Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 8, 2025

A COMPARATIVE STUDY OF MACHINE LEARNING ALGORITHMS FOR CARDIOVASCULAR RISK PREDICTION: SUPPORT VECTOR MACHINE, GRADIENT BOOSTING, AND ROTATION FOREST

Muhammad Asad Ullah[1], Aftab Ullah[2], Malak Roman[*3], Masood Anwar[4], Muhsin Ul Mulk Siddiqi[5], Muhammad Hasnain Jaffar[6], Umer Farooq[7], Dr. Junaid Ali[8]

[1,2,6,7]*BS-Computer Science, Department of Computer Science, University of Chitral-KPK, Pakistan*
[*3,4,5]*Lecturer, Department of Computer Science, University of Chitral-KPK, Pakistan*
[8]*Resident Pediatric Surgeon, MTI-KTH Peshawar-KPK, Pakistan*

[*3]malak_5116@uoch.edu.pk

DOI: https://doi.org/10.5281/zenodo.16880668

Keywords
Cardiovascular Disorder, Machine Learning, Support Vector Machine, Gradient Boosting, Rotation Forest.

Article History
Received: 13 May, 2025
Accepted: 18 July, 2025
Published: 14 August, 2025

Corresponding Author: *
Malak Roman

Abstract
Cardiovascular heart disease is one of the most fatal problems in the world and is a major cause of deaths globally, reaching around 17.9 million deaths every year. Timely prediction of heart disease is critical for instant response to achieve favorable outcomes; therefore, it requires accurate diagnosis at the right time. Today, the healthcare field has a lot of data, but not much enough knowledge. Machine learning allows computer programs to learn from existing data, get better at doing tasks through experience without needing help from people, and then use what they have learned to make smart choices. There are many different methods and tools in data mining and machine learning that can be used to get useful information from databases and to apply that information for better and more accurate diagnosis. In this research, we compared three machine learning algorithms–Support Vector Machine, Gradient Boosting, and Rotation Forest–to find out which one works best for predicting heart diseases on time. We looked at how accurate each method was, and both Rotation Forest and Gradient Boosting were the most accurate.

INTRODUCTION

Data mining is termed as acquiring useful and valuable information from complex data using a combination of statistical and computational tools. With the help of these tools, organizations can make predictions about future developments in complex data and facilitate intelligent multi-dimensional decision-making in numerous fields such as telecommunications, finance, transport, and insurance [1]. Machine Learning (ML) methods enhance diagnostic accuracy, make treatment plans more personalized, and facilitate more appropriate resource utilization in clinical environments [3]. ML techniques have shown great potential in the field of healthcare applications, especially in heart disease prediction and classification. In real-world settings, the poor quality and format of medical data hinder accurate and timely decision-making. While ML has been extensively used for the prediction of heart disease, little comparative literature is available to study the strengths and weaknesses of various models [9]. Considering advances in artificial intelligence (AI), computing capacity, and data storage, healthcare analytics is becoming more efficient and scalable [4]. Several methods of preprocessing and ML have been applied to convert raw

healthcare data into usable information for decision-making, specifically the prediction of heart disease [5]. Cardiovascular diseases (CVDs) are the leading cause of death globally, with an estimated 17.9 million deaths every year as per the World Health Organization [2]. The underutilization of rich medical data by healthcare systems continues to delay improvements in early disease detection and preventive interventions. This shortfall leads to the urgent requirement for intelligent systems that can interpret medical information for timely diagnoses and treatment [4]. The performance of three supervised machine learning methods—Support Vector Machine (SVM), Rotation Forest, and Gradient Boosting—is evaluated and compared in this study for predicting early-stage of heart disease. These methods work on labeled datasets to find complex relationships and assist clinical decision-making. SVM works to construct optimal hyperplanes to classify between healthy and diseased patients [6]. As an ensemble method, Rotation Forest improves prediction by creating varied decision trees through feature extraction [7]. This research addresses the gap by analyzing Support Vector Machine (SVM), Rotation Forest, and Gradient Boosting classifiers. Through their performance evaluation, we seek to determine the most appropriate model for real-time, accurate heart disease diagnosis. The goal of this research is to aid clinical decision-making with a strong assessment of machine learning classifiers, thereby contributing to timely and efficient treatment measures.

CARDIOVASCULAR DISEASES:

Cardiovascular diseases (CVDs) constitute a variety of heart and blood vessel disorders such as coronary heart disease, rheumatic heart disease, and disorders of peripheral arteries. Four out of five heart disease deaths are due to CVDs, and one-third of these happen in individuals aged under 70 years old [8]. The primary cause of most heart conditions is myocardial ischemia, resulting from atherosclerosis, which narrows or blocks arteries that curb blood and oxygen supply to tissues [6]. Reduced blood and oxygen flow can result in heart failure, often showing signs like breathlessness, leg swelling (edema), and sudden weight gain. Arrhythmias may present with dizziness, palpitations, or fainting. Valvular heart disease typically involves heart murmurs and similar signs to heart failure [9]. According to the World Health Organization, significant cardiovascular disease risk factors are [10].

- Alcohol consumption
- Use of Tobacco
- Increasing blood pressure (hypertension)
- Elevated cholesterol
- High blood glucose (diabetes)
- Restfulness (Inert)
- Obesity
- Unhealthy eating habits

LITERATURE REVIEW

Ahmed, I. [12] proposed a predictive model for detection of heart disease using an array of machine learning algorithms. UCI repository was used to obtain the dataset which contains 303 instances. After data cleaning, the following ML algorithms were applied: NB, DT, SVM, Bagging, Boosting, and RF. The result shows that RF (RandomForest) performed best with 89.4% accuracy.

Patel, J. et al. [13] conducted research on predicting heart disease using machine learning algorithms. They utilized the UCI repository's dataset for training and testing the classifier models, i.e., NN, SVM, and Random Forest models. Among the classifiers, the Support Vector Machine had the greatest accuracy at 84%, only slightly better than the Neural Networks at 83% and the Random Forest at 80%.

Mohan, S. et al. [13] employed heart disease data primarily from Cleveland's UCI repository. In their research work, a mixed approach was applied. For data preprocessing, multi-class and binary classification techniques were applied. The Random Forest classifier, along with a linear approach, was used as a machine learning algorithm and yielded an accuracy of 88.7%.

In 2024, Hajiarbabi, M. [14] extended an extensive review by integrating the outcomes of different studies in predicting heart disease for the period 2015–2024. Kaggle repository databases were recommended for data collection, and the

accuracy of multiple machine learning algorithms was analyzed. Among the analyzed algorithms, XGBoost, Naïve Bayes, and K-Nearest Neighbor classifiers were recommended on the basis of their performance.

Rodriguez, J. J. et al., [16] investigated cancer risk predictions through machine learning. They gathered multiple datasets from UCI and the National Lung Screening Trial (NLST). Support Vector Machine and Rotation Forest were implemented as machine learning classifier models. The proposed models were trained with specific dataset and then tested on breast, lung, prostate, colon, and leukemia cancer datasets. Results revealed that Rotation Forest performed excellently for breast cancer, with an accuracy rate of 96.49%.

Hassan, C. A. U. et al., [17] explored coronary heart disease prediction in 2022 through various machine learning classifiers, using a dataset of 303 samples and 14 features after data cleaning and removal of noisy data. The Gradient Boosted Tree, RF, and Multilayer Perceptron models were trained and tested in terms of accuracy to recognize the best classifier. Results revealed that the Random Forest-based structure outperformed other methods with an accuracy of 96.28%.

Wisaeng, K., [18] examined K-Nearest Neighbor and Multilayer Perceptron for classification and forecasting heart diseases. The obtained dataset, having 14 attributes, used back-propagation for data preprocessing, but with the use of feature selection techniques, the models showed poor performance. The differences in accuracy in the testing dataset between 13 attributes is 93% and 8 attributes is 90%.

In 2023, Al-Batah, M. S. et al., [19] built an intelligent system for heart disease forecasting for use in Jordanian hospitals. They utilized a dataset having 1,025 patients' recorded data with 14 features. Various models were tested, such as RF, Neural Networks, NB, and Logistic Regression. Random Forest delivered an accuracy of 98.4%, which was the best among those investigated.

Roman, M. et al., [20] investigated the prediction of stroke disease and constructed K-Nearest Neighbor and Decision Tree algorithm models. They gathered data from hospitals located in Peshawar, which had 12 data attributes. Genetic Search and Chi-Square were used for optimal feature selection to achieve better prediction. Based on their results, KNN, along with Genetic Search, yielded an optimal accuracy of 97.5%, better than Decision Tree models.

RESEARCH METHODOLOGY:

In our research work, an intelligent integrated model was developed for forecasting CVD diseases. After data preprocessing, the dataset is split into two: One portion of dataset were used for training and the other portion were used for model testing. SVM, Rotation Forest (RF), and Gradient Boosting classification models were applied individually using a Python environment, as shown in Figure 1. First of all, the models were trained based on the training portion of data and after training the test data was used to calculate the accuracy of the proposed integrated model.

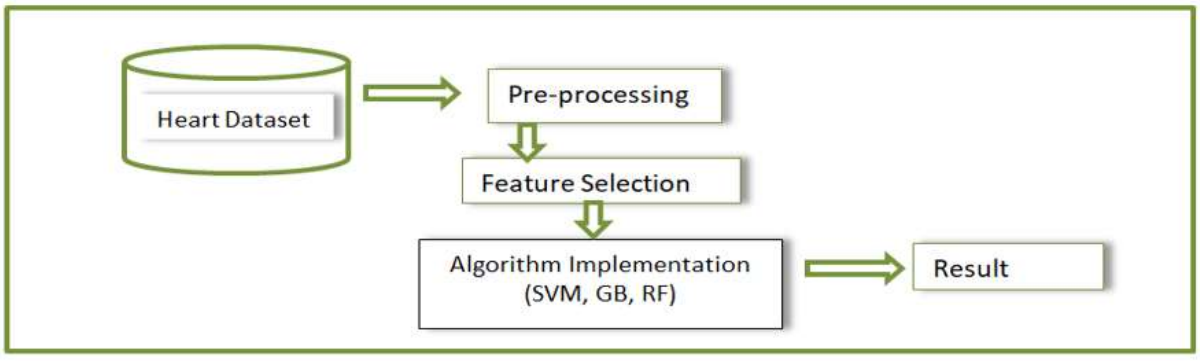

Fig 1: Machine Learning Models Architecture

The above model was executed with the PyCharm IDE. Pandas and NumPy libraries were imported and used to handle data, and Matplotlib was used to visualize outcomes. Additionally, other machine learning libraries were imported, such as Scikit-Learn for models, preprocessing methods, and performance metric measurements. All the above libraries were collectively used to perform data loading and handling, data processing, model building, and measuring performance.

Data Collection:
The dataset was sourced from Kaggle, comprising 1,025 patient entries with 14 features, which were used to evaluate the risk of having heart disease. Table 1 presents a summary of the collected data, listing key value attributes.

S. No	Attribute	Description	S. No	Attribute	Description
1	Age	Age in years	8	Thalach	Maximum heart rate achieved
2	Sex	1= male 0=female	9	Exang	1=yes, 0=no
3	CP	1=typical angina, 2=atypical angina, 3=non-Anginal pa, 4=asymptomatic	10	Slope	1=up sloping, 2=flat, 3=down sloping
4	Trestbps	Resting blood pressure (in mm Hg)	11	Oldpeak (Represent Change in ST segment)	0: No ST segment 1: Slight ST segment 2: Moderate ST 3: Severe ST segment 4: Extremely severe ST
5	Chol	Serum cholesterol in mg/dl	12	Thal (Thalassemia index)	3: Normal 6: Fixed Defect 7: Reversible Defect
6	Fbs	Fasting blood sugar > 120 mg/dl: 1=True 0=False	13	Ca (Coronary Arteries)	0: No blockage. 1: one blockage. 2: Two blockages. 3: Three or more
7	Restecg	0=normal 1=having ST-T wave 2=showing probable	14	Target	0: No CVD 1: CVD Present

Table 1: Attributes Description

Data Preprocessing:
This stage helps to enhance the data reliability. Data cleaning, integration, reduction, and removal of outlier tasks are performed at this stage. Feature subset evaluation is used to identify the least number of parameters. The selected attributes presented results similar to those attained with all attributes [21]. To pre-process the data for machine learning models, the dataset is first checked for outliers and missing values. Then, the separation of features and labels is done, where the dataset is divided into features (x) and the target variable (y).

MODEL IMPLEMENTATION:
The construction of a model that can be used to classify a group of items, which will later be used to assign class labels or attributes to yet unknown objects in the future, is known as classification [21]. In the organized intelligent integrated model, the support vector machine, Rotation Forest, and Gradient Boosting classifier were used for heart disease prediction.

Support Vector Machine:
Support Vector Machine (SVM) is a type of machine learning algorithm (Supervised) that is used mainly for classification tasks [23]. Its ultimate focus and goal is to identify the best

hyperplane that clearly separates data points from different classes. SVM transforms the input data into a higher-dimensional space using a kernel function and then builds a hyperplane that creates the widest possible margin between classes.

The algorithm works as follows:

- Mapping each point to a higher-dimensional space using a kernel function.
- Finding the hyperplane that excellently splits the mapped points into their corresponding classes.
- Passing the new data point through the same kernel function for applicable prediction.

Optimization problem used to find the hyper plane can be expressed as in equation 1:

$$Max\frac{1}{2}\|w\|\wedge 2-\sum\_i\alpha\_iy\_i(w\wedge Tx\_i+b)$$

$$Subject\alpha\_i>=0and\sum\_i\alpha\_iy\_i=0$$

(Equation 1: Optimization Problem in SVM)

Whereas,

w is the weight vector,

b is the bias term, α_i is a Lagrange multiplier,

iy is the label for the ith training example, and

ix is the ith training example mapped to the higher-dimensional space.

Rotation Forest Algorithm:

It is an ensemble learning approach that improves decision tree performance using feature extraction algorithms such as Principal Component Analysis (PCA) to generate rotated feature spaces for each of the base classifiers [23]. The training data is divided into feature subsets using PCA. These transformed features are aggregated to create another training set. Several decision trees are trained using these rotated datasets, and predictions from each of them are combined using majority voting, as shown in figure 2 [24].

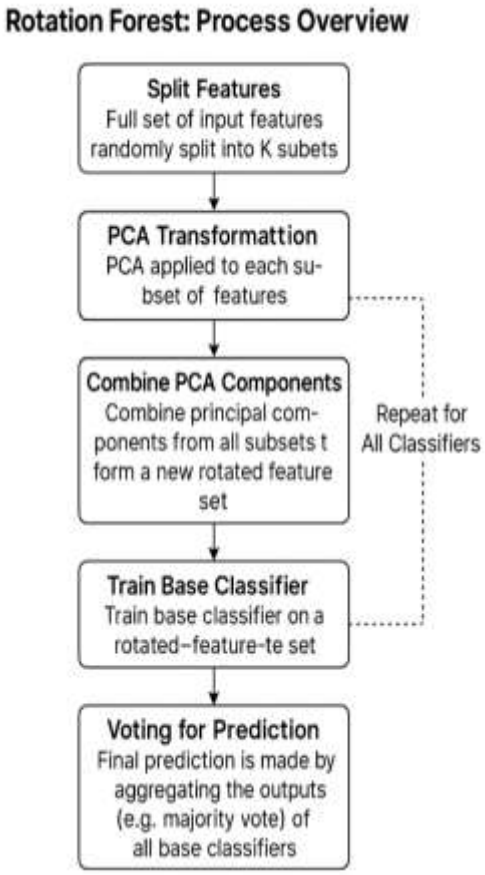

Fig 2: Working of Rotation Forest.

Gradient Boosting:

Gradient Boosting constructs models sequentially using gradient descent to optimize a loss function. It begins with an initial weak learner (usually a decision tree), and in each step, another model is trained to predict the residuals (errors) of the previous model. These additional models are used to reduce the overall error of prediction [25]. Figure 3 shows the working flow chart of the Gradient Boosting algorithm.

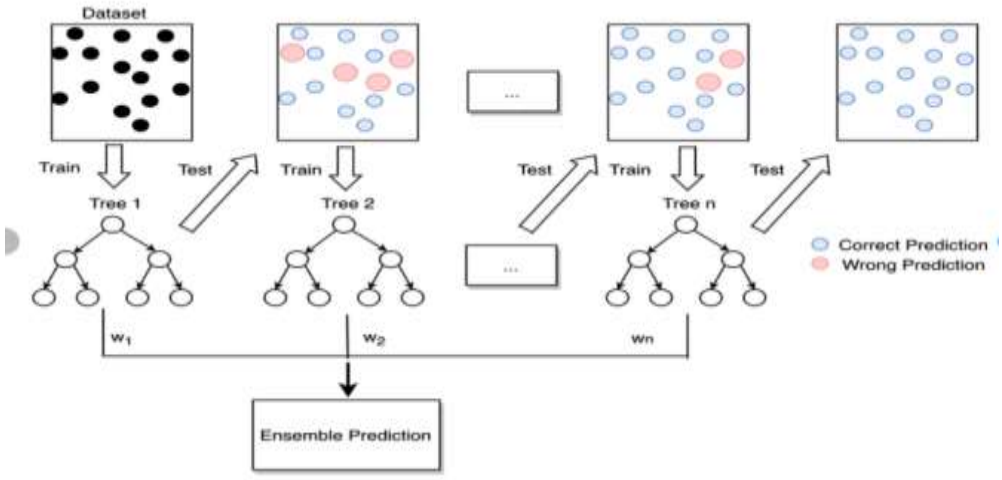

Fig 3: Working Flow Chart of Gradient Boosting Algorithm

RESULTS AND DISCUSSION

In the context of healthcare, particularly in the early detection of CVD, the precision of a diagnostic model is not merely a technical matter; it can define the trajectory of a patient's life. A false negative (FN) could indicate a missed identification, while a false positive (FP) could lead to avoidable psychological and medical stress. With this life-saving responsibility in mind, this work evaluated the performance of three machine learning classifiers: Support Vector Machine, Rotation Forest, and Gradient Boosting, which were applied to a CVD dataset. The proposed models were implemented in PyCharm IDE, as shown in Figure 4.

```
models = {
    "SVM": SVC(random_state=42),
    "Gradient Boosting": GradientBoostingClassifier(random_state=42),
    "Rotation Forest": RotationForest(random_state=42)
}
```

Fig 4: Models Implementation in PyCharm IDE.

Each model was trained using the training dataset and tested using the testing dataset. We computed three essential performance metrics: accuracy, precision, recall and F1 score, as discussed in Table 2 and implemented in fig 5.

```
# Train, Predict and Evaluate ---
results = []
conf_matrices = {}

for model_name, model in models.items():
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)

    acc = accuracy_score(y_test, y_pred)
    prec = precision_score(y_test, y_pred)
    f1 = f1_score(y_test, y_pred)

    results.append({
        "Model": model_name,
        "Accuracy": acc,
        "Precision": prec,
        "F1 Score": f1
    })
    conf_matrices[model_name] = confusion_matrix(y_test, y_pred)
    print(f"\nClassification Report for {model_name}:\n")
    print(classification_report(y_test, y_pred))
```

Fig 5: Performance Metrics: accuracy, precision, and F1 score

These metrics give a balanced insight into how each model generalizes, especially when dealing with datasets that have an imbalanced distribution.

Model	Accuracy	Precision	Recall	F1-Score
SVM	64.9%	0.68	0.755556	0.715789
Gradient Boosting	100%	1.00	1.000000	1.000000
Rotation Forest	100%	1.00	1.000000	1.000000

Table 2: ML Models Results Comparison

The Rotation Forest and Gradient Boosting classifiers demonstrated excellent performance across all metrics. The accuracy achieved is 100%. These outputs show the superiority of ensemble methods in classification problems, particularly when handling complex, high-dimensional data such as that in medical diagnoses. The Rotation Forest's approach to applying PCA-based feature transformation for each of its base learners provides an added advantage in identifying multivariate patterns in data, resulting in superior generalization and predictive power.

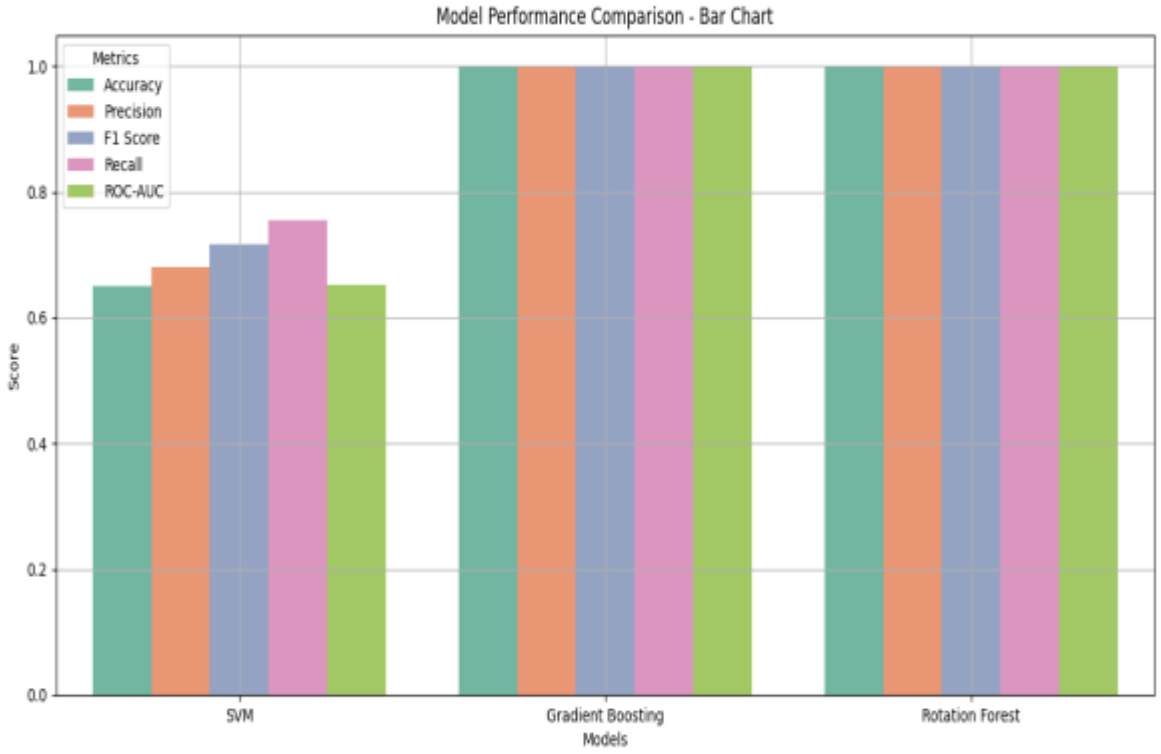

Fig 6: Visual representation of ML algorithms Performance

Figure 6 graphically presents a comparative assessment of classification in terms of accuracy, precision, F1-score, recall, and ROC-AUC across three machine learning algorithms: SVM, RF, and GB models. Mistakes and Explanations:

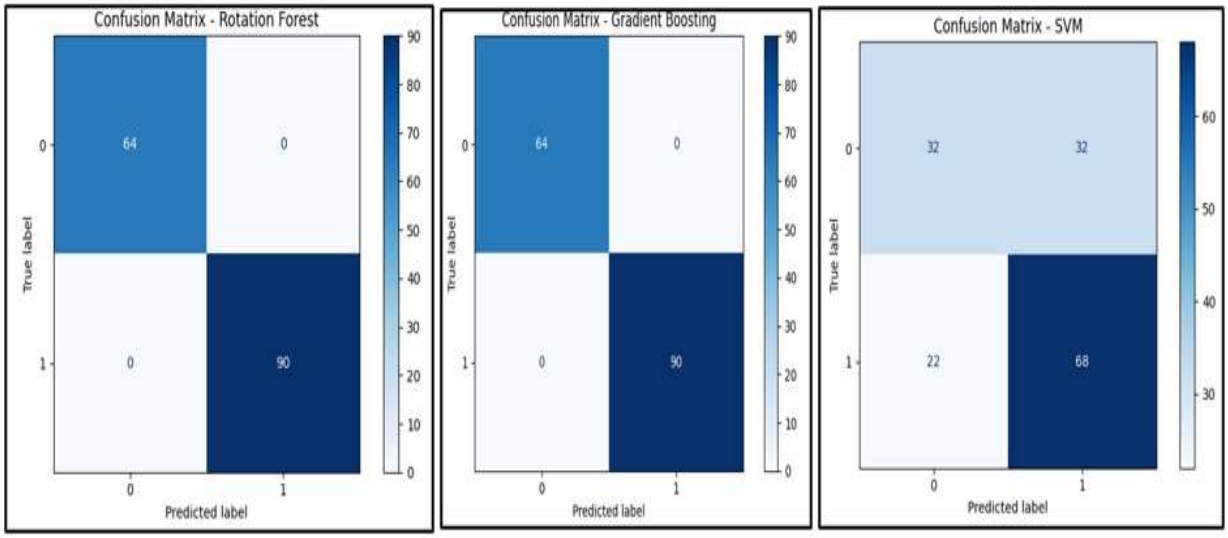

Fig 7: Models Confusion Matrix Evaluation

In Figure 7, the classifiers were assessed using standard evaluation metrics, including accuracy, precision, recall (sensitivity), false positive rate (FPR), and false negative rate (FNR) [26]. These metrics were computed using the following formulas:

$$Accuracy = \frac{No\ of\ correct\ predictions}{Total\ no\ of\ predictions} \#(equ.2)$$

$$Precision = \frac{True\ Positive}{True\ positive + False\ Positive} \#(equ.3)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \#(equ.4)$$

$$F1_{Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \#(equ.5)$$

The outcome strongly suggests that Rotation Forest and Gradient Boosting performed better than SVM in almost all their evaluated measures. They recorded their best in accuracy (100%), perfect precision (100%), and a robust F1-Score (100%), showcasing not just their success in detecting actual positive cases, but also their reliability in minimizing false positives. The SVM algorithm, despite its extensive use in various classification problems, recorded the poorest performance among the classifier algorithms.

FUTURE SCOPE

Future research could leverage data mining techniques to enhance CVD prediction by discovering hidden patterns in large-scale medical

datasets. Feature selection and association rule mining could identify key risk factors, while clustering methods may reveal patient subgroups for personalized treatment. Temporal data mining could track disease progression, and anomaly detection might flag early warning signs. Integrating these data mining approaches with machine learning could improve both prediction accuracy and clinical interpretability for better decision-making in CVD care.

REFERENCES

[1] Roman, M., Nawab, H. U., Ahmad, S., & Khan, I. A. (2022). K-Nearest Neighbor and Fuzzy K-Nearest Neighbor Algorithm Performance Analysis for Heart Disease Classification. *Webology (ISSN: 1735-188X)*, *19*(1).

[2] Khan, S. U., Bashir, Z. S., Khan, M. Z., Khan, M. S., Gulati, M., Blankstein, R., & Michos, E. D. (2020). Trends in cardiovascular deaths among young adults in the United States, 1999 to 2018. *American Journal of Cardiology*, *128*, 216-217.

[3] Rumsfeld, J. S., Joynt, K. E., & Maddox, T. M. (2016). Big data analytics to improve cardiovascular care: promise and challenges. *Nature Reviews Cardiology*, *13*(6), 350-359.

[4] Johnson, K. W., Torres Soto, J., Glicksberg, B. S., Shameer, K., Miotto, R., Ali, M., & Dudley, J. T. (2018). Artificial intelligence in cardiology. *Journal of the American College of Cardiology*, *71*(23), 2668-2679.

[5] Alom, Z., Azim, M. A., Aung, Z., Khushi, M., Car, J., & Moni, M. A. (2022). Early stage detection of heart failure using machine learning techniques. In *Proceedings of the International Conference on Big Data, IoT, and Machine Learning: BIM 2021* (pp. 75-88). Springer Singapore.

[6] Owusu, E., Boakye-Sekyerehene, P., Appati, J. K., & Ludu, J. Y. (2021). Computer-Aided Diagnostics of Heart Disease Risk Prediction Using Boosting Support Vector Machine. *Computational Intelligence and Neuroscience*, *2021*(1), 3152618.

[7] Johnson, K. W., Torres Soto, J., Glicksberg, B. S., Shameer, K., Miotto, R., Ali, M., & Dudley, J. T. (2018). Artificial intelligence in cardiology. *Journal of the American College of Cardiology*, *71*(23), 2668-2679.

[8] Khan, S. U., Bashir, Z. S., Khan, M. Z., Khan, M. S., Gulati, M., Blankstein, R., & Michos, E. D. (2020). Trends in cardiovascular deaths among young adults in the United States, 1999 to 2018. *American Journal of Cardiology*, *128*, 216-217.

[9] Theerthagiri, P., & Vidya, J. (2022). Cardiovascular disease prediction using recursive feature elimination and gradient boosting classification techniques. *Expert systems*, *39*(9), e13064.

[10] Nowbar, A. N., Gitto, M., Howard, J. P., Francis, D. P., & Al-Lamee, R. (2019). Mortality from ischemic heart disease: Analysis of data from the World Health Organization and coronary artery disease risk factors From NCD Risk Factor Collaboration. *Circulation: cardiovascular quality and outcomes*, *12*(6), e005375.

[11] Ahmed, I. (2022). A study of heart disease diagnosis using machine learning and data mining.

[12] Patel, J., Khaked, A. A., Patel, J., & Patel, J. (2021, May). Heart disease prediction using machine learning. In *Proceedings of Second International Conference on Computing, Communications, and Cyber-Security: IC4S 2020* (pp. 653-665). Springer Singapore.

[13] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, *7*, 81542-81554.

[14] Hajiarbabi, M. (2024). Heart disease detection using machine learning methods: a comprehensive narrative review. *Journal of Medical Artificial Intelligence*, *7*.

[15] Rodriguez, J. J., Kuncheva, L. I., & Alonso, C. J. (2006). Rotation forest: A new classifier ensemble method. *IEEE transactions on pattern analysis and machine intelligence*, *28*(10), 1619-1630.

[16] Hassan, C. A. U., Iqbal, J., Irfan, R., Hussain, S., Algarni, A. D., Bukhari, S. S. H., & Ullah, S. S. (2022). Effectively predicting the presence of coronary heart disease using machine learning classifiers. *Sensors*, *22*(19), 7227.

17] Wisaeng, K. (2014). Predict the diagnosis of heart disease using feature selection and k-nearest neighbor algorithm. *Applied Mathematical Sciences*, *8*(83), 4103-4113.

[18] Al-Batah, M. S., Alzboon, M. S., & Alazaidah, R. (2023). Intelligent Heart Disease Prediction System with Applications in Jordanian Hospitals. *International Journal of Advanced Computer Science and Applications*, *14*(9).

[19] Roman, M., Naz, I., Luqman, M. A., Ali, J., Jan, M. S., & Nawab, H. U. (2024). Stroke Disease Prediction Using K-Nearest Neighbor and Decision Tree Algorithms with Machine Learning Pre-Processing Techniques. *Migration Letters*, *21*(S4), 2015-2027.

[20] Roman, M., Ullah, A., Ullah, M. A., Hussain, F., Shams, S. S., Bint-e-Meraj, A., & Ali, S. (2025). Predicting Academic Success: A Machine Learning Approach Using Decision Tables and Random Forests Algorithms. *Spectrum of Engineering Sciences*, 3(5), 205-213

[21] Mining, M. D. Data Mining: Concepts and Techniques (2nd.edition)

[22] Rindhe, B. U., Ahire, N., Patil, R., Gagare, S., & Darade, M. (2021). Heart disease prediction using machine learning. *Heart Disease*, *5*(1).

[23] Tang, X., Yang, X., Liu, W., Deng, T., Cao, R., Tu, Z., & Hu, X. (2024, May). RFR-ABROF: A Multi-Strategy Collaborative Classification Prediction Model Based on Rotation Forest for PM2. 5. In *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)* (pp. 2209-2214). IEEE.

[24] Chen, W., Wang, C., Zhao, X., Bai, L., He, Q., Chen, X., & Ilia, I. (2025). Optimizing landslide susceptibility mapping using integrated forest by penalizing attributes model with ensemble algorithms. *Earth Science Informatics*, *18*(2), 225.

25] Wekalao, J., Srinivasan, G. P., Patel, S. K., & Al-zahrani, F. A. (2025). Optimization of graphene-based biosensor design for haemoglobin detection using the gradient boosting algorithm for behaviour prediction. *Measurement*, *239*, 115452.

[26] Anwar, M., Rahman, T., & Roman, M. (2025). Voice-activated smart environments: deep learning approach for pashto speech command processing. *Spectrum of Engineering Sciences*, *3*(5), 541-550.