Coronary Artery Disease Prediction: A Novel Approach Using Principal Component Analysis and Enhanced Support Vector Machine Classifier

Swarnasmita Roy[1], Kaustubh Kalpathy[2], Akshaya Gayathri K[3], Indhusree Reddy M[4], Geetha K N[5,a], Shali S[6]

[1,2,3,4,6] *Department of Mechanical Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Bengaluru 560035, India*
[5] *Department of Mathematics, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Bengaluru 560035, India*

[a)] Corresponding author: kn_geetha@blr.amrita.edu

Abstract. Coronary artery disease continues to be one of the leading causes of mortality in the modern world necessitating an effective diagnosis timely. This research focuses on developing a classification model that will be based on machine learning in order to determine the presence of this disease by making use of a dataset of health-related attributes. Reduction of dimensionality of the dataset that provides reasonable computational efficiency without losing essential information is done by principal component analysis. The dataset with 14 features is reduced to 8 principal components. The work assesses the effectiveness of support vector machine on the dimensionally reduced dataset, which labelled 2 classes: class 0- heart disease, class 1- no heart disease. The hypothetical model has achieved an accuracy of 92.21%. It is further evaluated using metrics such as precision, recall and F1-score. For class 0: precision- 0.91, recall- 0.94, and F1-score- 0.93. For class 1: precision- 0.94, recall- 0.90, and F1-score- 0.92. These experimental outcomes show that the model is quite efficient as a disease-diagnosis tool. The focus of this project is to demonstrate the capabilities of machine learning including principal component analysis and support vector machine in disease diagnosis, ensuring its accurate and effective use in a scalable manner to assist clinicians with improving patient outcome and resource management.

I. INTRODUCTION

Coronary Artery Disease (CAD), also known as coronary arteriosclerosis or ischemic heart disease, is a pathological condition wherein the coronary arteries constrict or occlude due to its inability to transport adequate amounts of oxygenated blood, nutrients, and essential elements to the myocardium. Atherosclerosis is primarily responsible for constriction, and is a pathological process wherein cholesterol, lipid deposits, and other substances combine inside the inner walls of arteries and form plaques. These plaques become hard and cause loss of elasticity in arteries or even rupture to form clots that prevent the blood from flowing further [1].

CAD is one of the most common global health issues that contributes significantly to disability, disease burden, and mortality which is increasing with an interplay of genetic, lifestyle, and environmental factors. Common risk factors include smoking, sedentary behavior, unhealthy dietary patterns, excessive alcohol consumption, and chronic stress. CAD also has a high association with advanced age, genetic predisposition, obesity, hypertension, diabetes, and elevated levels of low-density lipoprotein (LDL) cholesterol.

In the initial period, the symptoms of the disease can be silent or not apparent, and for the same reason, this "silent" but progressive disease finds its place. When symptoms have appeared, angina, shortness of breath, feeling fatigued, palpitation, dizziness, and in extreme cases myocardial infarction occurs. Besides adversely affecting the quality of life, CAD greatly exercises economic and social pressures on various health care systems in nearly all countries around the globe [2].

During the last two years, progress in big data analytics and machine learning has opened up novel avenues to overcome the challenges involved with CAD. Such technology allows the incorporation of large datasets to be reviewed, providing early diagnosis, risk assessment, and prognosis. The paper discusses Principal Component Analysis (PCA) and Support Vector Machine (SVM) methodologies, improving understanding and predictions of CAD. PCA reduces the number of features by retaining the relevant features and minimizing redundancy, and SVM forms an efficient prediction model.

This study employs a dataset containing fundamental predictors for heart disease, such as resting blood pressure, cholesterol concentrations, age, and a variety of clinical characteristics. The focus here is to develop an effective model that can confidently predict the chances of having CAD ensuring early intervention and further improving results for patients.

II. PROBLEM STATEMENT

Coronary Artery Disease is a popular area of healthcare concern as the result of numerous causalities and by-products that are usually associated with this illness [1]. While improvements in diagnostic medicine make it easier to deliver diagnostic information, the nature and amount of clinical information make it difficult to detect CAD in an efficient manner. For the most part, traditional diagnostic processes are characterized by sequential analytical steps that are time-consuming and might not yield the correct results [2].

The current research employs SVM along with PCA to achieve a better solution. After conducting feature extraction by using PCA for the dimensionality reduction procedure, an SVM classifier is used to predict the presence of CAD. Such an approach tries to reduce the length of the diagnostic stack and enhance the prediction accuracy to optimize results and the way of delivering healthcare services.

III. LITERATURE REVIEW

Coronary Artery Disease (CAD) is caused by the deposition of a plaque in coronary arteries causing limited blood supply to the heart and can result in heart attacks, strokes, and other cardiovascular problems. Detecting and predicting CAD is of utmost importance to prevent bad consequences, thus CAD has remained at the epicenter in clinical practice and research.

A. Traditional Methods

Some of the traditional clinical tests used in the detection of CAD include Echocardiograms, Electrocardiograms (ECGs), and Angiography. These frequently require well-trained doctors, costly equipment, and a lot of time. Even with all that, they may not always be accurate, especially in the early stages of disease onset [3].

B. Machine Learning in Healthcare

The Machine Learning (ML) techniques have been given tremendous attention in the health care sector over the last few years because they could deal with large datasets as well as handle patterns that are very complex to be detected by a human. The various models of Machine Learning - Support Vector Machines, Neural Networks, and Random Forest - have been applied for medical diagnosis purposes like cancer detection, disease prediction, and classification of medical images [4].

C. PCA for Dimensionality Reduction

Dimensionality reduction is an important step in the workflow of most machine learning applications, especially dealing with large datasets that contain many features. PCA has been applied in medical research to reduce the complexity of clinical data and retain important information [4].

D. Combination of SVM and PCA for CAD prediction

Several studies have been done on the combination of PCA with SVM for disease predictions. This hybrid combination is advantaging as PCA filters out the redundant variables and reduces the influence of irrelevant features, thus leaving the SVM the process of learning patterns underlying data.

A study done by Rahbre Islam (2024) has applied PCA and SVM using R studio environment to Cardiovascular Disease (CVD) related dataset collected from Rohilkhand Hospital, Uttarpradesh, India with an accuracy of 83.80% and precision of 97.91% [5]. Applying dimensionality reduction helped an SVM classifier to work more effectively with the model for the prediction of disease likelihood with an improvement in performance and efficiency in computations. dataset

IV. DATA PREPARATION

A. Loading and Preprocessing of Data

The heart disease dataset used in this study has been collected from a reliable public platform, Kaggle. For reasons like integrity in the dataset, rigorous pre-processing was carried out including correcting inconsistencies like existence of missing values, outliers, and also noisy data which tend to affect the performance accuracy or the machine learning model produced. The dataset can be cleaned and standardized by pre-processing so that it is well presented for both dimensionality reductions and classification tasks. By adequately preparing data, meaningful insight and increase in the validity of the findings of the research were obtained.

In MATLAB, the dataset is imported from CSV file using readtable( ) function with appropriate imported options. Then the data is converted into a numerical array using table2array command to carry out further analysis.

B. Feature Selection

The dataset consists of data of 1025 patients collected from Cleveland Medical Hospital, Ohio. The clinical parameters considered for dataset contribute to the diagnosis and risk evaluation in patients with CAD.

14 features are chosen based on the level of risk each feature contributes. They are:

1. Age
2. Sex
3. Chest pain type (cp)
4. Resting blood pressure (trestbps)
5. Serum cholesterol in mg/dl (chol)
6. Fasting blood sugar (fbs)
7. Resting electrocardiographic result (restecg)
8. Maximum heart rate achieved (thalach)
9. Exercise induced angina (exang)
10. ST depression induced by exercise relative to rest (oldpeak)
11. Slope of the peak exercise ST segment (slope)
12. Number of major vessels colored by fluoroscopy (ca)
13. Inherited blood disorder, Thallasemia (thal)
14. Buildup of plaque in the walls of blood vessels (target)

C. Data Standardization

The data is standardized using zscore( ) so that all features contribute equally, irrespective of their scale, by setting it to have mean of 0 and standard deviation of 1. This is extremely important because features might have different units or scales.

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak
-0.26831	0.66118	-0.91531	-0.37745	-0.65901	-0.41867	0.89082	0.82092	-0.71194	-0.060859
-0.15808	0.66118	-0.91531	0.47887	-0.83345	2.3862	-1.0036	0.25584	1.4032	1.7263
1.7158	0.66118	-0.91531	0.76432	-1.3956	-0.41867	0.89082	-1.0482	1.4032	1.3008
0.72373	0.66118	-0.91531	0.93558	-0.83345	-0.41867	0.89082	0.51665	-0.71194	-0.91188
0.83395	-1.511	-0.91531	0.3647	0.93037	2.3862	0.89082	-1.8741	-0.71194	0.70506
0.39305	-1.511	-0.91531	-1.8047	0.038765	-0.41867	-1.0036	-1.1786	-0.71194	-0.060859
0.39305	0.66118	-0.91531	-1.0054	1.3956	-0.41867	2.7852	-0.39617	-0.71194	2.8326
0.062372	0.66118	-0.91531	1.6206	0.83345	-0.41867	-1.0036	-0.17883	1.4032	-0.23106
-0.92966	0.66118	-0.91531	-0.66289	0.058148	-0.41867	-1.0036	-0.2223	-0.71194	-0.23106
-0.047854	0.66118	-0.91531	-0.54872	0.77531	-0.41867	-1.0036	-1.4394	1.4032	1.8114
1.826	-1.511	-0.91531	-1.1196	-1.8801	-0.41867	0.89082	-1.0482	-0.71194	0.44976
-1.2603	-1.511	-0.91531	0.022167	1.8414	2.3862	-1.0036	-0.57004	1.4032	1.6412
-2.2524	-1.511	0.055904	-0.77707	-0.69778	-0.41867	0.89082	1.8641	-0.71194	-0.31617
-0.37853	0.66118	-0.91531	0.47887	1.0079	-0.41867	0.89082	-1.1786	1.4032	2.6624

TABLE 1. Standardized data

V. PCA- IMPLEMENTATION AND INTERPRETATION

A. Computing the Covariance Matrix

Once the data is standardized, the covariance matrix is computed to capture the relationship between the features. The covariance matrix is significant for determining how the information varies and correlates.

covMatrix = *14×14*

1.0000	-0.1032	-0.0720	0.2711	0.2198	0.1212	-0.1327	-0.3902	0.0882	0.2081	-0.1691 ...
-0.1032	1.0000	-0.0411	-0.0790	-0.1983	0.0272	-0.0551	-0.0494	0.1392	0.0847	-0.0267
-0.0720	-0.0411	1.0000	0.0382	-0.0816	0.0793	0.0436	0.3068	-0.4015	-0.1747	0.1316
0.2711	-0.0790	0.0382	1.0000	0.1280	0.1818	-0.1238	-0.0393	0.0612	0.1874	-0.1204
0.2198	-0.1983	-0.0816	0.1280	1.0000	0.0269	-0.1474	-0.0218	0.0674	0.0649	-0.0142
0.1212	0.0272	0.0793	0.1818	0.0269	1.0000	-0.1041	-0.0089	0.0493	0.0109	-0.0619
-0.1327	-0.0551	0.0436	-0.1238	-0.1474	-0.1041	1.0000	0.0484	-0.0656	-0.0501	0.0861
-0.3902	-0.0494	0.3068	-0.0393	-0.0218	-0.0089	0.0484	1.0000	-0.3803	-0.3498	0.3953
0.0882	0.1392	-0.4015	0.0612	0.0674	0.0493	-0.0656	-0.3803	1.0000	0.3108	-0.2673
0.2081	0.0847	-0.1747	0.1874	0.0649	0.0109	-0.0501	-0.3498	0.3108	1.0000	-0.5752
⋮										

Covariance between 2 features is represented by the covariance matrix if each of the matrix is computed and diagonal entries are the variance of individual features.

- Diagonal Entries (Variance)- All diagonal elements are 1 because of standardization which means each feature has unit variance.
- Off- Diagonal Entries (Covariance)- Positive values indicates that the features increase/decrease together (Positive Correlation). Negative values indicates that a feature increases while the other decreases (Negative Correlation)

B. Eigenvalues and Variance

Eigenvalues and eigenvectors are computed from the covariance matrix using the eig( ) function. The Eigenvalues represent the amount of explained variance by each principal component, whereas eigenvectors define the direction of these components.

The eigenvalues are ordered in descending , and the eigenvectors are rearranged with respect to the corresponding eigenvalue. This is necessary to keep the components that explain the most variance.

eigenValuesSorted = *14×1*

3.3137
1.5882
1.2304
1.1793
0.9992
0.9727
0.8764
0.7680
0.7338
0.6334
⋮

The larger values donate components with more information, while the smaller ones carry less, which will help prioritize how many to retain in dimensional reduction.

The explained variance helps to indicate how much each component extracts from the dataset.

23.6696
11.3443
8.7889
8.4238
7.1372
6.9481
6.2602
5.4859
5.2417
4.5244
3.7724
3.1059
2.6616
2.6360

The first few components explain the most variance. This can be used to reduce the dimensions while preserving most of the data's information.

Scree plot is created to visually assess the explained variance for each principal component. This helps to decide how many components to keep for further analysis.

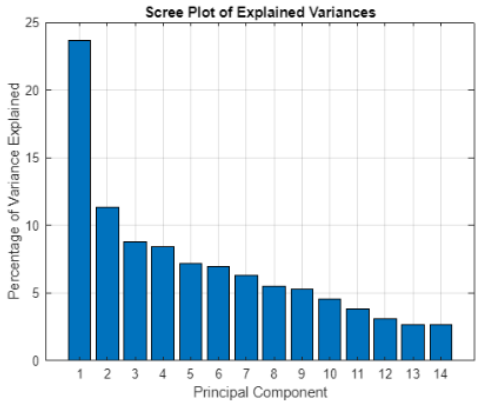


FIGURE 1. Scree Plot

In the scree plot shown in **FIG 1**, each bar height represents the percentage of variance that each principal component explains, helping to visualize how much information each component contributes.

D. Projection of Data

The data is projected onto the principal components that is the selected eigenvectors to reduce its dimensionality of data. The new subspace is defined by using the top k eigenvectors. Loadings for the top k principal components are gotten from the sorted eigenvectors. These loadings are put in a table that associates each feature with its contribution to each principal component.

PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	Feature
-0.25205	0.4347	0.037196	0.078528	0.29112	0.20012	0.2496	-0.22893	{'age' }
-0.11262	-0.39753	-0.49836	-0.27938	-0.053534	-0.027449	0.1844	-0.099757	{'sex' }
0.28192	0.24459	-0.068129	-0.42916	-0.18522	0.23581	0.21616	0.1243	{'cp' }
-0.14308	0.44536	-0.12251	-0.17058	-0.24017	0.14411	-0.31034	-0.62732	{'trestbps'}
-0.10219	0.36839	0.0034681	0.53691	-0.30666	0.0048188	-0.061769	0.4127	{'chol' }
-0.060366	0.3123	-0.3407	-0.34897	0.23077	-0.28528	-0.51806	0.36502	{'fbs' }
0.11269	-0.23758	0.28582	-0.055665	0.27914	0.6364	-0.51121	0.076563	{'restecg' }
0.36777	0.018254	-0.27871	0.023163	-0.33294	0.06078	-0.16516	0.13249	{'thalach' }
-0.33773	-0.20368	0.075175	0.10972	-0.021375	-0.31921	-0.38615	-0.13039	{'exang' }
-0.37337	0.0031423	0.22398	-0.28809	-0.25922	0.16711	0.054699	0.2108	{'oldpeak' }
0.32634	-0.018579	-0.37106	0.39013	0.23438	0.057125	-0.085846	-0.25502	{'slope' }
-0.25452	0.10354	-0.36492	0.05594	0.43087	0.29287	0.17409	0.27348	{'ca' }
-0.21589	-0.17826	-0.31566	0.1861	-0.4196	0.41647	-0.10027	0.017689	{'thal' }
0.43639	0.14307	0.16977	-0.061736	-0.0060274	-0.020432	-0.00049478	0.010183	{'target' }

TABLE 2. Projected Data

The major 8 components captures the most significant variance in the data.

E. Scatter Plot

A scatter plot is used to visualize the data after projection onto the first two principal components. Visualization of data distribution in the first two principal components is done with the help of scatter plot.

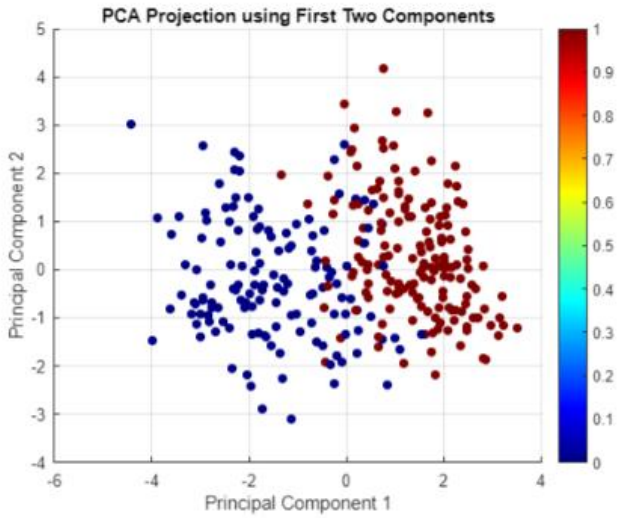


FIGURE 2. Scatter Plot

In the scatter plot shown in **FIG 2**, each point is a sample from the data, and colors denote different class labels.

F. Biplot for Component Loadings Interpretation

A biplot is generated to visualize both the data projections and the loadings of the features on the principal components. This indicates how each feature contributes to the elaboration of each principal component. The biplot combines the scores of observations with loadings of variables so that one can try to understand clustering, variable contributions, and data structure.

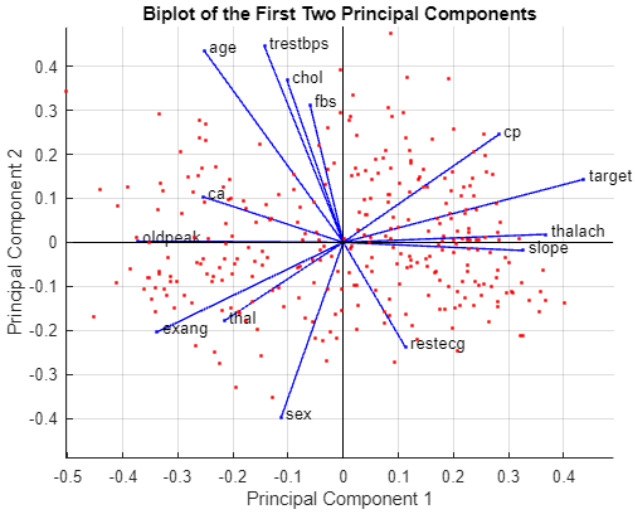


FIGURE 3. Biplot of first 2 principal components

In the biplot shown in **FIG 3,** points that lie close together refer similar patterns, outliers differ. Variable vectors reflect contributions to principal components: the direction indicates influence, and the length shows the strength. The angle between the vectors expresses the correlations among variables: acute for positive, right for none and obtuse for negative ones. The long vectors help interpret the components like risk factors for health for example, age, cholesterol, blood pressure etc.

Overall, this biplot has effectively summarized data structure showing how well the first two components capture variance.

G. Cumulative Variance Plot

A cumulative plot is formed to show the amount of total variance explained by increasing numbers of principal components.

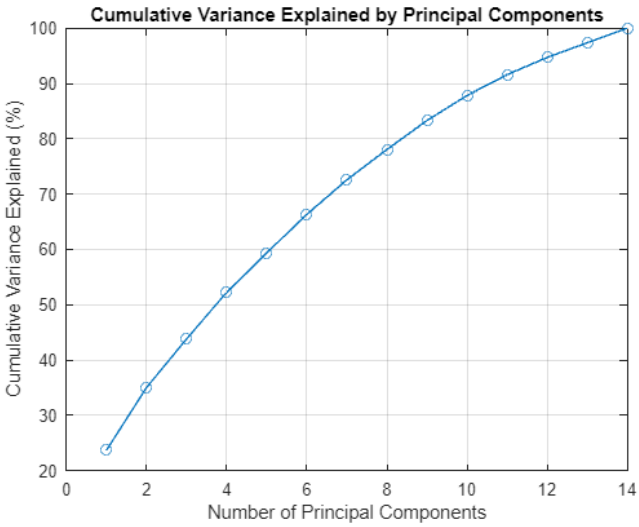


FIGURE 4. Cumulative Variance Plot

The cumulative plot shown in **FIG 4,** is used to decide how many principal components to keep , balancing simplicity and retention of information. In this case, a variance threshold of 75% is considered and the number of components required to reach that level is 8.

H. Major Features from 8 Components

```
Top features for Principal Component 1:
target (Loading: 0.4364)
oldpeak (Loading: -0.3734)
Top features for Principal Component 2:
trestbps (Loading: 0.4454)
age (Loading: 0.4347)
Top features for Principal Component 3:
sex (Loading: -0.4984)
slope (Loading: -0.3711)
Top features for Principal Component 4:
chol (Loading: 0.5369)
cp (Loading: -0.4292)
Top features for Principal Component 5:
ca (Loading: 0.4309)
thal (Loading: -0.4196)
Top features for Principal Component 6:
restecg (Loading: 0.6364)
thal (Loading: 0.4165)
Top features for Principal Component 7:
fbs (Loading: -0.5181)
restecg (Loading: -0.5112)
Top features for Principal Component 8:
```

The output shown above, infers that the most important features that can be considered as major major health factors are :

- Trestbps
- chol
- chest pain
- age
- sex

- thalassemia
- old Peak
- restecg

Overall, PCA effectively reduced the dimensions of the data, retained the important features. 14 features of the dataset got reduced to 8 principal components, preserving a variance of 75%. PCA helped in delivering important conclusions of major factors, leading to a more efficient and accurate analysis for predicting heart disease.

Further in SVM,1 feature is selected out of the 14 features to act as a decision boundary to determine whether the person have heart disease or not.

VI. SVM- IMPLEMENTATION AND INTERPRETATION

A. Choosing the Kernel

The choice of the kernel function in Support Vector Machines is quite crucial, especially to the performance of the model in the event that there is no linear separability on the initial feature space of the data. Kernel functions assist in mapping the data to a higher dimensional space so that even complicated datasets may realize an achievement of a separate hyperplane [6].

In this study, the linear and RBF kernels are used. They fulfill their respective purposes, considering the nature of the dataset and the preprocessing performed on the data.

1. Linear Kernel- The lightest and the least computationally costly of the methods, particularly with already linearly separable data. In this case, the dataset has some dimensionality reduction through PCA, which means it can have the original form but the complexity can be reduced without losing the characteristics originally present. The preprocessing achieves straightforwardness through the use of a linear kernel while avoiding too much complexity, which could otherwise contribute to unwanted complexity in data classification.
   The linear kernel computes a simple dot product between feature vectors, making it both interpretable and computationally efficient in terms of resource usage.
   This method is very useful when the data has already been preprocessed to eliminate any non-linearity, and this paper does that by using PCA.

2. Radial Basis Function (rbf) Kernel- This kernel is very versatile and highly used because it outperforms others in the high-dimensional data space with handling nonlinear relationships. It maps the data points to an infinite-dimensional space and SVM captures subtle patterns as well as complex relationships that the linear kernel would miss.
   Though the rbf kernel is more computationally expensive, but is quite flexible and can deal with very high non-linearity data distributions [7].

B. Training SVM

Concept of training and testing data [8]:

- The training data is approximately 70–80% of the set, can be used to train the model with adjustments to weights or parameters. Its aim is to reduce errors and help the model generalize patterns effectively. Direct interaction happens with the model in its training phase, and metrics such as training accuracy and loss can be observed at this point to monitor learning progress.
- The test data, typically 20–30% of the overall dataset, is further used to validate the model's goodness in unseen data. The model does not have access to such data while training; this eliminates biasing of an evaluation. The common metrics used to check how well a model generalizes include accuracy, precision, recall, F1-score, and the confusion matrix.

Splitting of data:

The heart disease dataset has 1,025 entries. This dataset splits into 70% training and 30% testing. Therefore, train-test splitting ratio is 70:30

- Training Set Size

$$70/100 \times 1025 = 717.5 \approx 717 \text{ entries}$$

- Testing Set Size

$$30/100 \times 1025 = 307.5 \approx 308 \text{ entries}$$

This split ensures that the model gets enough data for training but also has a separate set to test the performance of the model.

Tuning of Hyperparameters:

The hyperparameters used in this study are:

- C Parameter- This C parameter (also known as regularization parameter) does the job to balance between achieving low error on training set and keeping a simple generalized decision boundary. A smaller C allows the model to have some misclassifications in training data as it pushes model towards a simpler more generalized one. As opposed, a high C will make the model care about classifying all the training data correctly which could result in overfitting should the model start picking up on noise as well.

- Gamma parameter- gamma controls the influence of single data points. A lower gamma indicates that points no matter how distant can still influence the decision boundary, leading to more generalized but less accurate classification. On the other hand, gamma of high provide data points more power locally, so allows the model to learn complicated pattern in the dataset. However, this increased of complexity may also bring overfitting where the model memorizes the training data.

The interaction of C and gamma is a really important one when it comes to SVM with the RBF kernel. Setting both too high the model will overfit the training data, and those with low values for both can regress since the model will fail to fit the underlying patterns [8].

Tuning of hyperparameter is performed using GridSearchCV technique in python. Then the best model is selected using grid_search_rbf.best_estimator command and is trained using the fit( ) method. The trained model is used to predict the labels on the test set, and the accuracy of the model is computed.

C. Model Evaluation

- Accuracy score- The accuracy score for all models is calculated by comparing the predicted labels with the labels in the test set.
- Classification Report- For each model, a detailed classification report is created that includes precision, recall, F1-score, and support for each class.
- Confusion Matrix- For each model, a confusion matrix is printed to visualize the performance for true positives, false positives, etc.
- Receiver Operating Curve (ROC)- The ROC curve is plotted to assess the model's balance between true positive rate (TPR) and false positive rate (FPR).
- Precision-Recall Curve- This curve is plotted to find out how correctly the model is identifying heart disease patients.
- Decision Boundary for linear SVM- The decision boundary plot is first created for the linear kernel SVM by using a meshgrid.
- Decision Boundary for rbf SVM- Similarly, the decision boundary plot is created for the rbf kernel SVM by using a meshgrid.
- Accuracy Comparison Bar Plot- The accuracy comparison bar plot is plotted to compare the accuracy of each model.

VII. RESULT AND DISCUSSION

The SVM classifier on the heart disease dataset may lead to the following conclusions on model performance:

A. Accuracy Achieved

Accuracy indeed represents a basic metric of consideration which is concerned with the number of test units correctly predicted that were either diseased or healthy. An accuracy level of 85%, for instance, displays the proportion of test data average error of 15%.

- Reliable performance metrics imply that ignoring the goal of accuracy more than 85% reflects a deep understanding of the SVM model patterns within the data and ability to perform heart diseased patients classification.
- Loss on accuracy may imply that the features present in the dataset do not provide strong enough classification power or the level of data preprocessing or feature engineering or feature extraction is inadequate.

In this study, the accuracy on the reduced dataset after applying SVM (rbf kernel) is 92.21% which means that the result is highly reliable.

B. Classification Report

The classification report offers the precision, recall, and F1 score for each class: Heart Disease or No Heart Disease [9]. These measurements will give insight into which areas your model is particularly strong/weak:

- Precision: The ratio of correct positive predictions out of all the positive predictions. High precision means that when the model predicts you have heart disease, it is likely correct.
  (i) Class 0 (No heart disease) Precision = 0.91 That is, 91% of predictions for no disease are correct.
  (ii) Class 1 (Heart disease) Precision = 0.94 That is, 94% of predictions for disease are correct.

- Recall (Sensitivity): The ratio of true positive predictions to all the actual positive cases. High recall implies that the model has caught most of the real instances of heart disease.
  (i) Class 0: Recall = 0.94 (94% of true no-disease cases are correctly classified).
  (ii) Class 1: Recall = 0.90 (90% of true disease cases are correctly classified).

- F1 Score: Harmonic mean of precision and recall. A high F1 score indicates high precision and recall together, making it a good measure in absolute terms. Classification performance.
  (i) For class 0: F1-score = 0.93
  (ii) For class 1: F1-score = 0.92

- Support: The actual number of occurrences of each class in the data set:
  (i) 159 no disease samples of class 0
  (ii) 149 disease samples of class 1

Overall Performance

- Macro Average: This average is over all precision, recall, and F1-score by considering each class equally. Here, it is 0.92 for all metrics, which means it is balanced over both classes.
- Weighted Average: The weighted average considers the support associated with each class while averaging the metrics. Again, it is 0.92 for all metrics, and this says that the classifier is performing consistently over the two classes.

C. Confusion Matrix

The Confusion matrix indicates how many instances it has correctly predicted of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

- True Positives (TP): Prediction of right cases of heart disease. In this study, there are 150 cases which are correctly predicted
- False Positives (FP): Healthy patients are predicted to be diseased. In this study, there are 15 healthy patients who was misdiagnosed
- True Negatives (TN): Right prediction of healthy persons. In this study, there are 134 rightly predicted cases

- False Negatives (FN): Heart disease patients are wrongfully predicted as healthy. In this study, there are 9 patients who are misdiagnosed as healthy.

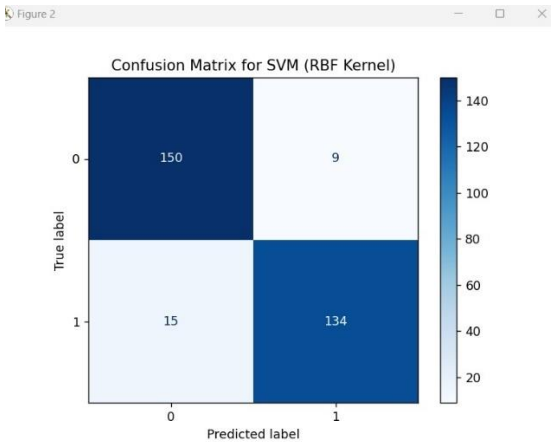


FIGURE 5. Confusion Matrix

D. Receiver Operating Curve

The Receiver Operating Characteristic (ROC) curve graphs the True Positive Rate (TPR) versus the False Positive Rate (FPR) [10]. Area Under the Curve (AUC) would then give you an idea about how well the model is able to differentiate between the two classes (heart disease vs. no disease).

- AUC = 0.5 means the model is no better than random guessing.
- AUC = 1.0 means perfect classification.
- AUC > 0.7 usually means a reasonably good model.

The closer the ROC curve is to the top-left corner, the better the model is at distinguishing between patients with heart disease and healthy individuals.

In this study, AUC – 0.95 which indicates that the model is reasonably good.

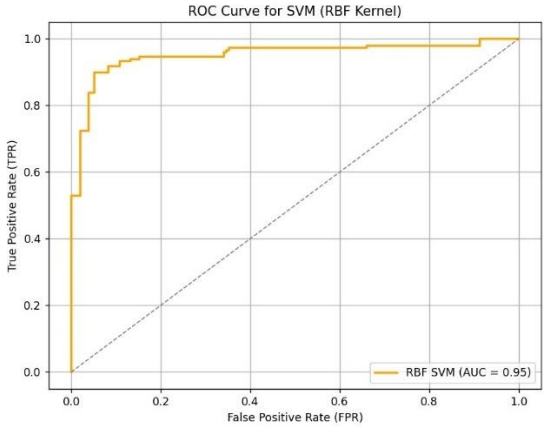


FIGURE 6. ROC Curve

E. Precision-Recall Curve

Precision-Recall Curve is very useful for datasets which does not have a good balance, where once class is represented much more frequent than the other (e.g., the no. of patients with heart disease can be significantly larger than the number of patients with heart disease).

- High average precision means the model is good at identifying heart disease patients without generating too many wrongly predicted values
- If the given dataset contains a class imbalance, this curve would be even more relevant as the emphasis is on the performance in relation to the minority class of heart diseases.

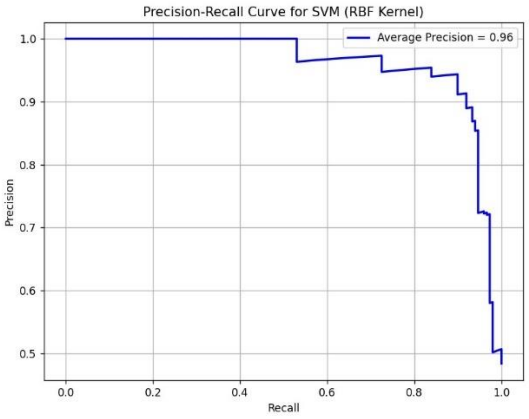
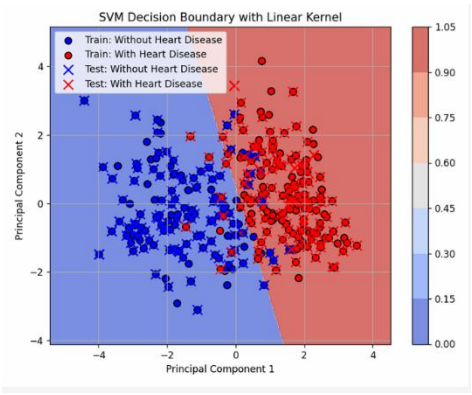


FIGURE 7. Precision-Recall Curve

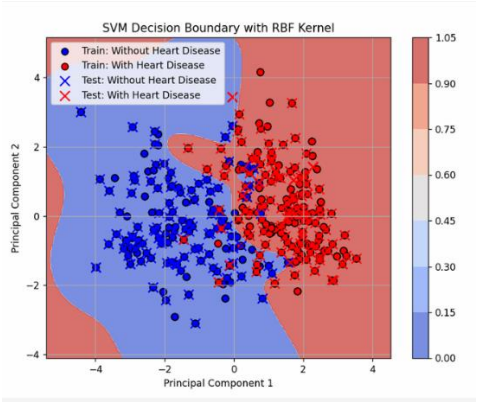
The graph shown in **FIG 7,** gives the value of average precision as 0.96 which means that the model is highly efficient at identifying patients with heart disease.

F. Visualization of Decision Boundary

Visualizing the decision boundary provides the level of detail about how the SVM classifier is partitioning the data [11]. 2 major principal components are chosen from the data, this visualization depicts how the SVM decision function is splitting the two classes (with heart disease and without heart disease).



(a)



(b)

FIGURE 8. Decision boundary with linear kernel. **FIGURE 9.** Decision boundary with rbf kernel

Comparison between **FIG 8** and **FIG 9** infers that SVM Decision Boundary with rbf kernel can separate the data more efficiently.

G. Comparison of SVM (linear, rbf) and KNN Accuracies

Concept of KNN: The k-Nearest Neighbors algorithm is one of the simple and non-parametric approaches to handle classification and regression tasks with machine learning. It works on finding k nearest data points neighbors

closest to any query point based on some distance metric, and one of the most common distance metrics used is Euclidean distance. The algorithm assigns the query point to the class most common among its k neighbors in classification. For regression, it predicts the output as the average of the values of its k neighbors.

In KNN all the calculation is deferred until a prediction is needed, rather than building a model during the training phase. KNN is easy to implement and interpret, but could be computationally expensive on larger datasets and sensitive to the choice of k and to the scaling of the features. Proper preprocessing, such as normalization, is important for optimal performance.

Here is a comparison to check which method is more efficient to categorize the data.

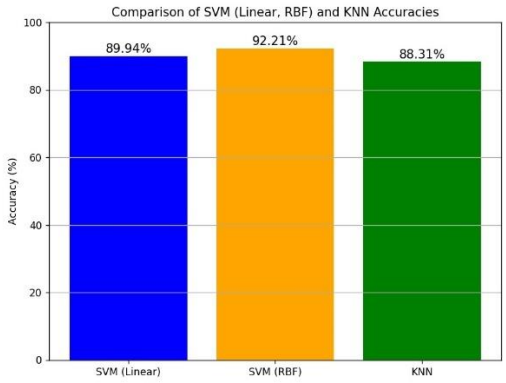


FIGURE 10. Comparison bar plot of SVM (linear, rbf) and KNN

The comparison in **FIG 10**, shows that SVM (rbf) behaves most efficiently and also excels the accuracy of KNN.

VI. ADVANTAGES

PCA and SVM have many benefits when are combined together, especially for high-dimensional data (used in our implementation) [12].

1) Dimensionality Reduction- PCA reduces the number of features while keeping most of the variance. This makes the data less complex and reduces overfitting especially if there are a small number of samples in the training set.
2) Noise Reduction and Redundancy Removal- Inevitably, the SVM could be considered to focus on only the most essential patterns for classification by denoising and removing redundancy intrinsically via projecting data in lower-dimensional space through PCA.
3) Improved Efficiency- As the computational complexity increases with dimensions, SVM run faster and more efficient in the small feature space.
4) Robustness Against Multicollinearity- PCA is a solution for high-dimensional datasets that often encounter multicollinearity (correlated features) by creating orthogonal principal components on which regression can now be applied
5) Improved visualization and interpretation ability- Visualization of the separation of classes with a 2D decision boundary after reducing the dataset to two or three dimensions; this is what was done in this implementation
6) Improved Generalization of Model- At the same time, PCA serves to make sure that SVM is able to generalize well on unseen data since principal components are those directions in feature space which usually represent patterns which tend to persist on test data as well.

This Python implementation of PCA and SVM under their respective framework combines their strengths to produce a classifier that is computationally efficient and interpretable yet not merely astronomically accurate.

VII. FUTURE SCOPE

1. Integration with Advanced Technologies
    - AI and Deep Learning : PCA and SVM can integrate deep learning algorithms such as Convolutional Neural Networks which would further enhance the CAD diagnosis from medical images electrocardiograms (ECGs).
    - Real-Time Decision Support: The integration of CAD diagnostic models into AI-healthcare platforms enables physicians to make faster decisions and find at-risk patients in real time.
    - Precision Medicine: The diagnosis for CAD can eventually go into precision medicine, where the most important features for which PCA and SVM suggested provide the maximum personalized care plan based on specific risk factors [14].

2) Big Data and Enhanced Model Training
    - Increased availability of health data will provide for large-scale data to be used in the training of deeper models. Electronic health record data, and even genetic data and monitoring systems, may help constitute a better view of CAD risk.
    - Transfer Learning: The knowledge learned from the deep datasets in one region can be applied to operate in a different region or patient population, and hence leads to higher generalizability. More redundant CAD risk diagnostic models based on PCA and SVM can be trained on deeper cross-region, cross-population datasets for better redundancy and scalability.
    - Further refinement of the best features can be achieved with additional data modalities, such as genetic and lifestyle information.

3) Ethical Implications and Reducing Bias
    - There is an absolute need to avoid bias in the SVM model across any demographical parameters. As models are developed, ensuring fair distribution of various groups of populations should not be overlooked. When AI is increasingly incorporated into healthcare, increasing interpretability and the ability to explain models to clinicians will be important to the trusted adoption of AI [15].

VIII. CONCLUSION

In this study, implementation of Principal Component Analysis (PCA) and Support Vector Machine (SVM) on heart disease dataset is shown successfully. PCA enabled the reduction of dimensionality of dataset while keeping most of its variance, which improved the visualization and gave us computational efficiency. The reduced dataset has allowed the SVM model to properly accomplish its learning with an accuracy rate of 92.21%.

Analyzing the model further with the confusion matrix confirmed that model was robustly able to correctly classify data points, its strengths and weaknesses. Further evaluation metrics in the form of precision, recall and F1 score validated the classifier's ability to separate diseased from physically non diseased cases with few cases wrongly classified.

Combination of PCA and SVM for predictive modeling in clinical applications is possible as these findings suggest. The model could be further improved in future work by integrating advanced techniques such as deep learning frameworks and robustness testing on larger, more diverse datasets for scale and generalization to the true greybox environment of many real world healthcare applications.

IX. ACKNOWLEDGEMENT

The authors would like to extend their deepest gratitude to Dr. Geetha K N Ma’am and Dr. Shali S Ma’am for their guidance in mathematics throughout the semester. Their cooperation and guidance were instrumental in steering this project towards completion. In every phase of this project, their supervision and guidance played an important role in shaping this report to be completed perfectly.

X. REFERENCES

1. Karen Okrainec, Devi K Banerjee, Mark J Eisenberg,Coronary artery disease in the developing world American Heart Journal,Volume 148, Issue 1,2004.
2. Ganeshkumar, M., V. Sowmya, E. A. Gopalakrishnan, and K. P. Soman. "Disease prediction mechanisms on large-scale big data with explainable deep learning models for multi-label classification problems in

healthcare." Healthcare Big Data Analytics: Computational Optimization and Cohesive Approaches, Amrita School of Engineering, 10 (2024): 207.
3. McCullough, Peter A.. Coronary Artery Disease. Clinical Journal of the American Society of Nephrology 2(3):p 611-616, May 2007.
4. Santhanam, T., & Ephzibah, E. P. (2013). Heart disease classification using PCA and feed forward neural networks. In Mining Intelligence and Knowledge.
5. Rahbre Islam., Safdar Tanweer, Imran Hussain., & Shahazad, A. B. (2024, March). Prognosis of Cardiovascular Disease using machine learning. In 2024 Advances in Science and Engineering Technology International Conferences (ASET) (pp. 1-6). IEEE.
6. Guido, R.; Ferrisi, S.; Lofaro, D.; Conforti, D. An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: Review. Information 2024, 15, 235.
7. Shawe-Taylor, J., & Cristianini, N. (2004). Kernel methods for pattern analysis. Cambridge University Press
8. Evgeniou, Theodoros & Pontil, Massimiliano. (2001). Support Vector Machines: Theory and Applications. 2049. 249-257. 10.1007/3-540-44673-7_12.
9. Dun, B., Wang, E., & Majumder, S. J. C. S. (2016). Heart disease diagnosis on medical data using ensemble learning. Comput. Sci, 1(1), 1-5.
10. Narkhede, S. (2018). Understanding auc-roc curve. Towards Data Science, 26(1), 220- 227.
11. Reddy, O. C., I. D. Kumar, P. Sathvika, V. V. Sajith Variyar, V. Sowmya, and R. Sivanpillai. "Effect of Hyperparameters on DEEPLABV3+ Performance to Segment Water Bodies in RGB Images." The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 48, Amrita School of Engineering. (2023): 203-209.
12. Jaswanth, K., Nikhil, M. T., Siddhartha, M. S. S., & Jayan, S.Comparative Analysis of Covid Detection using Chest X-rays by SVM-PCA and Deep Learning Techniques. In 2023 International Conference on Advances in Electronics, Communication, Computing and Intelligent Information Systems (ICAECIS) (pp. 188-193). IEEE. (2023, April).
13. Phogat, Divith, Dilip Parasu, Arun Prakash, and V. Sowmya. "Selective Kernel Networks for Lung Abnormality Diagnosis Using Chest X-rays." In International Conference on Information, Communication and Computing Technology, pp. 937-950. Amrita School of Engineering, 2023.
14. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297.
15. Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2(2), 121-167.