Enhancing Cardiovascular Disease Risk Prediction Using Resampling and Machine Learning

Ayesha Kiran[1], Muhammad Kashir Khan[1], Muhammad Daniyal Khan[1], Farrukh Liaquat[2]
[1]Department of Software Engineering, Lahore Garrison University, Lahore, Pakistan
[2] School of Systems and Technology, University of Management and Technology, Lahore, Pakistan
***Corresponding Author:** Ayesha Kiran Email: ayeshakiran@lgu.edu.pk

Citation | Kiran. A, Khan. M. K, Khan. M. D, Liaquat. F, "Enhancing Cardiovascular Disease Risk Prediction Using Resampling and Machine Learning", IJIST, Vol. 6 Issue. 2 pp. 514-531, May 2024

Received | April 28, 2024, **Revised |** May 11, 2024, **Accepted |** May 17, 2024, **Published |** May 28, 2024.

Cardiovascular Disease (CVD) remains a critical health concern around the globe, requiring precise risk prediction approaches for timely intervention. The primary motive of this study is to enhance CVD risk prediction through innovative techniques, just like resampling the imbalanced datasets using random oversampling and employing advanced Machine Learning (ML). In this study, different robust ML algorithms such as Random Forest Classifier, Decision Tree Classifier, XGBoost Classifier and Logistic Regression were trained on a diverse dataset encompassing demographic, clinical and lifestyle factors related to CVD. By addressing class imbalance through oversampling, the models showed significant performance improvements, showcasing the effectiveness of our ML algorithms in accurately forecasting CVD risks. Specifically, the Random Forest model with an accuracy score of 96% and AUC-ROC score of 99%. This study emphasizes the potential of modern approaches to improve CVD risk assessment by leveraging cutting-edge technologies for enhanced healthcare outcomes. Enfolding these approaches and tools, it becomes easy to pave the way for more personalized risk assessment and early intervention strategies, eventually aiming to alleviate the global burden of CVD.

Keywords: Machine Learning, Cardiovascular Disease, Risk Prediction, Resampling, Random Forest Model

Introduction:

CVD stands as leading cause of death worldwide, posing serious threat to both public health and health infrastructure. There is a dire need of timely and effective prediction of CVD, to address this increasing burden and optimize resource utilization. Multiple factors such as hypertension and hyperlipidemia contribute to the development of CVD by causing inflammation of the heart and blood vessels [1]. Similarly, a sedentary lifestyle, poor dietary choices, and tobacco use exacerbate inflammation and expedite CVD progression [2] [3]. The symptoms of cardiovascular disease encompass a spectrum of conditions including acute coronary syndrome, heart failure, valvular disorders, stroke, rhythm abnormalities, and peripheral vascular disease [4]. Moreover, studies have associated depressive symptoms with an elevated risk of cardiovascular conditions like peripheral artery disease, atrial fibrillation/flutter, coronary artery disease, ischemic stroke, and heart failure [5]. However, conventional risk assessment tools often falter in capturing the intricate interplay of multifaceted factors underlying disease development. There are several diagnostic approaches for heart disease. In addition to routine blood tests and chest X-rays, tests to diagnose heart disease can include the following; ECG (Quick test that records electrical signals of the heart) [6], Echocardiogram (This examination uses sound waves to create detailed images of the heart in motion and it is noninvasive), Heart CT Scan and Cardia MRI [7]. The economic toll of cardiovascular disease is substantial, impacting healthcare systems globally. For instance, the economic cost of CVD in the Russian Federation in 2016 was 2.7 trillion ₽, or 3.2% of GDP. Similarly, CVD treatment imposes significant financial burdens on households, with nearly 50% of Indian households facing catastrophic medical expenses due to CVD-related treatments. These statistics highlight the need for increased investments in the prevention and treatment of CVD to alleviate the economic burden on healthcare systems [8].

Treatment cost for cardiovascular disease varies significantly across different regions and countries. A global review of cost estimates around the world, revealed that there was a wide variation in the average cost of cardiovascular events, with the US being more expensive than the EU [9]. Another analysis calculated that 54 countries saw 4.4 million cardiovascular disease-related mortalities in 2018, with a productivity loss of €62 billion consequently. Notably, 47% of the costs associated with cardiovascular disease were attributable to deaths from coronary heart disease [10]. Addressing this financial strain on healthcare systems necessitates the adoption of highly cost-effective and health-promoting interventions. Iran's IraPEN program emerged as a cost-effective strategy for all cardiovascular disease risk groups, offering substantial potential for both cost savings and improved health outcomes [11].

Heart disease describes a range of conditions that affect your heart, accounting for 17.9 million deaths annually, according to the World Health Organization [12]. As reported by the World Health Organization (WHO), almost 17 million people pass away yearly due to CVD, which accounts for about 31% of global deaths [13]. In 2012, 7.4 million people died from CVD and 6.7 million from stroke. Notably, CHD (Coronary Heart Disease) is a subtype of cardiovascular disease and accounts for 64% of cases, affecting both men and women significantly [14]. The consequences of CVDs are profound, with 17.9 million deaths worldwide. The World Health Organization (WHO) predicts that CVD will continue to be a cause of mortality, posing a serious threat to human life in the future, conceivably beyond 2030 [15]. Preventing and effectively treating heart disease hinges significantly on lifestyle modifications. While the heart attacks and strokes can be quite fatal and can lead to imminent death, early heart disease prediction is essential, and can lead to timely intervention by the careful application of machine learning techniques. The use of Machine Learning (ML) techniques can unearth hidden patterns in data and process diverse types of data leading to early prediction and detection of heart disease. These predictive models, trained on historical data, can make close to accurate predictions on unseen data.

Machine learning models outperforms traditional CVD risk prediction methods but challenges still persist, particularly concerning imbalanced data. The evaluation metrics such as accuracy can give false results and predict the majority class frequently. This study utilizes sampling techniques to tackle the uneven distribution of CVD data. Furthermore, the study provides early detection of diseases for individuals identified as high risk for CVD. This proactive approach enables the implementation of early preventive measures and targeted treatment strategies, contributing significantly to better health outcomes. It uses several resampling techniques to address class imbalance issues. Furthermore, it ensures the inclusion of both positive and negative samples in the training set. Real world CVD datasets consist of imbalanced classes leading to occurrence of one outcome less frequently than others, posing challenges for ML models in predicting future data. To resolve this problem, multiple sampling approaches have been used in this research such as under sampling and oversampling to balance the target class for better model training. In addition to sampling, data cleaning, normalization, and feature engineering have also been used to improve the model performance.

Objectives

This study aims to improve Cardiovascular Disease (CVD) risk prediction through advanced machine learning models and address class imbalance problem. The study includes a comparative analysis of the various machine learning algorithms such as Random Forest, Logistic Regression, Decision Tree and XGBoost classifier. The thorough analysis of these models enabled the authors to identify the algorithm that accurately predicts heart disease. The comparison was done by the using evaluation metrics such as accuracy, precision, recall, AUC-ROC curve, and F1-score. Furthermore, the performance of these models has been visualized for better understanding of the model performance. Through the use of state-of-the-art ML techniques, this research aims to lessen the prevalence of heart disease around the world by offering early detection.

Literature Review

Over the years, a number of contributions have been made by researchers for the diagnosis and risk prediction of CVD. This section encompasses some of the latest research that utilizes machine learning algorithms such as Random Forest, Support Vector Machine (SVM), K Nearest Neighbor (KNN), Naïve Bayes, Multi-layer Perceptron and Logistic Regression. The subjects that these research touch upon are data preprocessing, feature engineering, accurate diagnosis and optimization techniques such as hyperparameter tuning. In [16], the researchers utilized supervised learning for long-term risk prediction of cardiovascular diseases. This study utilized several ML models such as support vector machine, logistic regression and random forest and used performance metrics such as accuracy, sensitivity and AUC. A 10-fold cross-validation technique was used for training these models. It concluded that logistic regression outperformed the rest of the models with an accuracy of 72.06%.

In another research [17], the researchers examined cardiovascular disease risk factors in individuals with fatty liver disease. A principal component analysis technique was used in which a multiple regression classifier was fitted to ten principal components. The model achieved an impressive AUC of 0.86 and was fine-tuned using the top 15 discriminative features. The most accurate learning algorithm identified 79.17% of patients at low risk while 85.11% of patients at high risk of CVD. A study [18] also utilized several machine learning models, concluding that the ensemble model performed best with 87.8% accuracy score, 88% precision score, 88.3% recall score, and AUC score of 98.2% respectively. It applied the Synthetic Minority Oversampling Technique (SMOTE) with 10-fold-cross validation. This study emphasized the importance of the even distribution and preprocessing of data for building improved ML models.

The authors in [19] applied six distinct machine learning models to various datasets to compare classification performance. Techniques such as data improvement, feature scaling, outlier treatment, and ensemble approaches were utilized in this study to improve model performance. For model assessment and optimization of hyperparameters, GridSearchCV and Cross-validation were utilized. By employing the above-mentioned techniques, the authors were able to implement an ensemble voting classifier. This classifier combined all the six ML algorithms which achieved an improved accuracy of 93.44% for one of the datasets. In another study [20], several ML models were used in which random forest outperformed and obtained a high accuracy of 91.8%. The performance of this model was compared with several other models such as decision tree, SVM, logistic regression, and KNN. The authors in [21] proposed a hybrid approach in a 2023 study that combined a support vector machine and a modified particle swarm optimization model for heart and liver disease prediction. The accuracy and recall metrics of the proposed model were compared with other models based on the UCI datasets. It focuses on the significance of the hybrid approaches which can vastly enhance diagnosis accuracies and play a huge role in the early detection and prevention of such diseases.

In [22], SHAP and LIME techniques were utilized to develop an enhanced CVD prediction model. SVM and XGBoost models achieved similar performance results with an f1-score of 88%. SHAP (Shapley Additive exPlanations) visualization identified essential variables in the forecast process. Furthermore, LIME (Local Interpretable Model-agnostic Explanations) was used to explain the classification of every data point. The literature review of this section highlighted a diverse range of effective algorithms for cardiovascular disease (CVD) prediction. While Convolutional Neural Network (CNN) algorithms were prominently featured, ensemble approaches also demonstrated high accuracy rates. Other techniques that work well include K-nearest neighbors (KNNs), boosting algorithms, Support Vector Machines (SVMs), and Recurrent Neural Networks (RNNs). The conclusion of this study suggests that researchers utilizing deep learning and machine learning techniques for forecasting cardiovascular illnesses would benefit from these findings [23]. Table 1 presents the summary of the literature.

Table 1: Other Related Studies

Authors	Findings
Moshawrab et al. (2023) [24]	Smart wearables for the detection of CVD, funded by the Natural Sciences and Engineering Research Council of Canada (NSERC). Models used are CNN, SVM, KNN, GNB, MLP, LSTM, ANN etc. Emphasis on future enhancements.
Khan et al. (2023) [25]	Utilized data from Lady Reading Hospital and Khyber Teaching Hospital in Pakistan. Employed algorithms are Decision Tree, Random Forest, Logistic Regression, Naïve Bayes, and SVM. Random Forest demonstrated the highest accuracy of prediction (85.01%), sensitivity (92.11%), and recursive operative characteristic curve (87.73%) for cardiovascular disease.
Rani et al. (2021) [26]	Developed a hybrid decision support system, the techniques include (MICE), (GA), (RFE), (SMOTE), and Standard Scalar. Algorithms include SVM, Random Forest, AdaBoost etc.
Gupta et al. (2022) [27]	UCI repository dataset utilized ML techniques SVM, KNN, GNB etc. To improve accuracy and efficiency in disease prediction, these results highlight the importance of using ensemble methods and different classifiers in network intrusion detection.
Mijwil et al. (2024) [28]	UCI repository dataset used models are KNN, SVM, MLP (Superior with 88% accuracy), Random Forest, Decision Tree, Logistic Regression, and Naïve Bayes. Highlighting the importance of accuracy, algorithm performance, dataset

	significance, personalized treatment, and future directions for leveraging machine learning in healthcare.

In another study based on Kaggle dataset published by Mirza Hasnine [29], Artificial intelligence (AI) methods were explored, including machine learning algorithms like Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest, along with Artificial Neural Networks (ANNs) like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Through the integration of multiple data sources, optimized feature selection, and the utilization of ensemble methods such as boosting and bagging, AI models demonstrated the potential for personalized and efficient screening, early detection, and continuous monitoring of heart conditions. Hyperparameter tuning approaches were utilized to further optimize the performance of these AI systems and use data-driven insights to improve patient outcomes and public health management. A dataset of 70,000 patient records with 12 characteristics was used in the investigation, and additional features including MAP and BMI were incorporated. K-mode clustering was used as a preprocessing step on the dataset to enhance its scalability and convergence. The clustered dataset was subjected to various methods, including XGBoost classifier, random forest, decision tree, and multilayer perceptron to assess the performance of the models using the area under the ROC curve, F1 score, recall, accuracy, and precision [30].

Materials and Method:

Dataset

CVD is a significant medical disorder that impacts the heart, leading to increased mortality rates, especially among middle-aged individuals. The dataset (https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease) under consideration contains 319,796 patient records spanning various age groups, see Table 2.

Table 2: Dataset Features

Sr. No.	Attribute Description	Distinct Values
1.	Heart Disease	Yes or No
2.	BMI	Continuous Values
3.	Smoking	Yes or No
4.	Alcohol Drinking	Yes or No
5.	Stroke	Yes or No
6.	Physical Health	0 – 30
7.	Mental Health	0 – 30
8.	Diff Walking	Yes or No
9.	Sex	Male or Female
10.	Age Category	24 – 80+
11.	Race	White, Hispanic, Asian, Black
12.	Diabetic	Yes, No, Borderline, High
13.	Physical Activity	Yes or No
14.	Gen Health	Good, V. Good, Fair, Excellent and Poor
15.	Sleep Time	Continuous Values
16.	Asthma	Yes or No
17.	Kidney Disease	Yes or No
18.	Skin Cancer	Yes or No

The dataset offers extensive information on age, BMI, drinking and smoking patterns, history of stroke, physical and mental health conditions, and mobility issues. With its 18 medical features, the dataset enables the identification of individuals at risk of developing heart disease. The dataset, which is divided into training and testing subsets, consists of 319,796 rows and 18 columns, where each row corresponds to a single patient record. This dataset serves as a valuable resource for studies aiming to understand and predict the prevalence and risk factors associated with heart disease. Figure 1, Figure 2, Figure 3, Figure 4, Figure 5, Figure 6, and Figure 7 represent BMI Data Points, Diabetic Categories, GenHealth Categories, Gender Categories, Stroke History Categories, Smoking Categories, and Physical Health Categories respectively.

Proposed Methodology:

In this study, we have employed different machine learning models (Logistic Regression, Decision Tree, Random Forest, XGBoost) with resampling techniques for the prediction of CVD risk. The main contribution lies in that a balanced dataset is considered for effective and reliable results. Subsequently, data cleaning was conducted to remove inconsistencies and outliers, while maintaining dataset integrity. The refining and cleaning of the dataset was carried out in the preprocessing phase. To begin with the preprocessing, the data was in a raw form and inconsistencies were observed in the data. This led to the removal of these inconsistencies. The duplications were detected and then eradicated so that the models could train well and generalize better on the new data. The instances containing null values were eradicated. Four prominent machine learning models were built and evaluated: Random Forest, Logistic Regression, Decision Tree and XGBoost Classifier. Each model underwent rigorous training and was evaluated using diverse metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. Figure 8 provides the general or abstract flow of methodology. Through this study approach, the goal was to study the potential of machine learning in CVD risk prediction while ensuring the transparency and reliability of our methodology.

Data Refining and Cleaning

During the preprocessing phase, the dataset underwent refining and cleaning processes. Initially, the raw data revealed inconsistencies that needed addressing. These inconsistencies were identified and removed to ensure data integrity. The duplications were detected and then eradicated, in order to train the model and generalize the new data. The instances containing null values, were eradicated.

Sampling

During the initial data examination, the most prominent finding was the class imbalance in the target variable [18]. The original shape of the target column revealed a significant imbalance, with a higher proportion of instances without cardiac disease compared to those with the condition. To address this issue and ensure accurate model performance and generalization on unseen data, a resampling technique was employed to balance the target variable class. This involved oversampling, specifically using "random oversampling" to increase the instances of the minority class. Although SMOTE (Synthetic Minority Oversampling Technique) was considered, it was not chosen due to its synthetic sample generation and unsatisfactory results upon analysis.

Additionally, random under sampling was attempted but proved ineffective as it reduced instances of the majority class, compromising crucial data points and diminishing the models' generalization capacity. Figure 9 depicts the dataset samples before sampling and after sampling.

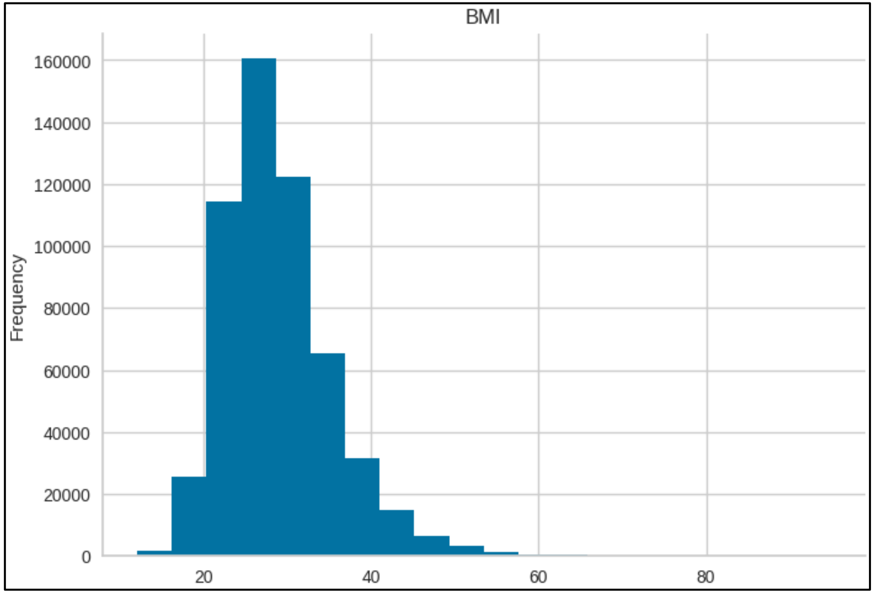

Figure 1: BMI Data Points

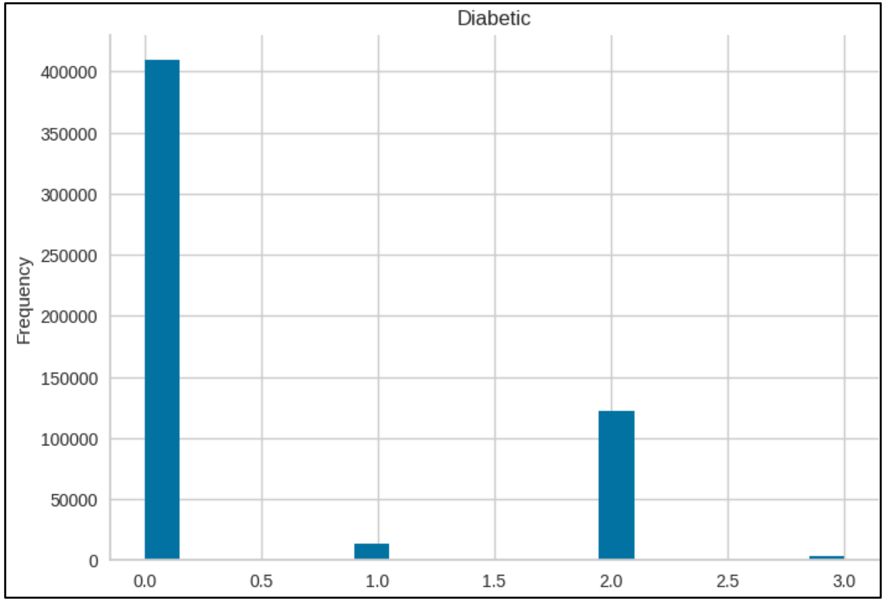

Figure 2: Diabetic Categories

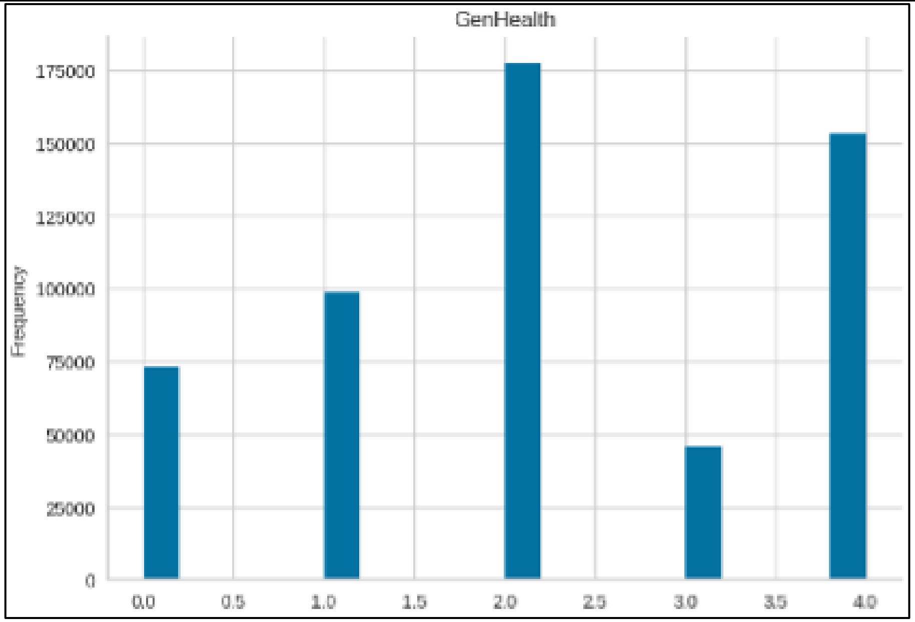

Figure 3: Gen Health Categories

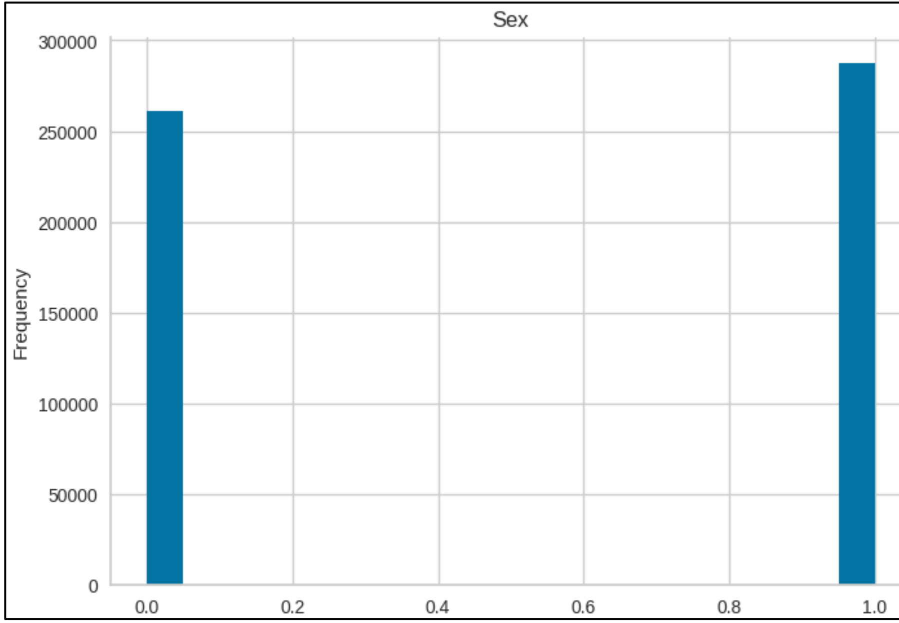

Figure 4: Gender Categories

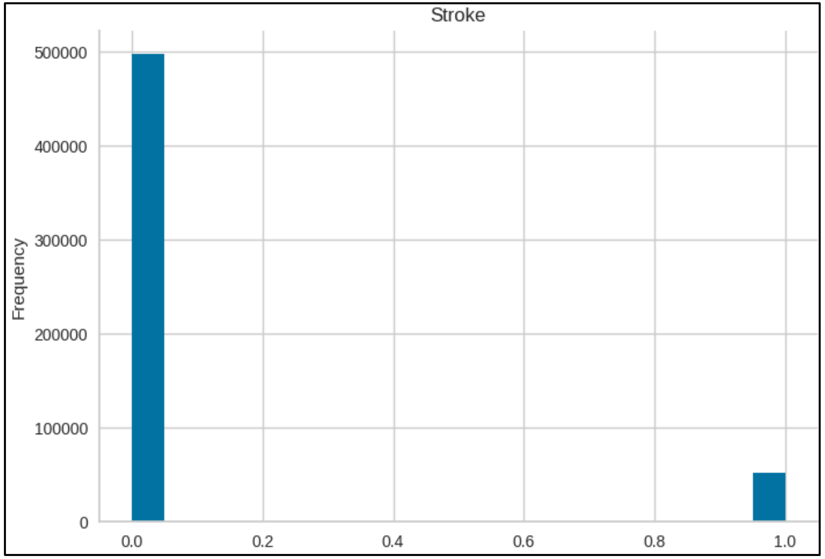

Figure 5: Stroke History Category

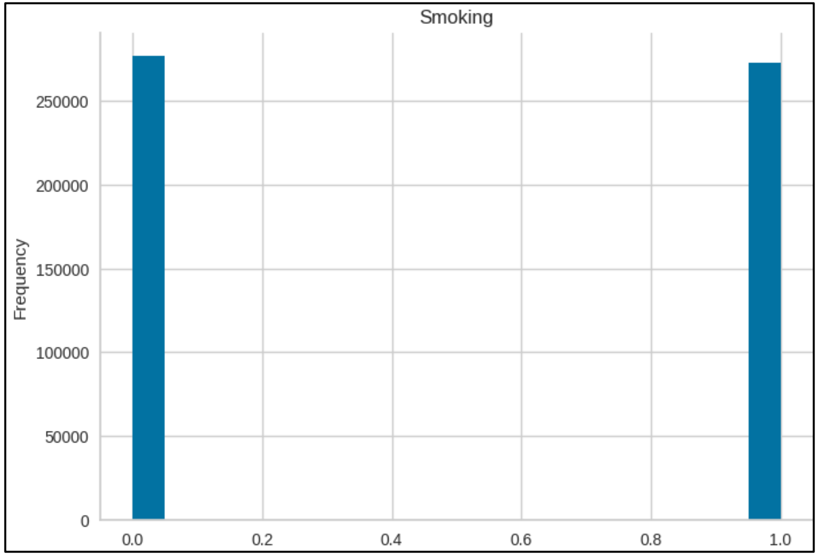

Figure 6: Smoking Category

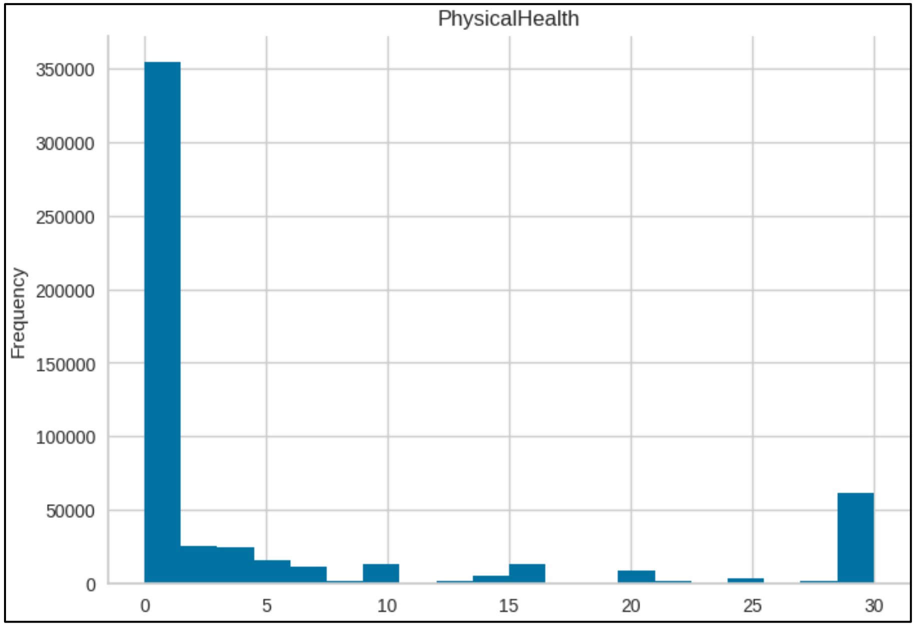

Figure 7: Physical Health Categories

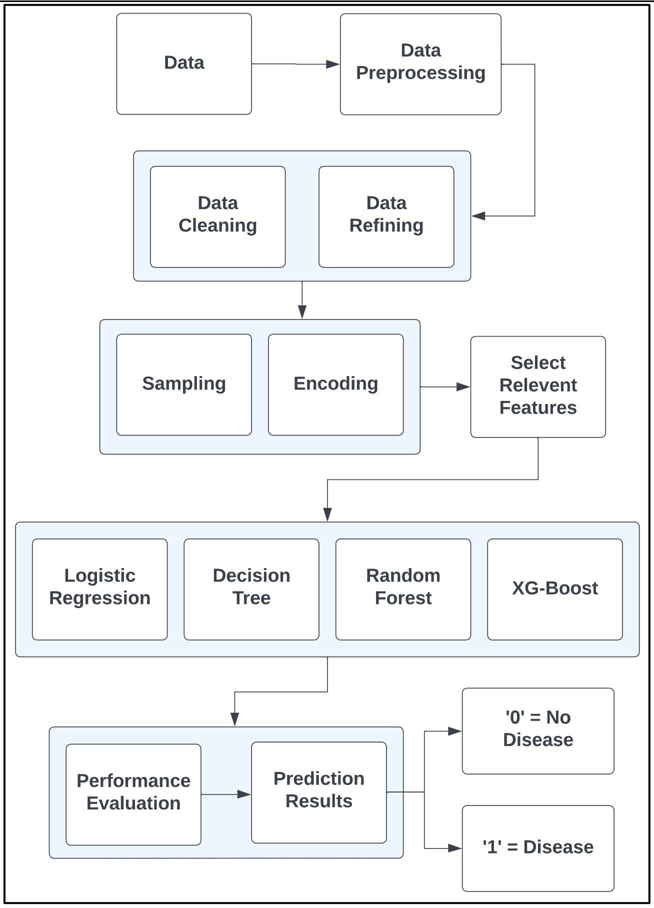

Figure 8: Methodological Abstract Diagram

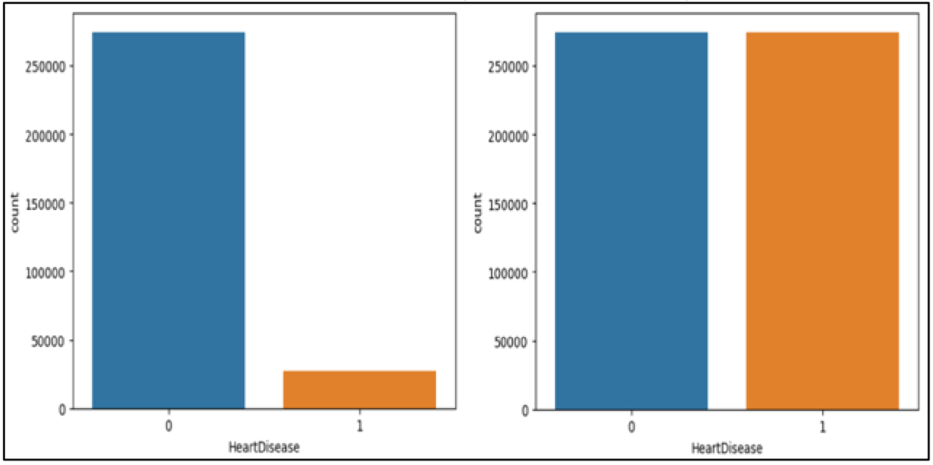

Figure 9: a) Imbalanced Data before Sampling. (b) Balanced Data after Sampling

Encoding:

Any type of machine learning method needs the dataset to be formatted so that the models can generate accurate predictions. One method to maintain uniformity in dataset is encoding. Label encoding, for instance transforms all of the categorical variables into numerical values to facilitate prediction by the model. For example, consider a "Diabetes" feature with categories "High," "Normal," and "Low." Through label encoding, these categories could be represented numerically as follows: Low: 2, Normal: 1, High: 0. This encoding process is applied to all features with categorical values, assigning them numerical representations for model compatibility.

Feature Relevance

Determining most important variables for predicting CVD risk is a challenging task [31]. The system needs to identify which clinical criteria are more suggestive, as not all variables contribute to risk in the same manner [32]. The use of a standard correlation matrix, which depicts how the features are linked together or what is the impact of a particular feature on the other, helps us choose the features. In order to gather information on feature relationships and possible relevance, a correlation matrix was used. It is extracted using '.corr' built-in method of sklearn library. The Figure 10 represents the correlation matrix regarding features.

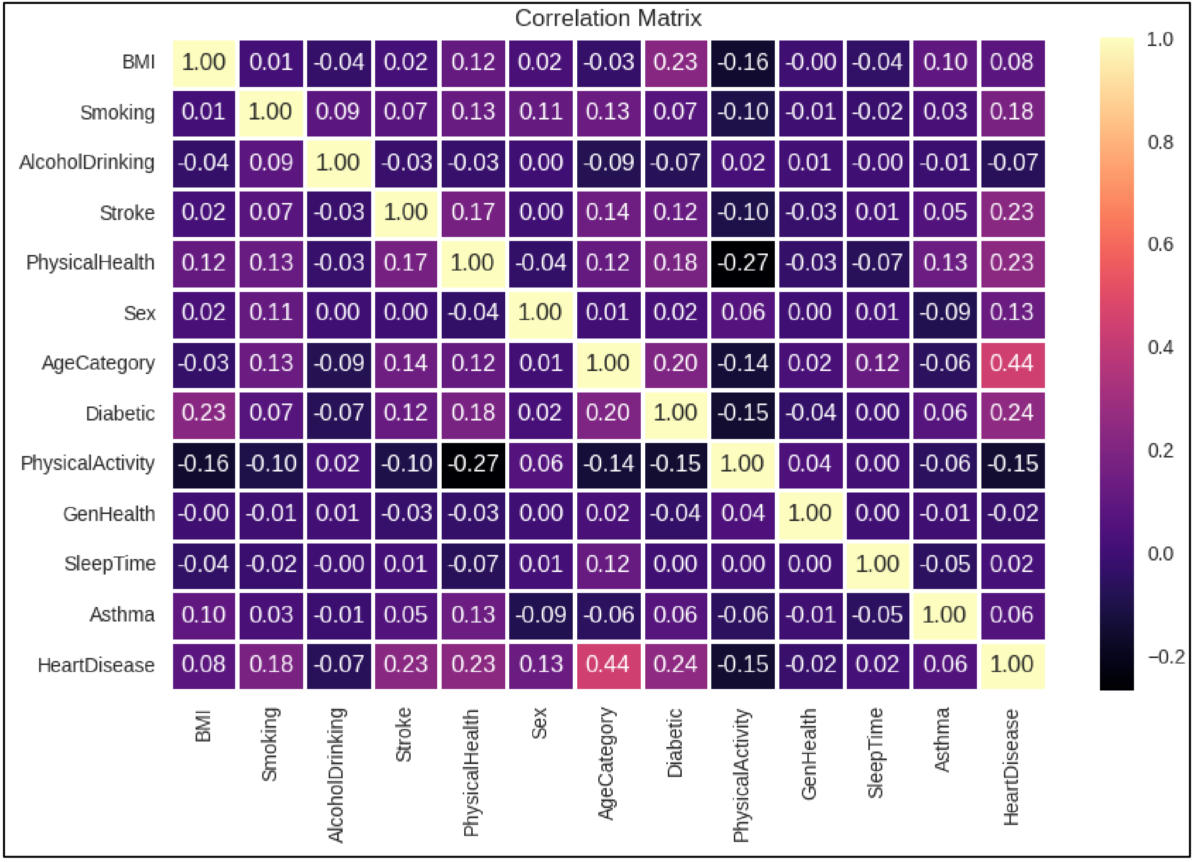

Figure 10: Correlation Matrix

Each cell in this table represents the correlation coefficient between any two features. 0 indicates no correlation, +1 indicates a strong positive correlation, and -1 indicates a strong negative correlation (as one attribute rises, the other falls).

Models

The Random Forest Classifier, Logistic Regression, Decision Tree Classifier, and XGBoost Classifier are the four well-known machine learning models that were developed and assessed. These models were selected for their diverse algorithmic approaches, aiming to capture different aspects of data patterns and enhance the robustness of cardiovascular disease risk prediction. Every model was put through a rigorous training process and evaluated using a variety of metrics, including F1-score, AUC-ROC, accuracy, precision, and recall. With a random state of '42', the train-to-test split ratio was 80:20, with '80' used for training and '20' for testing. Nonetheless, to improve the models' performance, hyperparameter tuning was done, with 'max_depth, max_leaf_nodes, max_iter', and other parameters that are supplied as inputs during the model-building phase.

Logistic Regression: Regression analysis is used by a logistic regression classification model [33] to determine and forecast the parameters in a given dataset. The likelihood of binary categorization serves as the foundation for both learning and prediction procedures. It is used when there are two alternative classes for the outcome variable. To produce results in the range of 0 to 1, logistic regression employs a logistic function, commonly known as the sigmoid function. S(x) = 1/1 + e^(-x), where "e" is the natural logarithm's base, is the sigmoid function. It is a simple, computationally efficient approach, but it is only applicable to binary classification, and it is not immune to outliers.

Random Forest Classifier: This model is also referred to as a supervised learning algorithm since it is a member of the classification family. First, a forest of manifold random trees is produced by this model [34]. For example, if the dataset has "x" number of aspects, it initially chooses a feature called "y" at random. When all features are used, or 'y,' the optimal rift technique generates nodes. Furthermore, by repeating the earlier phases, the algorithm builds an entire forest. The algorithm attempts to chain the trees using the voting procedure and projected outcome during the projection phase, aiming to identify the tree with the highest predicted accuracy and enhance predictions for future data. The voting mechanism is crucial for consolidating predictions from the random trees within the forest.

Decision Tree Classifier: A supervised machine learning approach for classification and regression issues is presented in the form of a decision tree [35]. Recursively splitting the data into subgroups based on the input feature values is how it works. During the decision-making process, a series of questions regarding the input features are asked to form a tree-like assembly where each internal node represents a decision based on a feature, each branch indicates the decision's outcome, and each leaf node represents the final anticipated outcome. Decision trees do not require feature scaling, manage non-linearity, and are interpretable.

XGBoost Classifier: A promising development in machine learning, the XGBoost Classifier uses the strength of numerous decision trees to produce reliable predictions. XGBoost functions as a group of professional weight-guessers, where "expert" (a decision tree) initially provides an estimate [36]. Then, XGBoost pinpoints their errors and assigns fresh specialists to focus on those areas. XGBoost demonstrates its collaborative learning power by achieving a substantially more accurate final weight prediction by integrating the improved estimations. Furthermore, it makes sure that its predictions hold up in a variety of scenarios by employing clever strategies to avoid overfitting.

Performance Evaluation: The notable models employ Accuracy, Precision, Recall, F1-Score, and AUC-ROC Curves as performance estimate metrics. The fraction of accurate predictions produced relative to all forecasts made is measured using the accuracy metric. Recall is the number of real positive cases that were accurately predicted, whereas precision is the number of anticipated positives that were positive. The F1-Score is the harmonic mean of precision and recall. In Table 3 as shown below, there are four evaluation metrics, accuracy, precision, recall, and f1-score. These are calculated as of the formula equations below.

Table 3: Performance Evaluation Metrics

Evaluation Metric	Description
$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$	The accuracy metric is used for the measurement of the proportion of correct predictions made over the total number of predictions made.
$\text{Precision} = \frac{TP}{TP+FP}$	Precision refers to as how many predicted positives were actually positive.
$\text{Recall} = \frac{TP}{TP+FN}$	Recall refers to as how many of the actual positive instances were correctly predicted.
$\text{F1-Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$	F1-Score is a measure that considers both precision and recall. It is the harmonic mean of the precision and recall.

Table 4: Accuracy Scores

Classifier	Accuracy Level
Random Forest Classifier	95.9 → 96%
Decision Tree Classifier	94.6 → 95%

Logistic Regression	75.3 → 75%
XGBoost Classifier	75.8 → 76%

Figure 11 shows a comparison of the test and training accuracies of the several models that were employed. Decision Tree and Random Forest perform better than the other models.

AUC-ROC:

A Receiver Operating Characteristic (ROC) curve visually represents the performance of a binary categorization model across various classification thresholds. It illustrates how changing the discriminating threshold affects the trade-off between the true positive rate and the false positive rate. Figure 12 shows a comparison of the ROC curves for all the machine learning techniques that were utilized in this study. The total performance of classifiers is measured by AUC, or Area under the Curve. The classifier is better the higher it's AUC. Out of the four models provided, it demonstrates that Random Forest performs the best, whereas Logistic Regression performs the poorest.

Table 5: Testing and Training Accuracies

Classifier	Testing Accuracy	Training Accuracy
Random Forest Classifier	96%	98%
Decision Tree Classifier	95%	98%
Logistic Regression	75%	75%
XGBoost Classifier	76%	76%

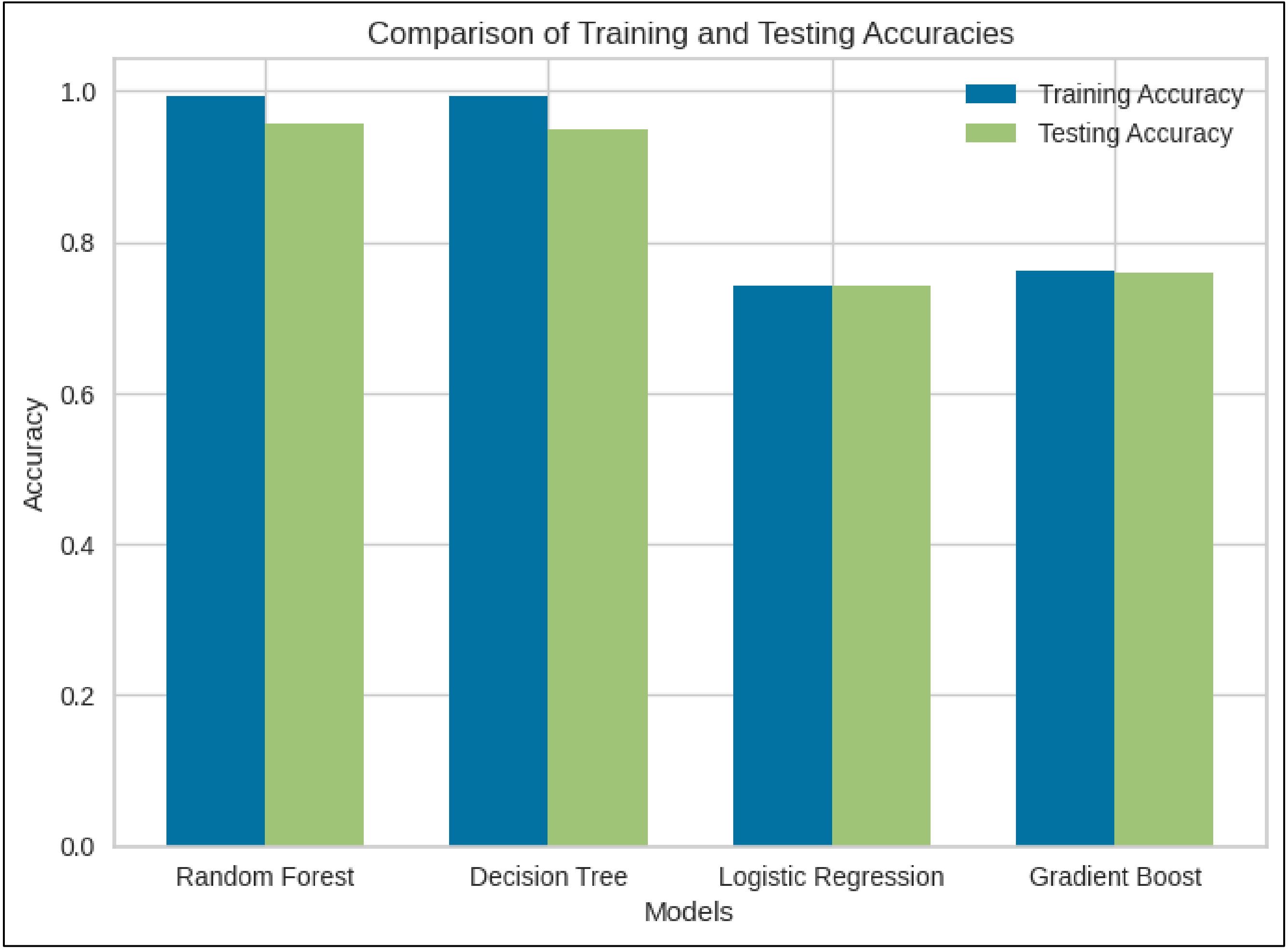

Figure 11: Test-Train accuracies of the models

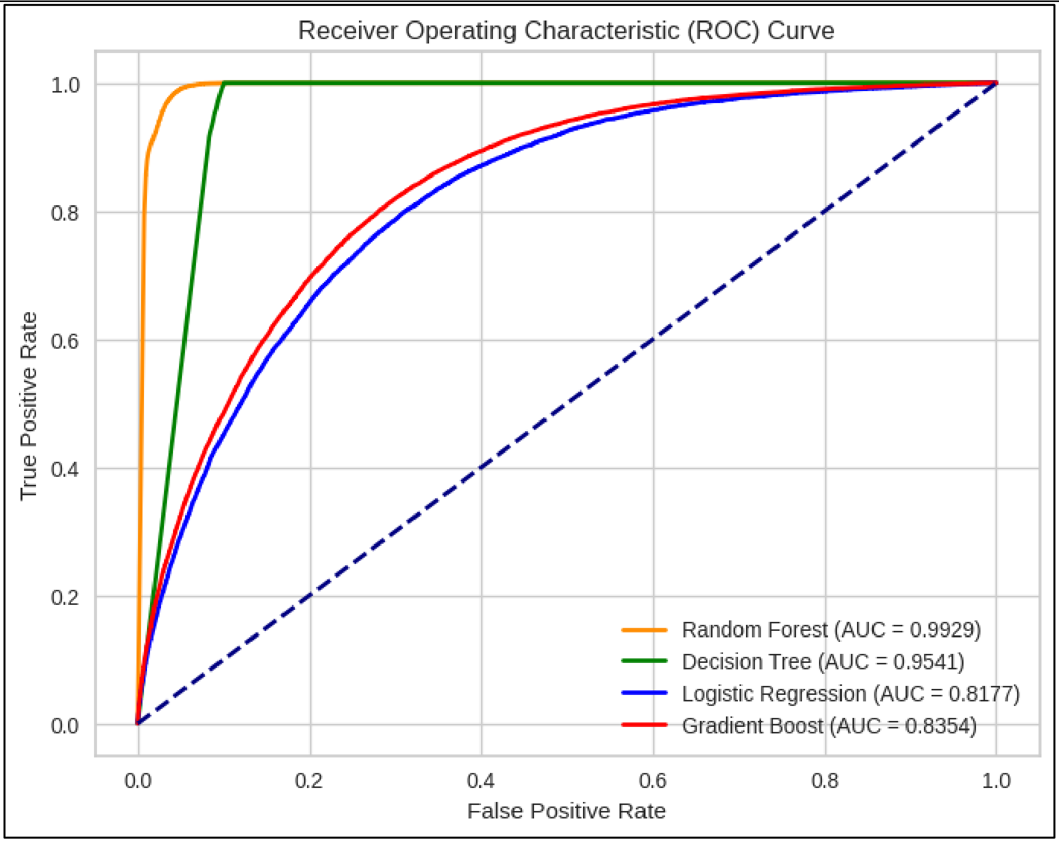

Figure 12: Comparative ROC Curve of Machine Learning Models

The classifiers' comparative AUC-ROC Curve is shown in Figure 12. Here, at a given threshold, the classification performance of all four models is displayed. Plotting the True Positive Rate (TPR) on the y-axis and the False Positive Rate (FPR) on the x-axis is known as the ROC curve. When TPR=1, the Random Forest model's ROC is at its best. This indicates that no negative cases (zero false positives) and all positive cases (zero false negatives) would be incorrectly classified by the model. As a result, the Random Forest Classifier has the highest AUC. The model's classification performance would be better the higher the AUC.

The representations of the performance measures are provided in Figures 13, 14, 15, and 16. The real and expected negatives as well as the actual and predicted positives are displayed in the appropriate models' confusion matrix. The classifiers' precision, recall, and f1-score metrics concerning the classes (positive and negative) are then displayed in the classification reports. The researchers can easily assess the models' performance by visualizing the performance measures below.

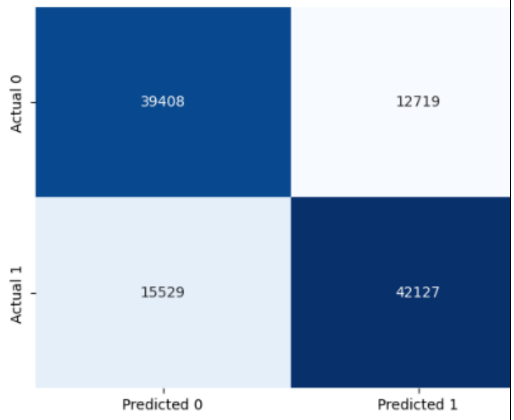

Figure 13: (a) Confusion Matrix

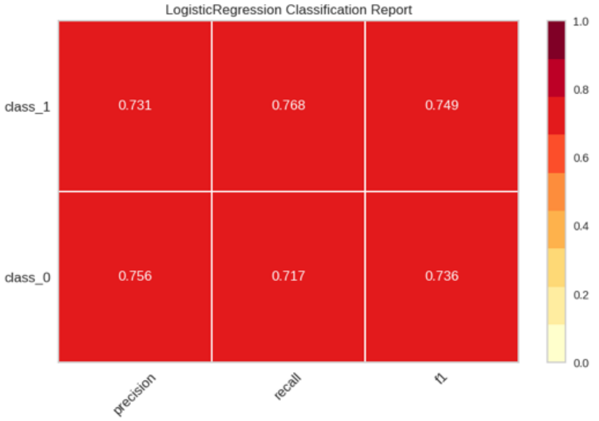

Figure 13: (b) Classification Report of Logistic Regression

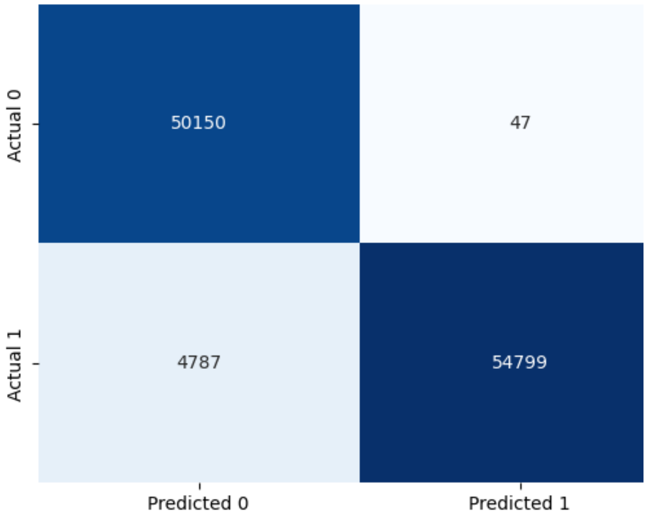

Figure 14: (a) Confusion Matrix

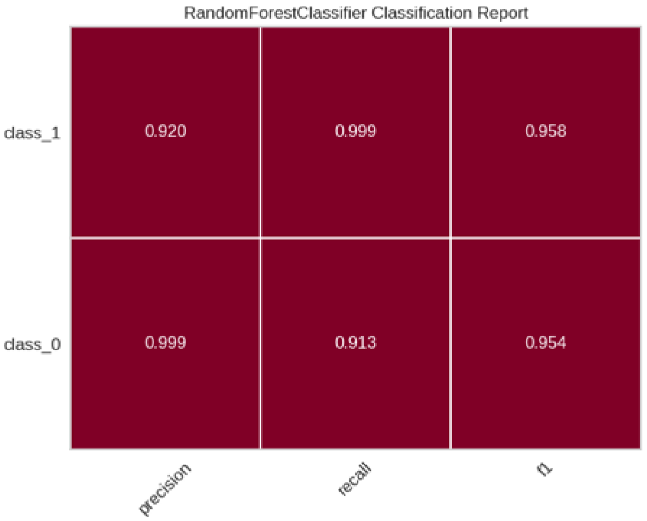

Figure 14: (b) Classification Report of Random Forest

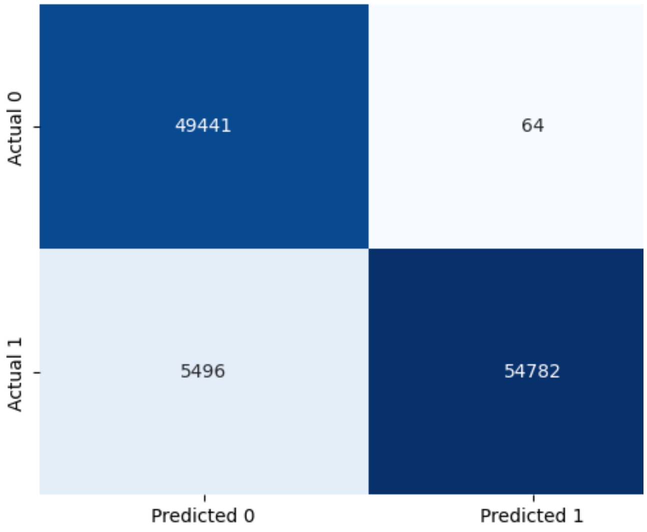

Figure 15: (a) Confusion Matrix

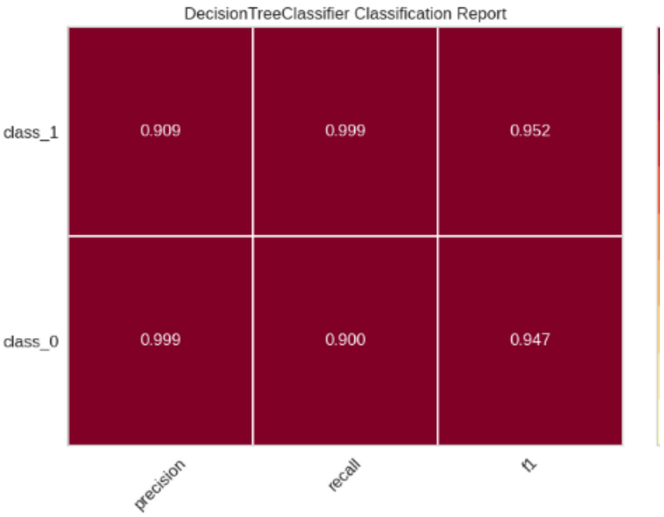

Figure 15: (b) Classification Report of Decision Tree

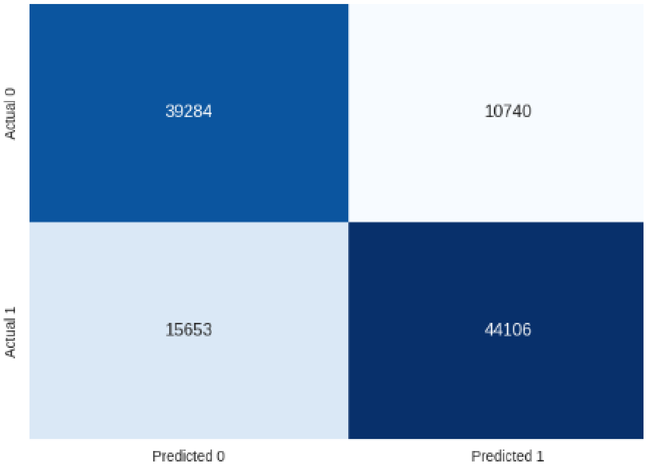

Figure 16: (a) Confusion Matrix

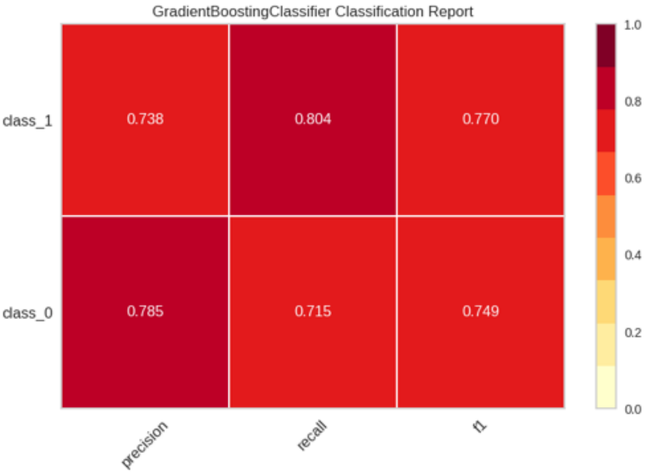

Figure 16: (b) Classification Report of XGBoost Classifier

Results:

The methodology of this study primarily focused on identifying the algorithm with the highest accuracy in comparison to others. In this study, the methods described earlier were implemented. Initially, the logistic regression algorithm was utilized in the experiment, but its accuracy did not meet expectations compared to other approaches. XGBoost, on the other hand, demonstrated superior precision and outperformed logistic regression. Surprisingly, the decision tree classifier, despite being an alternative algorithm, exhibited explicit accuracy, ultimately being selected as the algorithm with the highest accuracy in this context. However, a more refined attempt was made again to identify the robustness of the machine learning models, this time employing the random forest classifier, whose observed accuracy was marginally better than the decision tree's. However, when evaluating considered machine learning models' performance using the AUC-ROC curve, Random Forest demonstrated the best overall capacity to distinguish between those with and without CVD risk, with an AUC of 99% and an accuracy of 96%. The outstanding performance of the random forest model can be attributed to its ensemble learning approach, which combines multiple decision trees to make predictions, and its well-known effectiveness in handling imbalanced datasets.

Moreover, the capacity and ability of the random forest model to analyze and capture complex feature interactions played a pivotal role in its success. Random forest model is considered robust for its ability to generalize well on the new data which is an attribute to essential health care applications where reliability is crucial for assessments. Now, in the context of reviewing the objectives of the system, the random forest model was chosen as the heir of the dominion. Subsequently, the outcomes, the observations, and the objective analysis guided the significance of the findings, using the random forest as the superior classifier.

Table 6: Comparison with the Previous Studies

Models Used	Techniques	Accuracy	Our Results
Random Forest [32][37]	ANOVA	84%	**96%**
Decision Tree [32]	Chi-Square	82%	**95%**
Logistic Regression [32][37]	Relief	82.76%	**75%**
K-Nearest Neighbor [32][15]	RFE	60.34%	**-**
Support-Vector Machine [32][15]	Cross-Val	72%	**-**
Multi-Layer Perceptron [32]	Cross-Val	79.31%	**-**
Gradient Boost Classifier [32]	-	84%	**76%**
Neural Network [32]	Cross-Val	72%	**-**

A thorough comparison is made between the findings of this study and the similar findings of the earlier investigations in Table 6 above. The excerpts are drawn from several research that are part of the literature review. In one investigation, the Random Forest model yielded findings that were 84% accurate; in this study, however, the model's accuracy is 96%. Comparably, the Decision Tree has also been applied with an accuracy of 82%; nevertheless, the results of this study indicate that the Decision Tree has an accuracy of 95%.

Discussion

Amongst the machine learning models used in this study, the Random Forest model was successful in surpassing all the other ML models in terms of performance. The Random Forest model excelled at accurately predicting cardiovascular disease with a high accuracy of 96% and an AUC-ROC score of 99%. This model was successfully able to handle complex feature interactions and significant feature analysis which played a major role in its success.

The results of the proposed methodology were compared with the previous studies, and it was revealed that Random Forest and Decision tree models outperformed the models used in previous studies as they were able to achieve a significant improvement in prediction accuracies as depicted in Table 6. The improved performance results can equip medical professionals to create timely and effective treatment strategies that can lessen the burden on the healthcare system [38], [39]. With the help of these prediction models, it can become easy to understand the factors that affect CVD risk assessment.

Conclusion

This research study aimed at enhancing cardiovascular disease risk prediction by employing several essential machine learning algorithms. It utilizes a methodological approach which is composed of data preprocessing, model creation, and result analysis and evaluation. As the dataset used in this study had an uneven distribution of data, therefore the major aspect of the preprocessing phase involved balancing the data of the two classes by using a resampling approach. The obtained results highlight the importance of accurate feature selection and data preprocessing which leads to better model performance. The utilization of several ML models brought forward important details about their strengths and weaknesses for predicting cardiovascular disease. Moreover, it highlights that the choice of the suitable machine learning algorithm vastly depends on the unique characteristics of the dataset which must be carefully analyzed before deciding on an algorithm.

References:

[1] M. Bakhtiyari et al., "Contribution of obesity and cardiometabolic risk factors in developing cardiovascular disease: a population-based cohort study," Sci. Reports 2022 121, vol. 12, no. 1, pp. 1–10, Jan. 2022, doi: 10.1038/s41598-022-05536-w.

[2] M. Y. Henein, S. Vancheri, G. Longo, and F. Vancheri, "The Role of Inflammation in Cardiovascular Disease," Int. J. Mol. Sci. 2022, Vol. 23, Page 12906, vol. 23, no. 21, p. 12906, Oct. 2022, doi: 10.3390/IJMS232112906.

[3] S. Tiwari et al., "Lifestyle factors as mediators of area-level socio-economic differentials in cardiovascular disease risk factors. The Tromsø Study," SSM - Popul. Heal., vol. 19, p. 101241, Sep. 2022, doi: 10.1016/J.SSMPH.2022.101241.

[4] C. Y. Jurgens et al., "State of the Science: The Relevance of Symptoms in Cardiovascular Disease and Research: A Scientific Statement From the American Heart Association," Circulation, vol. 146, no. 12, pp. E173–E184, Sep. 2022, doi: 10.1161/CIR.0000000000001089.

[5] S. Ghorashi et al., "Leveraging Regression Analysis to Predict Overlapping Symptoms of Cardiovascular Diseases," IEEE Access, vol. 11, pp. 60254–60266, 2023, doi: 10.1109/ACCESS.2023.3286311.

[6] M. Jafari et al., "Automated diagnosis of cardiovascular diseases from cardiac magnetic resonance imaging using deep learning models: A review," Comput. Biol. Med., vol. 160, p. 106998, Jun. 2023, doi: 10.1016/J.COMPBIOMED.2023.106998.

[7] H. R. H. Al-Absi, M. T. Islam, M. A. Refaee, M. E. H. Chowdhury, and T. Alam, "Cardiovascular Disease Diagnosis from DXA Scan and Retinal Images Using Deep Learning," Sensors, vol. 22, no. 12, p. 4310, Jun. 2022, doi: 10.3390/S22124310/S1.

[8] G. Nicholson, S. R. Gandra, R. J. Halbert, A. Richhariya, and R. J. Nordyke, "Patient-level costs of major cardiovascular conditions: a review of the international literature," Clin. Outcomes Res., vol. 8, pp. 495–506, Sep. 2016, doi: 10.2147/CEOR.S89331.

[9] R. Luengo-Fernandez et al., "Cardiovascular disease burden due to productivity losses in European Society of Cardiology countries," Eur. Hear. J. - Qual. Care Clin. Outcomes, vol. 10, no. 1, pp. 36–44, Jan. 2024, doi: 10.1093/EHJQCCO/QCAD031.

[10] S. Mendis, I. Graham, and J. Narula, "Editorial: Reducing cardiovascular disease mortality and morbidity: implementing cost-effective and sustainable preventive interventions," Front. Cardiovasc. Med., vol. 10, p. 1236210, Jun. 2023, doi: 10.3389/FCVM.2023.1236210/BIBTEX.

[11] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning," Comput. Intell. Neurosci., vol. 2021, 2021, doi: 10.1155/2021/8387680.

[12] M. Pal, S. Parija, G. Panda, K. Dhama, and R. K. Mohapatra, "Risk prediction of cardiovascular disease using machine learning classifiers," Open Med., vol. 17, no. 1, pp. 1100–1113, Jan. 2022, doi: 10.1515/MED-2022-0508/MACHINEREADABLECITATION/RIS.

[13] S. Subramani et al., "Cardiovascular diseases prediction by machine learning incorporation with deep learning," Front. Med., vol. 10, p. 1150933, Apr. 2023, doi: 10.3389/FMED.2023.1150933/BIBTEX.

[14] A. Ogunpola, F. Saeed, S. Basurra, A. M. Albarrak, and S. N. Qasem, "Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases," Diagnostics 2024, Vol. 14, Page 144, vol. 14, no. 2, p. 144, Jan. 2024, doi: 10.3390/DIAGNOSTICS14020144.

[15] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," IEEE Access, vol. 8, pp. 107562–107582, 2020, doi: 10.1109/ACCESS.2020.3001149.

[16] E. Dritsas, S. Alexiou, and K. Moustakas, "Cardiovascular Disease Risk Prediction with Supervised Machine Learning Techniques," Int. Conf. Inf. Commun. Technol. Ageing Well e-Health, ICT4AWE - Proc., pp. 315–321, 2022, doi: 10.5220/0011088300003188.

[17] K. Drożdż et al., "Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: a machine learning approach," Cardiovasc. Diabetol., vol. 21, no. 1, pp. 1–12, Dec. 2022, doi: 10.1186/S12933-022-01672-9/FIGURES/3.

[18] E. Dritsas and M. Trigka, "Efficient Data-Driven Machine Learning Models for Cardiovascular Diseases Risk Prediction," Sensors 2023, Vol. 23, Page 1161, vol. 23, no. 3, p. 1161, Jan. 2023, doi: 10.3390/S23031161.

[19] N. Chandrasekhar and S. Peddakrishna, "Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization," Process. 2023, Vol. 11, Page 1210, vol. 11, no. 4, p. 1210, Apr. 2023, doi: 10.3390/PR11041210.

[20] U. Kamdi, "Heart Disease Prediction Using Machine Learning," vol. 4, no. 12, 2023.

[21] M. P. Behera, A. Sarangi, D. Mishra, and S. K. Sarangi, "A Hybrid Machine Learning algorithm for Heart and Liver Disease Prediction Using Modified Particle Swarm Optimization with Support Vector Machine," Procedia Comput. Sci., vol. 218, pp. 818–827, Jan. 2023, doi: 10.1016/J.PROCS.2023.01.062.

[22] P. S. Asih, Y. Azhar, G. W. Wicaksono, and D. R. Akbi, "Interpretable Machine Learning Model For Heart Disease Prediction," Procedia Comput. Sci., vol. 227, pp. 439–445, Jan. 2023, doi: 10.1016/J.PROCS.2023.10.544.

[23] Z. K. D. Alkayyali, S. Anuar Bin Idris, and S. S. Abu-Naser, "A SYSTEMATIC LITERATURE REVIEW OF DEEP AND MACHINE LEARNING ALGORITHMS IN CARDIOVASCULAR DISEASES DIAGNOSIS," J. Theor. Appl. Inf. Technol., vol. 28, no. 4, 2023, Accessed: May 20, 2024. [Online]. Available: www.jatit.org

[24] M. Moshawrab, M. Adda, A. Bouzouane, H. Ibrahim, and A. Raad, "Smart Wearables for the Detection of Cardiovascular Diseases: A Systematic Literature Review," Sensors 2023, Vol. 23, Page 828, vol. 23, no. 2, p. 828, Jan. 2023, doi: 10.3390/S23020828.

[25] A. Khan, M. Qureshi, M. Daniyal, and K. Tawiah, "A Novel Study on Machine Learning Algorithm-Based Cardiovascular Disease Prediction," Health Soc. Care Community, vol. 2023, pp. 1–10, Feb. 2023, doi: 10.1155/2023/1406060.

[26] P. Rani, R. Kumar, N. M. O. S. Ahmed, and A. Jain, "A decision support system for heart disease prediction based upon machine learning," J. Reliab. Intell. Environ., vol. 7, no. 3, pp. 263–275, Sep. 2021, doi: 10.1007/S40860-021-00133-6/METRICS.

[27] C. Gupta, A. Saha, N. V. S. Reddy, and U. D. Acharya, "Cardiac Disease Prediction using Supervised Machine Learning Techniques.," J. Phys. Conf. Ser., vol. 2161, no. 1,

p. 012013, Jan. 2022, doi: 10.1088/1742-6596/2161/1/012013.

[28] M. M. Mijwil, A. K. Faieq, and M. Aljanabi, "Early Detection of Cardiovascular Disease Utilizing Machine Learning Techniques: Evaluating the Predictive Capabilities of Seven Algorithms," Iraqi J. Comput. Sci. Math., vol. 5, no. 1, pp. 263–276, Feb. 2024, doi: 10.52866/IJCSM.2024.05.01.018.

[29] "View of A Detailed Analysis of Detecting Heart Diseases Using Artificial Intelligence Methods." Accessed: May 20, 2024. [Online]. Available: https://imiens.org/index.php/imiens/article/view/43/24

[30] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[31] S. Ahmed et al., "Prediction of Cardiovascular Disease on Self-Augmented Datasets of Heart Patients Using Multiple Machine Learning Models," J. Sensors, vol. 2022, 2022, doi: 10.1155/2022/3730303.

[32] K. Dissanayake and M. G. M. Johar, "Comparative study on heart disease prediction using feature selection techniques on classification algorithms," Appl. Comput. Intell. Soft Comput., vol. 2021, 2021, doi: 10.1155/2021/5581806.

[33] H. Arghandabi and P. Shams, "A Comparative Study of Machine Learning Algorithms for the Prediction of Heart Disease," vol. 8, 2020, doi: 10.22214/ijraset.2020.32591.

[34] A. O. Salau, T. A. Assegie, E. D. Markus, J. N. Eneh, and T. I. Ozue, "Prediction of the risk of developing heart disease using logistic regression," Int. J. Electr. Comput. Eng., vol. 14, no. 2, pp. 1809–1815, Apr. 2024, doi: 10.11591/IJECE.V14I2.PP1809-1815.

[35] M. A. Naser, A. A. Majeed, M. Alsabah, T. R. Al-Shaikhli, and K. M. Kaky, "A Review of Machine Learning's Role in Cardiovascular Disease Prediction: Recent Advances and Future Challenges," Algorithms 2024, Vol. 17, Page 78, vol. 17, no. 2, p. 78, Feb. 2024, doi: 10.3390/A17020078.

[36] N. Nissa, S. Jamwal, and M. Neshat, "A Technical Comparative Heart Disease Prediction Framework Using Boosting Ensemble Techniques," Comput. 2024, Vol. 12, Page 15, vol. 12, no. 1, p. 15, Jan. 2024, doi: 10.3390/COMPUTATION12010015.

[37] "Machine Learning-Based Cardiovascular Disease Detection Using Optimal Feature Selection", [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10416957

[38] "Machine Learning for High Risk Cardiovascular Patient Identification", [Online]. Available: https://www.researchgate.net/profile/Muhammad-Asad-Arshed/publication/363541183_Machine_Learning_for_High_Risk_Cardiovascular_Patient_Identification/links/63221052071ea12e36327f6e/Machine-Learning-for-High-Risk-Cardiovascular-Patient-Identification.pdf

[39] M. Hussain, A. Shahzad, F. Liaquat, M. A. Arshed, S. Mansoor, and Z. Akram, "Performance Analysis of Machine Learning Algorithms for Early Prognosis of Cardiac Vascular Disease," Tech. J., vol. 28, no. 02, pp. 31–41, Jun. 2023, Accessed: Dec. 26, 2023. [Online]. Available: https://tj.uettaxila.edu.pk/index.php/technical-journal/article/view/1778