Using the hybrid model, the final record was 98.33% test accuracy, 98.35% F1-score, 96.55% precision and 100% recall (Figure 17). There were the smallest generalization gaps for all of the metrics except precision (Figure 18), as the precision-recall trade-off showed its perfect balance (Figure 19). It means that with the help of a hybrid model, you can create a better classifier and one that is more stable than a single excellent classifier.

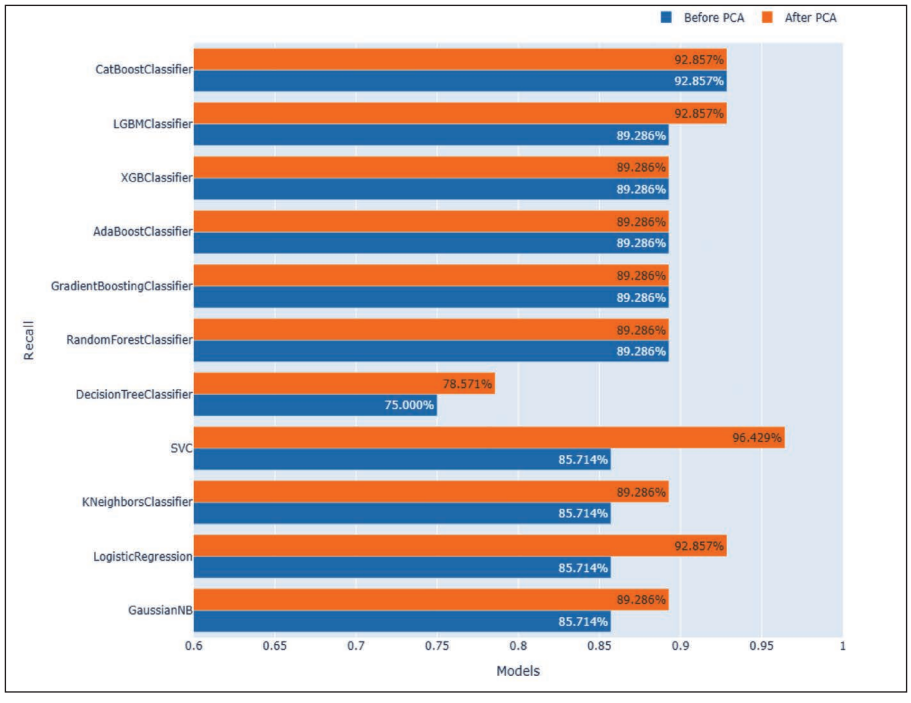

Figure 13. Recall Comparison Before vs After PCA.

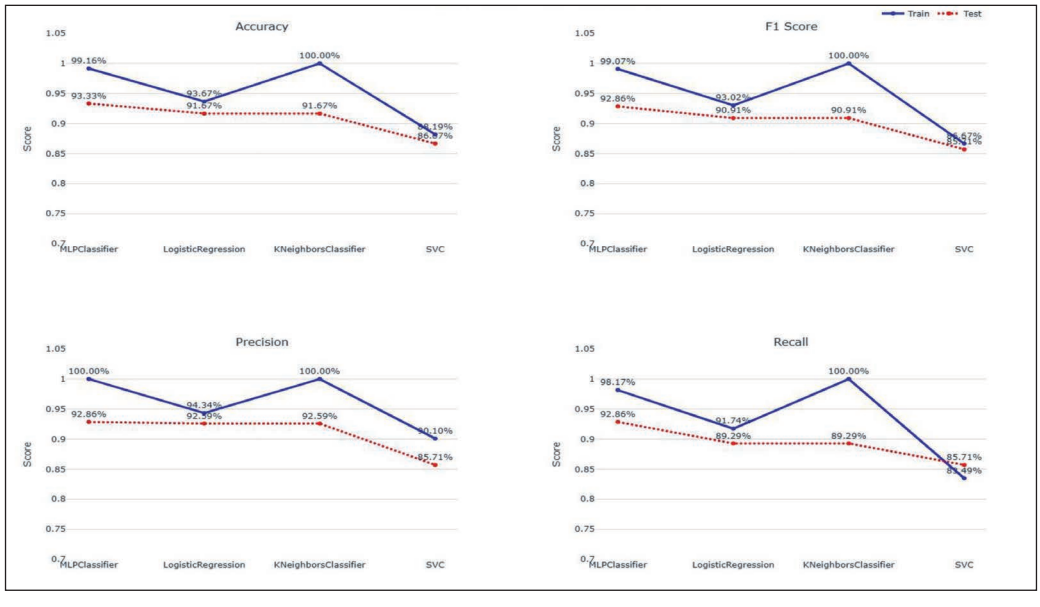

Figure 14. HyperTuned Models - Performance Curves.

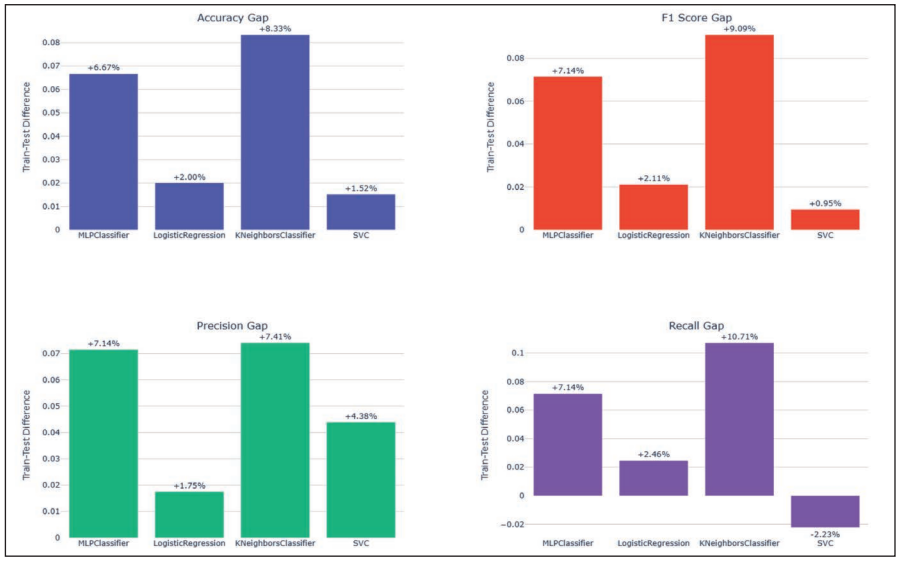

Figure 15. Hyper Tuned Models - Performance Gaps.

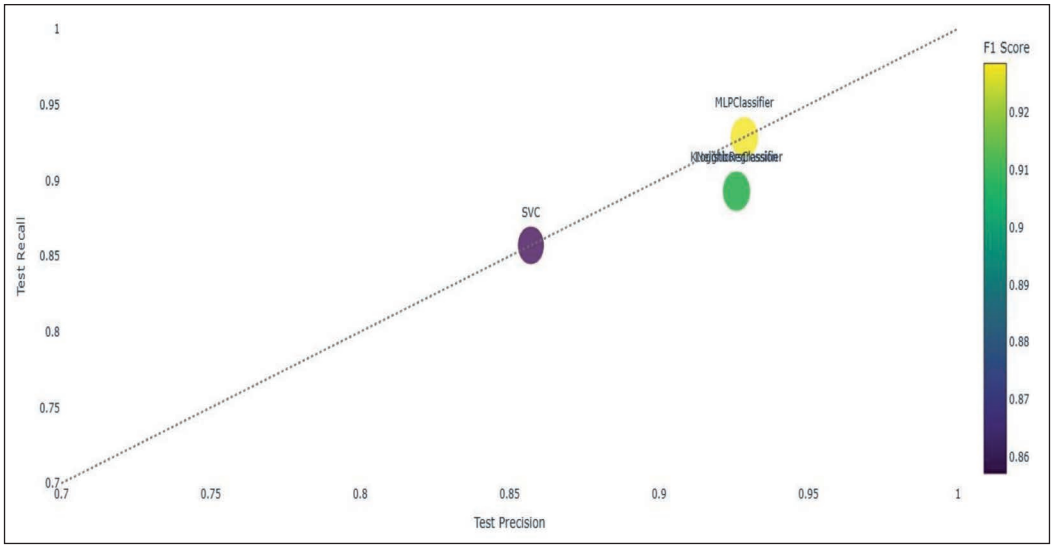

Figure 16. Hyper Tuned Models - Precision-Recall Tradeoff.

Learnings and Outcomes

The findings support the notion that being able to apply a model is just as important as accurately predicting cardiovascular diseases. Since they don't easily overfit, Support Vector Classifiers and Voting Classifiers are better and more reliable for actual real-world use. With the help of a hybrid model, medical AI can perform even better diagnostics.

Comparative Analysis with Previous Work

The achievements in this study are measured by comparing the results of the Voting Classifier with those of the best previous study that used the Support Vector Classifier. Following are the key improvements made by this study:

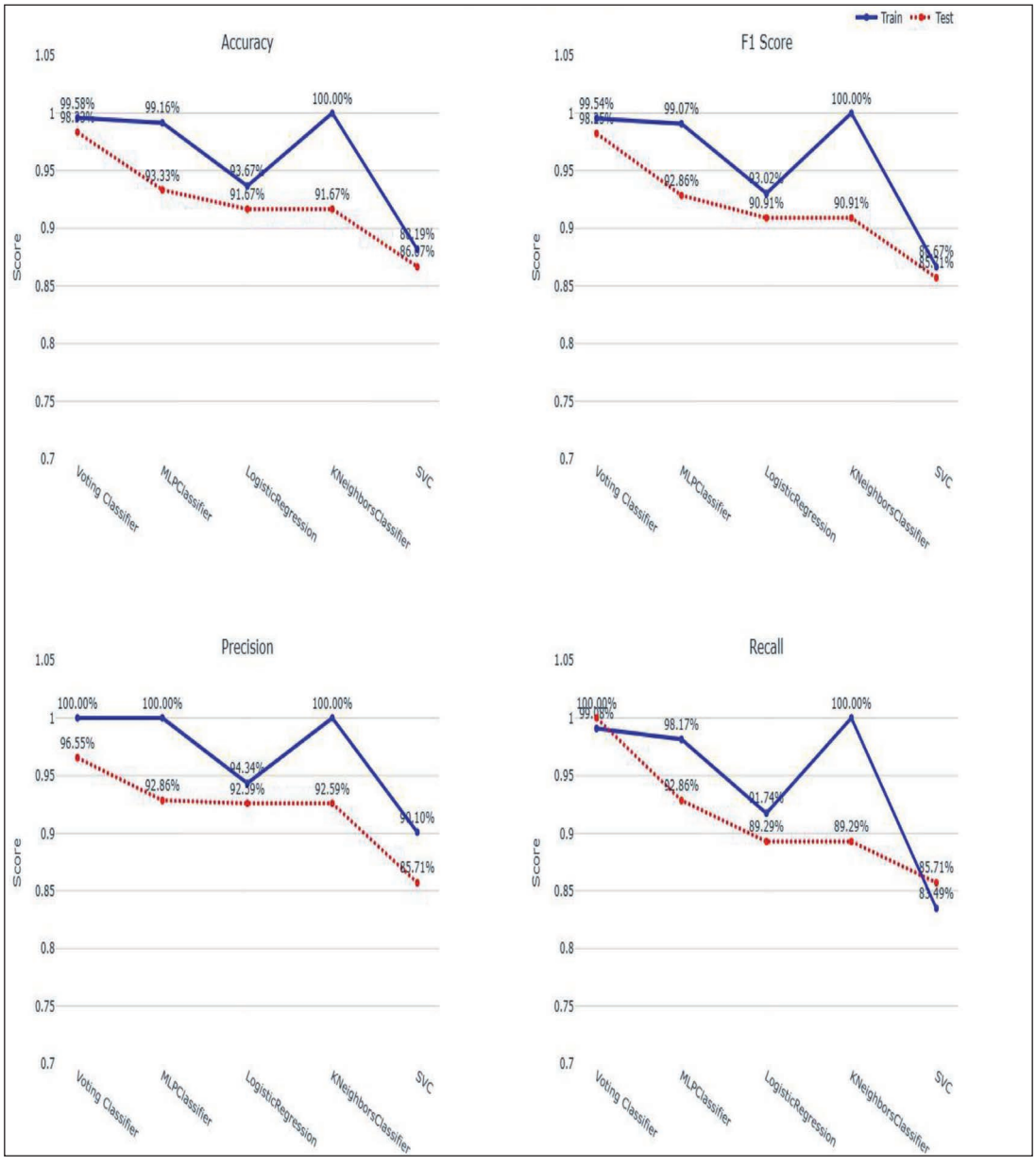

Figure 17. Final Comparison - Performance Curve.

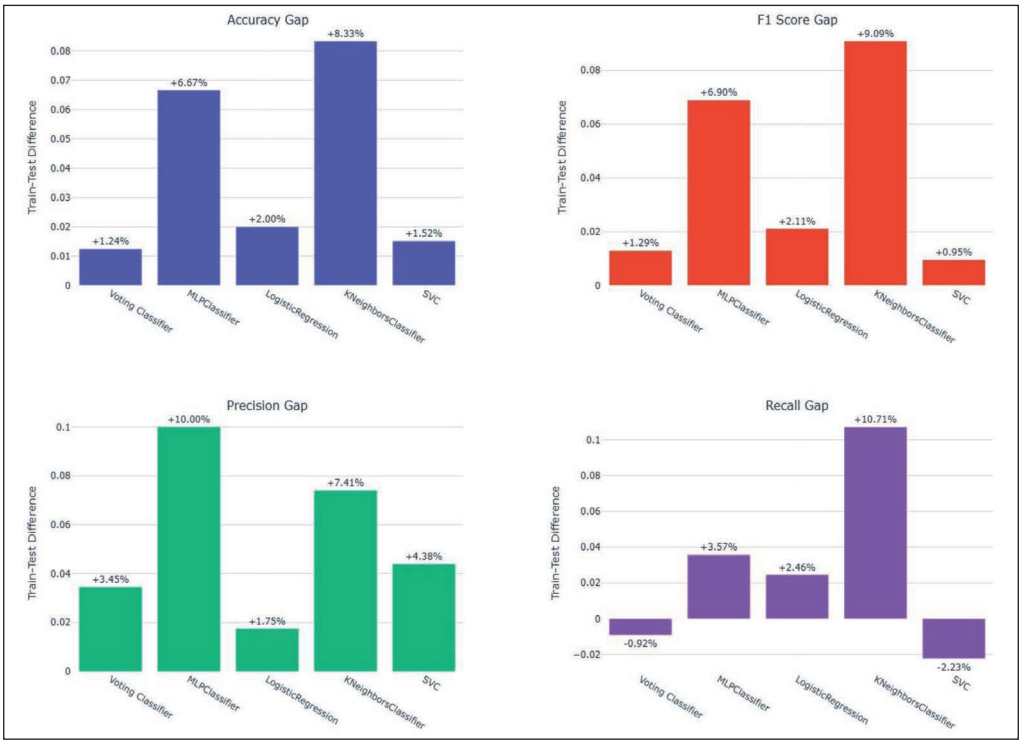

Figure 18. Final Comparison - Performance Gaps.

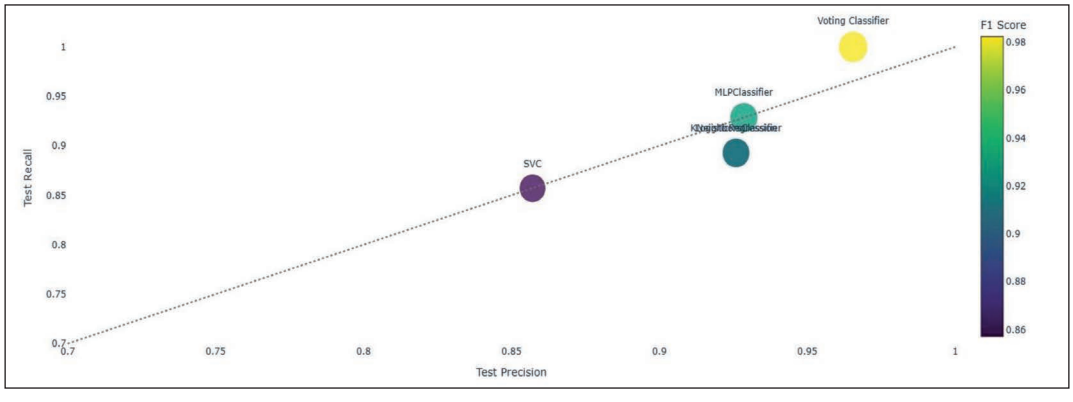

Figure 19. Final Comparison - Precision-Recall Tradeoff.

- The percentage of training accuracy increased from 95.781% using Support Vector Classifier to 99.58% using a Voting Classifier.
- The improvement in generalizability is shown by the increase in Testing Accuracy from 96.67% to 98.33%.
- There is no doubt that Voting Classifier offers much higher performance and better balance through a better F1 score.

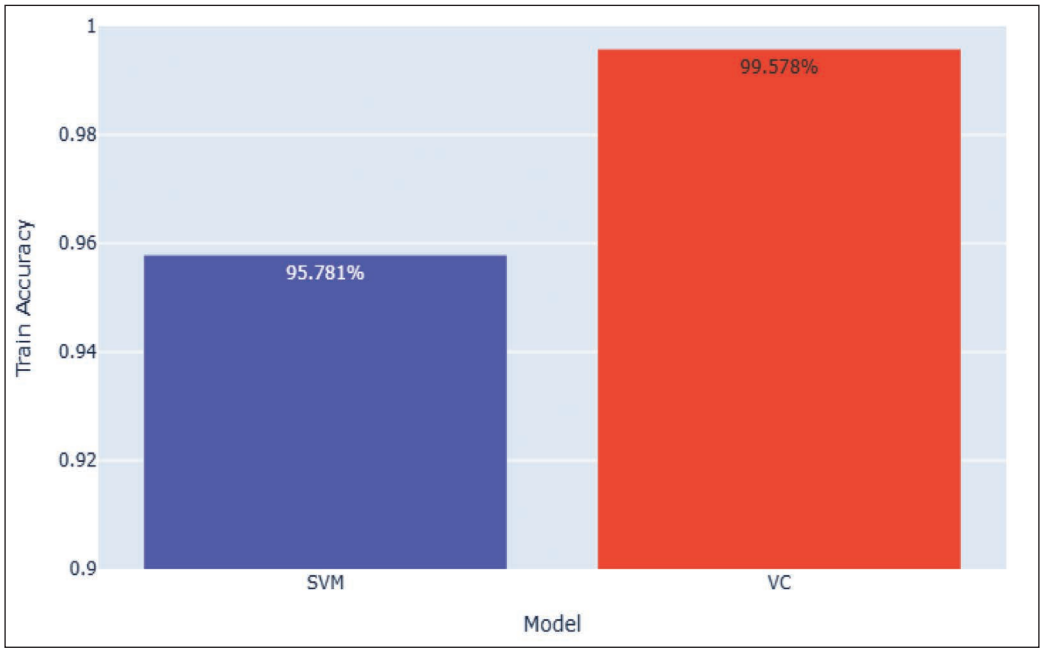

Figure 20. Previous vs Current - Training Accuracy.

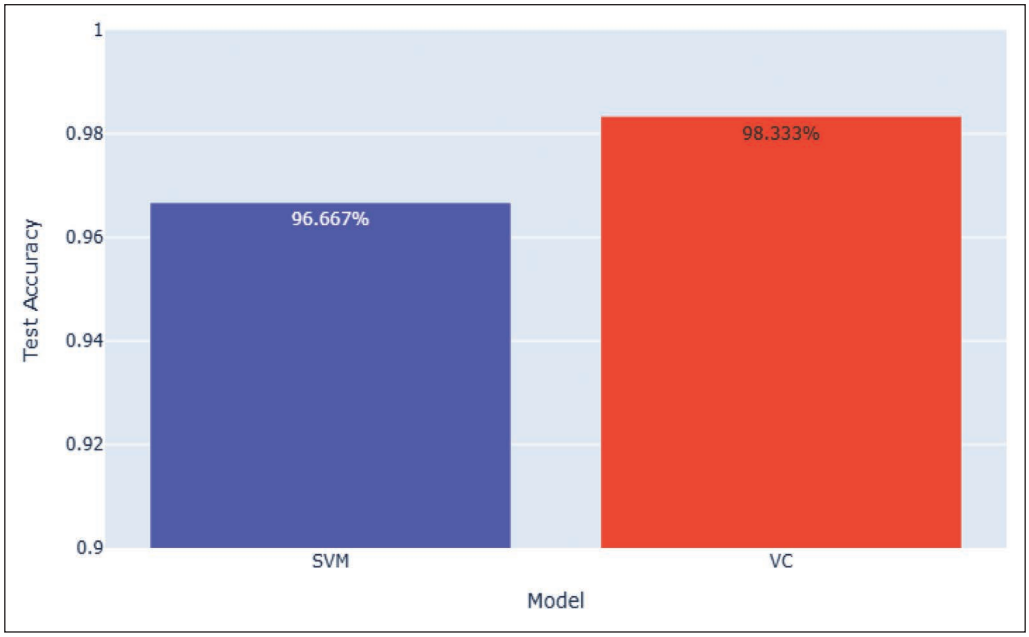

Figure 21. Previous vs Current - Testing Accuracy.

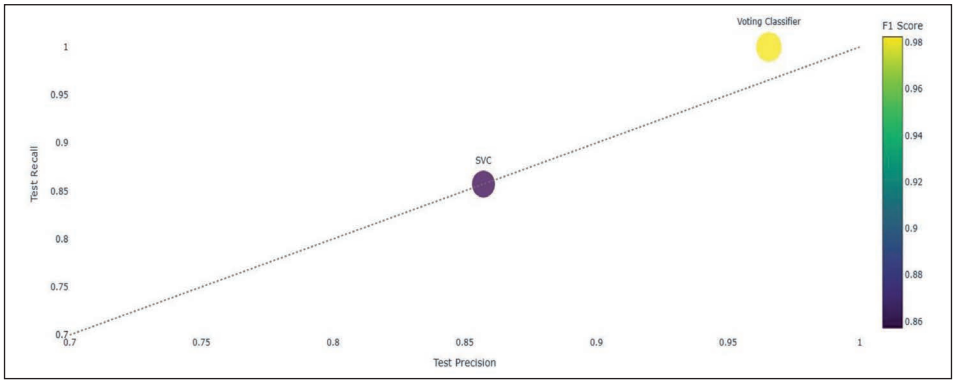

Figure 22. Previous vs Current - Precision-Recall Tradeoff.

We can see by the results that applying Principal Component Analysis (PCA), techniques for model optimization and ensembling to the algorithm led to a model that performs better on training data and, most importantly, also generalizes well on unknown data.

Conclusion

Applying the steps described in this research substantially improves cardiovascular disease modeling by reducing the dataset's features, fine-tuning the models and combining them. The study kept the original importance of features by transforming the data with Principal Component Analysis (PCA) and filled in the details only when appropriate.

Out of all baseline classifiers, the Support Vector Classifier (SVC) performed best, reaching high levels of accuracy (96.67%) and F1-score (96.43%) without much overfitting. Nevertheless, after adjusting the hyperparameters and combining different models, Voting Classifier achieved better results than any other model. A test accuracy of 98.33%, F1-score of 98.25%, clearly showing that it is reliable and effective enough for a medical diagnostic system.

Compared to the previous winner model (Support Vector Classifier), the new model shows notable improvements on all metrics. It's worth noting that the Voting Classifier performed better on a smaller number of different examples (lower overfitting) and achieved much higher accuracy, precision and recall. We can see the change in results by plotting gap analysis charts and precision-recall tradeoff plots.

In essence, this research both uses the knowledge gained from previous research and improves the overall cardiovascular disease detection approach. Extending this study would require the use of current clinical data, a check on bias across demographics and perhaps testing the hybrid model in clinical decision support.

ORCID iD

Shailendra Chaurasiaa https://orcid.org/0009-0009-4698-1434

References

1. World Health Organization (WHO). Cardiovascular Diseases (CVDs) Fact Sheet (11 June 2021). (Describes global CVD mortality: 17.9 million deaths in 2019, need for early detection.)
2. World Heart Federation. World Heart Report 2023: Confronting the World's Number One Killer (2023). heart-report23.world-heart-federation.org. (Provides statistics on CVD death trends, 12.1 million in 1990 to 18.6 million in 2019.)
3. Absar N., Das E.K., Shoma S.N., Khandaker M.U., Miraz M.H., Faruque M.R.I., Tamam N., Sulieman A. and Pathan R.K., 2022, June. The efficacy of machine-learning-supported smart system for heart disease prediction. In *Healthcare* (*Vol. 10*, No. 6, p. 1137). MDPI.
4. Alwakid G., Ul Haq F., Tariq N., Humayun M., Shaheen M. and Alsadun M., 2025. Optimized machine learning framework for cardiovascular disease diagnosis: a novel ethical perspective. *BMC Cardiovascular Disorders*, *25*(1), p.123.
5. Rehman M.U., Naseem S., Butt A.U.R., Mahmood T., Khan A.R., Khan I., Khan J. and Jung Y., 2025. Predicting coronary heart disease with advanced machine learning classifiers for improved cardiovascular risk assessment. *Scientific Reports*, *15*(1), p.13361.
6. Ganie S.M., Pramanik P.K.D. and Zhao Z., 2025. Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets. *Scientific reports*, *15*(1), p.13912.
7. Rimal, Y., Sharma N., Paudel S., Alsadoon A., Koirala M.P. and Gill S., 2025. Comparative analysis of heart disease prediction using logistic regression, SVM, KNN, and random forest with cross-validation for improved accuracy. *Scientific Reports*, *15*(1), p.13444.
8. Hambali M.A., Gbolagade M.D. and Olasupo Y.A., 2023. Heart disease prediction using principal component analysis and decision tree algorithm. *Journal of Computer Science and Engineering (JCSE)*, *4*(1), pp.1–14.
9. Boukhatem C., Youssef H.Y. and Nassif A.B., 2022, February. Heart disease prediction using machine learning. *In 2022 Advances in Science and Engineering Technology International Conferences (ASET) (pp. 1-6)*. IEEE.
10. Dangare C.S. and Apte S.S., 2012. Improved study of heart disease prediction system using data mining classification techniques. *International Journal of Computer Applications*, *47*(10), pp.44–48.
11. Mienye I.D., Sun Y. and Wang Z., 2020. An improved ensemble learning approach for the prediction of heart disease risk. *Informatics in Medicine Unlocked*, *20*, p.100402.
12. Sharma S. and Parmar M., 2020. Heart diseases prediction using deep learning neural network model. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, *9*(3), pp.2244–2248.
13. Chaurasia S. and Kamble M. (2024). An Effective Framework for Early Detection and Classification of Cardiovascular Disease (CVD) Using Machine Learning Techniques. In: Sharma H., Shrivastava V., Tripathi A.K., Wang L. (eds) *Communication and Intelligent Systems. ICCIS 2023. Lecture Notes in Networks and Systems*, *vol 969*. Springer, Singapore. https://doi.org/10.1007/978-981-97-2082-8_2
14. CVD Prediction Kaggle Dataset – Cleveland. Kaggle, 2018. (Online repository description of the Cleveland Heart Disease dataset; 303 observations, 13 features, widely used in heart disease prediction research.)
15. World Health Statistics, 2022 https://cdn.who.int/media/docs/default-source/gho-documents/world-health-statistic-reports/worldhealthstatistics_2022.pdf
16. Ekle F.A., Shidali V., Ochogwu R.E. and Igoche I.B., 2024. Machine Learning models for heart disease prediction and dietary lifestyle change therapy recommendation: a systematic review. *Discover Artificial Intelligence*, *4*(1), p.113.
17. López-Saynes J.L., Escobar-Gómez E.N., Trujillo S.V., Pérez C.V.D.C., Marroquín-Cano S.F., Chandomi-Castellanos E. and Hernández-Gutiérrez C.A., 2024, July. Analysis of Physiological Parameters for Assessing the Risk Level of Cardiovascular Diseases Using Machine Learning Algorithms. *In 2024 10th International Conference on Control, Decision and Information Technologies (CoDIT)* (pp. 2763–2768). IEEE.
18. Pathan M.S., Nag A., Pathan M.M. and Dev S., 2022. Analyzing the impact of feature selection on the accuracy of heart disease prediction. *Healthcare Analytics*, *2*, p.100060.

19. Ashraf M., Rizvi M.A. and Sharma H., 2019. Improved heart disease prediction using deep neural network. *Asian Journal of Computer Science and Technology*, *8*(2), pp.49–54.
20. WHO. Global Health Estimates: Leading Causes of Death, 2020. Available at: Leading causes of death
21. Bouqentar M.A., Terrada O., Hamida S., Saleh S., Lamrani D., Cherradi B. and Raihani A., 2024. Early heart disease prediction using feature engineering and machine learning algorithms. *Heliyon*, *10*(19).
22. Sarra R.R., Dinar A.M., Mohammed M.A. and Abdulkareem K.H., 2022. Enhanced heart disease prediction based on machine learning and $\chi 2$ statistical optimal feature selection model. *Designs*, *6*(5), p.87.
23. Menshawi A., Hassan M.M., Allheeib N. and Fortino G., 2023. A Hybrid Generic Framework for Heart Problem diagnosis based on a machine learning paradigm. *Sensors*, *23*(3), p.1392.
24. Maach A., Elalami J., Elalami N. and Mazoudi E.H.E., 2022. An intelligent decision support ensemble voting model for coronary artery disease prediction in smart healthcare monitoring environments. arXiv preprint arXiv:2210.14906.
25. World Heart Federation. World Heart Day 2023 – Know Your Heart. Available at: World Heart Day 2023: Know Your Heart for Better Health
26. El-Sofany H.F., 2024. *Predicting heart diseases using machine learning and different data classification techniques*. IEEE Access.
27. Singh A., 2020. Prediction of heart disease using machine learning. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol*, pp.150–166.
28. Bani Hani S.H. and Ahmad M.M., 2023. Machine-learning algorithms for ischemic heart disease prediction: a systematic review. *Current cardiology reviews*, *19*(1), pp.87–99.
29. Shrestha D., 2024. Advanced Machine Learning Techniques for Predicting Heart Disease: A Comparative Analysis Using the Cleveland Heart Disease Dataset. *Applied Medical Informatics*, *46*(3).
30. Chaurasia S., Kamble M. (2024). Implementation Of A Heart Disease Risk Prediction Model Using Machine Learning. DOI : https://doi.org/10.62441/nano-ntp.vi.5467
31. Karthick K, Aruna SK, Samikannu R, Kuppusamy R, Teekaraman Y, Thelkar AR, Implementation of a Heart Disease Risk Prediction Model Using Machine Learning, 2022, https://doi.org/10.1155/2022/6517716