Heart Disease Prediction Using Machine Learning Techniques

A Dissertation submitted
for the partial fulfillment of the degree of
Bachelor of Engineering in
Information Technology
(Session 2023 -2024)

Guided By:
Dr. Jagdish Raikwal

Submitted By:
Prateek Gupta (20I8141)
Urwashi Kumrawat (20I8163)
Vatsal Gupta (20I8165)

Department of Information Technology
Institute of Engineering & Technology
Devi Ahilya Vishwavidyalaya, Indore (M.P.)
(www.iet.dauniv.ac.in)

12/2023

Dissertation Approval Sheet

The dissertation entitled **"Heart Disease Prediction Using Machine Learning Techniques"** submitted by **Prateek Gupta, Urwashi Kumrawat, Vatsal Gupta** is approved as partial fulfillment for the award of **Bachelor of Engineering in Information Technology** degree by **Devi Ahilya Vishwavidyalaya, Indore**.

Internal Examiner **External Examiner**

Director
Institute of Engineering & Technology
Devi Ahilya Vishwavidyalaya,
Indore (M.P.)

Recommendation

The dissertation entitled **"Heart Disease Prediction Using Machine Learning Techniques"** submitted by **Prateek Gupta, Urwashi Kumrawat, Vatsal Gupta** is a satisfactory account of the bonafide work done under my supervision is recommended towards the partial fulfillment for the award of **Bachelor of Engineering in Information Technology** degree by **Devi Ahilya Vishwavidyalaya, Indore.**

Date:

Dr. Jagdish Raikwal
Project Guide

Endorsed By:

Head,
Department of Information Technology

Candidate Declaration

We hereby declare that the work which is being presented in this project entitled “Heart Disease Prediction using Machine Learning Techniques” in partial fulfillment of degree of Bachelor of Engineering in Information Technology is an authentic record of our own work carried out under the supervision and guidance of **Dr. Jagdish Raikwal**, Assistant Professor in Department of **Information Technology**, Institute of Engineering & Technology, Devi Ahilya Vishwavidyalaya, Indore

We are fully responsible for the matter embodied in this project in case of any discrepancy found in the project and the project has not been submitted for the award of any other degree.

Date:

Place:

Prateek Gupta

Urwashi Kumrawat

Vatsal Gupta

Acknowledgements

First and foremost, I would like to extend my heartfelt gratitude to my guide and mentor **Dr. Jagdish Raikwal**, Assistant Professor, Department of Information Technology, Institute of Engineering & Technology, Devi Ahilya Vishwavidyalaya, Indore for his invaluable advices, guidance, encouragement, and for sharing his broad knowledge. His wisdom, knowledge and commitment to the highest standards inspired and motivated me. He has been very generous in providing the necessary resources to carry out my research. He is an inspiring teacher, a great advisor, and most importantly a nice person.

I am thankful to respected **Dr. Mrs. Vrinda Tokekar,** Head of Department, Information Technology, Institute of Engineering & Technology, Devi Ahilya Vishwavidyalaya, Indore for providing me inspiration and guidance.

I am thankful to respected **Dr. Sanjiv Tokekar** Director, Institute of Engineering & Technology, Devi Ahilya Vishwavidyalaya, Indore for providing me inspiration and guidance.

I am greatly indebted to all my friends, who have graciously applied themselves to the task of helping me with ample moral supports and valuable suggestions. I also place on record, my sense of gratitude to one and all, who directly or indirectly, have lent their hand in this venture.

Abstract

Heart disease is one of the most significant causes of mortality in the world today. Prediction of cardiovascular disease is a critical challenge in the area of clinical data analysis. Machine learning (ML) has been shown to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry. We have also seen ML techniques being used in recent developments in different areas of the Internet of Things (IoT). Various studies give only a glimpse into predicting heart disease with ML techniques. In this paper, we propose a novel method that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. The prediction model is introduced with different combinations of features and several known classification techniques. We produce an enhanced performance level with an accuracy level of 90% through the prediction model for heart disease with the random forest.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Overview and Issues Resolved

It is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate and many other factors. Various techniques in data mining and neural networks have been employed to find out the severity of heart disease among humans. The severity of the disease is classified based on various methods like K-Nearest Neighbor Algorithm (KNN), Decision Trees (DT), Genetic algorithm (GA), and Naive Bayes (NB) [11], [13]. The nature of heart disease is complex and hence, the disease must be handled carefully. Not doing so may affect the heart or cause premature death. The perspective of medical science and data mining are used for discovering various sorts of metabolic syndromes. Data mining with classification plays a significant role in the prediction of heart disease and data investigation.

In this work, we will compare different machine learning techniques and choose the best one. The experiment results show that our proposed method i.e., Random Forest has stronger capability to predict heart disease compared to existing methods.

1.2 Statement of Problem

The primary problem addressed here is the difficulty in identifying heart disease, given its complex nature and various contributing risk factors. The main objective of this research is to improve the performance accuracy of heart disease prediction.

1.3 Significance of Problem

The significance lies in the potential consequence of not accurately identifying heart disease, leading to heart-related complications or premature death. Various methods have been used for knowledge abstraction by using known methods of data mining for prediction of heart disease. Many studies have been conducted that-results in restrictions of feature selection for algorithmic use. In contrast, the random forest method uses all features without any restrictions of feature selection. Here we conduct experiments used to identify the features of a machine learning algorithm with a random forest model.

1.4 Assumptions

This includes the effectiveness of data mining and neural networks in prediction heart disease severity. Various methods have been used for knowledge abstraction by using known methods of data mining for prediction of heart disease. In this work, we are mainly focused on random forest using hyperparameter tuning and comparing it with several different methods. When working on heart disease prediction using Random Forest with hyperparameter tuning, several assumptions should be considered. Firstly, it is assumed that the dataset used for training and testing the model is representative of the population it aims to predict for, ensuring that it contains diverse and accurate information related to heart disease attributes. Furthermore, it is presumed that the hyperparameters chosen for tuning, such as the number of trees in the forest, the maximum depth of the trees, and the minimum number of samples required to split a node, are optimized effectively to enhance the model's predictive performance without overfitting. Moreover, it is assumed that the evaluation metrics used to assess the model's performance, such as accuracy, precision, recall, and F1-score, are appropriate for the task at hand and provide meaningful insights into the model's effectiveness. Overall, these assumptions play a crucial role in developing a robust and reliable heart disease prediction model using Random Forest with hyperparameter tuning.

1.5 Proposed Solution

The proposed solution for heart disease prediction in this research paper is centered on leveraging machine learning techniques, specifically Random Forest with hyperparameter tuning. Random Forest is chosen for its ability to handle large datasets with high dimensionality, making it suitable for analyzing the diverse range of attributes associated with heart disease. By employing Random Forest, the model can effectively capture complex relationships and interactions between various risk factors, thus enhancing predictive accuracy. Moreover, hyperparameter tuning is incorporated to optimize the performance of the Random Forest model. Through careful adjustment of parameters such as the number of trees, tree depth, and minimum samples per leaf, the model's ability to generalize to unseen data is enhanced while mitigating overfitting. This proposed solution aims to develop a robust and reliable predictive model for heart disease, contributing to early detection and prevention strategies, ultimately improving patient outcomes and healthcare management. Here's a brief explanation using 13 independent attributes (features) and 1 dependent attribute (target variable) in the context of predicting heart disease, for example:

- **Random Forest (RF)**

  The RF component involves constructing an ensemble of decision trees based on subsets of the 13 attributes. Each tree in the forest is trained on a different subset of the data, promoting diversity and reducing overfitting. It's particularly effective in capturing complex, non-linear relationships within the data.

- **Benefits**

  - Enhanced Predictive Accuracy: By leveraging Random Forest, which aggregates predictions from multiple decision trees, the model can capture complex relationships and interactions among various risk factors associated with heart disease. This results in improved predictive accuracy compared to traditional linear models.

  - Robustness to Overfitting: Hyperparameter tuning allows for fine-tuning of model parameters to optimize performance while preventing overfitting. This ensures that the model generalizes well to unseen data, reducing the risk of making erroneous predictions on new patient data.

  - Interpretability: Random Forest models provide feature importance scores, allowing clinicians and researchers to understand which attributes contribute most to the prediction of heart disease. This interpretability aids in identifying key risk factors and understanding the underlying mechanisms of the disease.

  - Flexibility and Adaptability: Random Forest models can accommodate different types of data and are less sensitive to outliers and missing values compared to some other machine learning techniques. This flexibility makes them adaptable to diverse healthcare datasets and patient populations.

- **Training and Evaluation**

  The model is trained on a dataset containing examples of the 13 attributes and the corresponding target variable. Evaluation metrics such as accuracy, precision, recall, or F1 score are used to assess the model's performance on a separate test dataset.

CHAPTER 2
LITERATURE SURVEY

2.1 Existing Papers

A quiet Significant amount of work related to the diagnosis of Cardiovascular Heart disease using Machine Learning algorithms has motivated this work. This paper contains a brief literature survey. An efficient Cardiovascular-disease prediction has been made by using various algorithms some of them include Logistic Regression, KNN, Random Forest Classifier Etc. It can be seen in Results that each algorithm has its strength to register the defined objectives [2].

This [1] paper addresses the critical challenge of predicting cardiovascular disease using machine learning techniques. It introduces a novel method that aims to improve prediction accuracy by identifying significant features. Neural networks are generally regarded as the best tool for prediction of diseases like heart disease and brain disease. The proposed method which we use has 13 attributes for heart disease prediction.

The risk factors of Coronary Heart-Disease or atherosclerosis is identified by McPherson et al.,[3] using the inbuilt implementation algorithm using uses some techniques of Neural Network and were just accurately able to predict whether the test patient is suffering from the given disease or not. Diagnosis and prediction of Heart-Disease and Blood Pressure along with other attributes using the aid of neural networks was introduced by R. Subramanian et al.,[4].

A deep Neural Network was Built incorporating the given attributes related to the disease which were able to produce an output which was carried out by the output perceptron and almost included 120 hidden layers which is the basic and most relevant technique of ensuring a accurate result of having heart disease if we use the model for Test Dataset. The supervised network has been advised for diagnosis of heart diseases [5].

When the testing of the model was done by a doctor using an unfamiliar data, the model used and trained from the previous learned data and predicted the result thereby calculating the accuracy of the given model. The severity of the disease is classified based on various methods like K-Nearest Neighbor Algorithm (KNN), Decision Trees (DT), Genetic algorithm (GA), and Naive Bayes (NB) [6], [7].
For experimental validation, we use the well-known Cleveland dataset which is collected from a UCI machine learning repository. We will see later on how our results prove to be prominent when compared to some of the known supervised learning techniques [8], [9].

The most powerful evolutionary algorithm Particle Swarm Optimization (PSO) is introduced and some rules are generated for heart disease. The rules have been applied randomly with encoding techniques which result in improvement of the accuracy overall [10]. Heart disease is predicted based on symptoms namely, pulse rate, sex, age, and many others. The ML algorithm with Neural Networks is introduced, whose results are more accurate and reliable as we have seen in [11], [12].

ML algorithms on network traffic data has been shown to provide accurate identification of IoT devices connected to a network. Meidan et al. collected and labeled network traffic data from nine distinct IoT devices, PCs and smartphones. Using supervised learning, they trained a multi-stage meta classifier. In the first stage, the classifier can distinguish between traffic generated by IoT and non-IoT devices. In the second stage, each IoT device is associated with a specific IoT device class. Deep learning is a promising approach for extracting accurate information from raw sensor data from IoT devices deployed in complex environments [13-16].

2.2 Technologies And Research Tools Used

2.2.1 K-Nearest Neighbor Algorithm, Decision Trees and Naive Bayes

These classification algorithms are used for assessing the severity of heart disease based on different methods. The choice of multiple classification algorithms enables a comprehensive evaluation of the data. Each algorithm may capture different aspects of the dataset, contributing to a more robust prediction model.

2.2.2 Dataset from UCI Machine Learning Repository

The Cleveland dataset is utilized for experimental validation of heart disease prediction. It includes symptoms like pulse rate, sex, age, etc. The Cleveland dataset is a well-known and widely used dataset in the field of heart disease prediction, providing a standard benchmark for evaluating the proposed research.

2.2.3 Machine Learning (ML) Algorithms

ML algorithms are applied to network traffic data to accurately identify IoT devices connected to a network. Deep learning is used for extracting information from raw sensor data from IoT devices. ML algorithms offer a powerful approach for analyzing network traffic data, particularly in the context of IoT. Deep learning is chosen for its ability to extract meaningful patterns from complex sensor data.

CHAPTER 3
DESIGN

3.1 Description of Research Design and Procedures Used

The research design for this study on heart disease prediction utilizing machine learning techniques encompasses a meticulously structured approach, designed to ensure robustness and reliability in the analysis and interpretation of the data. The procedure employed follows a systematic framework tailored to leverage the strengths of various machine learning algorithms, culminating in the identification of the most effective predictive model.

3.1.1 Data Pre-Processing

Heart disease data is pre-processed after collection of various records. The initial phase involves the acquisition of data from the UCI dataset. The dataset contains a total of 303 patient records, where 6 records are with some missing values. Those 6 records have been removed from the dataset and the remaining 297 patient records are used in pre-processing. The multiclass variable and binary classification are introduced for the attributes of the given dataset. The multi-class variable is used to check the presence or absence of heart disease. In the instance of the patient having heart disease, the value is set to 1, else the value is set to 0 indicating the absence of heart disease in the patient. The pre-processing of data is carried out by converting medical records into diagnosis values. The results of data pre-processing for 297 patient records indicate that 137 records show the value of 1 establishing the presence of heart disease while the remaining 160 reflected the value of 0 indicating the absence of heart disease.

3.1.2 Feature Selection And Reduction

From among the 13 attributes of the data set, two attributes pertaining to age and sex are used to identify the personal information of the patient. The remaining 11 attributes are considered important as they contain vital clinical records. Clinical records are vital to diagnosis and learning the severity of heart disease. As previously mentioned in this experiment, several machine learning techniques are used namely, Naïve Bayes, Logistic Regression, K-Nearest Neighbor, Decision Tree, Random Forest and Support Vector Machine. The experiment was repeated with all the ML techniques using all 13 attributes. The feature selection and modeling keep on repeating for various combinations of attributes. The performance of each model generated based on 13 features and ML techniques used for each iteration and performance are recorded.

3.1.3 Model Selection and Evaluation

Various machine learning algorithms, including logistic regression, naive Bayes, k-nearest neighbors (KNN), support vector machine (SVM), random forest, and decision tree, are evaluated. Each algorithm is trained and tested using appropriate evaluation metrics such as accuracy, precision, recall, and F1-score to assess its performance in predicting heart disease.

3.1.4 Performance Measures

Several standard performance metrics such as accuracy, precision and error in classification have been considered for the computation of performance efficacy of this model. Accuracy in the current context would mean the percentage of instances correctly predicting from among all the available instances. Precision is defined as the percentage of corrective prediction in the positive class of the instances. Classification error is defined as the percentage of accuracy missing or error available in the instances. To identify the significant features of heart disease, three performance metrics are used which will help in better understanding the behavior of the various combinations of the feature-selection. ML technique focuses on the best performing model compared to the existing models. We introduced Random Forest with hyperparameter tuning, which produces high accuracy and less classification error in the prediction of heart disease. The performance of every classifier is evaluated individually and all results are adequately recorded for further investigation.

3.1.5 Random Forest using Hyperparameter Tuning

The Random Forest with hyperparameter tuning, offering a robust and accurate framework. It uses the leveraging Random Forest's ability to capture complex relationships among various risk factors to enhances predictive accuracy. Hyperparameter tuning optimizes the model's performance, ensuring effective generalization to unseen data while preventing overfitting. Additionally, Random Forest's interpretability facilitates the identification of key contributors to heart disease prediction, aiding clinicians in understanding underlying factors. This scalable solution accommodates diverse datasets and contributes to early detection and personalized treatment strategies, ultimately improving patient outcomes in healthcare settings.

3.1.6 Classification Modelling

In this phase of our research project, we focus on constructing and evaluating classification models aimed at predicting the presence or absence of heart disease based on the dataset provided. Our primary goal is to develop a predictive model that can effectively

differentiate between individuals with and without heart disease, thus aiding in early detection and intervention.

To commence our modeling endeavor, we first familiarize ourselves with the dataset's attributes. These attributes encompass a wide range of factors potentially associated with heart health, including age, gender, chest pain type, blood pressure, cholesterol levels, and electrocardiographic results, among others. Each attribute contributes valuable information that may influence the presence or absence of heart disease, forming the basis for our predictive modeling approach.

We embark on our modeling journey by exploring a diverse array of machine learning algorithms suited for classification tasks. Among the algorithms considered are logistic regression, naive Bayes, k-nearest neighbors (KNN), support vector machine (SVM), random forest, and decision tree. Each algorithm brings its unique strengths and characteristics to the table, offering various approaches to classify patients into distinct categories based on their attribute profiles.

Following the selection of candidate algorithms, we proceed to evaluate their performance using rigorous validation techniques. Metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve are employed to assess the models' predictive capabilities comprehensively. Through this process, we gain insights into the strengths and limitations of each algorithm, facilitating informed decision-making regarding model selection.

In our quest to optimize model performance, we employ hyperparameter tuning techniques to fine-tune the parameters of selected algorithms. Hyperparameter tuning involves systematically exploring different parameter configurations to identify the optimal settings that maximize predictive accuracy and generalization. Grid search and randomized search are commonly employed strategies for hyperparameter tuning, enabling us to refine the models and enhance their robustness.

Upon thorough experimentation and evaluation, we identify the Random Forest algorithm as the most promising candidate for our heart disease prediction task. Leveraging its ability to handle complex datasets and mitigate overfitting, we further refine the Random Forest model through hyperparameter tuning. By optimizing the algorithm's parameters, we aim to create a highly accurate and reliable predictive model capable of effectively identifying individuals at risk of heart disease. Thus, the proposed model for our research project is the Random Forest classifier

with optimized hyperparameters, poised to deliver actionable insights for proactive cardiovascular health management.

The research design and procedure adopted in this study exemplify a systematic and comprehensive approach aimed at developing a robust predictive model for heart disease. From data collection and preprocessing to model selection and evaluation, each step is meticulously executed to ensure the reliability and validity of the findings, ultimately advancing the field of medical diagnostics through innovative machine learning methodologies.

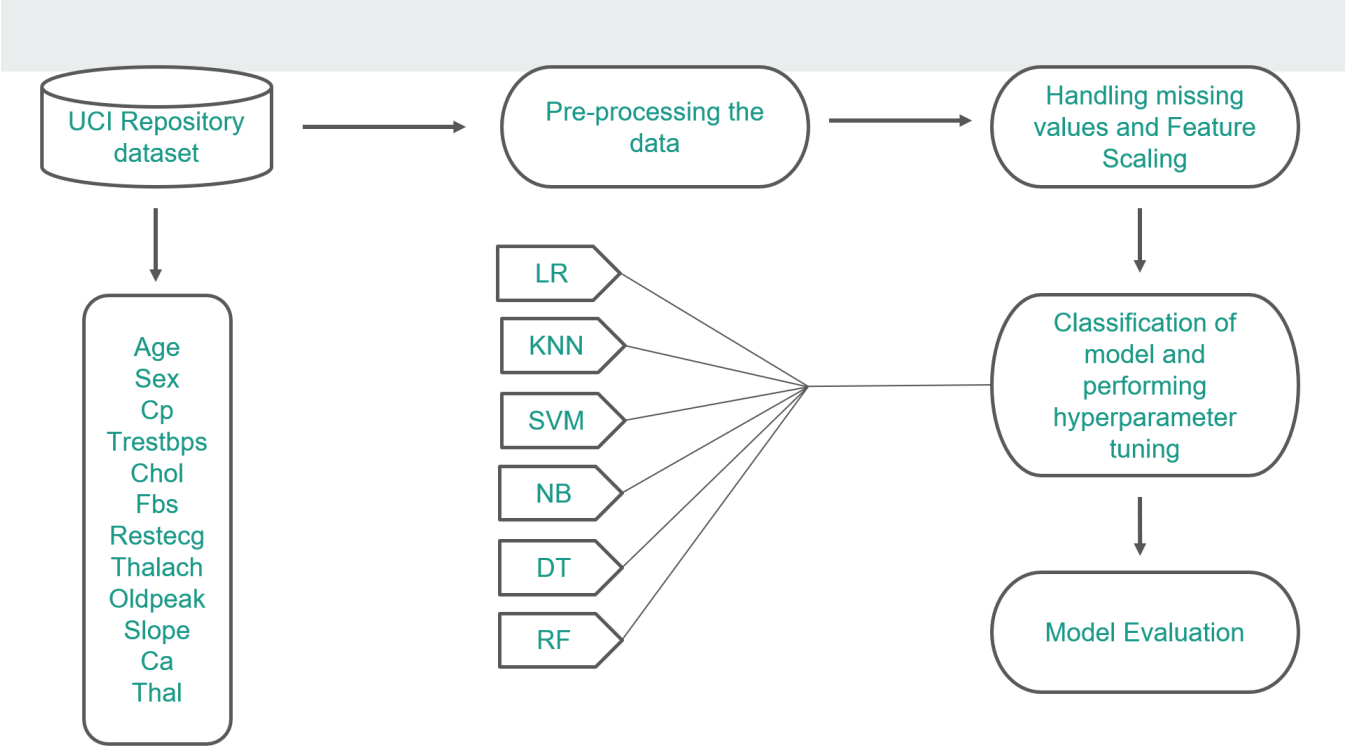

Fig. 3.1 Prediction of Heart Disease using Various ML Techniques

3.2 Data Sets

Heart disease data was collected from the UCI machine learning repository. There are four databases (i.e. Cleveland, Hungary, Switzerland, and the VA Long Beach). The Cleveland database was selected for this research because it is a commonly used database for Machine Learning researchers with comprehensive and complete records. The dataset contains 303 records. Although the Cleveland dataset has 76 attributes, the data set provided in the repository furnishes information for a subset of only 14 attributes. The data source of the Cleveland dataset

is the Cleveland Clinic Foundation. Table 3.1 shows the UCI dataset detailed information with the attributes used.

Within our dataset, we encapsulate a comprehensive array of attributes meticulously curated to capture various facets of cardiovascular health and associated risk factors. Each attribute serves as a key building block in our quest to develop a robust predictive model for heart disease detection. Let's delve into the essence of these attributes:

1. **Age**: This fundamental demographic feature denotes the age of each patient in years, providing crucial insight into the impact of aging on cardiovascular health and disease prevalence.

2. **Sex**: Gender plays a pivotal role in heart disease risk assessment, with males and females exhibiting divergent susceptibility patterns. This binary attribute distinguishes between male (1) and female (0) patients, facilitating gender-specific analysis.

3. **Chest Pain Type (cp):** Characterizing the nature of chest pain experienced by patients, this attribute categorizes chest pain into distinct types, offering valuable diagnostic clues indicative of underlying cardiac conditions.

4. **Resting Blood Pressure (trestbps):** Reflecting the baseline blood pressure levels of patients at rest, this metric serves as a vital indicator of hypertension, a significant risk factor for cardiovascular disease.

5. **Serum Cholesterol** (**chol**): Elevated serum cholesterol levels constitute a major cardiovascular risk factor, contributing to atherosclerosis and coronary artery disease. This attribute quantifies the concentration of cholesterol in the patient's bloodstream.

6. **Fasting Blood Sugar (fbs):** The fasting blood sugar level serves as a marker for diabetes mellitus, a metabolic disorder closely linked to cardiovascular complications. Values exceeding 120 mg/dl signify elevated blood sugar levels, warranting further investigation.

7. **Resting Electrocardiographic Results (restecg):** Electrocardiographic findings obtained during rest provide valuable insights into cardiac function and potential abnormalities. This attribute encompasses different electrocardiogram (ECG) patterns, aiding in the diagnosis of myocardial ischemia and other cardiac conditions.

8. **Maximum Heart Rate Achieved (thalach):** The maximum heart rate achieved during exercise stress testing serves as a reliable indicator of cardiovascular fitness and autonomic function, offering prognostic implications for heart disease risk.

9. **Exercise Induced Angina (exang):** The presence of exercise-induced angina signifies myocardial ischemia and coronary artery disease, prompting further evaluation of cardiac function and ischemic burden.

10. **ST Depression Induced by Exercise Relative to Rest (oldpeak):** Exercise-induced ST segment depression on electrocardiography is indicative of myocardial ischemia, providing valuable diagnostic information regarding the severity and extent of coronary artery disease.

11. **The Slope of the Peak Exercise ST Segment (slope):** The morphology of the ST segment during peak exercise provides additional diagnostic insights into myocardial ischemia, aiding in risk stratification and therapeutic decision-making.

12. **Number of Major Vessels Colored by Fluoroscopy (ca):** Fluoroscopy-assisted visualization of major coronary vessels enables the assessment of coronary artery disease severity and extent, informing treatment strategies and prognostic considerations.

13. **Thalassemia (thal):** Thalassemia, a hereditary blood disorder, may influence cardiovascular health through various mechanisms, warranting inclusion as a potential confounding factor in our analysis.

14. **Target**: The target attribute serves as the focal point of our predictive modeling endeavor, indicating the presence (1) or absence (0) of heart disease in each patient based on comprehensive clinical assessment and diagnostic criteria.

Collectively, these attributes encapsulate a wealth of clinical information essential for accurate risk assessment, diagnosis, and prognosis in the context of cardiovascular health. Fig. 3.2 shows bar plots of some of the attributes. By leveraging the rich insights embedded within our dataset, we endeavor to develop a predictive model capable of discerning subtle patterns and nuances indicative of underlying heart disease, thus empowering proactive intervention and personalized patient care.

Table 3.1 UCI Dataset Attributes detailed information

Attribute	Description	Type
Age	Patient's age in completed years	Numeric
Sex	Patient's Gender (male represented as l and female as 0)	Nominal
Cp	The type of Chest pain categorized into 4 values: 1. typical angina, 2. atypical angina, 3. non-anginal pain and 4. symptomatic	Nominal
Trestbps	Level of blood pressure at resting mode (in mm/Hg at the time of admitting in the hospital)	Numeric
Chol	Serum cholesterol in mg/dl	Numeric
FBS	Blood sugar levels on fasting > 120 mg/dl; represented as 1 in case of true, and 0 in case of false	Nominal
Resting	Results of electrocardiogram while at rest are represented in 3 distinct values: Normal state is represented as Value 0, Abnormality in ST-T wave as Value 1, (which may include inversions of T-wave and/or depression or elevation of ST of > 0.05 mV) and any probability or certainty of LV hypertrophy by Estes' criteria as Value 2	Nominal
Thali	The accomplishment of the maximum rate of heart	Numeric
Exang	Angina induced by exercise. (0 depicting 'no' and I depicting 'yes')	Nominal
Oldpeak	Exercise-induced ST depression in comparison with the state of rest	Numeric
Slope	ST segment measured in terms of the slope during peak exercise depicted in three values: 1. unsloping, 2. flat and 3. downsloping	Nominal
Ca	Fluoroscopy coloured major vessels numbered from 0 to 3	Numeric
Thal	Status of the heart illustrated through three distinctly numbered values. Normal numbered as 3, fixed defect as 6 and reversible defect as 7	Nominal
Num	Heart disease diagnosis represented in 5 values, with 0 indicating total absence and 1 to 4 representing the presence in different degrees	Nominal

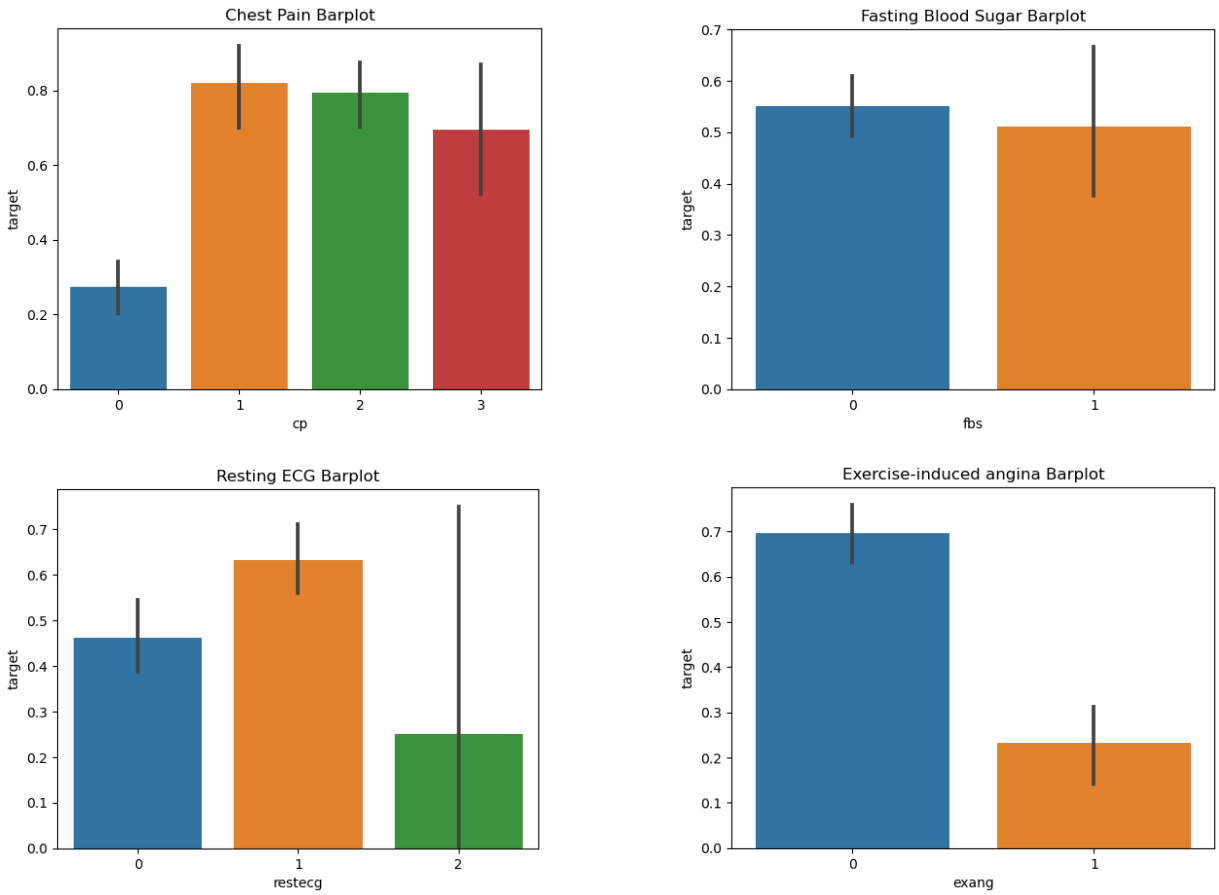

Fig. 3.2 Bar Plot of Attributes

There are 13 attributes that feature in the prediction of heart disease, where only one attribute serves as the output or the predicted attribute to the presence of heart disease in a patient. The Cleveland dataset contains an attribute named num to show the diagnosis of heart disease in patients on different scales, from 0 to 4. In this scenario, 0 represents the absence of heart disease and all the values from 1 to 4 represent patients with heart disease, where the scaling refers to the severity of the disease (4 being the highest). and Table 3.2 shows the data type and range of values.

Table 3.2 UCI dataset range and datatype

AGE	Numeric [29 to 77; unique=41; mean=54.4; median=56]
SEX	Numeric [0 to I; unique=2; mean=0.68; median=1]
CP	Numeric [1 to 4; unique=4; mean=3.16; median=3]
TRESTBPS	Numeric [94 to 200; unique=50; mean=131.69; median=130|
CHOL	Numeric [126 to 564; unique= 152; mean=246.69; median=241]
FBS	Numeric [0 to 1; unique=2; mean=0.15; median=01
RESTECG	Numeric [0 to 2; unique=3; mean=0.99; median=1]
THALACH	Numeric [71 to 202; unique=91; mean=149.61; median=153]
EXANG	Numeric [0 to 1; unique=2; mean=0.33; median=0.00]
OLDPEAK	Numeric [0 to 6.20; unique=40; mean=1.04; median=0.80]
SLOPE	Numeric [1 to 3; unique=3; mean=1.60; median=2]
CA	Categorical [5 levels]
THAL	Categorical [4 levels]
TARGET	Numeric [0.00 to 4.00; unique=5; mean=0.94; median=0.00]

3.3 Sampling of data

In this research, the sampling of data is a critical step in ensuring the representative nature of the dataset used for heart disease prediction. The Cleveland dataset from the UCI Machine Learning Repository is selected for experimental validation. This dataset is well-established in the field, providing a diverse set of symptoms, including pulse rate, sex, and age. The use of this dataset ensures a comprehensive representation of the population under study. The sampling process involves a thorough exploration of various attributes, as detailed in Table 1, to capture the nuances of heart disease manifestations. The iterative approach of feature selection based on decision tree entropy allows for the inclusion of relevant attributes without restrictions, ensuring a holistic understanding of the dataset. The utilization of Jupyter Notebook enhances the sampling process by providing a visual representation of the dataset, facilitating effective data exploration, and contributing to the overall accuracy of the heart disease prediction model. By employing a well-established dataset and advanced tools for sampling, the research aims to enhance the robustness and reliability of the predictive analytics conducted in the subsequent stages of the study.

3.4 Statistical Treatment

3.4.1 Logistic Regression

Logistic regression is a statistical method used for binary classification tasks, predicting the probability of a binary outcome based on one or more predictor variables. It models the relationship between the independent variables and the log-odds of the dependent variable using the logistic function, ensuring that the predicted probabilities fall between 0 and 1. The model estimates coefficients for each predictor, indicating the strength and direction of their influence on the outcome. Logistic regression was evaluated using statistical metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. Cross-validation techniques were employed to obtain reliable estimates of model performance.

Table 3.3 Classification Report of Logistic Regression

STATISTICAL→ MEASURES ↓ /PREDICTION	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.85	0.81	0.83	27
1	0.86	0.88	0.87	34
ACCURACY			0.85	61
MACRO AVG	0.85	0.85	0.85	61
WEIGHTED AVG	0.85	0.85	0.85	61

3.4.2 Decision Trees

A decision tree is a hierarchical tree-like structure used for both classification and regression tasks. It recursively splits the data based on the value of features, aiming to maximize information gain or minimize impurity at each node. Decision trees are interpretable and can capture complex relationships between features and the target variable. They're robust to outliers and missing values, making them suitable for datasets with heterogeneous data types. However, decision trees are prone to overfitting, especially when

the tree depth is not controlled. Techniques like pruning and ensemble methods like Random Forest address this issue, improving the generalization ability of decision trees.

For training samples of data D, the trees are constructed based on high entropy inputs. These trees are simple and fast constructed in a top-down recursive divide and conquer (DAC) approach. Tree pruning is performed to remove the irrelevant samples on D.

$$\text{Entropy} = -\sum_{j=1}^{m} p_{ij} \log_2 p_{ij} \qquad \text{eq. (1)}$$

Decision Tree's performance was evaluated using similar statistical metrics. Pruning techniques and hyperparameter tuning, such as controlling the tree's maximum depth or minimum samples per leaf, were employed using statistical methods.

Table 3.4 Classification Report of Decision Tree

STATISTICAL → MEASURES /PREDICTION ↓	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.79	0.81	0.80	27
1	0.85	0.82	0.84	34
ACCURACY			0.82	61
MACRO AVG	0.82	0.82	0.82	61
WEIGHTED AVG	0.82	0.82	0.82	61

3.4.3 Support Vector Machine

Support Vector Machine (SVM) is a supervised learning algorithm used for both classification and regression tasks. It aims to find the hyperplane that best separates the classes in the feature space while maximizing the margin between them. SVM is effective in high-dimensional spaces and for datasets with complex relationships. It can handle linear and non-linear decision boundaries by using different kernel functions. SVM is robust to overfitting, thanks to the margin concept, and is less affected by outliers compared to other algorithms. However, SVM can be computationally expensive, especially for large datasets, and requires careful selection of hyperparameters for optimal performance.

Let the training samples having dataset Data = $\{y_i, x_i\}$; i = 1, 2, …, n where $x_i \in R^n$ represent the i^{th} vector and $y_i \in R^n$ represent the target item. The linear SVM finds the optimal hyperplane of the form $f(x) = w^T x + b$ where w is a dimensional coefficient vector and b is an offset. This is done by solving the subsequent optimization problem:

$$\text{Min}_{w,b,\xi i}\ (1/2)\ w^2 + C \sum_{i=1}^{m} \xi i \qquad \text{eq. (2)}$$

$$\text{s.t. } y_i\ (w^t x_i + b) >= 1 - \xi i,\ \xi i \geq 0, \qquad \forall i \in \{1, 2, \ldots, m\} \qquad \text{eq. (3)}$$

SVM's performance was evaluated using standard statistical metrics. Additionally, since SVM allows for different kernel functions, statistical tests or cross-validation were used to select the best-performing kernel.

Table 3.5 Classification Report of SVM

STATISTICAL MEASURES → /PREDICTION ↓	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.83	0.74	0.78	27
1	0.81	0.88	0.85	34
ACCURACY			0.82	61
MACRO AVG	0.82	0.81	0.81	61
WEIGHTED AVG	0.82	0.82	0.82	61

3.4.4 Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees during training and outputs the mode of the classes (classification) or the average prediction (regression) of the individual trees. It reduces overfitting and increases robustness by combining predictions from multiple models. Random Forest randomly selects subsets of features and data points for each tree, further improving generalization. It's computationally efficient and can handle high-dimensional data. Random Forest is less sensitive to noise and outliers compared to individual decision trees, making it suitable for noisy datasets.

This ensemble classifier builds several decision trees and incorporates them to get the best result. For tree learning, it mainly applies bootstrap aggregating or bagging. For a given data, X = {x1, x2, x3, ..., xn} with responses Y = {x1, x2, x3, ..., xn} which repeats the bagging from b = 1 to B.

The unseen samples x∈0 P is made by averaging the predictions B b=1 fb(x 0) from every individual trees on x 0 :

$$j = (1/B) \sum_{b=1}^{B} fb(x') \qquad \text{eq. (4)}$$

The uncertainty of prediction on this tree is made through its standard deviation.

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B} (fb(x')-f^{\wedge})^2}{(B-1)}} \qquad \text{eq. (5)}$$

Random Forest's performance was assessed using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. Hyperparameter tuning, particularly tuning the number of trees and maximum depth, was performed using statistical techniques like grid search or randomized search.

Table 3.6 Classification Report of Random Forest

STATISTICAL → MEASURES /PREDICTION ↓	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.92	0.85	0.88	27
1	0.89	0.94	0.91	34
ACCURACY			0.90	61
MACRO AVG	0.90	0.90	0.90	61
WEIGHTED AVG	0.90	0.90	0.90	61

3.4.5 Naïve Bayes

Naive Bayes is a simple yet powerful probabilistic classifier based on Bayes' theorem. It operates under the naive assumption that all features are conditionally independent given the class label. Despite this simplification, Naive Bayes often performs well, especially in text

classification tasks such as spam filtering, sentiment analysis, and document classification. It's computationally efficient, requires a small amount of training data to estimate parameters, and can handle both categorical and continuous data. Naive Bayes is widely used in real-world applications due to its simplicity, scalability, and ability to handle large datasets with high-dimensional feature spaces. This learning model applies Bayes rules through independent features. Every instance of data D is allotted to the class of highest subsequent probability. The model is trained through the Gaussian function with prior probability P (X_f) = priority ∈ (0:1)

$$P(X_{f1}, X_{f2}, \ldots\ldots, X_{fn}|c) = \prod_{i=1}^{n} P(X_{fi}|c) \quad \text{eq. (6)}$$

$$P(X_{fi}|c) = \frac{P(c_i|X_f)\, P(X_f)}{P(c_i)}, \quad c \in \{\text{benigh, malignant}\} \quad \text{eq. (7)}$$

Table 3.7 Classification Report of Naïve Bayes

STATISTICAL → MEASURES /PREDICTION ↓	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.88	0.78	0.82	27
1	0.84	0.91	0.87	34
ACCURACY			0.85	61
MACRO AVG	0.86	0.84	0.85	61
WEIGHTED AVG	0.85	0.85	0.85	61

3.4.6 K-Nearest Neighbor

K-Nearest Neighbors (k-NN) is a simple yet effective classification algorithm that classifies data points based on the majority class of their k nearest neighbors in the feature space. It's a non-parametric method that doesn't make assumptions about the underlying data distribution. K-NN is intuitive and easy to understand, making it suitable for beginners. However, its performance can degrade with high-dimensional data and when the dataset is imbalanced. The choice of the hyperparameter k is critical, as it affects the bias-variance trade-off.

Despite its limitations, k-NN is widely used in recommendation systems, anomaly detection, and pattern recognition tasks. It extract the knowledge based on the samples Euclidean distance function d (xi, xj) and the majority of k-nearest neighbors.

$$d(x_{i,\,xi,}) = \sqrt{(x_{i,1} - x_{j,1})^2 + \ldots\ldots + (x_{i,m} - x_{j,m})^2} \qquad \text{eq. (8)}$$

Table 3.8 Classification Report of KNN

STATISTICAL→ MEASURES ↓ /PREDICTION	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.62	0.67	0.64	27
1	0.72	0.68	0.70	34
ACCURACY			0.67	61
MACRO AVG	0.67	0.67	0.67	61
WEIGHTED AVG	0.68	0.67	0.67	61

3.4.6 Experimental Setup for Evaluation

In the first step, the UCI dataset is loaded and the data becomes ready for pre-processing. The subset of 13 attributes (Age, sex, cp, treetops, chol, FBS, restecg, thalach, exang, olpeak, slope, ca, that, target) is selected from the pre-processed data set of heart disease. The six existing models for heart disease prediction (DT, RF, LR, KNN, NB, SVM) are used to develop the classification. The evaluation of the model is performed with the confusion matrix. Totally, four outcomes are generated by confusion matrix, namely TP (True Positive), TN (True Negative), FP (False Positive) and FN (False Negative). The following measures are used for the calculation of the accuracy, sensitivity, specificity.

Accuracy = (TN+TP) / (TN+TP+FN+FP) =107+155/295 = 0.9018

Sensitivity = (TP/TP+FN) =155/155+12 = 92.8

Specificity = (TN/TN+FP) =105/105+22 = 82.6

Precision = TP /TP+FP = 155/155+22==87.5

F-Measure = 2TP/ 2TP+FP+FN = 310 /310+22+12 = 0.90

CHAPTER 4
IMPLEMENTATION AND TESTING

4.1 Subsystem and Their Dependencies

The architecture of the research project comprises several interconnected subsystems, each playing a crucial role in the overall workflow of heart disease prediction.

4.1.1 Data Pre-processing Subsystem

Responsible for cleaning and organizing the raw Cleveland dataset. Input from the Cleveland dataset, processes influenced by data mining techniques.

4.1.2 Feature Selection Subsystem

Utilizes decision tree entropy for iterative feature selection. Output from Data Pre-processing Subsystem, influences the attributes considered for heart disease prediction.

4.1.3 Classification Subsystem with ML Techniques

Implements various ML algorithms (e.g., KNN, DT, NB, RF, SVM, LR) for heart disease severity assessment. Input from Feature Selection Subsystem, influences the choice of ML techniques for accurate classification.

4.1.4 Random Forest using Hyperparameter Tuning

Introduces a hybrid method for heart disease prediction, avoiding feature selection restrictions. Output from Feature Selection Subsystem, influences the overall accuracy improvement in heart disease prediction.

- **Architecture:**

- Utilizes the ensemble learning technique of Random Forest, comprised of multiple decision trees.
- Each decision tree independently predicts the target variable (heart disease severity) based on a subset of features.
- The final prediction is made by aggregating the predictions of all decision trees in the forest.

- **Components:**

- Decision Trees: Fundamental building blocks of the Random Forest model, each making individual predictions.

- Ensemble Method: Aggregates predictions from multiple decision trees to improve overall accuracy and generalization.
- Hyperparameter Tuning: Fine-tunes model parameters such as tree depth and number of trees to optimize performance.

4.2 Test Cases

Here are the results of these test cases which include dataset integrity and pre-processing, model training and classification, comparison with existing models' sensitivity to feature changes, interpretability, generalization across various diverse populations, real-world clinical validation, handling of imbalance data when tested with different algorithms.

Table 4.1 Result of various models with proposed model

Models	Accuracy	Precision	Recall	F1-score
Logistic regression	0.85	0.85	0.85	0.85
Decision Tree	0.82	0.79	0.81	0.80
Naïve Bayes	0.85	0.85	0.85	0.85
K-Nearest Neighbor	0.67	0.62	0.67	0.64
Random Forest	0.90	0.92	0.85	0.88
Support Vector Machine	0.82	0.83	0.74	0.78

4.3 Comparative Analysis

The prediction models are developed using 13 features and the accuracy is calculated for modeling techniques. The best classification methods are given above in snapshots. These snapshots compare the accuracy, classification error, precision, F-measure, sensitivity and specificity. The highest accuracy is achieved by Random Forest classification method in comparison with existing methods with consideration of time.

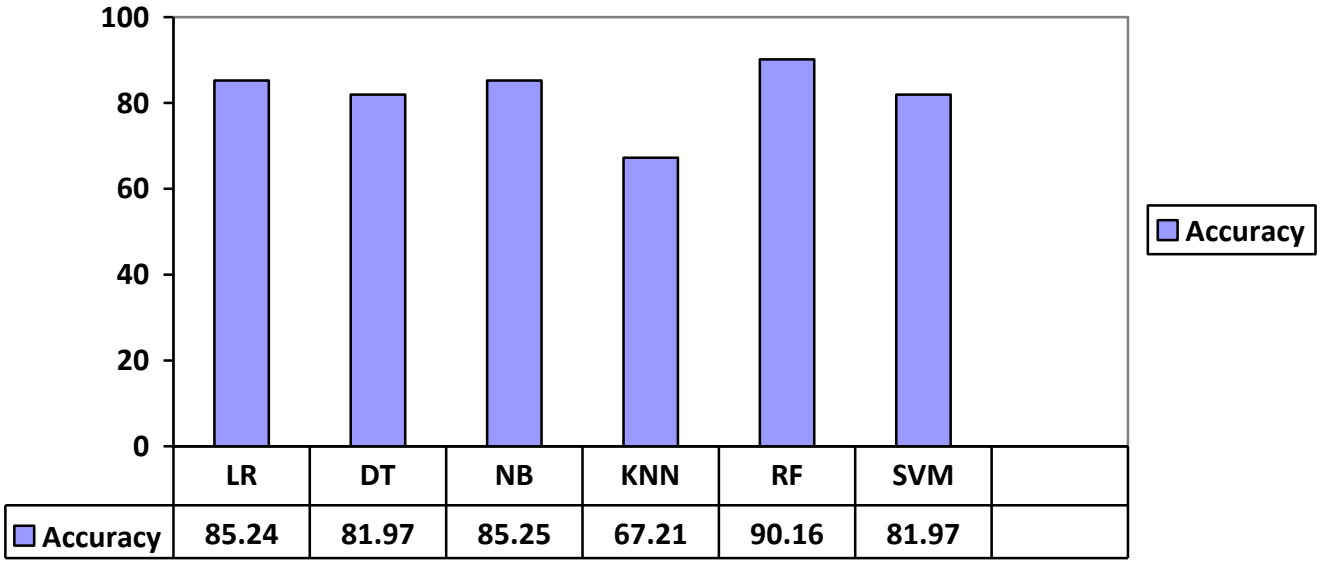

Fig 4.1 Accuracy of various ML models used

The UCI dataset is further classified into 8 types of datasets based on classification rules. The classification rules are listed in Table 4.2. Each dataset is further classified and processed by R Studio Rattle. The results are generated by applying the classification rule for the dataset. The classification rules generated based on the rule after data pre-processing is done. After pre-processing, the data's three best ML techniques are chosen and the results are generated. The various datasets with DT, RF, LR, SVM, KNN, NB are applied to find out the best classification method. The above snapshots shows that results of existing and proposed methods. The results show that RF is the best. Those snapshots also show the results of the proposed method.

CHAPTER 5
CONCLUSION

Identifying the processing of raw healthcare data of heart information will help in the long-term saving of human lives and early detection of abnormalities in heart conditions. Machine learning techniques were used in this work to process raw data and provide a new and novel discernment towards heart disease. Heart disease prediction is challenging and very important in the medical field. However, the mortality rate can be drastically controlled if the disease is detected at the early stages and preventative measures are adopted as soon as possible. Further extension of this study is highly desirable to direct the investigations to real-world datasets instead of just theoretical approaches and simulations. The research paper has presented a comprehensive framework for heart disease prediction using Random Forest with hyperparameter tuning and achieving an accuracy of 90.18%. Through the utilization of ensemble learning techniques and fine-tuning of model parameters, the proposed solution offers a robust and accurate predictive model. The architecture leverages the inherent strengths of Random Forest, allowing for the capture of complex relationships among diverse risk factors associated with heart disease. Hyperparameter tuning optimizes model performance, striking a balance between predictive accuracy and generalization to unseen data. Moreover, the interpretability of the model facilitates the identification of key contributing factors, aiding clinicians in understanding the underlying mechanisms of heart disease. Overall, this research contributes to the advancement of predictive analytics in healthcare, offering insights for early detection, personalized treatment strategies, and improved patient outcomes. Further research could explore additional optimization techniques and the integration of multimodal data sources to enhance the predictive capabilities of the proposed framework.

REFERENCES

[1] A. Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava. "Effective heart disease prediction using hybrid machine learning techniques." *IEEE access* 7 (2019): 81542-81554.

[2] R. Katarya and P. Srinivas, "Predicting Heart Disease at Early Stages using Machine Learning: A Survey," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 302-305, doi: 10.1109 / ICESC48915.2020.9155586.

[3] Jabbar M A, Deekshatulu B L & Chandra P (2013, March). Heart disease prediction using lazy associative classification. In 2013 International Mutli-Conference on Automation, Computing,Communication, Control and Compressed Sensing (iMac4s) (pp. 40- 6). IEEE.

[4] Kiyasu J Y (1982). U.S. Patent No. 4,338,396. Washington, DC: U.S. Patent and Trademark Office.

[5] Raihan M, Mondal S, More A, Sagor M O F, Sikder G, Majumder M A & Ghosh K (2016, December). Smartphone based ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design. In 2016 19th International Conference on Computer and Information Technology (ICCIT) (pp. 299-303). IEEE.

[6] M. Durairaj and V. Revathi, ‘‘Prediction of heart disease using back propagation MLP algorithm,’’ Int. J. Sci. Technol. Res., vol. 4, no. 8, pp. 235–239, 2015.

[7] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, ‘‘Prediction of heart disease using machine learning,’’ in Proc. 2nd Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA), Mar. 2018, pp. 1275–1278.

[8] P. K. Anooj, ‘‘Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules,’’ J. King Saud Univ.-Comput. Inf. Sci., vol. 24, no. 1, pp. 27–40, Jan. 2012. doi: 10.1016/j.jksuci.2011.09.002.

[9] V. Krishnaiah, G. Narsimha, and N. Subhash, ‘‘Heart disease pre□diction system using data mining techniques and intelligent fuzzy approach: A review,’’ Int. J. Comput. Appl., vol. 136, no. 2, pp. 43–51, 2016.

[10] A. H. Alkeshuosh, M. Z. Moghadam, I. Al Mansoori, and M. Abdar, ‘‘Using PSO algorithm for producing best rules in diagnosis of heart disease,’’ in Proc. Int. Conf. Comput. Appl. (ICCA), Sep. 2017, pp. 306– 311.

[11] H. A. Esfahani and M. Ghazanfari, ‘‘Cardiovascular disease detection using a new ensemble classifier,’’ in Proc. IEEE 4th Int. Conf. Knowl.- Based Eng. Innov. (KBEI), Dec. 2017, pp. 1011–1014.

[12] M. Gandhi and S. N. Singh, ‘‘Predictions in heart disease using techniques of data mining,’’ in Proc. Int. Conf. Futuristic Trends Comput. Anal. Knowl. Manage. (ABLAZE), Feb. 2015, pp. 520–525.

[13] J. Wu, S. Luo, S. Wang, and H. Wang, ‘‘NLES: A novel lifetime extension scheme for safety-critical cyber-physical systems using SDN and NFV,’’ IEEE Internet Things J., no. 6, no. 2, pp. 2463–2475, Apr. 2019.

[14] J. Wu, M. Dong, K. Ota, J. Li, and Z. Guan, ‘‘Big data analysis-based secure cluster management for optimized control plane in software-defined networks, IEEE Trans. Netw. Service Manag., vol. 15, no. 1, pp. 27–38, Mar. 2018.

[15] J. Wu, M. Dong, K. Ota, J. Li, and Z. Guan, ‘‘FCSS: Fog computing-based content-aware filtering for security services in information centric social networks,’’ IEEE Trans. Emerg. Topics Comput., to be published. doi: 10.1109/TETC.2017.2747158.

[16] G. Li, J. Wu, J. Li, K. Wang, and T. Ye, ‘‘Service popularity-based smart resources partitioning for fog computing-enabled industrial Internet of things,’’ IEEE Trans. Ind. Informat., vol. 14, no. 10, pp. 4702–4711, Oct. 2018.