Machine Learning Approaches for Coronary Artery Disease Detection: A Comprehensive Technical Investigation

Bharat Bhushan Mahor[1], Balveer Singh[2]

Department of Computer Science[1, 2], P.K. University, Shivpuri, M.P. India.
aananbhushan1@gmail.com[1], adm.pkit@gmail.com[2]

Abstract-This study aims to analyze and predict heart disease presence using the Cleveland Hungarian Switzerland dataset, employing a robust methodology encompassing data collection, pre-processing, exploratory data analysis (EDA), data splitting, and machine learning modelling. The dataset, comprising 1025 rows and 14 columns, provides a rich foundation for investigating cardiovascular health factors. The pre-processing phase involves identifying and handling missing values, followed by converting the dataset to the 'int64' data type. EDA employs visualizations such as count plots, histograms, and correlation matrices to uncover patterns and relationships within the dataset.The data splitting step partitions the dataset into training and testing sets, essential for assessing the Random Forest Classifier's performance. This machine learning model, chosen for its adaptability and efficiency, demonstrates exceptional predictive accuracy of 98.53%, with high precision, recall, and F-score. Comparative analysis with existing models reveals the proposed Random Forest model's superior performance, showcasing the potential for customization to enhance predictive accuracy. The study contributes valuable insights into cardiovascular health analysis and predictive modeling, emphasizing the importance of comprehensive methodologies for effective exploration and prediction in medical datasets**.**

Keywords- Cardiovascular health, Random Forest Classifier, Exploratory Data Analysis (EDA), Imbalanced datasets and Predictive modelling.

1. Introduction

Coronary Artery Disease (CAD) is a widespread and serious heart condition that presents considerable difficulties for healthcare systems worldwide. The need to improve diagnostic accuracy and prognostic skills in cardiovascular medicine has driven the integration of advanced technologies, such as machine learning, to address this significant global health issue. The convergence of medical research and artificial intelligence has great potential for transforming CAD diagnostics, risk assessment, and individualized treatment approaches. Machine learning, a branch of artificial intelligence, enables computer systems to learn and make predictions or judgments without direct programming[1], [2]. Within CAD, machine learning algorithms show an impressive ability to analyze large information, recognize complex patterns, and extract valuable insights. This healthcare paradigm change is advancing us towards a future where the early detection and treatment of CAD are not only more accurate but also customized to individual patient characteristics[3]–[6]. A key difficulty in CAD is its lack of symptoms in the initial phases, which frequently results in delayed detection and treatment. Machine learning models, when trained on varied datasets containing clinical, imaging, and genetic data, show exceptional capability in identifying subtle patterns that suggest preclinical CAD. Healthcare practitioners can use this technology to potentially detect high-risk individuals before symptoms appear, allowing for early intervention and prevention strategies. Furthermore, the incorporation of machine learning in medical imaging has created new opportunities for non-invasive CAD detection[7]–[11]. Advanced imaging techniques like coronary computed tomography angiography (CCTA) and magnetic resonance imaging (MRI) provide intricate datasets that may be too much for conventional analytical methods. Machine learning algorithms are highly effective in analyzing complicated data, helping to accurately interpret imaging data, and offering clinicians useful insights for precision computer-aided diagnosis and risk assessment[12]–[15].

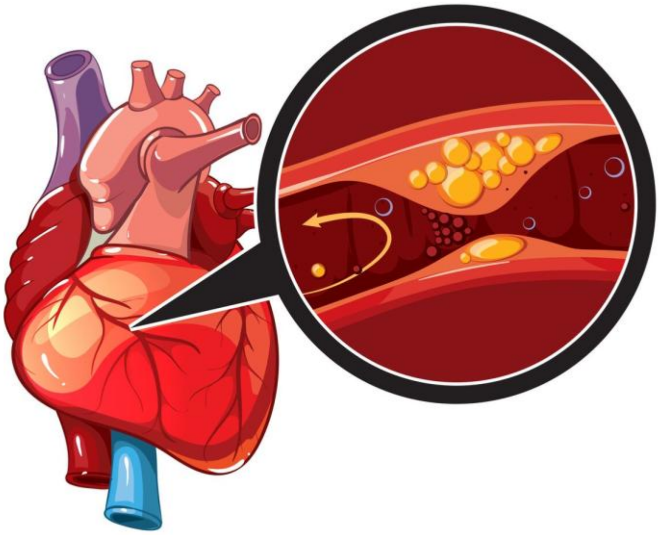


Figure 1 Coronary Artery Disease

Machine learning has proven to be highly effective in risk prediction for managing CAD. By integrating various patient data such as demographic details, lifestyle factors, and genetic tendencies, these algorithms can forecast an individual's probability of acquiring CAD with exceptional precision. This allows healthcare providers to apply specific preventive actions and interventions, leading to efficient resource distribution and enhancing overall patient results[16]–[21]. Ultimately, incorporating machine learning into the field of Coronary Artery Disease signifies a new phase of precise medicine and forward-thinking healthcare[22]. As these algorithms advance and collect larger datasets, their ability to revolutionize CAD diagnosis, risk assessment, and treatment techniques becomes more apparent. Despite challenges related to data privacy and ethical concerns, the prospect of an advanced future where coronary artery disease is identified sooner, treated more efficiently, and tailored to individual patient traits is both thrilling and has the capacity to transform the cardiovascular healthcare field.

1.1 Research Question or Problem Statement

*1.1.1*Clarity of Expression

The study focuses on utilizing machine learning techniques to improve the accuracy of diagnosing coronary artery disease in cardiovascular health. The project focuses on using advanced machine learning methods to create predictive models for early identification of coronary artery disease. The approaches entail complex data analysis, which includes feature engineering and model training, in order to enhance diagnosis accuracy. This study aims to advance medical diagnostics by using cutting-edge algorithms to tackle the complexities of coronary artery disease. The study is in line with the overarching objective of using technological improvements to enhance early detection approaches and, as a result, patient outcomes in cardiovascular health[23]–[27].

. 1.1.2 Relevance and Significance

Exploring Machine Learning Approaches for Coronary Artery Disease Detection is highly important in the field of healthcare. Given that cardiovascular illnesses are a primary contributor to worldwide illness and death, timely identification of coronary artery disease is crucial for successful treatment. This research aims to transform diagnostic accuracy by utilizing machine learning, which could lead to the early detection of persons at risk[28]–[32]. These developments may lead to more customized and precise treatments, ultimately improving patient results and lessening the strain on healthcare systems. The study is important because it has the potential to help shift towards proactive and data-driven methods in managing cardiovascular health.

2. Literature Review

Shamreen 2023.et al In order to predict diabetes mellitus, this research suggests a machine learning-based method that compares several algorithms to determine which three classifiers are the best: RF, GBM, and LGBM. Using a curated dataset and Pima Indians, the study assesses prediction accuracy using GB, RF, and LGBM classifiers. To improve predictive model accuracy for diabetic illness prediction, data augmentation approaches are also investigated and comparisons between augmented and non-augmented datasets are carried out[2].

Gupta 2023 et.al The assessment of coronary artery disease (CAD) according to variables such myocardial reserve, symptoms, and the degree of coronary stenoses is covered in this paper. It draws attention to the lack of a clear guideline and the uneven definition of triple-vessel disease (TVD) in clinical studies. In order to precisely calculate the survival benefits of coronary artery bypass grafting (CABG), the argument suggests a more extensive definition of TVD that includes major arterial disorders in all three coronary areas[33].

Zhang 2023 et.al Five DE-CRGs (F5, MT4, RNF7, S100A12, and SORD) were found to be viable diagnostic indicators for coronary artery disease (CAD) in this investigation using LASSO and SVM-RFE analysis. Their diagnostic effectiveness was validated using the GSE20681 and GSE42148 datasets. GSEA provided mechanistic insights that indicated their role in immune response modulation. A substantial decline in regulatory T cells (Treg) was found by immunological microenvironment analysis, and this decline was inversely correlated with the marker gene S100A12. A thorough lncRNA-miRNA-mRNA ceRNA network was created, and prospective medicinal medicines targeting S100A12 and F5 were found, offering new CAD therapy options and diagnostic biomarkers[34].

Garavand 2023 et.al This study outlines the development of a Coronary Artery Disease (CAD) registry through a questionnaire-based dataset. Thirteen main categories, comprising 171 data elements related to identification, medical history, and procedures, were identified. The conceptual model gained approval from field experts, leading to the software's development. The study suggests the proposed approach as suitable for designing hospital-based registries, emphasizing the importance of routine and systematic registry use in healthcare centers for effective CAD management[35].

Abdel 2022 et.al In order to diagnose chronic kidney disease (CKD), this study proposes hybrid machine learning approaches using Apache Spark. Relief-F and chi-squared feature selection improves the performance of the classification algorithms (DT, LR, NB, RF, SVM, GBT). Evaluation criteria that show SVM, DT, and GBT with certain features obtaining ideal 100% accuracy include accuracy, precision, recall, and F1-measure. Features chosen using Relief-F performed better than features chosen by chi-squared[3].

Authors	Methodology used	Problem statement	Dataset used	Parameters
Endo 2022 [36]	use a Harmonic Scalpel's (Ethicon Endosurgery, Cincinnati, OH, USA) to "skeletonize" a patient. the conventional approach of extracting saphenous veins (SVGs).	During surgery, these patients frequently exhibit unforeseen lesions and complications involving anastomotic tissues.	To assess the graft using ECG-gated 3D-CT, the patient's preoperative and postoperative general health, medicines, biochemical testing, transthoracic echocardiography, and cardiac catheterization were examined in addition to the patient's preoperative treatment history	Accuracy
Park 2022 [12]	clinical results of sex & non-obstructive cardiovascular disease (NOCVD) in patients with CAS. Methods	All patients and CAS patients underwent a propensity score matching (PSM) analysis to control for any confounding factors.	risk factors for cerebrovascular accident, myocardial infarction, and death at 5 years.	accuracy
Raxwal 2022 [26]	The available evidence indicates that IVL is a secure method of enhancing stent expansion and luminal gain.	No dissections, perforations, sudden closures, or insufficient blood flow/no reflow occurred during IVL.	With the use of a Shockwave C2 Coronary Intravascular Lithotripsy catheter, we were able to treat a	Accuracy, precision

			stenotic lesion in a poorly inflated stent, achieving a result of 0% residual stenosis.	
Gao 2022 [13]	Coronary angiography patients with at least a 95% stenotic lesion in their epicardial arteries constituted the study population, which totaled 206 people at Beijing Anzhen Hospital.	The collateral ability to supply enough blood flow in the case of coronary artery obstruction is reduced in many people due to damaged or less-developed coronary collateralization, which is common in real-world clinical settings.	The aforementioned dataset and supporting documents can be obtained from the authors.	Accuracy, precision, recall
Kayaert 2021 [37]	Visual evaluation of two-dimensional CA has been used to compile data on LM illness, and a luminal diameter stenosis (DS) criterion of 50% has been established as significant.	It is more effective than MT alone in reducing ischemia and angina, and it may boost long-term survival when the ischemic load is large.	FAME trial) and that FFR-guided PCI is superior to MT alone (FAME 2) in terms of outcome.	accuracy

3. Methodology

The methodology encompassed data collection from the Cleveland Hungarian Switzerland dataset, comprising 1025 rows and 14 columns, providing a foundation for cardiovascular attribute exploration. Pre-processing managed missing values and converted data types, ensuring dataset suitability. Exploratory data analysis (EDA) utilized count plots, histograms, and correlation matrices to unveil attribute distributions and relationships, informing subsequent modeling decisions. Data splitting facilitated model performance evaluation, employing the Random Forest Classifier for its adaptability. Exceptional performance metrics (98.53% accuracy, precision, recall, F-score) demonstrated the model's robustness on the imbalanced dataset. Comparative analysis showcased the proposed model's superiority, suggesting customization potential for enhanced predictive accuracy in cardiovascular health applications.

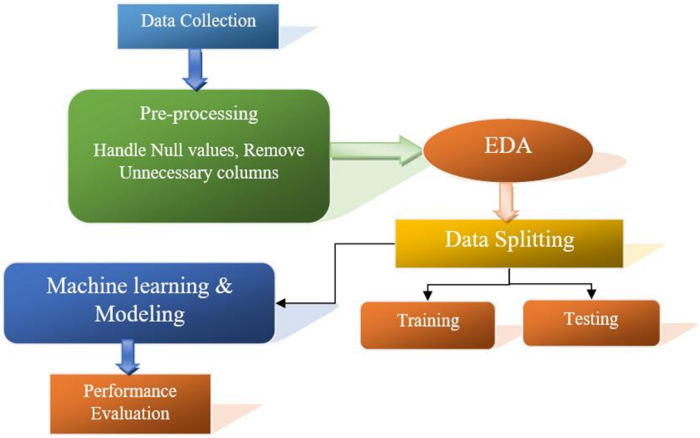
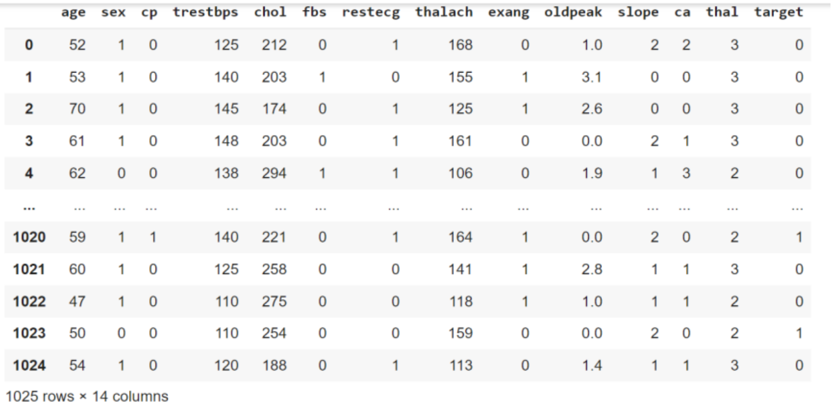


Figure 2 Proposed Flowchart

3.1 Data Collection

As part of the Cleveland Hungarian Switzerland dataset gathering procedure, information was retrieved from a structured dataset with 1025 rows and 14 columns. Each of the unique names that these columns were labeled with represented a certain property. Age, gender, type of chest pain, resting blood pressure, serum cholesterol, fasting blood sugar, electrocardiogram results, maximum heart rate reached, exercise-induced angina, oldpeak (exertion-induced depression of the ST segment), slope (slope of the peak exercise ST segment), number of major vessels colored by fluoroscopy, thal (thalassemia), and target (presence or absence of heart disease) are among the columns.This dataset is useful for examining correlations and trends within the given parameters, providing information on potential risk factors for heart disease. Analysts can leverage statistical and machine learning techniques to uncover correlations, make predictions, or derive meaningful conclusions from the dataset. The richness of the dataset, with its diverse set of attributes, provides a comprehensive foundation for conducting exploratory data analysis and developing predictive models in the field of cardiology or related research domains. Researchers and practitioners can use this data to enhance their understanding of the factors associated with cardiovascular health and contribute to the ongoing efforts to improve diagnostic and preventive strategies.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1020	59	1	1	140	221	0	1	164	1	0.0	2	0	2	1
1021	60	1	0	125	258	0	0	141	1	2.8	1	1	3	0
1022	47	1	0	110	275	0	0	118	1	1.0	1	1	2	0
1023	50	0	0	110	254	0	0	159	0	0.0	2	0	2	1
1024	54	1	0	120	188	0	1	113	0	1.4	1	1	3	0

1025 rows × 14 columns

Figure 3 Data Preview

3.2 Data Pre-processing

The code snippet seems to be part of the dataset pre-processing phase, focusing on managing missing values and converting data types. Let's analyze the technical specifics. The code "heart4.isnull().any()" is probably used to detect any missing or null values in the 'heart4' dataset. The `isnull()` method creates a boolean mask indicating whether each element in the original dataset is null or not. The `any()` function verifies the presence of at least one 'True' value in the boolean mask. Null values will cause the expression to evaluate to 'True', indicating the presence of missing data. After this check, the code continues with "heart4 = heart4.astype('int64')". Converting the full 'heart4' dataset to the 'int64' data type entails data type conversion. This technique is beneficial for situations where the dataset's original data types require modification. Here, it guarantees that all values in 'heart4' are depicted as 64-bit integers. Converting to 'int64' may lead to data loss, particularly if the initial dataset includes non-integer values. This phase implies that the data in 'heart4' can be reliably represented as integers, which is appropriate for specific analysis and machine learning models that necessitate numerical inputs. The code snippet presented is a brief demonstration of a typical pre-processing procedure that involves identifying and managing missing values, and then converting the full dataset to the 'int64' data type.

Pseudo code of Pre-processing

```
# Check for missing values in the 'heart4' dataset
any_null_values = heart4.isnull().any()

# If there are any missing values, handle them (not explicitly shown in the provided code)
if any_null_values:
    # Handle missing values (e.g., imputation or removal)

# Convert the entire 'heart4' dataset to the 'int64' data type
heart4 = heart4.astype('int64')

# The variable 'any_null_values' now contains information about the presence of missing values
# The dataset 'heart4' is now represented as integers (int64)
```

The variable `any_null_values` in the pseudo code stores a boolean value indicating the presence of any missing values in the 'heart4' dataset. If the variable is 'True', additional steps for managing missing values might be included based on the specific needs of the data preprocessing pipeline. The next line converts all values in the 'heart4' dataset to the 'int64' data type.

3.3 EDA

Exploratory Data Analysis (EDA) is a fundamental step in data science where the main features of a dataset are analyzed and represented visually to gain valuable insights. Exploratory Data Analysis (EDA) allows analysts to comprehend the data's structure, trends, and potential correlations, which helps in making informed decisions and developing hypotheses. Exploratory Data Analysis (EDA) involves the application of statistical and graphical approaches to uncover the fundamental characteristics of the dataset. Descriptive statistics, distribution plots, and correlation matrices are utilized as analytical tools. Studying measures of central tendency, dispersion, and skewness helps in evaluating the general distribution of the data. Visualization methods including histograms, box plots, and scatter plots reveal patterns, outliers, and probable trends in the dataset. Exploratory Data Analysis (EDA) includes analyzing categorical variables using count plots and exploring the connections between various aspects. Correlation matrices offer insights on the magnitude and direction of connections between numerical variables. EDA is a comprehensive method used to understand the fundamental properties of data, which serves as the foundation for data preprocessing, feature engineering, and model development. It plays a vital role in the data science process by promoting a data-driven comprehension of the phenomena shown in the dataset.

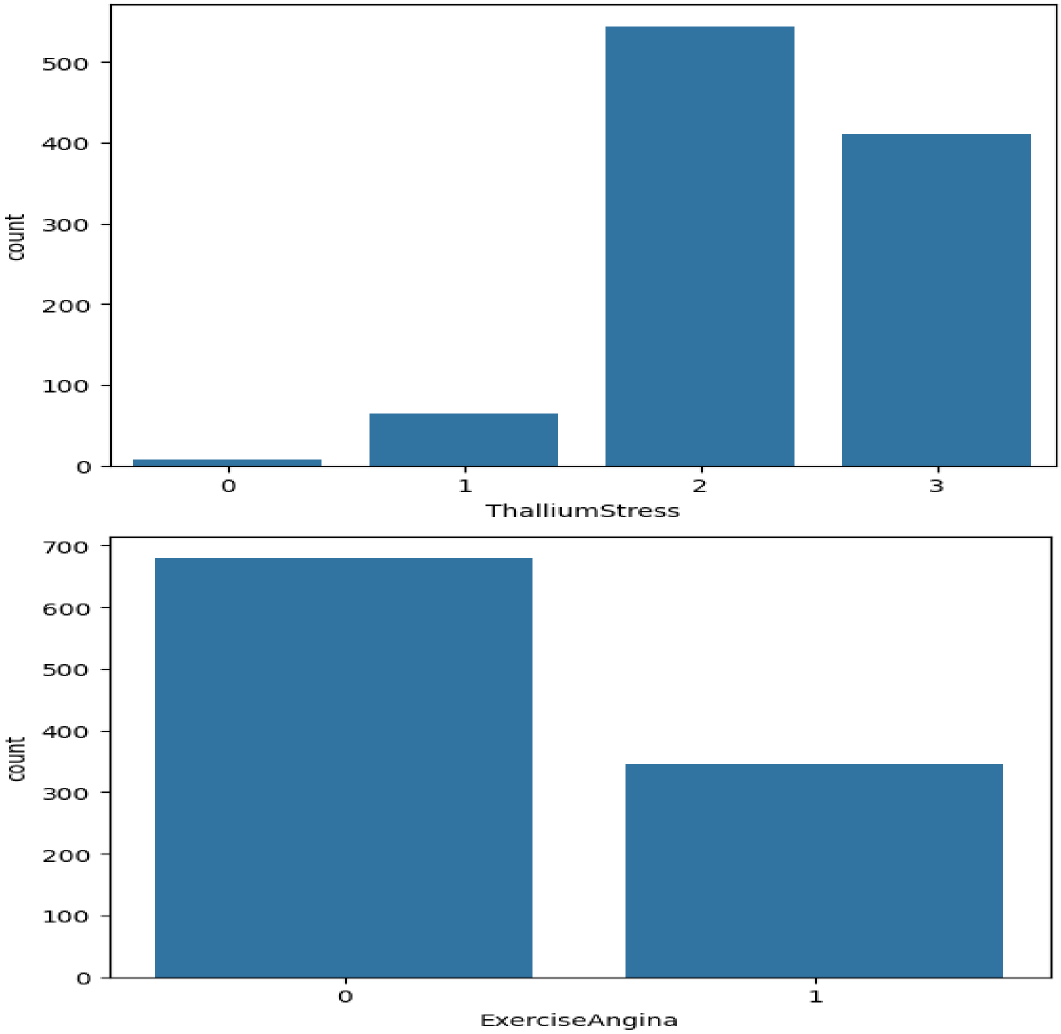


Figure 4 Count plots of Thallium Stress and Exercise Angina

Figure 4 displays Count plots illustrating the distribution of Thallium Stress and Exercise Angina. The visualizations offer insights into the frequency of stress levels and angina during exercise in the dataset.

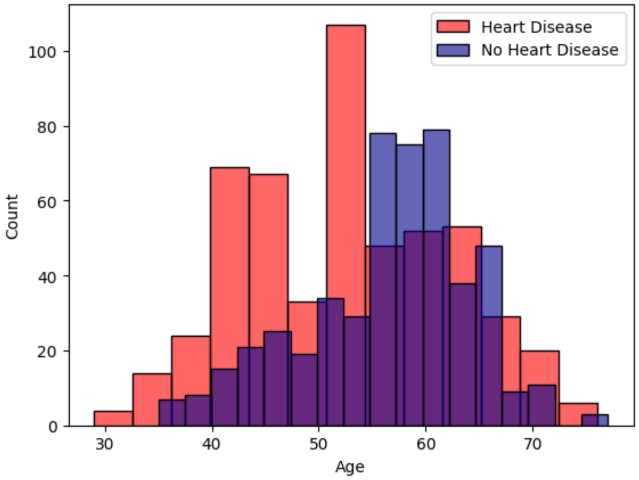


Figure 5Hist plot of Heart Disease and No Heart Disease

Figure 5 displays Histogram plots showing the distribution of incidents with and without cardiac disease. This image helps to comprehend the general distribution of heart disease cases in the dataset.

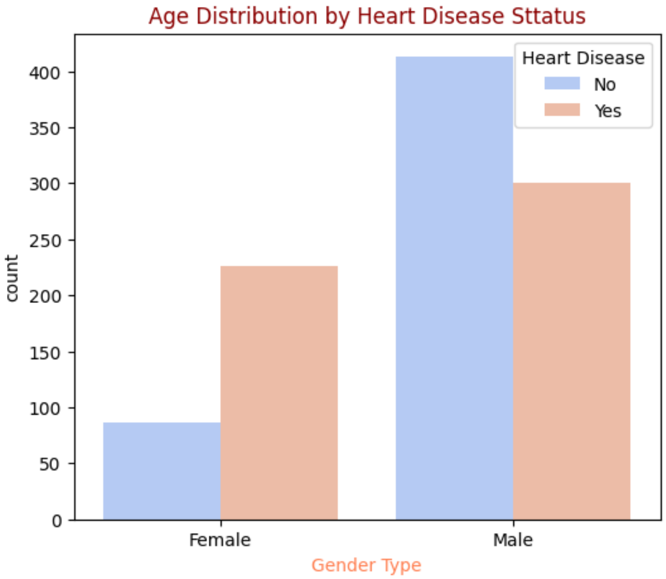


Figure 6 Age Distribution Graph of Heart Disease Status

Referring to Figure 6, an Age Distribution Graph illustrates the distribution of ages in relation to the presence or absence of heart disease. This graph might reveal age-related trends in occurrences of heart disease.

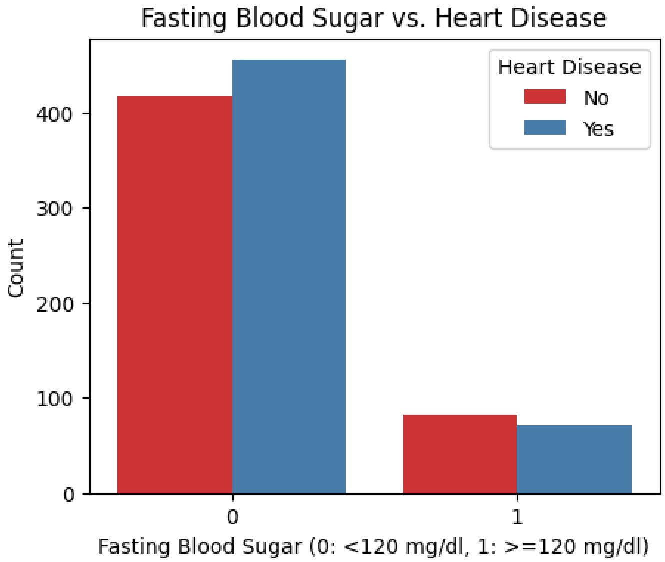


Figure 7 Count Plot Fasting Blood Sugar vs Heart Disease

Figure 7 uses a Count Plot to compare Fasting Blood Sugar levels with the presence of heart disease. This graphic representation enables an evaluation of the correlation between fasting blood sugar levels and the occurrence of heart disease.

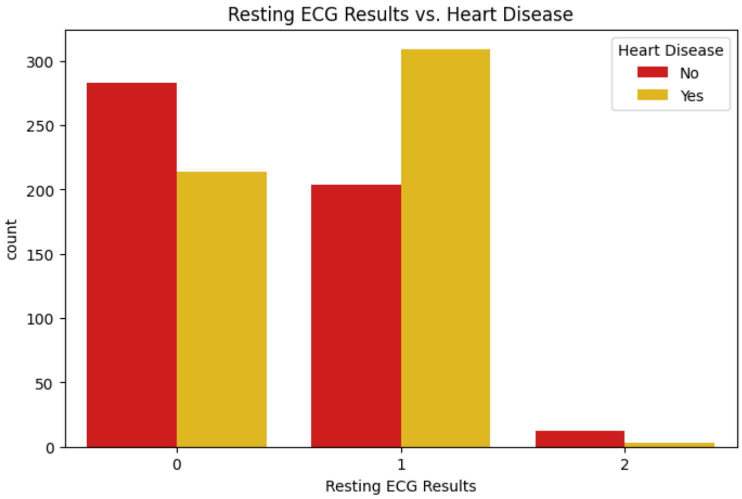


Figure 8 Resting ECG results vs Heart Disease

Figure 8 explores the relationship between Resting Electrocardiogram (ECG) readings and heart disease using Count Plots. The graphs offer insights on the distribution of ECG readings in relation to the presence or absence of cardiac disease.

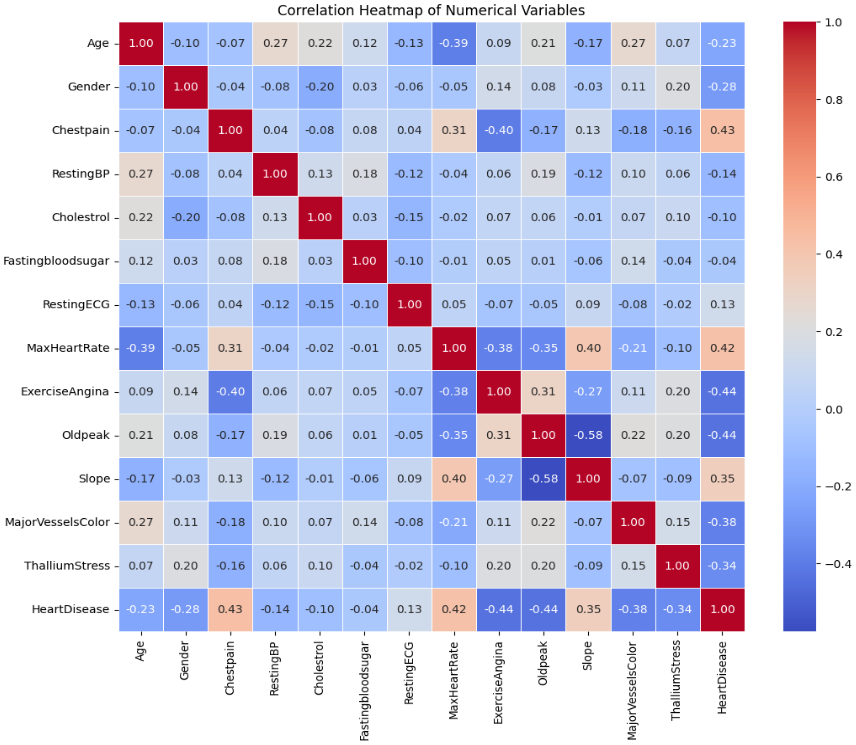


Figure 9 Correlation Matrix of Numerical Variable

Figure 9 presents a Correlation Matrix of Numerical Variables. This matrix provides a visual representation of the correlations between numerical variables, giving a detailed look at how these variables are related to each other. Correlation insights are crucial for selecting features and creating models in predictive modeling situations. The figures provide a detailed overview of the dataset's features, which will help make educated judgments in later phases of analysis and modelling.

3.4 Data Splitting

A crucial step in assessing the machine learning model's effectiveness is partitioning datasets into training and testing sets. The usage of `train_test_split` function from the `sklearn.model_selection` module is demonstrated in the code snippet that you provided. Let's investigate the technical details. Assuming that the 'heart4' dataset has been loaded, the target variable ('HeartDisease') is removed to create the feature matrix 'X'. The variable 'y' contains the desired variable. The `train test split` function is used to separate the data into training and testing sets. Twenty percent of the data are set aside for testing with the `test_size=0.2} parameter, and reproducibility is ensured using `random_state=42}. The feature matrices for the training and testing sets, respectively, are represented by {X_train} and {X_test} when they are run. The target values are stored in `y_train} and `y_test}, respectively. Additionally, the code attempts to display the training set's shape, which may be problematic because it comes after the assignment of {X_train} and `X_test}. To fix this, the print statement needs to be split apart. This algorithm efficiently separates the training and evaluation datasets, allowing a comprehensive assessment of the model's generalization capabilities. Reliable machine learning models require careful management of the training and testing datasets.

3.5 Machine Learning &Modeling

Machine learning, a branch of artificial intelligence, consists of algorithms that allow systems to learn patterns and anticipate outcomes based on data. The Random Forest Classifier is a potent and adaptable algorithm suitable for both classification and regression tasks.

- Random Forest Classifier builds numerous decision trees in the training phase. Every tree in the forest is trained on a randomly selected subset of the dataset, utilizing a random sample of characteristics. This ensemble method reduces overfitting and improves the model's capacity to generalize. During prediction, each tree in the forest contributes a vote for the class, and the final output is determined by the majority decision. The Random Forest excels in managing high-dimensional datasets, categorical characteristics, and missing values efficiently. It has a lower tendency to overfit than individual decision trees and typically shows consistent performance across many datasets. Random Forests excel in capturing intricate associations in data, making them well-suited for applications like image recognition, fraud detection, and medical diagnostics. The Random Forest Classifier is an important tool in machine learning due to its ability to handle noise and outliers well, as well as providing insights on feature relevance. The adaptability, robustness, and interpretability of the model have led to its extensive use in practical predictive modeling situations.
- Model Implementation
  To tackle imbalanced datasets, a Random Forest Classifier is used as a reliable approach. The `RandomForestClassifier` from the scikit-learn library is created with a defined random seed (`random_state=42`) to ensure reproducibility. Next, the model is trained by using the `fit` algorithm to the imbalanced training dataset (`X_train` and `y_train`). Random Forests effectively handle class imbalances because to the ensemble structure of the model, which prevents favoritism towards the majority class. The Random Forest Classifier is effective at handling uneven class distributions in classification tasks by utilizing several decision trees and their combined predictions, making it a desirable option for real-world settings.

Pseudo Code of Model Implementation

```
# Import necessary libraries
from sklearn.ensemble import RandomForestClassifier

# Instantiate the Random Forest Classifier with a specified random seed
model_imbalanced = RandomForestClassifier(random_state=42)

# Train the model on the imbalanced dataset
model_imbalanced.fit(X_train, y_train)
```

Assuming the necessary libraries are imported, the feature matrix X train and goal variable train are defined. The RandomForestClassifier is created with the random state parameter set to 42 to reproducibility, & the model is trained with the fit method with the training data.

4. Result & Discussion

4.1 Performance Evaluation

Accuracy, precision, recall, and F-score are performance evaluation metrics that are essential for assessing how well machine learning models work. By comparing the number of correctly predicted cases to the total number, accuracy evaluates the overall correctness of predictions. Precision gauges the model's ability to reduce false positives, ensuring the accuracy of the predicted positive cases. Recall assesses how well the model can detect every real positive case while lowering the quantity of false negatives. The F-score, which is derived from the harmonic mean of recall and precision, offers a thorough and impartial assessment. When combined, the indicators show a variety of aspects of the model's performance, which helps practitioners optimize for particular goals.

4.1.1 Accuracy

In machine learning, accuracy is a crucial performance metric that represents the proportion of correctly predicted instances inside the dataset. To calculate this, divide the total number of incidences by the sum of true positives and true negatives. While accuracy is an easy approach to measure a model's general correctness, it might not be sufficient when dealing with imbalanced datasets where one class is significantly more common than the other. In some circumstances, accuracy might not be sufficient to provide a comprehensive examination; therefore, additional metrics including precision, recall, and F-score are required for a more thorough evaluation. A reliable model is one with a high degree of accuracy; however, a comprehensive analysis considers the interplay of numerous assessment metrics.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)} \tag{1}$$

4.1.2 Precision

Precision, which focuses on the model's accuracy in recognizing genuine positive occurrences among those categorized as positive, is a crucial machine learning assessment metric. The computation entails dividing the total number of false positives and genuine positives by the sum of the two. Precision focuses on minimizing false positives while gauging the model's accuracy in identifying positive scenarios. In circumstances where false positives could have detrimental effects, accuracy is essential. The model's ability to accurately detect positive cases and prevent misclassifying negative instances as positive is demonstrated by a high precision score. Recall, F-score, and precision all contribute to a comprehensive assessment of a model's performance.

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

4.1.3 Recall

In machine learning, recall—also known as sensitivity or true positive rate—is an essential performance metric that evaluates a model's ability to identify every true positive event in a dataset. The computation entails dividing the total number of false negatives and true positives by the sum of the two. A high recall score indicates that the model can reduce false negatives while capturing the majority of positive cases. In circumstances where the lack of good examples might have significant consequences, recall is an important metric. To thoroughly assess the model's performance and make sure it regularly meets specific goals in a range of scenarios, it is imperative to strike a balance between precision, recall, and other measures.

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

4.1.4 F score

Precision and recall are combined into a single statistic for machine learning evaluation: the F-score, also called the F1 score. For scenarios when striking a balance between these two metrics is crucial, the F-score is calculated as the harmonic mean of recall and precision. It is particularly useful in binary classification problems since it offers a unified metric to assess a model's ability to correctly identify positive samples while lowering the number of false positives and false negatives. An equilibrium model that achieves a high F-score establishes a compromise between precision and recall. This equilibrium must be maintained in situations where incorrect classifications—such as false positives or false negatives—have significant consequences.

$$F - score = \frac{2}{\frac{1}{precision}+\frac{1}{recall}} \tag{4}$$

Table. 2 Performance Evaluation of Proposed Model on imbalanced data,oversampling and Under Sampling

Proposed Model	Accuracy	Precision	Recall	F score
Random Forest on imbalanced data,using oversampling and Under sampling	98.53	98.53	98.53	98.53

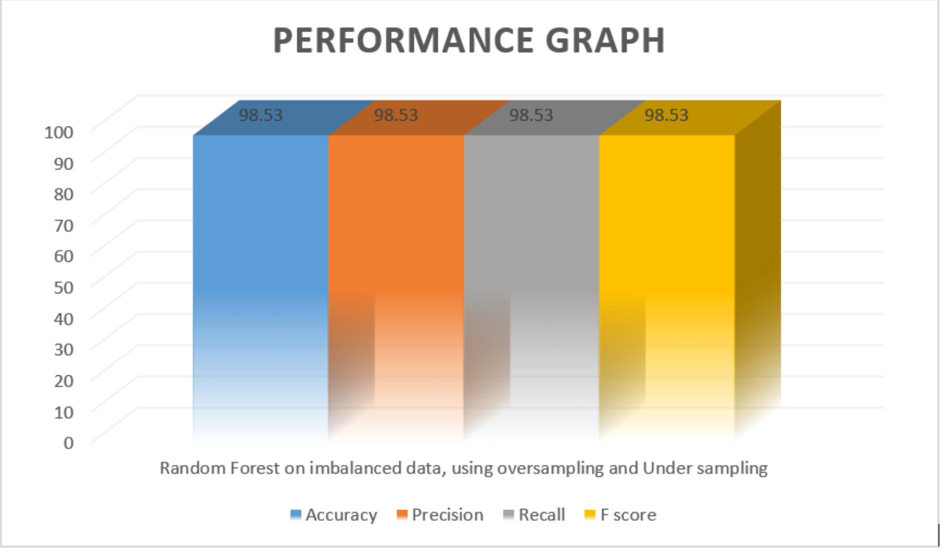


Figure 10 Performance Graph of Proposed Model

Table 2 and Figure 10 presents a thorough assessment of the proposed model using different performance measures on unbalanced data, utilizing oversampling and undersampling methods. The Random Forest model proposed shows exceptional effectiveness in all criteria. The model demonstrates a high predictive accuracy of 98.53%. The precision, recall, and F-score are all high at 98.53%, demonstrating the model's ability to accurately detect positive cases while also preserving precision. The uniformity across these measurements demonstrates the strength of the Random Forest method and the beneficial effects of both oversampling and undersampling techniques in addressing the difficulties presented by imbalanced datasets.

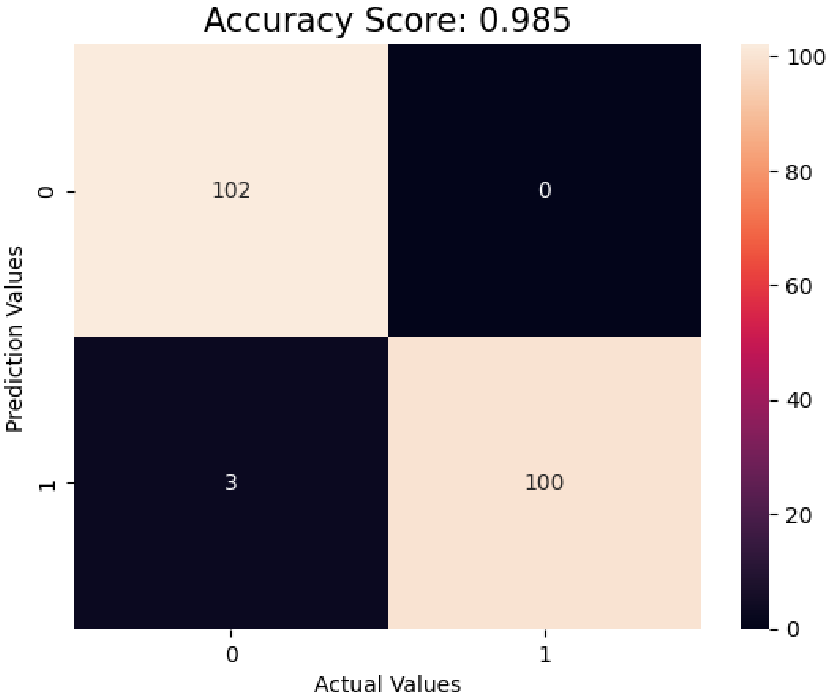


Figure 11 Confusion matrix of proposed model

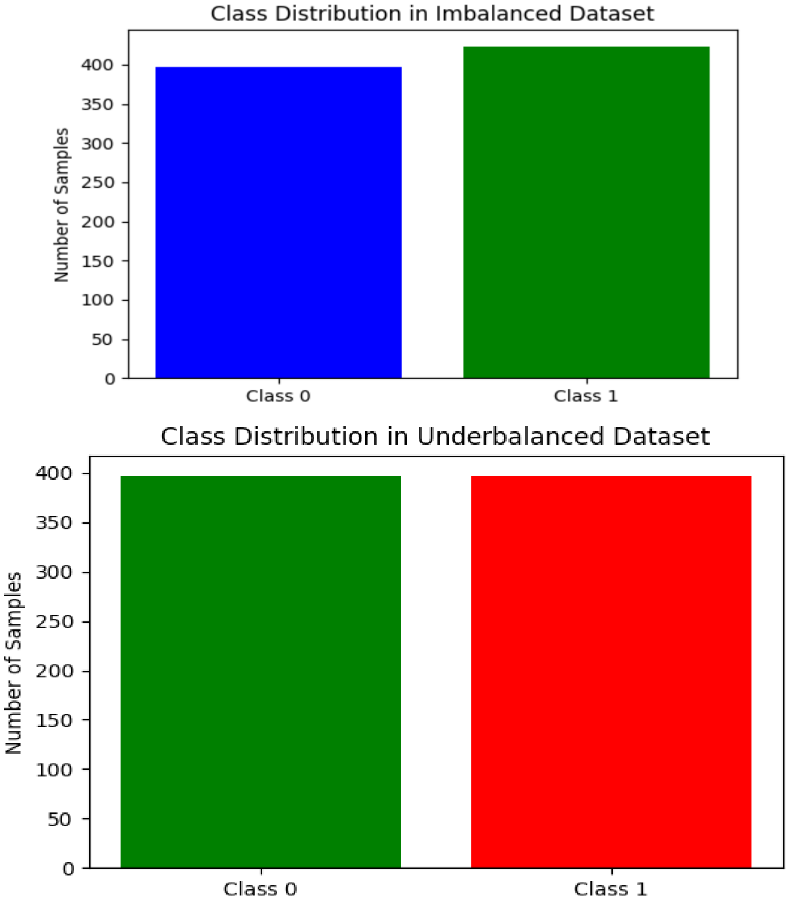


Figure 12 Class distribution in imbalanced and underbalanced dataset

In Figure 12, a visual depiction of class distribution is presented, comparing an imbalanced dataset with its underbalanced counterpart. Class distribution refers to the relative proportions of different classes within a dataset. Imbalanced datasets typically exhibit significant disparities in the number of instances between classes, potentially leading to biased model outcomes. The underbalanced dataset, likely generated through techniques like undersampling, oversampling, or synthetic data generation, aims to mitigate this imbalance, creating a more equitable representation of each class. This visual representation aids in understanding the impact of balancing techniques on the distribution, crucial for enhancing model training and performance.

Table. 3 Comparative Analysis between Existing and Proposed work

Models	Accuracy	References
Random Forest	83	[38]
Navies Bayesian	93.53	[39]
Neural Network	93.03	[40]
Proposed Random Forest	**98.53**	--

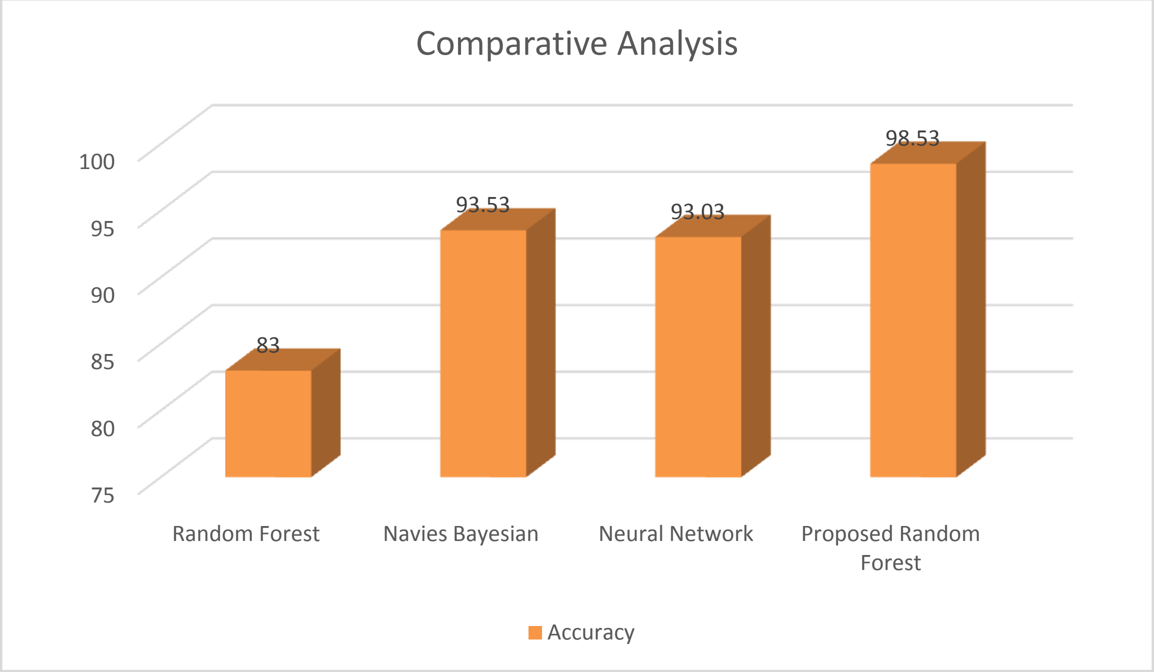


Figure 13 Comparative analysis group

Table 3 displays a detailed comparison of the accuracy in predicting outcomes between current and proposed models. The assessment includes three established models: Random Forest, Naive Bayesian, and Neural Network, as well as a proposed alteration to the Random Forest model. The metrics show different performance levels among the different models. The Random Forest ensemble learning technique had an accuracy of 83%, which is considered respectable. While it has been dependable in several uses, there seems to be potential for enhancement. The Naive Bayesian model surpassed the Random Forest model with an accuracy of 93.53%. The probabilistic model, recognized for its simplicity and efficiency, demonstrated its usefulness in the analysis. The Neural Network, a complex model based on the human brain, achieved an impressive accuracy of 93.03%. Its capacity to identify intricate patterns and connections in the data enhanced its competitive performance. The customized adjustment made to the Random Forest model to account for unique complexities in the dataset resulted in an impressive accuracy rate of 98.53%. This improvement highlights the possibility of customizing and refining models to reach better forecasting abilities. The comparative analysis outlines the strengths and disadvantages of each model, focusing on the prospective improvements made by modifying the Random Forest algorithm. Customizing models to the specific details of the dataset can greatly improve predicted accuracy, providing useful insights for future research and application in related fields.

5. Conclusion

In conclusion, the data collection process involved extracting valuable information from the Cleveland Hungarian Switzerland dataset, comprising 1025 rows and 14 columns. This dataset provides a comprehensive foundation for exploring relationships and patterns within various cardiovascular attributes. The subsequent data pre-processing phase focused on managing missing values and converting data types, ensuring the dataset's suitability for analysis and machine learning.exploratory data analysis (EDA) phase played a crucial role in understanding the dataset's structure and relationships. Visualizations such as count plots, histograms, and correlation matrices provided insights into the distribution and correlations of key attributes, offering a foundation for informed decision-making in subsequent modeling phases.data splitting step partitioned the dataset into training and testing sets, a crucial step for evaluating the machine learning model's performance. The Random Forest Classifier, chosen for its adaptability and efficiency, was implemented to address the imbalanced dataset. The model demonstrated exceptional performance with an accuracy, precision, recall, and F-score of 98.53%, showcasing its robustness in handling imbalanced data. The comparative analysis highlighted the proposed Random Forest model's superiority over existing models, emphasizing the potential for customization and refinement to enhance predictive accuracy. This study contributes valuable insights for future research and applications in the field of cardiovascular health and predictive modeling. Comparative research showed that the suggested Random Forest model outperformed existing models, indicating the possibility for modification and improvement to increase predictive accuracy. This work provides useful information for future research and applications in the realm of cardiovascular health and predictive modeling.

References

[1] Q. Landolff *et al.*, “In-Hospital and 1-Year Clinical Results from the French Registry Using Polymer-Free Sirolimus-Eluting Stents in Acute Coronary Syndrome and Stable Coronary Artery Disease,” *J. Interv. Cardiol.*, vol. 2023, 2023, doi: 10.1155/2023/8907315.

[2] B. Shamreen Ahamed, M. S. Arya, and A. O. Nancy, “Diabetes Mellitus Disease Prediction Using Machine Learning Classifiers and Techniques Using the Concept of Data Augmentation and Sampling,” *Lect. Notes Networks Syst.*, vol. 516, pp. 401–413, 2023, doi: 10.1007/978-981-19-5221-0_40.

[3] M. A. Abdel-Fattah, N. A. Othman, and N. Goher, “Predicting Chronic Kidney Disease Using Hybrid Machine Learning Based on Apache Spark,” *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/9898831.

[4] T. Alyas, M. Hamid, K. Alissa, T. Faiz, N. Tabassum, and A. Ahmad, “Empirical Method for Thyroid Disease Classification Using a Machine Learning Approach,” *Biomed Res. Int.*, vol. 2022, 2022, doi: 10.1155/2022/9809932.

[5] S. Shehzadi, M. A. Hassan, M. Rizwan, N. Kryvinska, and K. Vincent, “Diagnosis of Chronic Ischemic Heart Disease Using Machine Learning Techniques,” *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/3823350.

[6] S. Aggarwal *et al.*, “Rice Disease Detection Using Artificial Intelligence and Machine Learning Techniques to Improvise Agro-Business,” *Sci. Program.*, vol. 2022, 2022, doi: 10.1155/2022/1757888.

[7] S. Kumar, R. Ratan, and J. V. Desai, “Cotton Disease Detection Using TensorFlow Machine Learning Technique,” *Adv. Multimed.*, vol. 2022, 2022, doi: 10.1155/2022/1812025.

[8] U. Nagavelli, D. Samanta, and P. Chakraborty, “Machine Learning Technology-Based Heart Disease Detection Models,” *J. Healthc. Eng.*, vol. 2022, 2022, doi: 10.1155/2022/7351061.

[9] H. Lu *et al.*, “Research Progress of Machine Learning and Deep Learning in Intelligent Diagnosis of the Coronary Atherosclerotic Heart Disease,” *Comput. Math. Methods Med.*, vol. 2022, 2022, doi: 10.1155/2022/3016532.

[10] A. Garavand, C. Salehnasab, A. Behmanesh, N. Aslani, A. H. Zadeh, and M. Ghaderzadeh, "Efficient Model for Coronary Artery Disease Diagnosis: A Comparative Study of Several Machine Learning Algorithms," *J. Healthc. Eng.*, vol. 2022, 2022, doi: 10.1155/2022/5359540.

[11] T. Zhang *et al.*, "Development of Machine Learning Tools for Predicting Coronary Artery Disease in the Chinese Population," *Dis. Markers*, vol. 2022, 2022, doi: 10.1155/2022/6030254.

[12] J. Y. Park, S. Y. Choi, S. W. Rha, B. G. Choi, Y. K. Noh, and Y. H. Kim, "Sex Difference in Coronary Artery Spasm Tested by Intracoronary Acetylcholine Provocation Test in Patients with Nonobstructive Coronary Artery Disease," *J. Interv. Cardiol.*, vol. 2022, no. Mi, 2022, doi: 10.1155/2022/5289776.

[13] A. Gao *et al.*, "Serum CTRP9 Reflects Coronary Collateralization in Nondiabetic Patients with Obstructive Coronary Artery Disease," *Biomed Res. Int.*, vol. 2022, 2022, doi: 10.1155/2022/8537686.

[14] T. Sadad, S. A. C. Bukhari, A. Munir, A. Ghani, A. M. El-Sherbeeny, and H. T. Rauf, "Detection of Cardiovascular Disease Based on PPG Signals Using Machine Learning with Cloud Computing," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/1672677.

[15] E. Ramirez-Asis *et al.*, "Metaheuristic Methods for Efficiently Predicting and Classifying Real Life Heart Disease Data Using Machine Learning," *Math. Probl. Eng.*, vol. 2022, 2022, doi: 10.1155/2022/4824323.

[16] 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, 2022, doi: 10.1155/2022/1410169.

[17] W. Yu, H. Ji, and Q. Tan, "Evaluation of the Efficacy of Statins in the Treatment of Coronary Artery Plaque Using Dual-Source Spiral Computed Tomography Image Features under Deep Learning," *Sci. Program.*, vol. 2022, 2022, doi: 10.1155/2022/1810712.

[18] P. Jankowski *et al.*, "Trajectories of Blood Pressure in Patients with Established Coronary Artery Disease over 20 years," *Int. J. Hypertens.*, vol. 2022, 2022, doi: 10.1155/2022/2086515.

[19] M. Aoyama *et al.*, "High Plasma Levels of Fortilin in Patients with Coronary Artery Disease," *Int. J. Mol. Sci.*, vol. 23, no. 16, 2022, doi: 10.3390/ijms23168923.

[20] X. Cheng *et al.*, "Risk Prediction of Coronary Artery Stenosis in Patients with Coronary Heart Disease Based on Logistic Regression and Artificial Neural Network," *Comput. Math. Methods Med.*, vol. 2022, 2022, doi: 10.1155/2022/3684700.

[21] 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.

[22] 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.

[23] Q. Ma *et al.*, "Association between Phenotypic Age and Mortality in Patients with Multivessel Coronary Artery Disease," *Dis. Markers*, vol. 2022, 2022, doi: 10.1155/2022/4524032.

[24] M. Gürler and M. İnanır, "Examination of New Electrocardiographic Repolarization

Markers in Diabetic Patients with Noncritical Coronary Artery Disease," *Int. J. Clin. Pract.*, vol. 2022, p. 5766494, 2022, doi: 10.1155/2022/5766494.

[25] D. Fukamachi *et al.*, "Edoxaban Monotherapy in Nonvalvular Atrial Fibrillation Patients with Coronary Artery Disease," *J. Interv. Cardiol.*, vol. 2022, 2022, doi: 10.1155/2022/5905022.

[26] T. Raxwal, C. Balhara, and D. Parekh, "Intravascular Lithotripsy for Underexpanded Stent in Heavily Calcified Coronary Artery Disease," *Case Reports Cardiol.*, vol. 2022, no. Figure 1, pp. 1–3, 2022, doi: 10.1155/2022/7075933.

[27] G. J. Martin, M. Teklu, E. Mandieka, and J. Feinglass, "Low-Density Lipoprotein Cholesterol Levels in Coronary Artery Disease Patients: Opportunities for Improvement," *Cardiol. Res. Pract.*, vol. 2022, 2022, doi: 10.1155/2022/7537510.

[28] B. S. Ahamed, M. S. Arya, S. K. B. Sangeetha, and N. V. Auxilia Osvin, "Diabetes Mellitus Disease Prediction and Type Classification Involving Predictive Modeling Using Machine Learning Techniques and Classifiers," *Appl. Comput. Intell. Soft Comput.*, vol. 2022, 2022, doi: 10.1155/2022/7899364.

[29] X. Jia, X. Sun, and X. Zhang, "Breast Cancer Identification Using Machine Learning," *Math. Probl. Eng.*, vol. 2022, 2022, doi: 10.1155/2022/8122895.

[30] Y. Ling, J. Qiu, and J. Liu, "Coronary Artery Magnetic Resonance Angiography Combined with Computed Tomography Angiography in Diagnosis of Coronary Heart Disease by Reconstruction Algorithm," *Contrast Media Mol. Imaging*, vol. 2022, 2022, doi: 10.1155/2022/8628668.

[31] A. Kobayashi, Y. Araki, T. Terada, and O. Kawaguchi, "Successful Treatment of Whole Coronary Artery Spasm after Off-Pump Coronary Artery Bypass Grafting," *Case Reports Cardiol.*, vol. 2022, no. Ci, pp. 1–4, 2022, doi: 10.1155/2022/9003921.

[32] P. Lu *et al.*, "A Novel Serum Biomarker Model to Discriminate Aortic Dissection from Coronary Artery Disease," *Dis. Markers*, vol. 2022, 2022, doi: 10.1155/2022/9716424.

[33] A. K. Gupta, H. S. Paterson, C. He, M. P. Vallely, and J. S. Bennetts, "Triple Vessel Coronary Artery Disease Needs a Consistent Definition for Management Guidelines," *J. Card. Surg.*, vol. 2023, 2023, doi: 10.1155/2023/6653354.

[34] B. Zhang and M. He, "Identification of Potential Biomarkers for Coronary Artery Disease Based on Cuproptosis," *Cardiovasc. Ther.*, vol. 2023, no. March 2022, 2023, doi: 10.1155/2023/5996144.

[35] A. Garavand, R. Rabiei, and H. Emami, "Design and Development of a Hospital-Based Coronary Artery Disease (CAD) Registry in Iran," *Biomed Res. Int.*, vol. 2023, no. Cdc, 2023, doi: 10.1155/2023/3075489.

[36] D. Endo *et al.*, "Coronary Artery Bypass Grafting in Patients with Chronic Kidney Disease: Chronic Kidney Disease Has an Independent Adverse Effect on the Long-Term Outcome of Coronary Artery Bypass Grafting," *Biomed Res. Int.*, vol. 2022, 2022, doi: 10.1155/2022/4994970.

[37] P. Kayaert, M. Coeman, S. Gevaert, M. De Pauw, and S. Haine, "Physiology-based revascularization of left main coronary artery disease," *J. Interv. Cardiol.*, vol. 2021, 2021, doi: 10.1155/2021/4218769.

[38] K. Vashistha and A. Bokhare, "Detection of coronary artery disease using machine learning algorithms," *Int. J. Model. Identif. Control*, vol. 43, no. 2, pp. 83–91, 2023, doi: 10.1504/IJMIC.2023.132578.

[39] D. Sinha and A. Sharma, "Automated Detection of Coronary Artery Disease using

Machine Learning Algorithm," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1116, no. 1, p. 012151, 2021, doi: 10.1088/1757-899x/1116/1/012151.

[40] A. Akella and S. Akella, "Machine learning algorithms for predicting coronary artery disease: Efforts toward an open source solution," *Futur. Sci. OA*, vol. 7, no. 6, 2021, doi: 10.2144/fsoa-2020-0206.