Predicting Cardiovascular Disease Using Supervised Machine Learning Techniques: A Comparative Study

Nur Alam (1189508)

Department of Computer Science,

Lakehead University

Abstract—**Cardiovascular disease (CVD) remains one of the leading causes of death worldwide, presenting a major challenge to healthcare systems. Early prediction and prevention are critical to reducing mortality and improving patient outcomes. This study applies four supervised machine learning algorithms—Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)—to predict cardiovascular disease using a large-scale dataset of patient health records from Kaggle. Comprehensive data preprocessing, feature engineering, and dimensionality reduction through Principal Component Analysis (PCA) were performed to enhance model accuracy and efficiency. Experimental results show that SVM achieved the highest prediction accuracy of 71% before feature selection, while Random Forest and KNN both reached 69% after feature selection. The findings demonstrate the potential of machine learning techniques in improving early CVD detection and supporting data-driven preventive healthcare strategies.**

Keywords: Cardiovascular Disease, Machine Learning, Random Forest, Decision Tree, SVM, KNN, PCA, Feature Engineering, Classification.

I. INTRODUCTION

Machine learning (ML) has emerged as a powerful tool for advancing medical diagnostics and predictive analytics. In cardiovascular healthcare, ML enables data-driven identification of subtle risk factors that may not be captured through conventional statistical methods. The growing availability of large-scale electronic health records and patient monitoring data has opened opportunities to model complex relationships among physiological, behavioral, and demographic variables influencing cardiovascular disease (CVD) [1] [2].

Traditional diagnostic approaches often rely on limited or linear assumptions, which restrict their predictive capability in capturing nonlinear patterns inherent in medical data. In contrast, ML algorithms such as Decision Trees, Random Forests, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN) can process heterogeneous data and uncover underlying trends to support accurate clinical decision-making.

Recent research has demonstrated the importance of integrating feature engineering and dimensionality reduction techniques to improve model robustness, interpretability, and computational efficiency. Incorporating derived features such as Body Mass Index (BMI), blood pressure categories, and pulse pressure, along with dimensionality reduction using Principal Component Analysis (PCA), can enhance prediction accuracy and generalization.

This study applies four supervised ML algorithms—Decision Tree, Random Forest, SVM, and KNN—on a publicly available Kaggle dataset comprising 70,000 patient health records [3]. Through a systematic process involving data preprocessing, feature engineering, and PCA, the research aims to identify the most effective algorithm for predicting cardiovascular disease. The findings contribute to developing interpretable, data-driven models that can support healthcare professionals in early detection and preventive treatment strategies.

II. LITERATURE REVIEW

A growing body of research has demonstrated the potential of machine learning (ML) for early prediction and prevention of cardiovascular disease (CVD). Traditional statistical methods, such as logistic regression, have been applied to model CVD risk but often struggle to capture complex

nonlinear relationships among multiple risk factors, as discussed by Liu et al.[4]

Alizadehsani et al. [5] compared several classifiers, including Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbors (KNN), on a dataset of 303 patients, reporting that SVM achieved the highest accuracy of 89%. Their findings emphasized the importance of kernel optimization and feature normalization in improving predictive performance. Similarly, Detrano et al. [6] highlighted the predictive significance of coronary calcium scores when integrated with demographic and clinical attributes, demonstrating how combined feature sets improve diagnostic precision.

Tree-based models such as Decision Trees (DT) and Random Forests (RF) are widely used due to their interpretability and ability to handle heterogeneous datasets. Pandey et al. [7] noted that DTs provide clear visualization of decision rules but are prone to overfitting, particularly on small datasets. Thenmozhi et al. [8] demonstrated that RF mitigates this limitation by aggregating multiple trees, thereby improving generalization and predictive accuracy. Support Vector Machines (SVM) remain competitive in high-dimensional settings. Escamila et al. [9] showed that SVM with kernel tuning and proper feature preprocessing can achieve high accuracy in cardiovascular datasets. K-Nearest Neighbors (KNN) and Artificial Neural Networks (ANN) have also been explored; Jadhav et al. [10] highlight that KNN is sensitive to feature scaling, while Abhijit et al. [10] indicate that ANN can capture complex nonlinear relationships but require large datasets and significant computational resources.

Feature engineering is critical for enhancing model performance. Ramalingam et al. [11] emphasized incorporating physiological variables (BMI, blood pressure, glucose) alongside lifestyle factors (smoking, alcohol use, physical activity) to improve model discriminative power. Dimensionality reduction techniques such as Principal Component Analysis (PCA) have been widely applied. Shah et al. [12] demonstrated that PCA reduces high-dimensional datasets, mitigates multicollinearity, and preserves predictive information. Recent studies highlight the value of large-scale datasets and ensemble methods. Anna Karen et al. [9] applied machine learning to electronic health records for CVD risk prediction, emphasizing preprocessing and feature selection. Wajid et al. [13] showed that ensemble models combining machine learning and deep learning techniques improve predictive accuracy. Alim et al. [14] confirmed that ensemble approaches enhance classifier robustness for complex cardiovascular datasets. Building on these findings, this study leverages the large-scale Kaggle CVD dataset [3] to evaluate multiple supervised classifiers—Decision Tree, Random Forest, SVM, and KNN—augmented with feature engineering and PCA. This allows a systematic comparison of classifier performance while optimizing predictive accuracy through advanced feature transformations.

III. Methodology

The methodology of this study outlines the systematic approach used to develop and evaluate machine learning models for cardiovascular disease prediction. The workflow consists of five key phases: data acquisition, data preprocessing, feature engineering, dimensionality reduction, and model training and evaluation

A. Dataset

The dataset[3] used in this study contains comprehensive health information for 70,000 individuals, each represented by a unique identifier and 11 features, along with a target variable (cardio) indicating the presence (1) or absence (0) of cardiovascular disease (CVD).

Key attributes include:

- Age (in days) - converted to years for interpretability; a critical determinant of cardiovascular risk.
- Gender - categorical variable used to analyze potential health disparities. Height and Weight - used to calculate Body Mass Index (BMI), an important indicator of CVD risk related to obesity.
- Systolic (ap_hi) and Diastolic (ap_lo) Blood Pressure - essential indicators of hypertension, a major contributor to cardiovascular conditions.
- Cholesterol and Glucose Levels - categorical features reflecting metabolic health.
- Smoking (smoke) and Alcohol Consumption (alco) - binary lifestyle factors linked to elevated cardiovascular risk.
- Physical Activity (active) - denotes exercise engagement, a protective factor against CVD.

- Cardio - the target variable representing CVD presence or absence.

This dataset provides a balanced combination of demographic, clinical, and lifestyle-related variables, enabling a comprehensive analysis of factors influencing cardiovascular disease risk and supporting the development of predictive machine learning models for early diagnosis and prevention.

B. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a fundamental step in understanding the characteristics of a dataset prior to modeling. Its primary purpose is to reveal patterns, detect anomalies, examine variable distributions, and provide insights that guide subsequent preprocessing, feature engineering, and model selection. In this study, EDA was performed using descriptive statistics and visualizations to gain a comprehensive understanding of the cardiovascular dataset and its key attributes.

1) Descriptive Statistics of the Dataset: The dataset comprises 70,000 records, each containing demographic, clinical, and lifestyle-related features, along with a binary target variable indicating the presence or absence of cardiovascular disease (CVD). The 'age' attribute, originally recorded in days, has a mean of approximately 19,468.87 days with a standard deviation of 2,467.25 days, reflecting a moderately wide age distribution across the population. The 'gender' variable, encoded as 1 for female and 2 for male, shows a mean value of 1.35, indicating a slightly higher representation of female individuals. Physical measurements reveal a mean height of 164.36 cm and a mean weight of 74.21 kg, while systolic (ap_hi) and diastolic (ap_lo) blood pressure readings exhibit significant variability, suggesting potential outliers in the data. Cholesterol and glucose levels are predominantly in the normal range, and lifestyle indicators, including smoking, alcohol consumption, and physical activity, capture important behavioral factors associated with cardiovascular health. The target variable 'cardio' is fairly balanced across classes, facilitating reliable predictive modeling.

These statistics are illustrated in Fig. 1, which provides a visual summary of the distribution, central tendencies, and spread of each feature, laying the groundwork for identifying anomalies and guiding subsequent data preprocessing steps.

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
count	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000
mean	49972.419900	19468.865614	1.349571	164.359229	74.205690	128.817286	96.630414	1.366871	1.226457	0.088129	0.053771	0.803729	0.499700
std	28851.302323	2467.251667	0.476838	8.210126	14.395757	154.011419	188.472530	0.680250	0.572270	0.283484	0.225568	0.397179	0.500003
min	0.000000	10798.000000	1.000000	55.000000	10.000000	-150.000000	-70.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000
25%	25006.750000	17664.000000	1.000000	159.000000	65.000000	120.000000	80.000000	1.000000	1.000000	0.000000	0.000000	1.000000	0.000000
50%	50001.500000	19703.000000	1.000000	165.000000	72.000000	120.000000	80.000000	1.000000	1.000000	0.000000	0.000000	1.000000	0.000000
75%	74889.250000	21327.000000	2.000000	170.000000	82.000000	140.000000	90.000000	2.000000	1.000000	0.000000	0.000000	1.000000	1.000000
max	99999.000000	23713.000000	2.000000	250.000000	200.000000	16020.000000	11000.000000	3.000000	3.000000	1.000000	1.000000	1.000000	1.000000

Fig. 1: Statistics of Dataset

2) Data Quality Assessment: Ensuring the integrity and completeness of the dataset is crucial for robust model development. The dataset was examined for missing values and duplicates, and no such issues were detected. The absence of null entries confirms that all records are complete, while the lack of duplicates ensures that no artificial inflation or bias exists in the data. This clean dataset provided a reliable foundation for further analysis and model training.

3) Feature Transformation and Encoding: Certain variables required transformation to improve interpretability and compatibility with machine learning algorithms. The categorical 'gender' attribute was label-encoded into numerical values, with 1 representing male and 0 representing female, simplifying algorithmic processing. The 'age' attribute was converted from days to years by dividing by 365 and rounding, allowing for a more intuitive understanding of participants' ages. Additionally, the 'id' column, which served solely as a unique identifier, was removed from the dataset as it had no predictive value. These transformations enhanced the quality and usability of the dataset for modeling purposes.

4) Outlier Detection and Treatment: Outliers can significantly distort model performance and predictive accuracy. In this dataset, outliers were identified in key numerical features including height, weight, and blood pressure readings. Values beyond the 2.5th and 97.5th percentiles were removed, ensuring that the remaining data better represented the population distribution. Further exploratory grouping revealed gender-specific trends: on average, male individuals were taller than females, and alcohol consumption was higher among males. Recognizing such patterns provided additional context for feature engineering and interpretation of the data.

5) Correlation Analysis: To understand the relationships among variables, a correlation matrix was constructed, quantifying the pairwise correlations between all features. As illustrated in Fig. 2, systolic (ap_hi) and diastolic (a_lo) blood pressure exhibit a

strong positive correlation, indicating that individuals with high systolic pressure are likely to have elevated diastolic pressure. Cholesterol and glucose levels show a moderate positive correlation, suggesting a link between metabolic factors. A smaller positive correlation between gender and height reflects expected physiological differences. The correlation matrix (Fig. 2) provides valuable insights into interdependencies among features, helping to identify redundant variables and guiding subsequent dimensionality reduction and feature selection steps.
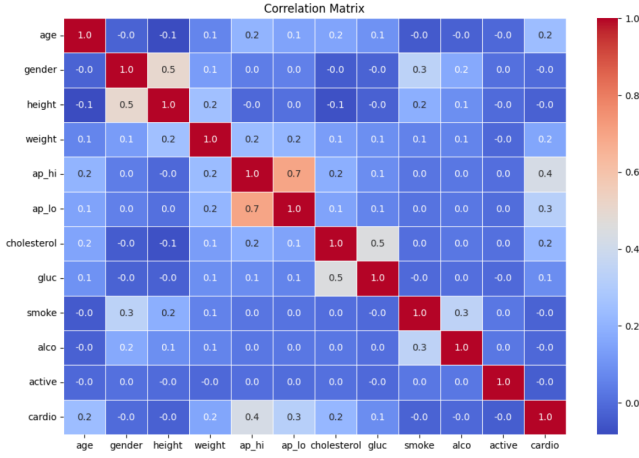

Fig. 2: Correlation matrix among the features

C. Feature Engineering (Feature Extraction)

Feature engineering plays a critical role in enhancing the quality and predictive power of a dataset by creating or transforming features into more informative representations. In this study, the dataset was initially divided into independent variables (X) and a dependent target variable (y), with 'cardio' serving as the target. The original 11 attributes formed the initial set of independent variables. To improve model performance and interpretability, four new features were derived from the existing data. The first is Body Mass Index (BMI), computed as an individual's weight in kilograms divided by the square of their height in meters. BMI serves as an important indicator of body fat and overall health status. To provide a more interpretable representation, BMI was further categorized into seven classes, ranging from Anorexic to Obesity Class III, allowing for nuanced classification of body weight levels. Next, blood pressure categories (bp_cat) were introduced by classifying combined systolic and diastolic blood

pressure readings according to standard clinical thresholds. These categories, ranging from normal to severe hypertension, provide a meaningful representation of cardiovascular risk. Additionally, pulse pressure (pulse_press) was computed as the difference between systolic and diastolic blood pressure, quantifying the variation in arterial pressure during a heartbeat cycle. Through these feature engineering steps, the dataset was enriched with clinically relevant and interpretable attributes, capturing critical information about individuals' cardiovascular health. After adding the derived features, the dataset now contains 15 independent variables, offering a more comprehensive basis for predictive modeling and enhancing the potential for accurate cardiovascular disease prediction

D. Dimensionality Reduction

After feature engineering, the dataset contained 15 independent variables. To improve computational efficiency and model interpretability, Principal Component Analysis (PCA) was applied [15]. The data was first standardized to ensure all features contributed equally, as PCA relies on the covariance structure. PCA transformed the 15-dimensional dataset into two principal components, each a linear combination of the original features. These orthogonal components retain most of the variance, with eigenvalues indicating their contribution. Reducing to two components preserved the essential information while simplifying the dataset. This reduction provides faster computation, easier visualization, and improved interpretability, offering a compact and informative input for subsequent predictive modeling on the cardiovascular dataset.

E. Model Training and Evaluation

1) Data Splitting: After dimensionality reduction, the dataset was divided into training (80%) and testing (20%) subsets. The variables x_train and y_train were used for model training, while x_test and y_test served for evaluation. This split ensures that models are tested on unseen data, providing an unbiased assessment of performance.

2) Classifier Implementation: Four supervised classifiers were applied: Random Forest, Decision Tree, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). Initially, SVM achieved the highest accuracy of 71%, followed by KNN at
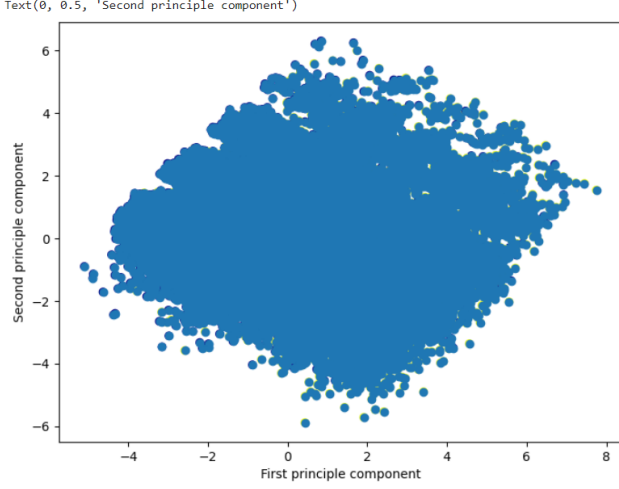
Fig. 3: Dimensionality reduction using PCA.

68%, Random Forest [16] at 66%, and Decision Tree at 62%. These results highlight the varying suitability of each algorithm for predicting cardiovascular disease.

3) Feature Selection and Its Impact: To improve model efficiency and accuracy, a feature selection [17] step was conducted using feature importance scores. The top eight features were retained, while four less influential ones (id, gender, height, alco) were removed. After retraining, KNN and Random Forest achieved the highest accuracy of 69%, Decision Tree improved to 63%, and SVM dropped to 50%. This demonstrates the significant impact of feature selection, emphasizing that the relevance of specific features can vary across algorithms.
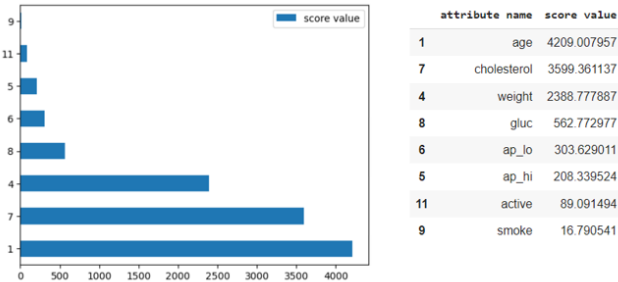

Fig. 4: Top features and bar chart using "Feature Selection Technique"

4) Classifier Characteristics: Random Forest combines multiple decision trees, handling both categorical and numerical features effectively, and provides inherent feature importance scores. Decision Tree models are interpretable and capture non-

Classifier	Accuracy (after using PCA)	Accuracy (after using Feature Selection)
Decision Tree	62%	63%
Random Forest	66%	69%
SVM	71%	50%
KNN	68%	69%

Fig. 5: Accuracy of the various classifiers

linear relationships, but can overfit without feature refinement. SVM excels in high-dimensional spaces and captures complex patterns through kernel functions; its performance decline after feature selection indicates reliance on specific features. KNN [18] relies on local data patterns and benefits from standardized features; its improved accuracy after feature reduction suggests that removing irrelevant features enhanced predictive capability.

5) Feature Treatment: Prior to modeling, the dataset underwent extensive preprocessing. EDA was conducted to explore distributions, detect outliers, and visualize patterns. The categorical gender variable was label-encoded. Feature engineering introduced BMI, BMI categories (bmi_class), blood pressure categories (bp_cat), and pulse pressure (pulse_press). PCA reduced the dataset from 15 dimensions to two principal components, preserving most variance while improving computational efficiency and interpretability. This iterative process of feature engineering, dimensionality reduction, and feature selection demonstrates that both the choice of features and the selection of classifiers critically influence predictive performance. The combined approach provides a robust framework for cardiovascular disease prediction.

IV. RESULT

The performance of machine learning classifiers for predicting cardiovascular disease (CVD) was evaluated using standard metrics derived from the confusion matrix and classification report, including accuracy, precision, recall, F1-score, and support. These metrics provide a comprehensive understanding of the predictive capabilities of each model, particularly in the context of imbalanced healthcare data.

1) Support Vector Machine (SVM) Performance: Among the classifiers tested, the Support Vector Machine (SVM) [19] achieved the highest overall accuracy of 71%, establishing it as the most effective model for this prediction task. The SVM model demonstrated strong discrimination between individuals with and without CVD, highlighting its suitability for high-dimensional and complex healthcare datasets.

2) Confusion Matrix Analysis: The confusion matrix summarizes model performance on the test dataset. Each row represents the actual class, and each column represents the predicted class (Fig. 6).
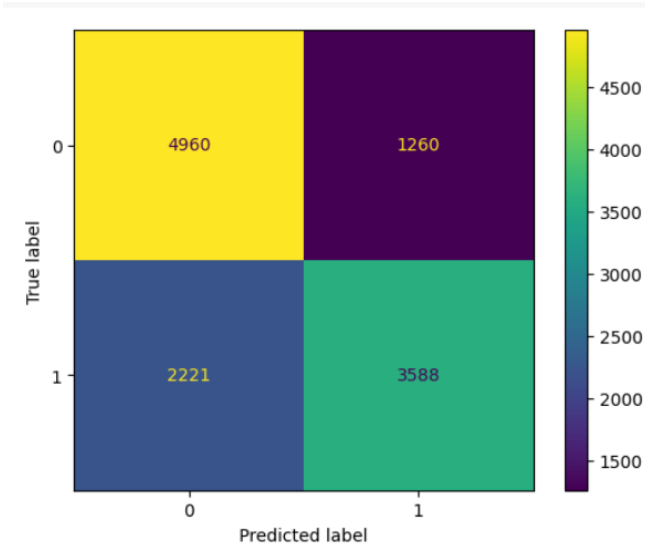

Fig. 6: Confusion Matrix for Support Vector Machine

- True Positive (TP) = 3,588 – correctly predicted CVD cases.
- False Positive (FP) = 1,260 – incorrectly predicted CVD cases.
- True Negative (TN) = 4,960 – correctly predicted non-CVD cases.
- False Negative (FN) = 2,221 – incorrectly predicted non-CVD cases.

This analysis indicates that while the model accurately identifies a majority of both positive and negative cases, some misclassifications remain, particularly in detecting all CVD cases.

3) Classification Report: Based on the confusion matrix, the following performance metrics were calculated (Fig. 7):

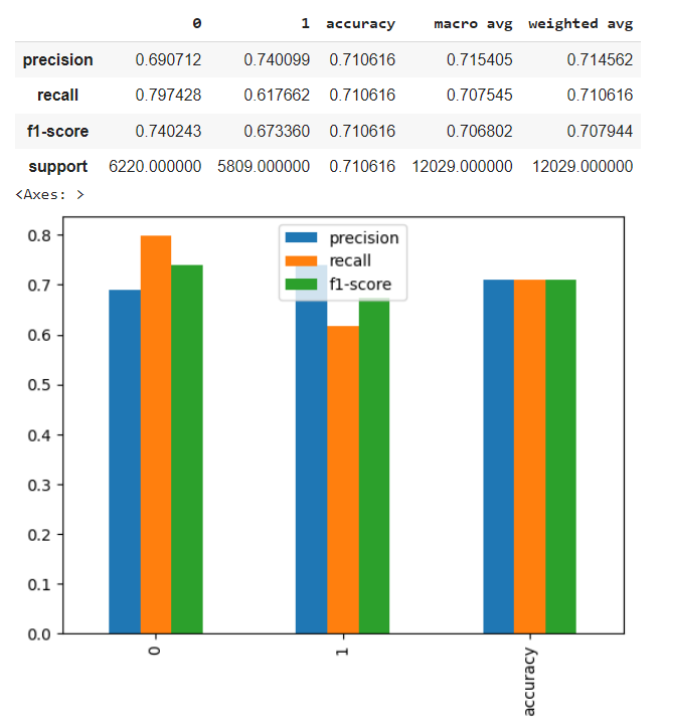|	0	1	accuracy	macro avg	weighted avg
precision	0.690712	0.740099	0.710616	0.715405	0.714562
recall	0.797428	0.617662	0.710616	0.707545	0.710616
f1-score	0.740243	0.673360	0.710616	0.706802	0.707944
support	6220.000000	5809.000000	0.710616	12029.000000	12029.000000

`<Axes: >`

Fig. 7: Classification Report

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = 0.711$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN},$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- Class 0 (No CVD): Precision = 0.691, Recall = 0.797, F1-score = 0.740
- Class 1 (CVD): Precision = 0.740, Recall = 0.618, F1-score = 0.673
- Macro Average: Precision = 0.715, Recall = 0.708, F1-score = 0.707
- Weighted Average: Precision = 0.715, Recall = 0.711, F1-score = 0.708
- Support: Class 0 = 6,220 instances, Class 1 = 5,809 instances

These metrics highlight the balanced performance of the SVM, with high precision for CVD cases indicating that positive predictions are reliable, and a reasonable recall ensuring a significant proportion of actual CVD cases are correctly identified. The F1-score provides a harmonic balance between precision and recall, which is particularly important in

healthcare applications where both false positives and false negatives carry significant consequences.

4) Interpretation of Metrics: The SVM classifier demonstrates a balanced and reliable performance. High precision ensures that predicted CVD cases are mostly correct, reducing unnecessary interventions. Recall indicates the model's effectiveness in capturing actual CVD cases, minimizing missed diagnoses. The F1-score provides a harmonic balance between precision and recall, which is particularly important for datasets with unequal class distributions. Support contextualizes the metrics by showing the number of samples in each class.

V. Conclusion

This study demonstrated the effective application of multiple supervised machine learning algorithms for predicting cardiovascular disease (CVD) using patient health record data. Among the evaluated models, the Support Vector Machine (SVM) achieved the highest accuracy of 71%, underscoring its strong potential for reliable CVD prediction. The integration of feature engineering and Principal Component Analysis (PCA) significantly enhanced model interpretability and computational performance. Nevertheless, the variations in accuracy following feature selection highlight the importance of fine-tuning model parameters and adapting algorithms to dataset-specific characteristics for optimal results. Future research will focus on hyperparameter optimization, the incorporation of larger and more diverse datasets, and the exploration of advanced ensemble and deep learning methods to further improve predictive accuracy and clinical applicability in cardiovascular risk assessment.

References

[1] Edgardo Olvera Lopez, Brian D Ballard, and Arif Jan. Cardiovascular disease. In *StatPearls [Internet]*. StatPearls Publishing, 2023.

[2] BD Ballard, A Jan, et al. Cardiovascular disease. 2018.

[3] Sulianova. Cardiovascular disease dataset, kaggle, 2017. [online].

[4] Tianyi Liu, Andrew Krentz, Lei Lu, and Vasa Curcin. Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis. *European Heart Journal-Digital Health*, 6(1):7–22, 2025.

[5] Roohallah Alizadehsani, Moloud Abdar, Mohamad Roshanzamir, Abbas Khosravi, Parham M Kebria, Fahime Khozeimeh, Saeid Nahavandi, Nizal Sarrafzadegan, and U Rajendra Acharya. Machine learning-based coronary artery disease diagnosis: A comprehensive review. *Computers in biology and medicine*, 111:103346, 2019.

[6] Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H Guppy, Stella Lee, and Victor Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology*, 64(5):304–310, 1989.

[7] Atul Kumar Pandey, Prabhat Pandey, KL Jaiswal, and Ashish Kumar Sen. A heart disease prediction model using decision tree. *IOSR Journal of Computer Engineering (IOSR-JCE)*, 12(6):83–86, 2013.

[8] K Thenmozhi and P Deepika. Heart disease prediction using classification with different decision tree techniques. *International Journal of Engineering Research and General Science*, 2(6):6–11, 2014.

[9] Anna Karen Gárate-Escamila, Amir Hajjam El Hassani, and Emmanuel Andrès. Classification models for heart disease prediction using feature selection and pca. *Informatics in Medicine Unlocked*, 19:100330, 2020.

[10] Abhijit Dnyaneshwar Jadhav and Santosh V Chobe. Risk assessment of cardiovascular diseases using knn and decision tree classifier. *Revue d'Intelligence Artificielle*, 36(1), 2022.

[11] VV Ramalingam, Ayantan Dandapath, M Karthik Raja, et al. Heart disease prediction using machine learning techniques: a survey. *International Journal of Engineering & Technology*, 7(2.8):684–687, 2018.

[12] Syed Muhammad Saqlain Shah, Safeera Batool, Imran Khan, Muhammad Usman Ashraf, Syed Hussnain Abbas, and Syed Adnan Hussain. Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis. *Physica A: Statistical Mechanics and its Applications*, 482:796–807, 2017.

[13] Ibtasam Wajid, Li Dan, and Qin Wang. Hybrid ensemble approaches for cardiovascular disease prediction: Leveraging interpretable ai for clinical insight. *Intelligence-Based Medicine*, page 100297, 2025.

[14] Muhammad Affan Alim, Shamsheela Habib, Yumna Farooq, and Abdul Rafay. Robust heart disease prediction: a novel approach based on significant feature and ensemble learning model. In *2020 3rd international conference on computing, mathematics and engineering technologies (iCoMET)*, pages 1–5. IEEE, 2020.

[15] Laurens Van Der Maaten, Eric O Postma, H Jaap Van Den Herik, et al. Dimensionality reduction: A comparative review. *Journal of machine learning research*, 10(66-71):13, 2009.

[16] Madhumita Pal and Smita Parija. Prediction of heart diseases using random forest. In *Journal of Physics: Conference Series*, volume 1817, page 012009. IOP Publishing, 2021.

[17] B Venkatesh and J Anuradha. A review of feature selection and its methods. *Cybern. Inf. Technol*, 19(1):3–26, 2019.

[18] Oliver Kramer. K-nearest neighbors. In *Dimensionality reduction with unsupervised nearest neighbors*, pages 13–23. Springer, 2013.

[19] Derek A Pisner and David M Schnyer. Support vector machine. In *Machine learning*, pages 101–121. Elsevier, 2020.