Journal on Artificial Intelligence

Tech Science Press

DOI: 10.32604/jai.2024.050277

ARTICLE

Cardiovascular Disease Prediction Using Risk Factors: A Comparative Performance Analysis of Machine Learning Models

Adil Hussain[1,*] **and Ayesha Aslam**[2]

[1]School of Electronics and Control Engineering, Chang'an University, Xi'an, 710000, China

[2]School of Information Engineering, Chang'an University, Xi'an, 710000, China

*Corresponding Author: Adil Hussain. Email: 2022032907@chd.edu.cn

Received: 01 February 2024 Accepted: 01 April 2024 Published: 21 May 2024

ABSTRACT

The diagnosis and prognosis of cardiovascular diseases are critical medical responsibilities that assist cardiologists in correctly classifying patients and treating them accordingly. The utilization of machine learning in the medical domain has witnessed a notable surge due to its ability to discern patterns from vast amounts of data. Machine learning algorithms that can categorize cases of cardiovascular illness may help doctors reduce the number of wrong diagnoses. This research investigates the efficacy of different machine learning algorithms in predicting cardiovascular disease in accordance with risk factors. This study utilizes a variety of machine learning models, including Logistic Regression, Random Forest, Decision Tree, Extra Trees classifier, Support Vector Machine (SVM), XGBoost (XGB), Light Gradient Boosting Machine (LGBM), GaussianNB, and Multilayer Perceptron (MLP). The machine learning models are applied to a concrete dataset acquired from Kaggle. The models underwent training using a dataset that was partitioned into an 80:20 ratio. Machine learning model evaluation involves the utilization of performance measurements such as Accuracy, Precision, Recall, and ROC curves. An exhaustive evaluation is carried out to gauge the efficacy of the models.

KEYWORDS

Cardiovascular disease; heart disease; disease prediction; machine learning; performance analysis

1 Introduction

Cardiovascular disease has been well recognized as a highly severe and potentially fatal condition in the human population. The rising prevalence of cardiovascular illnesses, characterized by a substantial mortality rate, poses a substantial risk and burden to healthcare systems on a global scale. The prevalence of cardiovascular illnesses is higher in males compared to females, particularly during middle or old age [1]. However, it is worth noting that children are also experiencing similar health conditions [2]. Based on data presented by the World Health Organization (WHO), it is evident that cardiovascular disease accounts for approximately one-third of all global mortality. Cardiovascular Diseases (CVDs) are responsible for the mortality of roughly 17.9 million individuals annually on a global scale, with a greater prevalence observed in the Asian population [3]. According to a European Cardiology Society (ESC) survey, the global prevalence of heart disease stands at 26 million adults, with an annual identification rate of 3.6 million individuals. Approximately 50% of patients who receive

a diagnosis of heart disease experience mortality within a relatively short timeframe of 1–2 years. Additionally, approximately 3% of the healthcare budget is devoted to treating heart disease [4].

To predict the occurrence of heart disease, it is imperative to do a series of different tests. The possibility of inaccurate forecasts stems from a deficiency in proficiency among healthcare professionals [1]. Identifying heart disease early can provide difficulties [2]. The surgical management of cardiovascular disease poses challenges, especially in developing countries characterized by a scarcity of skilled healthcare practitioners and restricted availability of diagnostic tools and essential resources required for the adequate treatment and diagnosis of individuals with cardiac ailments [3]. To increase patient safety and reduce the occurrence of serious heart attacks, a precise prediction of the probability of cardiac failure is necessary [4]. Machine learning algorithms have shown efficacy in disease detection when taught on suitable data [5]. Heart disease datasets that are accessible to the public enable the evaluation and comparison of prediction models. Researchers utilize machine learning and AI to construct highly accurate prediction models by leveraging vast databases.

Recent research has placed significant emphasis on addressing CVDs in both adults and children, with a particular focus on reducing death rates associated with these conditions. Given the uneven and redundant nature of the available clinical datasets, it is imperative to emphasize the significance of effective preprocessing as a critical step [6]. Predictive models rely on the identification of relevant qualities that might operate as risk factors. Careful consideration of feature and algorithm optimization is required to build accurate prediction models using machine learning [7]. The examination of risk variables that meet three particular criteria, namely substantial frequency in the majority of populations, independent significant impact on heart illnesses, and the capacity for management or therapy to lessen risks, is crucial and has a great deal of value. Various researchers have used diverse risk variables or characteristics while constructing models to predict CVDs. Research studies have incorporated various features in developing CVD prediction models [8]. To accurately predict cardiac disease, researchers have a difficult time combining these indicators with the necessary machine-learning algorithms. Machine learning algorithms achieve optimal effectiveness when taught on suitable datasets [9–12]. Since the algorithms depend on the consistency of the training and test data, feature selection procedures are important.

Machine learning models have been implemented in various fields recently including intrusion detection [13] and fraud detection [14] and others. Machine learning models are also being used for healthcare systems along with disease identification, disease detection, and diagnosis. Early intervention, personalized treatment plans, risk stratification, efficient resource allocation, and patient involvement are all made possible by accurate diagnosis and prognosis in cardiovascular care. Machine learning is essential in recognizing intricate patterns, producing individualized risk forecasts, assisting in therapeutic decision-making, and consistently enhancing diagnostic precision [15]. The integration of machine learning improves the effectiveness and efficiency of cardiovascular healthcare delivery by minimizing misdiagnoses and enhancing patient outcomes. Extensive research work has been performed for cardiovascular disease predictions using machine learning, however, the performance of the machine learning models varies. A comprehensive comparison of the performance of the models for disease prediction is required to understand the model's performance in terms of disease predictions using a public dataset. The primary objective and contribution of this study is to predict cardiovascular diseases using several risk factors from a real-world dataset through several supervised learning models, such as Logistic Regression, Random Forest, Decision Tree, Extra Trees, Support Vector Classifier (SVC), XGBoost, Light Gradient Boosting Machine (LGBM), Gaussian Naive Bayes (GaussianNB), and Multi-layer Perceptron (MLP) classifiers are applied. Furthermore, a

comprehensive performance comparison among these models is performed to evaluate the models using the dataset.

This paper is organized as follows: Section 2 contains related work, and Section 3 contains a methodology and an overview of the dataset. Section 4 contains the implementation and results. Section 5 is the discussion and analysis, followed by Section 6, the conclusion of this work.

2 Related Work

The increased precision and efficacy of machine learning and AI technologies in making predictions have led to their broad use in the last several years [16]. The ability to develop and choose models that show remarkable accuracy and performance is the key to the success of this area of study [17].

Maiga et al. [18] performed a comparative analysis of machine learning algorithms for predicting cardiovascular disease based on patients' cardiovascular risk factors. The data is sourced from Kaggle machine learning competitions and comprises 70,000 patient records. The study employs machine learning methods such as Random Forest, Naïve Bayes, KNN, and Logistic Regression. The comparison results indicate that the Random Forest algorithm achieves a classification accuracy of 73%, a specificity of 65%, and a sensitivity of 80%.

Shah et al. [19] examined the feasibility of employing machine learning techniques to construct a predictive model for cardiovascular disease. The data utilized for this objective were acquired from the Cleveland heart disease dataset, comprising of 303 occurrences and 17 attributes, and were retrieved from the UCI machine learning repository. The authors utilized several supervised classification techniques, such as Naïve Bayes, decision tree, random forest, and k-nearest neighbors (KKN). The study findings revealed that the KKN model demonstrated the best level of accuracy, reaching 90.8%. The work underscores the potential efficacy of machine learning methods in forecasting cardiovascular illness and underscores the significance of choosing suitable models and approaches to attain optimal outcomes.

Using a dataset provided by the Cleveland Clinic Foundation, Alotalibi [20] also applied machine learning (ML) techniques to forecast heart failure disease. To develop prediction models, the authors implemented a variety of ML algorithms, including decision tree, logistic regression, random forest, Naïve Bayes, and support vector machine (SVM). A 10-fold cross-validation method was utilized in the process of developing the model. The findings revealed that the decision tree algorithm exhibited the best level of accuracy in forecasting heart illness, achieving a rate of 93.19%. The SVM algorithm followed closely behind with a rate of 92.30%. This work offers valuable insights into the potential of machine learning techniques as a powerful tool for forecasting heart failure disease. It specifically highlights the decision tree algorithm as a promising choice for further research.

Hasan et al. [21] conducted a study to determine the most effective feature selection method for predicting cardiovascular disease by comparing several algorithms. The initial consideration involved three widely recognized feature selection methods: Filter, wrapper, and embedding. Subsequently, a feature subset was obtained by applying a Boolean process-based common "True" condition to the results of these three algorithms. This method consisted of extracting subsets of features in two stages. Multiple models, such as random forest, support vector classifier, k-nearest neighbors, Naïve Bayes, and XGBoost, were considered to determine their comparative accuracy and choose the best predictive analytics model. The artificial neural network (ANN) was utilized as a benchmark for comparing all attributes. The study revealed that the XGBoost classifier, when combined with the wrapper approach, yielded the most precise predictions for cardiovascular disease. XGBoost achieved an accuracy rate of 73.74%, whereas SVC achieved 73.18% and ANN achieved 73.20%.

The main limitation of the previous research is its restricted dataset, which increases the likelihood of overfitting. The models created may not be suitable for extensive datasets. On the other hand, a large size dataset on cardiovascular diseases that included 70,000 patients and 12 characteristics is utilized in this research, which helped decrease the risk of overfitting. Table 1 provides a brief review of studies on predicting cardiovascular disease.

Table 1: Related work

Refs.	Techniques	Results	Limitations
Maiga et al. [18]	-Random forest -Naïve Bayes -Logistic regression -KNN	Random Forest algorithm achieves a classification accuracy of 73%, a specificity of 65%, and a sensitivity of 80%.	Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes), using only 3 models
Shah et al. [19]	Naïve Bayes, decision tree, random forest, and k-nearest neighbors (KKN)	KKN model demonstrated the best accuracy of 90.8%.	Cleveland heart disease dataset of 303 occurrences and 17 attributes
Alotalibi [20]	Decision tree, logistic regression, random forest, Naïve Bayes, and support vector machine (SVM).	Decision tree algorithm exhibited the best level of accuracy rate of 93.19%.	Cleveland heart disease dataset of 303 occurrences and 17 attributes
Hasan et al. [21]	Random forest, support vector classifier, k-nearest neighbors, Naïve Bayes, and XGBoost	XGBoost achieved an accuracy rate of 73.74%, whereas SVC achieved 73.18% and ANN achieved 73.20%.	UCI cardiovascular dataset (303 patients, 14 attributes)

2.1 Machine Learning Models

2.1.1 Logistic Regression

Logistic Regression is a primary model used for classifying binary problems in several fields, such as healthcare. Its interpretability and ability to evaluate probabilities make it particularly valuable in diagnosing cardiovascular illness. Within clinical settings, it assists in evaluating risk and making informed decisions by offering valuable insights into the probability of particular outcomes.

2.1.2 Random Forest

The Random Forest algorithm is a popular ensemble learning technique that is extensively utilized in cardiovascular research due to its strong resilience and capacity to effectively handle data with a high number of dimensions. Through the process of combining forecasts from numerous decision trees, it reduces the risk of overfitting and effectively captures intricate relationships within cardiovascular datasets. Its usefulness in activities like risk prediction and disease classification is immeasurable.

2.1.3 Decision Tree

Decision trees are intuitive models that provide transparency and interpretability in the context of cardiovascular disease diagnosis. Data is partitioned based on distinctive qualities to construct a hierarchical structure resembling a tree, with each node representing a point of decision-making. Decision Trees are highly effective at finding significant predictor factors and are frequently utilized in clinical decision support systems for evaluating risk and developing treatment strategies.

2.1.4 Extra Trees

Extra Trees is an ensemble learning method that is similar to Random Forest, but it incorporates more randomness in the process of selecting features. The inherent unpredictability of Extra Trees contributes to the variety of tree structures, which in turn improves their ability to handle data with high levels of noise and prevents overfitting. Extra Trees is a valuable tool in cardiovascular research for creating accurate diagnostic and prognostic models, especially when dealing with complicated or noisy information.

2.1.5 Support Vector Machine (SVM)

SVM is a highly effective model used for both classification and regression tasks in the investigation of cardiovascular diseases. Support Vector Machines (SVM) classify data points by identifying the hyperplane that maximizes the distance between different classes in a space with many dimensions. The valuable applications of this tool include patient risk stratification and disease categorization, thanks to its capability to handle non-linear correlations and high-dimensional data.

2.1.6 Gaussian Naïve Bayes

Gaussian Naïve Bayes is a probabilistic model that is both straightforward and highly effective, and it is extensively employed in the diagnosis of cardiovascular disease. The assumption is that the features are independent and follow a normal distribution. This makes the method computationally efficient and resistant to the influence of irrelevant features. Gaussian Naïve Bayes is used in healthcare contexts to assist with risk assessment and classification tasks. It provides probabilistic predictions by analyzing patient data.

2.1.7 XGBoost

XGBoost is a gradient-boosting algorithm that is widely recognized for its outstanding performance in predictive modeling applications, specifically in the analysis of cardiovascular disease. The process involves constructing a series of decision trees, where a differentiable loss function is optimized at each step. The favored choice for risk prediction, disease categorization, and feature selection in cardiovascular research is XGBoost due to its capacity to handle big datasets, capture intricate correlations, and prevent overfitting.

2.1.8 Light GBM

Light GBM is a gradient-boosting framework that has gained recognition for its ability to process large-scale datasets rapidly and effectively. The method employed in this system is a unique tree-based learning approach that emphasizes the growth of leaves in a hierarchical manner, resulting in reduced computational requirements and enhanced training speed. Light GBM expedites model construction and allows for real-time analysis in cardiovascular research, hence aiding prompt decision-making and patient care.

2.1.9 Multi-Layer Perceptron (MLP)

MLP, or Multilayer Perceptron, is a crucial component of artificial neural networks. It has the ability to acquire intricate patterns from cardiovascular data through learning. The system is composed of numerous layers of interconnected nodes, or neurons, which enables it to effectively capture complex relationships and representations in high-dimensional spaces. MLPs are particularly effective in cardiovascular disease analysis, demonstrating superior performance in tasks such as predicting risk, classifying diseases, and extracting features from various data sources.

Each of these machine learning models has strengths that are useful for cardiovascular disease diagnosis and prognosis, providing a complete toolkit. Logistic Regression simplifies binary classification, while Random Forest and Decision Tree models capture complex interactions in high-dimensional data for feature selection and clinical decision-making. SVM can separate data points in high-dimensional domains, making it useful for risk stratification and disease categorization. The Gaussian Naïve Bayes algorithm is efficient and reliable for probabilistic classification problems, especially with big datasets. Ensemble approaches like Extra Trees and XGBoost reduce overfitting and capture complex cardiovascular data correlations to improve predicting performance. Light GBM's speed and efficiency enable real-time analysis and decision-making, while Multi-layer Perceptron learns complex patterns from diverse data sources to predict risk and classify diseases in cardiovascular research and clinical practice. These models can improve cardiovascular care through precise diagnosis, prognosis, and individualized treatment.

3 Methodology

The research analyses the machine learning models for Cardiovascular Disease Predictions using the Risk Factors dataset. Several Machine Learning algorithms are used in this research, including Logistic Regression, Random Forest, Decision Tree, Extra Trees, Support Vector Classifier (SVC), XGBoost, Light GBM, Gaussian Naïve Bayes, and Multi-layer Perceptron (MLP). The dataset preprocessing is performed, after which the feature extraction is performed to understand the data. Furthermore, the dataset analysis is performed using some of the extracted features. The dataset is divided into two parts, training, and testing, with 70% and 30% of the total, respectively, to implement the models, where the models are trained using the training set, and the performance of the models is evaluated using a test set through various evaluation metrics including accuracy, precision, recall, and F-1 score. The methodology of this research work is shown in Fig. 1.

3.1 Dataset

This dataset includes thorough information about cardiovascular disease risk factors [22]. It includes information of 70,000 patients with 12 distinct features including age, gender, height, weight, blood pressure, cholesterol, glucose, smoking habits, and alcohol use. Furthermore, it provides information regarding the individual's activity level and presence of any cardiovascular conditions. Additionally, it outlines if the person is active or not and if he or she has any cardiovascular diseases. The Age is in days, which is converted into years. To better understand cardiovascular disease and develop more effective preventative strategies, researchers use this dataset to investigate potential relationships between risk factors and the condition. It can be accomplished by applying modern machine learning techniques. The dataset attributes are shown in Table 2.

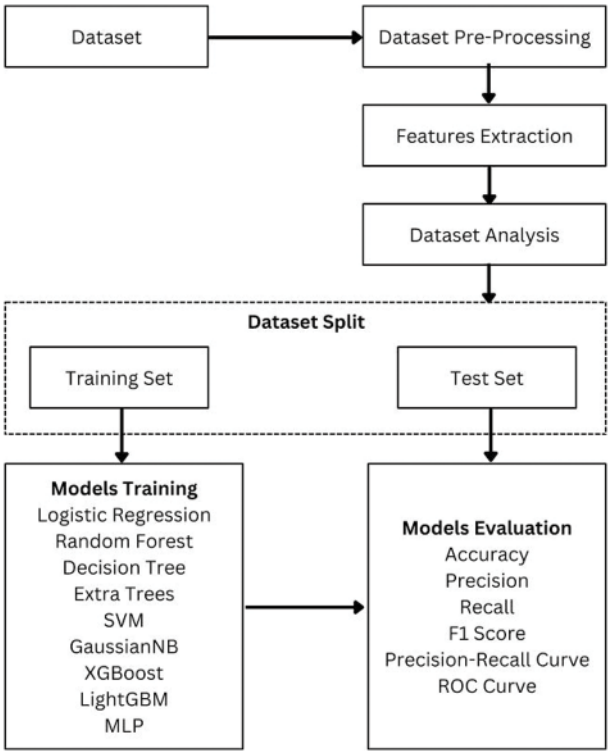

Figure 1: Methodology

Table 2: Dataset attributes

Feature	Variable	Min and max values
Age	Age	Min: 10,798 and max: 23,713 days
Height	Height	Min: 55 and max: 250
Weight	Weight	Min: 10 and max: 200
Gender	Gender	1: Female, 2: Male
Systolic blood pressure	ap_hi	Min: −150 and max: 16,020
Diastolic blood pressure	ap_lo	Min: −70 and max: 11,000
Cholesterol	Chol	Categorical value = 1(min) to 3(max)
Glucose	Gluc	Categorical value = 1(min) to 3(max)
Smoking	Smoke	1: Yes, 0: No
Alcohol intake	Alco	1: Yes, 0: No
Physical activity	Active	1: Yes, 0: No
Presence or absence of cardiovascular disease	Cardio	1: Yes, 0: No

Fig. 2 shows the dataset overview according to diseases.

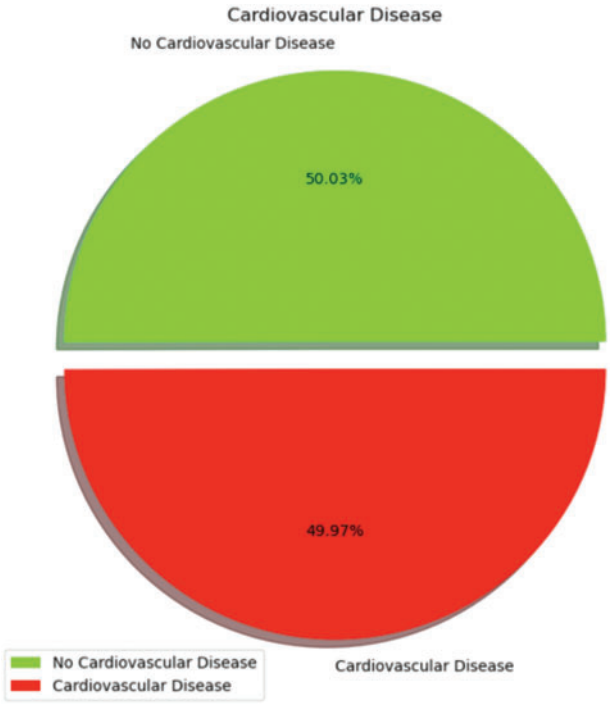

Figure 2: Dataset distribution

3.2 Feature Extraction

The feature extraction is performed using a correlation matrix, which shows the data available in the dataset, including age, gender, weight, and height, along with cholesterol, smoke, glucose, alcohol, and cardio, i.e., 0 or 1. Fig. 3 shows the correlation matrix.

3.3 Dataset Analysis

The dataset analysis uses various features, including age, height, weight, and gender. The dataset analysis is shown below. Figs. 4–7 show the risk factors analysis using Age, Weight, Height, and Gender, respectively.

3.4 Evaluation Metrics

The performance of various machine learning algorithms is analyzed using the metrics such as Accuracy, Precision, Recall, F1 Score, Precision-Recall Curve and ROC-AUC Curve.

3.4.1 Accuracy

Accuracy is the most common metric used to evaluate classification algorithms. Accuracy is defined as the ratio of correct predictions to the total number of samples. The primary criterion for assessing the effectiveness of supervised machine learning algorithms is accuracy.

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

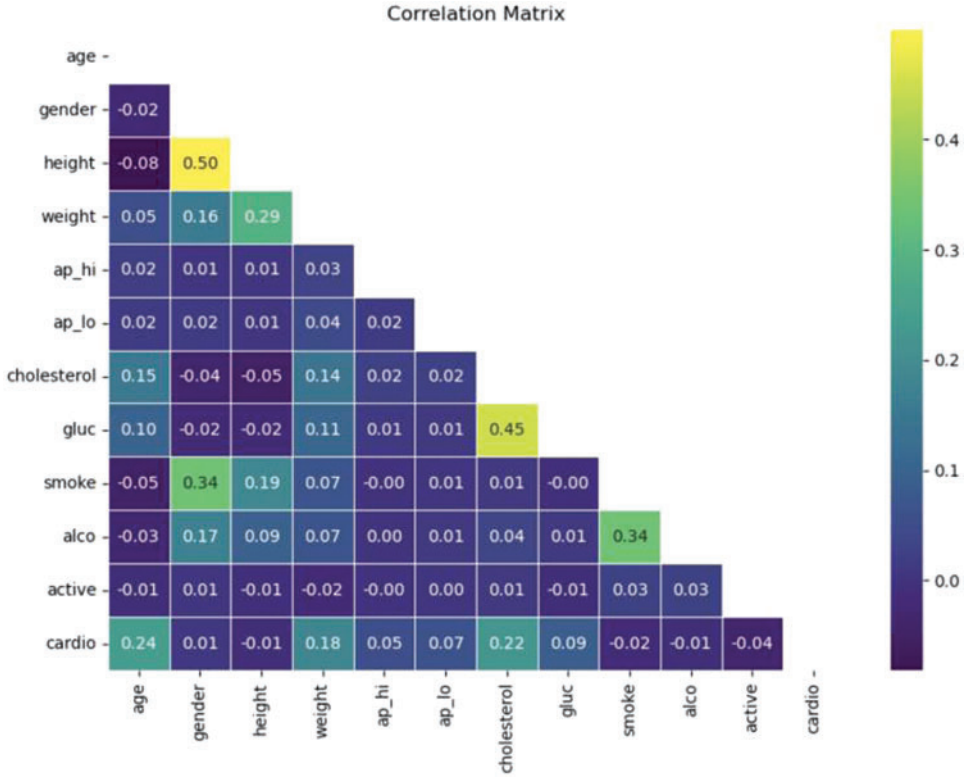

Figure 3: Correlation matrix

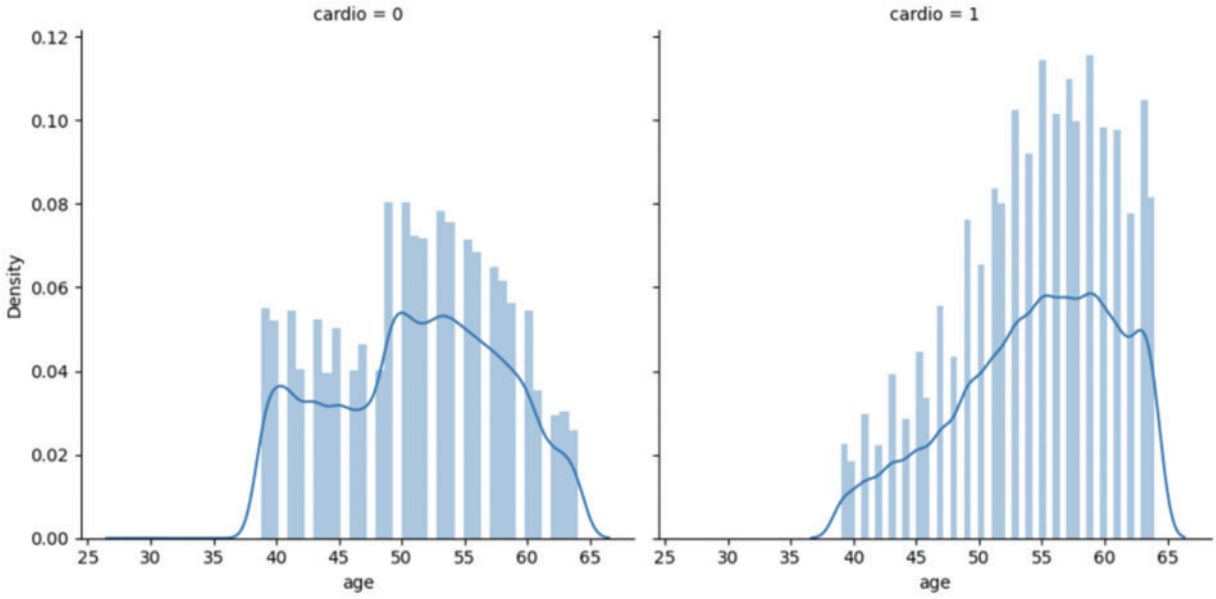

Figure 4: Risk factors analysis using age

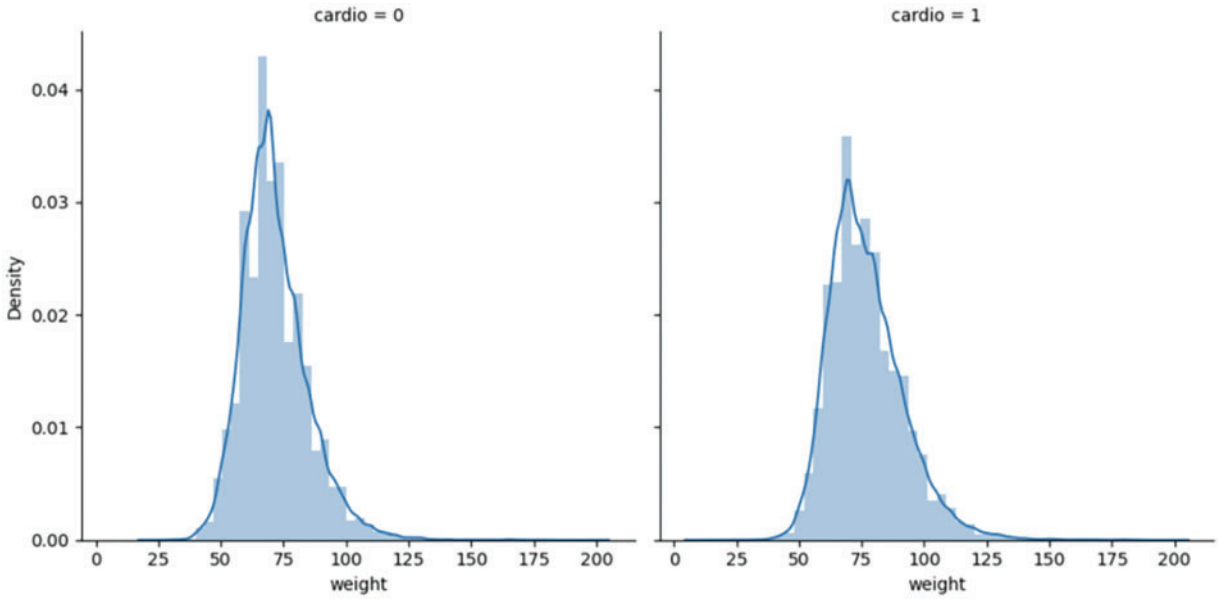

Figure 5: Risk factors analysis using weight

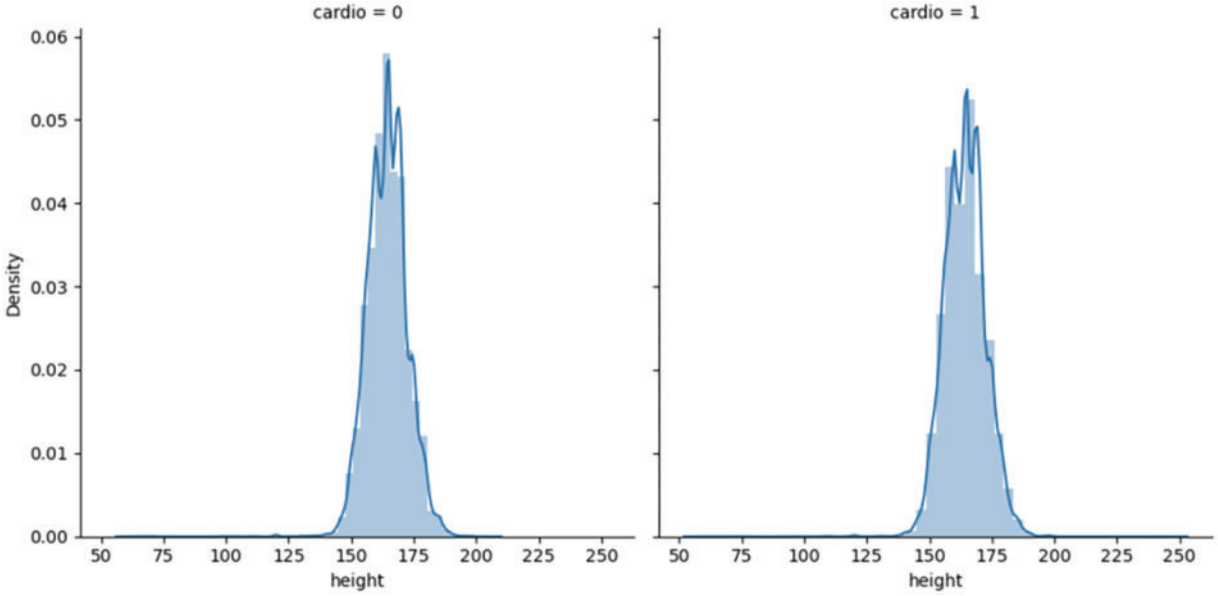

Figure 6: Risk factors analysis using height

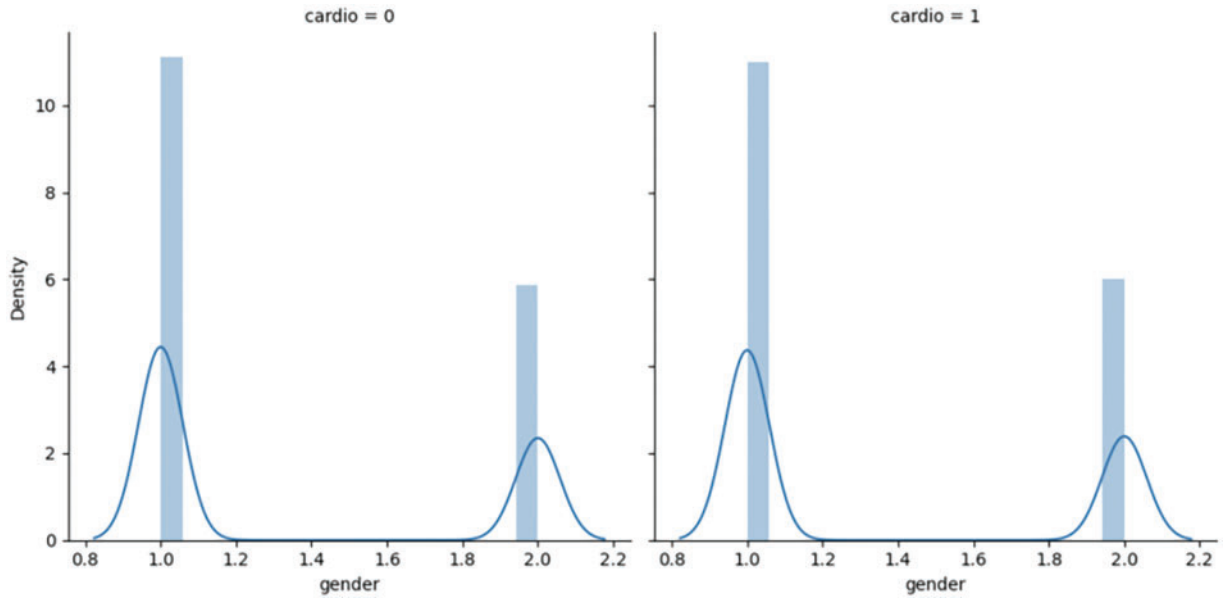

Figure 7: Risk factors analysis using gender

3.4.2 Precision

Precision is defined as the ratio of true positives to all positive predictions made by the algorithms. The precision is the percentage of relevant algorithm outcomes.

$$Precision = \frac{TP}{TP + FP}$$

3.4.3 Recall

It is the average probability of complete retrieval.

$$Precision = \frac{TP}{TP + FN}$$

3.4.4 F1 Score

The F1 Score is the combination of precision and recall. It is calculated by taking the harmonic mean of precision and recall.

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

3.4.5 Area Under the ROC Curve (AUC)

The area under the Receiver Operating Characteristic (ROC) Curve (AUC) is a performance metric mostly used for classification tasks. The AUC is calculated by plotting the True Positive Rate (TPR), also known as sensitivity or recall, against the False Positive Rate. The higher the AUC, the more accurate the model is at predicting heart disease. The ROC curve displays the performance trade-off between a classification model's TPR and False Positive Rate (FPR).

4 Implementation and Results

The main objective of this research is to implement different machine-learning models for predicting cardiovascular heart disease utilizing a dataset of risk variables. The machine learning models are applied to the Cardiovascular Risk Factors dataset. Based on the risk factors, the dataset predicts cardiovascular heart illnesses. The dataset is divided into two parts, training and testing, with 70% and 30% of the total, respectively, to implement the models. Following their training on the training set, the machine learning models are evaluated on the test set. The performance of the machine learning model is assessed by comparing it using several evaluation measures, such as precision, recall, and F1 Score. Additionally, the Precision-Recall and AUC curves are utilized, and the confusion matrix is displayed for each model output. In addition, the evaluation measures allow for a thorough comparison of the models' performance.

4.1 Evaluation Results

In this section, the performance of the models is highlighted, using several evaluation metrics and a confusion matrix.

4.1.1 Logistic Regression

With a precision of 0.73, recall of 0.66, accuracy of 0.71, and F1 Score of 0.70, the logistic regression model performs well in the prediction of cardiovascular disease. Fig. 8 displays the confusion matrix for the Logistic Regression model.

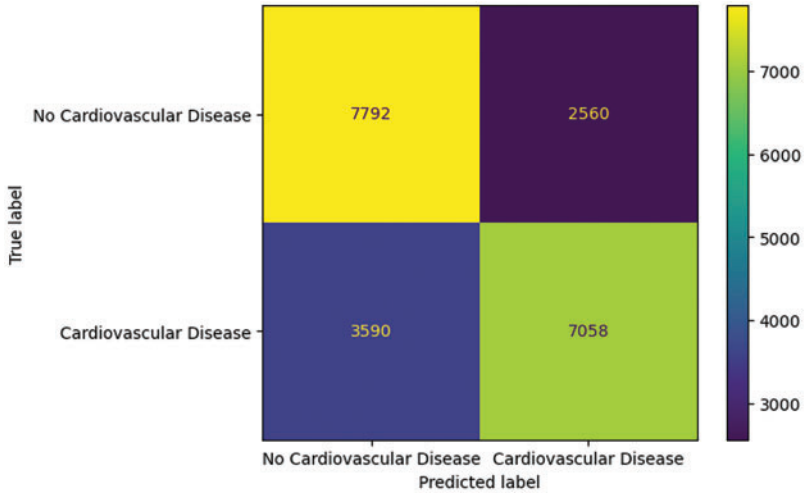

Figure 8: Confusion matrix of logistic regression

The performance of the model is shown in Table 3.

Table 3: Logistic regression performance

Metrics	Performance
Precision	0.73
Recall	0.66
Accuracy	0.71
F1 Score	0.70

4.1.2 Random Forest

With a precision of 0.72, recall of 0.70, accuracy of 0.72, and F1 Score of 0.71, the Random Forest has performed well. Fig. 9 displays the confusion matrix for the Random Forest model.

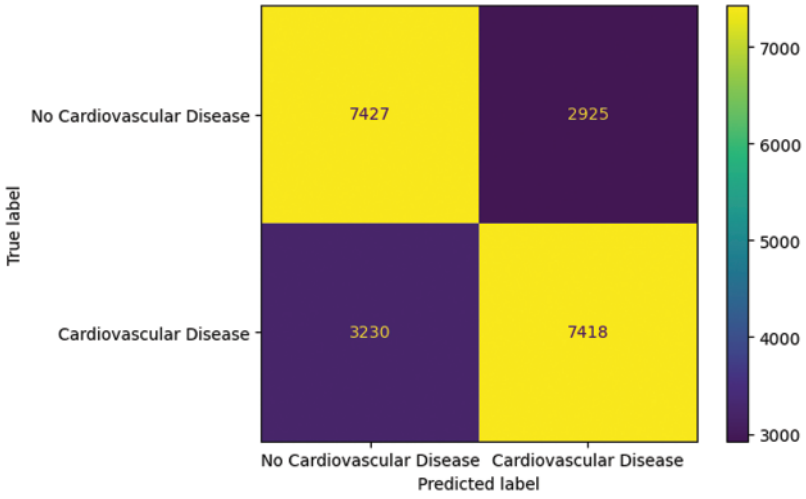

Figure 9: Confusion matrix of random forest

The performance of the model is shown in Table 4.

Table 4: Random forest performance

Metrics	Performance
Precision	0.72
Recall	0.70
Accuracy	0.72
F1 Score	0.71

4.1.3 Decision Tree

The Precision, Recall, Accuracy, and F1 Score for the Decision Tree are all 0.64, 0.62, 0.63, and 0.63, respectively. Fig. 10 displays the confusion matrix about the Decision Tree model.

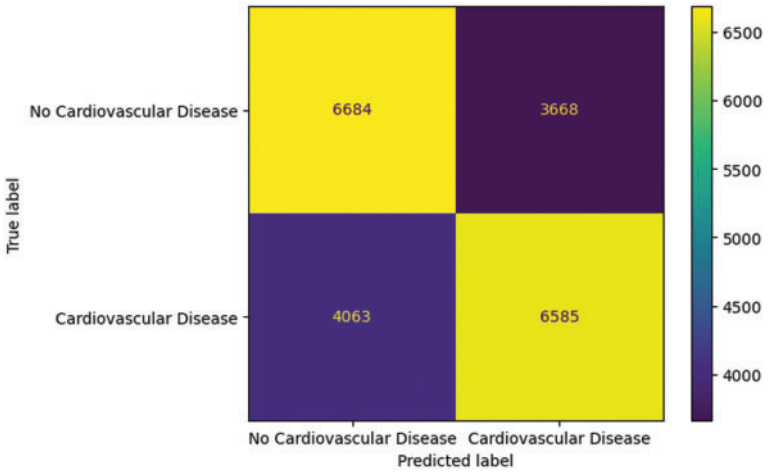

Figure 10: Confusion matrix of decision tree

The performance of the model is shown in Table 5.

Table 5: Decision tree performance

Metrics	Performance
Precision	0.64
Recall	0.62
Accuracy	0.63
F1 Score	0.63

4.1.4 Extra Trees Classifier

The Extra Trees Classifier has obtained a Precision of 0.70, Recall of 0.69, Accuracy of 0.69, and F1 Score of 0.69. The confusion matrix for the Extra Trees Classifier model is displayed in Fig. 11.

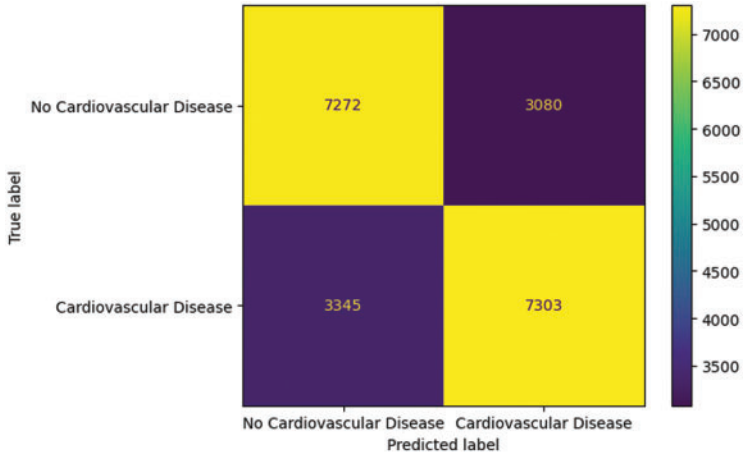

Figure 11: Confusion matrix of extra trees classifier

The performance of the model is shown in Table 6.

Table 6: Extra trees classifier performance

Metrics	Performance
Precision	0.70
Recall	0.69
Accuracy	0.69
F1 Score	0.69

4.1.5 Support Vector Classifier (SVC)

The SVC has obtained a Precision of 0.78, Recall of 0.62, Accuracy of 0.72, and F1 Score of 0.69. The confusion matrix for the SVC model is depicted in Fig. 12.

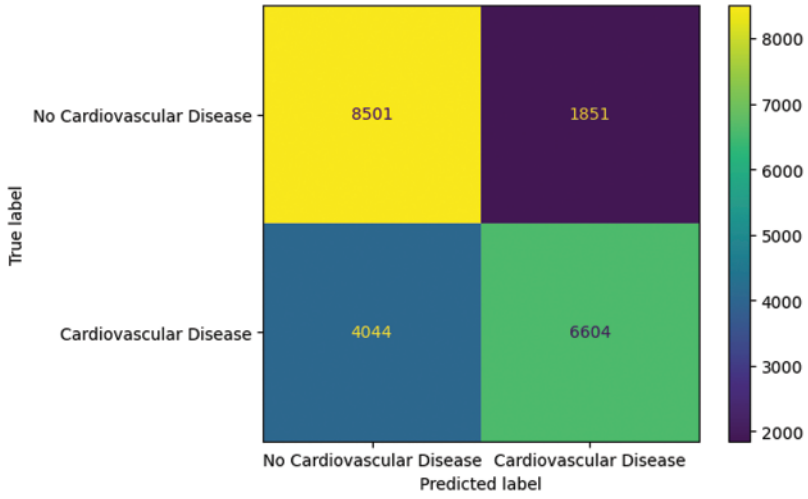

Figure 12: Confusion matrix of support vector classifier

The performance of the model is shown in Table 7.

Table 7: Support vector classifier performance

Metrics	Performance
Precision	0.78
Recall	0.62
Accuracy	0.72
F1 Score	0.69

4.1.6 Gaussian Naïve Bayes

The Gaussian Naïve Bayes model has attained a Precision of 0.74, Recall of 0.27, Accuracy of 0.58, and F1 Score of 0.40. The confusion matrix for the Gaussian Naïve Bayes model is displayed in Fig. 13.

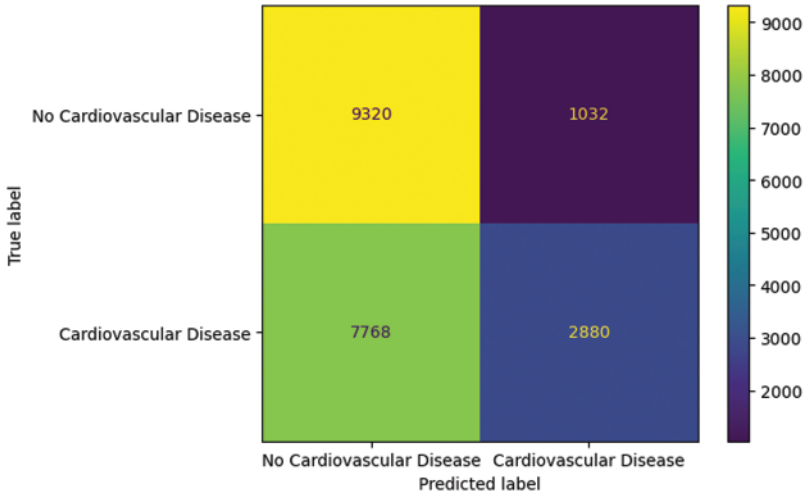

Figure 13: Confusion matrix of Gaussian Naïve Bayes

The performance of the model is shown in Table 8.

Table 8: Gaussian Naïve Bayes performance

Metrics	Performance
Precision	0.76
Recall	0.69
Accuracy	0.72
F1 Score	0.72

4.1.7 XGBoost Classifier (XGB)

The XGBoost Classifier has attained a Precision of 0.76, Recall of 0.69, Accuracy of 0.72, and F1 Score of 0.72. The confusion matrix of the XGBoost Classifier model is depicted in Fig. 14.

The performance of the model is shown in Table 9.

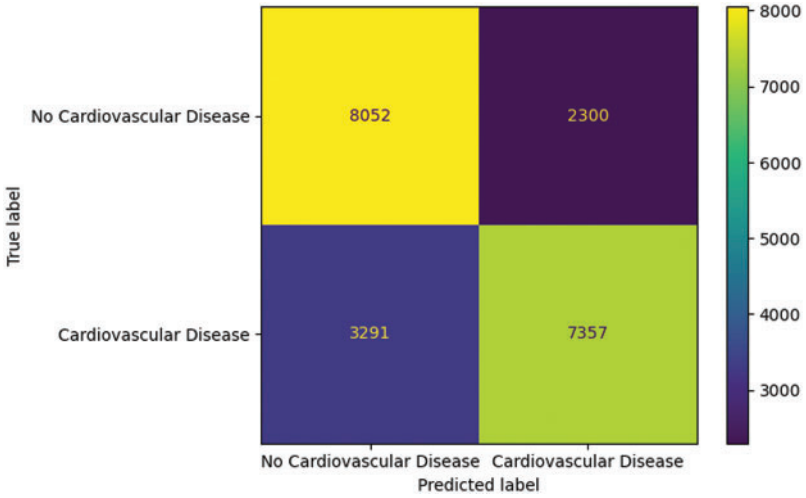

Figure 14: Confusion matrix of XBG classifier

Table 9: XBG classifier performance

Metrics	Performance
Precision	0.76
Recall	0.70
Accuracy	0.73
F1 Score	0.73

4.1.8 Light GBM Classifier (LGBM)

The Light GBM Classifier has attained a Precision of 0.76, Recall of 0.70, Accuracy of 0.73, and F1 Score of 0.73. The confusion matrix of the Light GBM Classifier model is displayed in Fig. 15.

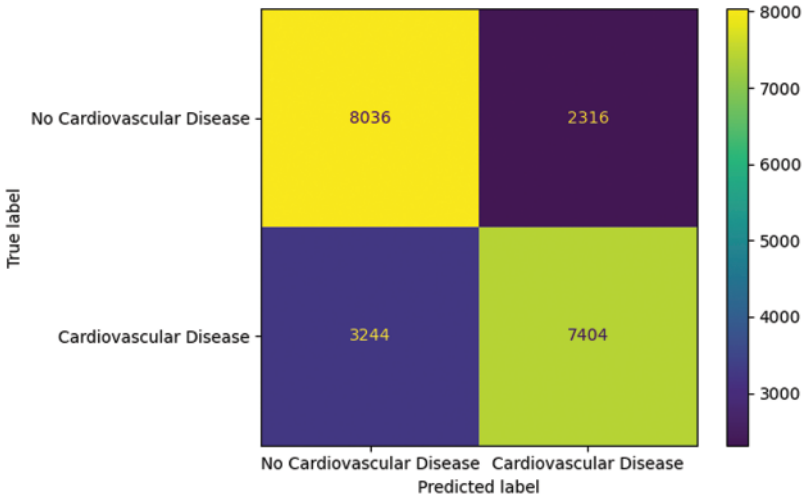

Figure 15: Confusion matrix of LGBM

The performance of the model is shown in Table 10.

Table 10: LGBM performance

Metrics	Performance
Precision	0.74
Recall	0.27
Accuracy	0.58
F1 Score	0.40

4.1.9 Multi-Layer Perceptron (MLP)

The Multi-Layer Perceptron has attained a Precision of 0.77, Recall of 0.64, Accuracy of 0.72, and F1 Score of 0.70. Fig. 16 displays the confusion matrix of the MLP model.

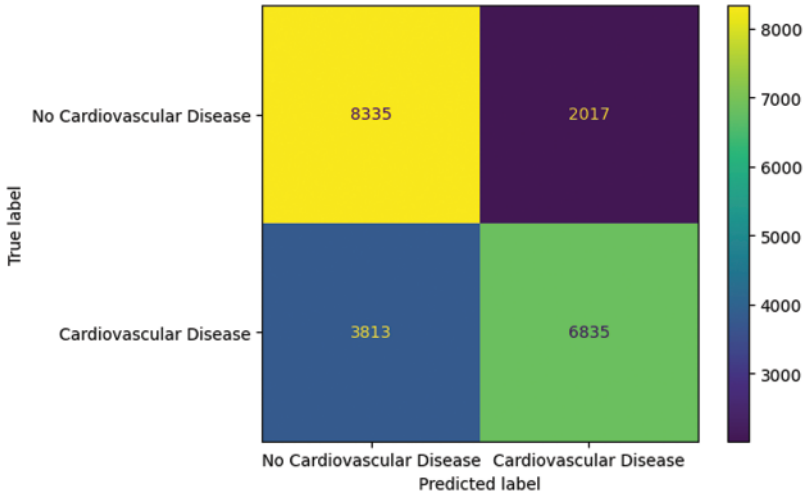

Figure 16: Confusion matrix of MLP

The performance of the model is shown in Table 11.

Table 11: MLP performance

Metrics	Performance
Precision	0.77
Recall	0.64
Accuracy	0.72
F1 Score	0.70

5 Discussion and Analysis

This section provides a performance comparison of machine learning models used for predicting cardiovascular illness. The comparison is based on performance metrics such as accuracy, precision, recall, and F1 Score. Additionally, the Precision-Recall and ROC curves are included. It offers an in-depth evaluation of the model's performance.

5.1 Performance Comparison

Among the models evaluated, XGBoost exhibits the most favorable overall performance, as seen by its superior F1 Score (0.73) and Accuracy (0.73). Additionally, it has a comparatively high precision of 0.76 and a recall of 0.70. The Support Vector Machine demonstrates the highest precision, achieving a value of 0.78. The Gaussian Naïve Bayes is closely behind, which has good precision at 0.76 and recall at 0.69. On the other hand, it is worth noting that Light GBM exhibits a significantly lower Recall value of 0.27, which consequently leads to a lower F1 Score of 0.40. The Decision Tree algorithm consistently demonstrates the lowest performance across all evaluation metrics, including Precision (0.64), Recall (0.62), Accuracy (0.63), and F1 Score (0.63). When selecting the optimal model, it is imperative to consider the performance indicators, namely Precision and Recall, along with other pertinent criteria such as computational efficiency and dataset characteristics. The performance of the machine learning models is presented in Table 12.

Table 12: Performance comparison of machine learning models

Models	Precision	Recall	Accuracy	F1 Score
Logistic regression	0.73	0.66	0.71	0.70
Random forest	0.72	0.70	0.72	0.71
Decision tree	0.64	0.62	0.63	0.63
Extra trees	0.70	0.69	0.69	0.69
Support vector machine	0.78	0.62	0.72	0.69
Gaussian Naïve Bayes	0.76	0.69	0.72	0.72
XGBoost	0.76	0.70	0.73	0.73
Light GBM	0.74	0.27	0.58	0.40
Multi-layer perceptron	0.77	0.64	0.72	0.70

The performance of the machine learning models is presented in Fig. 17.

The evaluation of different machine learning models using the given dataset of cardiovascular risk variables indicates that XGBoost is the most suitable solution for accurate prediction and classification. XGBoost exhibits outstanding performance in several measures such as precision, recall, accuracy, and F1 Score. It demonstrates a well-balanced and resilient approach in dealing with intricate data patterns that are inherent in cardiovascular risk factor research. Although Support Vector Machine and Gaussian Naïve Bayes have their own benefits in some measures, XGBoost stands out as the best option for achieving high overall predicting accuracy and dependability. These findings emphasize the significance of choosing the most appropriate model for the particular context and goals of cardiovascular disease research and clinical decision-making. XGBoost is a compelling choice for improving risk assessment and patient care in cardiovascular health.

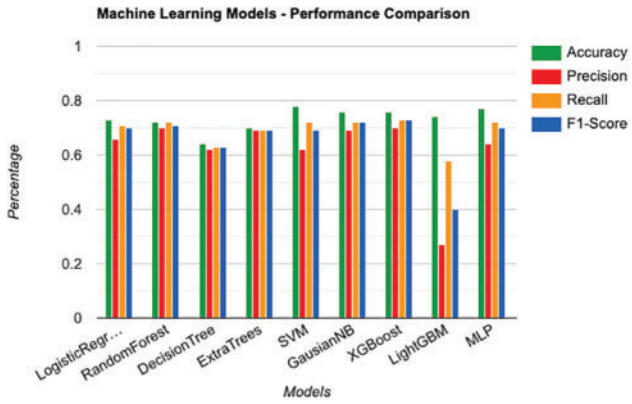

Figure 17: Performance comparison of machine learning models

5.2 Precision-Recall Curve

The Precision-Recall curves provide useful insights into the performance of binary classification models, especially when class imbalances exist. XGBoost, SVC, MLP, and Light GBM exhibit exceptional performance, as seen by their identical Average Precision (AP) ratings of 0.68. These models demonstrate a robust trade-off between precision and recall, efficiently discerning good examples from negative ones. The Random Forest algorithm demonstrates strong precision-recall properties, as seen by its AP value of 0.65. The Decision Tree and GaussianNB models, with AP ratings of 0.59 and 0.57, respectively, demonstrate acceptable albeit somewhat diminished performance, implying the possibility of enhancing their effectiveness. The occurrence of a negative AP score in Logistic Regression is atypical and necessitates additional examination, given that AP values generally fall within the range of 0 to 1. The Precision-Recall curves highlight the effectiveness of XGBoost, SVC, MLP, and Light GBM in attaining a desirable balance between precision and recall. This balance is particularly important when accurately detecting positive examples is significant. Fig. 18 shows the comparison of the precision-recall curve.

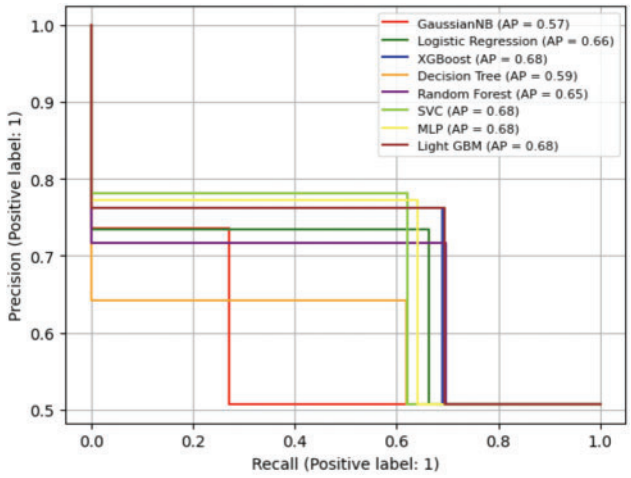

Figure 18: Comparison of precision-recall curve

5.3 ROC Curve

The ROC curves show the performance of binary classification models over different decision thresholds, highlighting significant differences among the models. The Light GBM model demonstrates superior performance, with a notable ROC-AUC score of 0.74. This value signifies the model's ability to differentiate between positive and negative cases effectively. The XGBoost, SVC, and MLP models exhibit AUC scores of 0.73 and 0.72, indicating strong discriminatory capabilities. Logistic Regression and Random Forest exhibit notable performance, as evidenced by their AUC values 0.71. The Decision Tree model, while achieving a somewhat lower discrimination capability of 0.63, nevertheless demonstrates a reasonable level of performance. The GaussianNB model demonstrates relatively inferior performance compared to the other models, as evidenced by its AUC value of 0.59. ROC-AUC highlights the effectiveness of Light GBM, XGBoost, SVC, and MLP in attaining elevated true positive rates while simultaneously keeping false positive rates at a minimum. This demonstrates their robust discriminatory ability in tasks involving binary classification. Fig. 19 shows the comparison of the precision-recall curve.

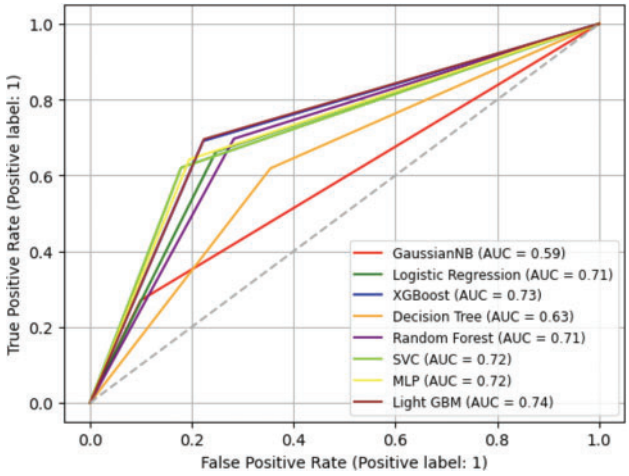

Figure 19: Comparison of ROC curve

5.4 Discussion

After evaluating the strengths and weaknesses of both the high-performing and low-performing models using the given metrics, it is evident that XGBoost emerges as the most effective performer. XGBoost shows the ability in managing complex information connection associated with cardiovascular risk factor analysis, as evidenced by its outstanding scores in accuracy, precision, recall, and F1 Score across all measures. The versatility and suitability of this technology for many sorts of datasets and activities is due to its exceptional ability to capture subtle patterns. Nevertheless, the dependence of XGBoost on computational resources and the requirement for thorough parameter tweaking might present difficulties, particularly in contexts with limited resources. SVM also exhibits high precision, suggesting its ability to properly identify affirmative cases. The great utility of this instrument extends beyond cardiovascular risk assessment due to its efficacy in effectively distinguishing data points in environments with a high number of dimensions.

Gaussian Naïve Bayes is notable for its equitable performance in both precision and recall, providing dependability in prediction problems. Despite its restricted expressiveness compared to more complicated models, its simplicity and computational efficiency make it well-suited for lightweight

categorization tasks. However, the assumption of feature independence made by Naïve Bayes may not be valid for all datasets, which could impact its performance in specific situations. at contrast, Light GBM exhibits rapidity and effectiveness, rendering it highly skilled at managing extensive datasets and doing real-time analysis. The capacity of the model to properly handle categorical features and missing data enhances its attractiveness in specific applications. Nevertheless, the comparatively inferior precision, recall, and accuracy of Light GBM in respect to other models give rise to doubts over its capacity to effectively capture intricate data associations. Moreover, the fact that it is prone to overfitting, especially when dealing with imbalanced datasets or noisy features, indicates that it may not be entirely reliable for some applications.

To summarize, XGBoost, SVM, and Gaussian Naïve Bayes demonstrate proficiency in precision, accuracy, and dependability. However, it is important to note that each model also has its own set of constraints. Although Light GBM is efficient in terms of speed and resource utilization, it falls short in performance measures, suggesting difficulties in effectively collecting intricate data patterns. When choosing the most appropriate model, it is important to take into account the specific needs and limitations of the application, while also finding a balance between performance, computational efficiency, and interpretability.

6 Conclusion

In cardiovascular heart disease prediction, the assessment of different machine learning models has yielded significant insights into their efficacy in identifying individuals who are susceptible to the condition. The performance of the machine learning models is evaluated using performance metrics, including Precision, Recall, F1 Score, and the Precision-Recall and AUC curve. The SVM demonstrates a significant precision value of 0.78, indicating a robust capacity to accurately classify positive instances, a critical factor in predicting heart disease. Nevertheless, the recall value of 0.62 suggests that although the SVM model has high accuracy, it may fail to identify certain positive cases. This highlights the need to consider both precision and recall in evaluating cardiovascular risk assessment. XGBoost and Gaussian Naïve Bayes have demonstrated considerable potential in heart disease prediction, exhibiting a favorable equilibrium between precision and recall. This is reflected in their notable F1 Score of 0.73 and 0.72, respectively. Achieving a balance between accurately identifying positive instances and reducing the occurrence of false negatives is of utmost importance within the healthcare domain. The ongoing development and refinement of models in the field of cardiovascular health offer significant potential for enhanced identification of patients at risk of heart disease, leading to more precise and timely therapies and, ultimately, better patient outcomes.

This research work implemented the machine learning models which show better performance for the dataset used, but it is still not enough for accurate predictions of cardiovascular disease. Also, the research uses only one dataset for the performance analysis, however, multiple datasets can be used for the performance comparisons. In the future, models can be implemented and analyzed using multiple datasets. Furthermore, advanced machine learning and deep learning-based algorithms can be applied to cardiovascular predictions.

Acknowledgement: None.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm their contribution to the paper as follows: Study conception and design: Adil Hussain; data collection: Ayesha Aslam; analysis and interpretation of results:

Adil Hussain, Ayesha Aslam; draft manuscript preparation: Ayesha Aslam. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available in Risk Factors for Cardiovascular Heart Disease at https://www.kaggle.com/datasets/thedevastator/exploring-risk-factors-for-cardiovascular-diseas.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] C. Trevisan, G. Sergi, and S. Maggi, "Gender differences in brain-heart connection," *Brain Heart Dynam.*, pp. 937–951, 2020. doi: 10.1007/978-3-030-28008-6.

[2] D. C. Yadav and S. Pal, "Prediction of heart disease using feature selection and random forest ensemble method," *Int. J. Pharm. Res.*, vol. 12, no. 4, pp. 56–66, 2020.

[3] K. Uyar and A. İlhan, "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks," *Procedia Comput. Sci.*, vol. 120, no. 3, pp. 588–593, 2017. doi: 10.1016/j.procs.2017.11.283.

[4] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mob. Inf. Syst.*, vol. 2018, no. 8, pp. 1–21, 2018. doi: 10.1155/2018/3860146.

[5] S. Pouriyeh, S. Vahid, G. Sannino, G. de Pietro, H. Arabnia and J. Gutierrez, "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease," in *2017 IEEE Symp. Comput. Commun. (ISCC)*, IEEE, 2017, pp. 204–207.

[6] J. Mourao-Miranda, A. L. Bokde, C. Born, H. Hampel, and M. Stetter, "Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional MRI data," *NeuroImage*, vol. 28, no. 4, pp. 980–995, 2005. doi: 10.1016/j.neuroimage.2005.06.070.

[7] S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, "Innovative artificial neural networks-based decision support system for heart diseases diagnosis," *J. Intell. Learn. Syst. Appl.*, vol. 5, no. 3, pp. 176–183, 2013. doi: 10.4236/jilsa.2013.53019.

[8] F. Amato, A. López, E. M. Peña-Méndez, P. Vaňhara, A. Hampl and J. Havel, *Artificial Neural Networks in Medical Diagnosis*. Warsaw, Poland, Elsevier, vol. 11, pp. 47–58, 2013.

[9] A. Saboor, M. Usman, S. Ali, A. Samad, M. F. Abrar and N. Ullah, "A method for improving prediction of human heart disease using machine learning algorithms," *Mob. Inf. Syst.*, vol. 2022, no. 15, pp. 1–9, 2022. doi: 10.1155/2022/1410169.

[10] H. Benhar, A. Idri, and J. Fernández-Alemán, "Data preprocessing for heart disease classification: A systematic literature review," *Comput. Methods Programs Biomed.*, vol. 195, no. 5, pp. 105635, 2020. doi: 10.1016/j.cmpb.2020.105635.

[11] N. Kausar, S. Palaniappan, B. B. Samir, A. Abdullah, and N. Dey, "Systematic analysis of applied data mining based optimization algorithms in clinical attribute extraction and classification for diagnosis of cardiac patients," *Appl. Intell. Optimiz. Biology Med.: Current Trends Open Prob.*, vol. 96, pp. 217–231, 2016. doi: 10.1007/978-3-319-21212-8.

[12] M. M. Alam *et al.*, "D-care: A non-invasive glucose measuring technique for monitoring diabetes patients," in *Proc. Int. Joint Conf. Comput. Intell.*, Springer, 2020, pp. 443–453.

[13] A. Hussain, A. Khatoon, A. Aslam, and M. A. Khosa, "A comparative performance analysis of machine learning models for intrusion detection classification," *J. Cyber Secur.*, vol. 6, no. 1, pp. 1–23, 2024.

[14] A. Aslam and A. Hussain, "A performance analysis of machine learning techniques for credit card fraud detection," *J. Artif. Intell.*, vol. 6, pp. 1–21, 2024. doi: 10.32604/jai.2024.047226.

[15] I. D. Mienye, Y. Sun, and Z. Wang, "An improved ensemble learning approach for the prediction of heart disease risk," *Inform. Med. Unlocked*, vol. 20, no. 8, pp. 100402, 2020. doi: 10.1016/j.imu.2020.100402.
[16] H. Wang, Z. Huang, D. Zhang, J. Arief, T. Lyu and J. Tian, "Integrating co-clustering and interpretable machine learning for the prediction of intravenous immunoglobulin resistance in kawasaki disease," *IEEE Access*, vol. 8, pp. 97064–97071, 2020. doi: 10.1109/ACCESS.2020.2996302.
[17] B. A. Tama, S. Im, and S. Lee, "Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble," *Biomed Res. Int.*, vol. 2020, pp. 1–10, 2020. doi: 10.1155/2020/9816142.
[18] J. Maiga and G. G. Hungilo, "Comparison of machine learning models in prediction of cardiovascular disease using health record data," in *Proc. ICIMCIS*, Jakarta, Indonesia, 2019, pp. 45–48.
[19] D. Shah, S. Patel, and S. K. Bharti, "Heart disease prediction using machine learning techniques," *SN Comput. Sci.*, vol. 1, no. 6, pp. 345, 2020. doi: 10.1007/s42979-020-00365-y.
[20] F. S. Alotaibi, "Implementation of machine learning model to predict heart failure disease," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, pp. 261–268, 2019. doi: 10.14569/issn.2156-5570.
[21] N. Hasan and Y. Bao, "Comparing different feature selection algorithms for cardiovascular disease prediction," *Health Technol.*, vol. 11, pp. 49–62, 2021.
[22] Kaggle, "Cardiovascular disease dataset," 2024. Accessed: Nov. 1, 2022. [Online]. Available: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset