Employing Machine Learning Techniques for Predicting Heart Attacks

Ahmed Mohamed Yousef Gad Al-Mawla, Mohamed Ezz El-Dean Ahmed Abd El-Samea

Abstract: ***Cardiovascular disease (CVD) encompasses various heart conditions and ranks among the top causes of global mortality, accounting for about 31% of all deaths. These conditions result from complex interactions among modifiable and non-modifiable risk factors, with many cases being preventable through lifestyle changes. Machine learning (ML) techniques have revolutionized predictive modeling in healthcare, aiding in the early detection of heart-related issues. Using extensive patient datasets, researchers employ diverse ML algorithms to predict heart attacks, aiming to support healthcare professionals. This study focuses on supervised ML classifiers, including Gradient Boosting, Decision Tree, Random Forest, Support Vector Machine, K Nearest Neighbor, and Logistic Regression, to predict Myocardial Infarction. Utilizing datasets from sources like the Framingham database and UCI Heart repository, the research aims to forecast heart attack probabilities. Despite using classifiers without optimizations, Decision Tree and Random Forest achieved the highest accuracy score of 99.7% in the UCI dataset. Key predictive attributes include chest pain type, cholesterol levels, heart rate, Thal, and age. Adherence to a healthy lifestyle, including diet, exercise, and hydration, can prevent around 80% of premature heart attacks. Monitoring blood pressure, cholesterol, and heart rate, along with meditation, further reduces major heart attack risks. Resampling imbalanced data and feature scaling with min-max Scaler library are additional considerations for model refinement.***

Manuscript received on 10 January 2023 | Revised Manuscript received on 22 January 2023 | Manuscript Accepted on 15 February 2023 | Manuscript published on 28 February 2023.

***Correspondence Author(s)**

Ahmed Mohamed Yousef, Student, Department of Artificial Intelligence, October 6 University, 6th of October City, Egypt.

Mohamed Ezz El Dean Ahmed, Student, Department of Artificial Intelligence, October 6 University, 6th of October City, Egypt

Keywords*: *Cardiovascular Disease, Framingham Model, Gradient Boosting, Machine Learning, Mayo-cardinal Infarction, Support Vector Machine, K Nearest Neighbor, UCI Model.

I. INTRODUCTION

The heart plays a crucial role in the human body by pumping blood throughout the circulatory system. This system is responsible for distributing essential substances like nutrients, oxygen, water, and minerals to various bodily tissues and organs. Any disruption in the heart's functioning can result in serious health complications, potentially leading to death. Cardiovascular diseases encompass a range of conditions affecting either the heart or the blood vessels, with coronary artery disease being the most prevalent form. Many instances of cardiovascular diseases stem from factors that can be altered through lifestyle modifications, making them preventable. However, there are also cases where risk factors are non-modifiable, presenting challenges for improvement.

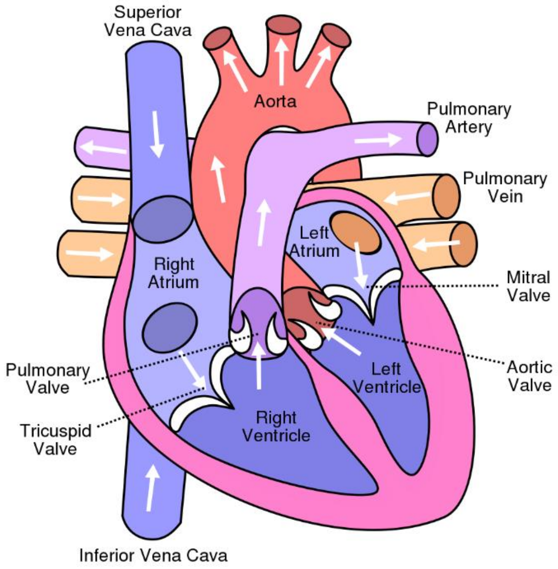

Figure 1: Human Heart Image

Historically, endeavors to decrease the occurrence of cardiovascular diseases have primarily centered on altering individual behaviors [2]. Factors such as physical inactivity, tobacco consumption, inadequate dietary habits, and obesity play pivotal roles as contributors to well-known risk factors associated with CVDs, which encompass conditions like diabetes, hypertension, and the onset of atherosclerosis. Regrettably, CVDs account for approximately 17.9 million fatalities annually, predominantly attributable to heart attacks and strokes. This translates to 31% of total global mortality [3]. CVDs constitute a cluster of ailments affecting both the heart and the blood vessels.

Retrieval Number: XXXX/2023©BEIESP
DOI: 10.35940/ijeat
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

A. Coronary artery disease

The formation of fat and cholesterol deposits, known as plaques, can lead to blockages in coronary arteries. When these plaques become damaged, they release platelets, initiating blood clotting, which may accumulate around the plaque and hinder blood flow. Consequently, this restriction in blood flow can result in damage to the heart muscle. Complete blockage of blood flow to a particular segment of the heart can exacerbate the extent of damage. Coronary artery disease arises from the accumulation of plaques within the arteries, leading to their narrowing or blockage. Atherosclerosis serves as the primary cause of this condition. The most severe complication associated with coronary artery disease is myocardial infarction, commonly referred to as a heart attack. While prevention is possible in the early stages of this disease, untreated cases can lead to serious health complications or even death. Environmental factors significantly influence the development and severity of cardiovascular diseases, albeit individuals often possess limited control over them. The cardiovascular system is susceptible to various environmental agents, including solvents, tobacco, pesticides, smoke, and other pollutants inhaled or ingested, as well as extremes in noise and temperature. Exposure to environmental pollutants occurs through inhalation, ingestion, and absorption. Air pollution has been identified as the primary environmental health risk by the World Health Organization. In 2012, outdoor air pollution was attributed to over 3.5 million deaths in individuals over 60 years old, with cardiovascular diseases accounting for 80% of these fatalities.

B. Heart Attack

Consequences stemming from various cardiovascular diseases can manifest in different types of heart attacks. One such type is STEMI, characterized by atherosclerosis-induced obstruction of blood flow to a significant portion of the heart, resulting in continuous damage to the cardiac muscle. This condition can escalate to complete heart failure and potentially lead to fatality, underscoring its critical nature that necessitates immediate attention. Another form is NSTEMI, arising from partial blockage of coronary arteries, causing substantial restriction in blood flow to a specific region of the heart. While less perilous than STEMI, NSTEMI still has the potential to inflict permanent damage on the affected heart area. Lastly, there exists coronary artery spasm, referred to as a silent heart attack. This variant occurs due to the contraction of heart-connected arteries, impeding blood flow to vital heart segments. Although less severe compared to other types of heart attacks, coronary artery spasm does not induce permanent heart damage.

- **Symptoms**

A heart attack presents a critical medical condition characterized by sensations of tightness or pain in the chest, neck, back, and arms, alongside symptoms like fatigue, dizziness, irregular heartbeat, and anxiety. Risk factors associated with heart attacks encompass both unmodifiable and modifiable elements. Unmodifiable factors include age, sex, and family medical history, while modifiable factors encompass smoking, elevated cholesterol levels, hypertension, obesity, inadequate dietary habits, sedentary lifestyle, and heightened stress levels. Treatment modalities commonly employed for heart attacks include medication, electrocardiography (ECG), and bypass surgery.

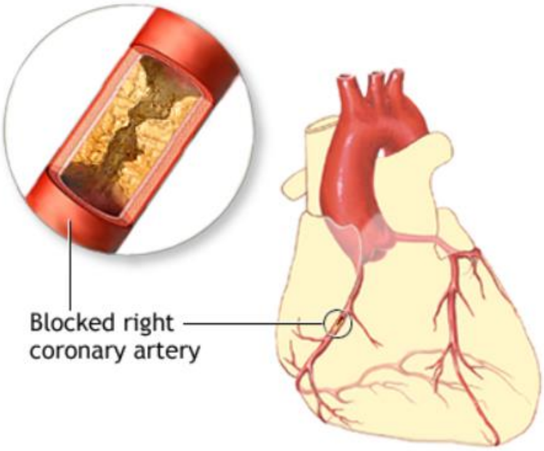

Figure 2: Blockage in Coronary Artery

- **Factors causing Heart Attacks**

The factors in our diet that can either assist or protect against the development of coronary heart disease resulting from atherosclerosis are still being studied. While the primary cause of atherosclerosis is unknown, one theory suggests that free radical damage to cholesterol in circulating low-density lipoproteins could be the primary event. There are seven dietary factors with two promoters and five protective features that may contribute to the development of coronary heart disease. Myocardial infarction, the most serious complication of coronary heart disease, is a combination of two effects of dietary factors. While some observational evidence suggests that lowering serum homocysteine containing folic acid could reduce the risk of cardiovascular diseases, drug treatment for preventing cardiovascular disease events and heart strokes has been limited to single risk factors that target a small percentage of the population with values in the upper range of the risk factor distribution. The goal is to reduce the risk factors to average population values.

C. Artificial Intelligence

Artificial intelligence is the capability of a computer program to emulate the human brain, thereby developing intelligence. This gives computers the ability to think and makes them more intelligent. Alan Mathison Turing first proposed the concept of artificial intelligence over sixty years ago. According to him, a machine is considered "intelligent" if it can provide answers, solutions or responses that cannot be distinguished from those given by a human.

In the last two decades, we've seen remarkable progress in the fields of artificial intelligence and robotics, with the promise of even more groundbreaking developments on the horizon. The term "artificial intelligence" was initially coined in 1956. At its core, AI revolves around the concept of machines endeavoring to replicate and perform tasks ranging from the straightforward to the exceedingly intricate, leveraging principles of human intelligence.

- **Machine Learning**

Machine Learning represents a subset of Artificial Intelligence, focusing on training computer programs to learn from data or input. It operates on two foundational principles: firstly, devising

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

methods by which a computer system can enhance its outcomes by learning from past experiences; and secondly, understanding the core statistical, computational, and information-theoretic principles that dictate learning systems, irrespective of whether they are human, computer-based, or organizational. The primary objective of machine learning is to develop intelligent systems capable of addressing specific challenges without explicit programming for each task. To be able to achieve this, programs undergo learning processes by analyzing extensive datasets to extract knowledge that facilitates predictive capabilities.

D. Classification

Classification serves as a fundamental function within Machine Learning, employing ML algorithms to assign class labels to cases within problem sets. It facilitates the prediction of labeled classes for given datasets. For instance, it can discern between male and female individuals, identify spam emails, or classify handwritten characters as known or unknown. Supervised learning algorithms form the backbone of machine learning, trained on labeled data. These algorithms exhibit high robustness when appropriately utilized and are primarily employed for predictive tasks. The overarching objective is to forecast or categorize the target outcome of interest, such as the presence or absence of heart problems.

▪ Decision Tree

A decision tree stands as a machine learning algorithm utilized to address classification and regression tasks. It operates as a non-parametric supervised algorithm, employing a decision support tool depicted in the form of a flowchart-like graph structure to aid in decision-making. Within the decision tree framework, each internal node represents a test conducted on a specific feature, while the branches delineate the classification rules leading to the outcome or class label, as defined by each leaf node. The decision tree functions by segmenting the dataset into subsets predicated on value tests conducted on attributes. This process iteratively recurses on each extracted subset until either the subset at a node possesses the same value as the target variable or contributes no additional predictive value upon further splitting.

▪ Logistic Regression

Logistic Regression is a type of supervised machine-learning algorithm used to address classification problems. It functions as a linear model and predicts the outcome of a categorical dependent variable based on a set of predictors or independent variables. Despite having the term "regression" in its name, Logistic Regression is primarily used for binary and linear classification problems. It is suitable for cases where the response variable consists of binary outcomes along with continuous explanatory variables. This model can also handle multi-class classification problems. Logistic Regression models the probability of classification problems with two possible outcomes for the target variable. It employs logistic equations to derive results within the range of 0 to 1. The logistic function, also known as the sigmoid function, is defined as:

$$\sigma(x) = 1 / (1 + e^{(-x)}).$$

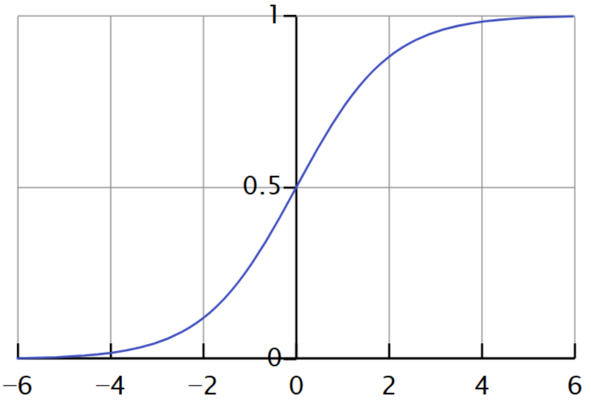

Figure 3: Classification of target value (y) w.r.t feature (x)

▪ Gradient Boosting Classifier

The Gradient Boosting Classifier stands as a form of Ensemble learning technique, which entails amalgamating multiple weak learning algorithms characterized by lower accuracy into a robust model capable of predicting with enhanced accuracy. Various types of gradient-boosting classifier algorithms have emerged, each presenting distinct proposals aimed at refining the model's accuracy. Within Gradient Boosting, each predictor endeavors to rectify the errors made by its predecessors' algorithms. This process entails fitting a new predictor to address the inaccuracies of prior algorithms while preserving overall accuracy. The methodology hinges on leveraging the errors generated by preceding algorithms to guide subsequent algorithms toward heightened accuracy. In essence, Gradient Boosting serves as a potent approach for constructing predictive models capable of achieving notable accuracy.

▪ Random Forest

The Random Forest algorithm is a type of supervised learning that utilizes multiple decision trees generated during training. It falls under the category of Ensemble learning techniques and can be used to effectively solve both classification and regression problems. The model's prediction is based on the class prediction of the tree that receives the highest number of "votes" or has the highest mode value among the classes. This approach involves creating several decision trees independently and then combining them to improve accuracy and prediction quality. One of the significant benefits of Random Forest is its ability to balance high variance and high bias by averaging their effects.

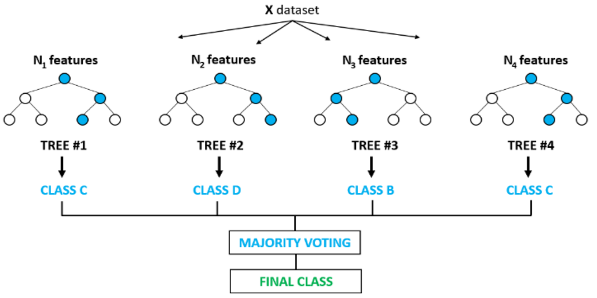

Figure 4: Basic Random Forest Model

- **Support Vector Machine**

Support Vector Machine (SVM) is a supervised learning algorithm commonly utilized for classification and regression tasks. As a member of Ensemble learning techniques alongside Random Forest and Gradient Boosting, SVM identifies an optimally separating hyperplane in an n-dimensional space. This hyperplane is defined by support vectors, the data points closest to the decision boundary. SVM maximizes the margin between support vectors of different classes, enhancing generalization. Known for its proficiency in both linear and non-linear classification, SVM employs kernel functions to transform input data into higher-dimensional space. By balancing classification error minimization with margin maximization, SVM exhibits robustness and efficacy across diverse applications, making it a vital tool in machine learning and data analysis.

- **K Nearest Neighbor**

K Nearest Neighbor (KNN) is a supervised learning algorithm that is capable of handling classification and regression tasks, like other Ensemble techniques and Support Vector Machines. It makes use of a proximity-based approach, where data points are classified based on the majority class of their nearest neighbors. The parameter "K" represents the number of neighbors considered. This method is effective in both linear and non-linear classification challenges, providing flexibility in complex decision boundary scenarios. Although KNN is simple and easy to interpret, its computational complexity increases linearly with the size of the dataset, which might affect scalability. Nevertheless, KNN remains a valuable tool in machine learning and data analysis, providing a straightforward yet effective solution to classification and regression tasks.

E. Problem

Cardiovascular diseases are responsible for 31% of all global deaths, with heart attacks accounting for approximately 85% of these fatalities, according to the World Health Organization (WHO). Surveys conducted in the United States reveal that around 48% of the population has some form of heart disease, and every 19 deaths in the US are caused by heart problems. In India, data from the National Crime Records Bureau (NCRB) show a 53% increase in deaths due to heart attacks over the past five years, with approximately 33% of all deaths in the country attributed to heart attacks and related issues. Despite the possibility of medical intervention saving patients' lives following minor or major attacks, many individuals succumb due to the lack of timely medical treatment. This can be attributed to the fact that patients may not seek help during an attack, and their families may be unfamiliar with emergency protocols. Consequently, seeking medical assistance post-attack may not always be a foolproof solution for preventing heart attacks.

F. Approaches

AI and ML have the potential to mitigate the risk of death by forecasting the probability of heart attacks and leveraging patient health data and medical records. While numerous ML models exist for heart disease prediction, specialized models for heart attacks are limited. Our project addressed this gap by deploying a range of ML algorithms on two distinct datasets, aiming to enhance accuracy in preemptively identifying heart attacks. By leveraging diverse algorithms and datasets, from simplistic to intricate, we aimed to achieve superior predictive performance. This proactive approach can preempt critical conditions, offering vital opportunities for early intervention and potentially saving lives.

G. Objective

For both datasets, the risk factors provided are linked to the likelihood of heart attacks in patients, aiding in determining their risk of cardiovascular problems in the future. Based on the datasets provided, our objectives include:

- Predicting the probability of a patient experiencing a heart attack in the future.
- Identifying the primary factors influencing heart attacks.
- Identifying cholesterol levels is associated with a higher risk of heart attacks.
- Identifying chest pain types with a higher likelihood of heart attacks.
- Offering recommendations for preventing or reducing the risk of heart attacks.

Through comprehensive analysis of the datasets, we aim to develop insights that inform proactive measures for mitigating the risk of heart attacks and promoting cardiovascular health.

H. Next Sections

II. Related Works
III. Datasets And Pre-Processing
IV. Methodology
V. Experimental Setup
VI. Result And Analysis
VII. Conclusion
References

I. Contribution

We successfully managed two datasets, the UCI dataset, and the Framingham dataset. For the UCI dataset, We utilized the min max scaler library to perform feature scaling and checked for imbalanced data. We also converted categorical features to numerical features and created a variety of visualizations. In the Framingham dataset, we thoroughly addressed null values and scaled the features. Additionally, we introduced two classifiers, support vector machine, and k nearest neighbor, which achieved impressive accuracy rates. Although

We did not use optimization in my approach, there are models.

within the Framingham dataset that yield high accuracy rates, more than Suraj Kumar Gupta, Aditya Shrivastava, Satya Prakash Upadhyayand Pawan Kumar Chaurasia, "A Machine Learning Approach for Heart Attack research paper. However, in the UCI dataset, certain models such as random forests and decision trees have great result, we recommend using my code because it got higher accuracies better than Suraj Kumar Gupta, Aditya Shrivastava, Satya Prakash Upadhyayand Pawan Kumar Chaurasia, "A Machine Learning Approach for Heart Attack research paper.

J. Compare Results

We compare by accuracy, precision, recall, f1 micro, and f1 macro.

A. Framingham Dataset

- Suraj Kumar Gupta, Aditya Shrivastava, Satya Prakash Upadhyayand Pawan Kumar Chaurasia Paper

	Logistic Regression	Decision Tree	Gradient Boosting	Random Forest
Accuracy	0.689	0.6774	0.7097	0.7419
Precision	0.2727	0.7059	0.6818	0.7143
Recall	0.6186	0.7059	0.8824	0.8824
F1_micro	0.3806	0.7059	0.7737	0.7913
F1_macro	0.6471	0.7036	0.6969	0.7334

Table 1 Results of Original Work of Framingham Dataset

- **Proposed Work**

	Logistic Regression	Decision Tree	Gradient Boosting	Random Forest	K Nearest Neighbor	Support Vector Machine
Accuracy	0.853774	0.756132	0.850472	0.850708	0.835613	0.847170
Precision	0.730684	0.237028	0.543959	0.563462	0.364162	0.246667
Recall	0.063675	0.267139	0.085380	0.058963	0.113324	0.007752
F1_micro	0.853774	0.758019	0.849764	0.845755	0.835613	0.847170
F1_macro	0.518527	0.551712	0.532291	0.511493	0.540262	0.466086

Table 2 Results of Original Work of Framingham Dataset

B. UCI Dataset

- **Suraj Kumar Gupta, Aditya Shrivastava, Satya Prakash Upadhyayand Pawan Kumar Chaurasia, Paper**

	Logistic Regression	Decision Tree	Gradient Boosting	Random Forest
Accuracy	0.8065	0.7722	0.8465	0.841
Precision	0.7619	0.1923	0.5	0.2
Recall	0.9412	0.1546	0.0309	0.0103
F1_micro	0.8451	0.1681	0.031	0.0203
F1_macro	0.7983	0.7521	0.7203	0.1659

Table 3 Results of Original Work of UCI Dataset

- **Proposed Work**

	Logistic Regression	Decision Tree	Gradient Boosting	Random Forest	K Nearest Neighbor	Support Vector Machine
Accuracy	0.844878	0.997073	0.963902	0.997073	0.840976	0.895610
Precision	0.820902	1.000000	0.966067	0.994495	0.861634	0.879405
Recall	0.893621	0.986667	0.963899	1.000000	0.825085	0.924061
F1_micro	0.844878	0.996098	0.963902	0.994146	0.840976	0.895610
F1_macro	0.843966	1.000000	0.963873	0.997066	0.840850	0.895280

Table 4 Results of Proposed Work of UCI Dataset

II. Related Works

In the realm of healthcare industries, machine learning (ML) holds promise for advancing predictive models, particularly in assessing cardiac risk. To explore this potential, various algorithms were evaluated to gauge their predictive efficacy compared to established models like the Framingham Score, a pivotal tool in clinical risk assessment for heart attacks. Despite the existence of several cardiovascular (CV) risk prediction algorithms, their performance remains a subject of concern. Notably, traditional scores such as the Framingham Score and Systematic Coronary Risk Evaluation often underestimate cardiac risk in patients. Recent studies have conducted comparative analyses of ML algorithms, including k-nearest neighbor, support vector machine, gradient boosting, logistic regression tree, and random forest, for CV risk prediction. Among these, Naive Bayes, SVM, and KNN emerged as promising classifiers for heart attack prediction. Additionally, research endeavors have explored novel approaches such as random partitioning of datasets and employing old data mining techniques like J48, REPTREE, Naïve Bayes, Bayes Net, and CART to predict cardiac infarction risk. In a study conducted in February 2021, an algorithm named under sampling-clustering-oversampling (UCO) algorithm was developed, leveraging random under-sampling, clustering, and oversampling techniques to balance training data for ML algorithms. This algorithm demonstrated remarkable predictive performance, achieving an accuracy of 70.29%, precision of 70.05%, 1-recall of 75.59%, and 0-recall of 63.95% using random forest. Other research efforts focused on early prediction of heart attacks, utilizing decision trees and random forest classifiers to analyze datasets containing chest pain and 24 other attributes. Similarly, another technique involved partitioning datasets randomly and creating homogeneous ensembles using various classification and regression tree models, leading to accurate risk assessment for heart disease.

Two datasets, Cleveland, and Framingham achieved classification accuracies of 93% and 91%, respectively. In a study conducted in July 2020, researchers aimed to develop a model predicting the risk of cardiovascular disease using two techniques. The Support Vector Machine (SVM) was meticulously trained and tuned for its parameters. After 1000 training iterations, the SVM model achieved an average accuracy of 96.5% with an average recall rate of 89.8%. Similarly, the recall rate using K-nearest neighbors reached 92.9%. With the burgeoning big data in biomedical and healthcare industries, substantial datasets offer opportunities for earlier detection of cardiovascular diseases or strokes. A proposed algorithm utilized a latent factor model to reconstruct missing data. This algorithm, known as a Convolutional Neural Network (CNN), based on a multimodal disease risk prediction approach, achieved a prediction accuracy of 94.8%. Concerns regarding data privacy in models predicting cardiac arrest prompted the incorporation of techniques such as masking encryption, dynamic data encryption, and granular access control to safeguard patient data. Acute myocardial infarction, or heart attack, is the deadliest cardiovascular-related disease. Big data analytics in healthcare offers potential for disease prevention, prediction, and treatment. National and international databases of heart patients were meticulously examined to identify various studies on big data analytics in healthcare, myocardial infarction prevention, and prediction. While various approaches have been explored for myocardial infarction prediction, researchers continue to seek optimal solutions. One approach involves using machine learning algorithms along with feature selection algorithms for myocardial infarction prediction. Various machine learning approaches with optimum parameters and different feature selection techniques were deployed, with the SVM algorithm using a linear kernel achieving the highest accuracy of 84.21%. Another approach utilized the Fuzzy C Means Classifier, an unsupervised machine learning algorithm, for predicting heart attacks using patient medical reports, yielding an accuracy of 92%. A hybrid intelligence system, Neuro-fuzzy, combining fuzzy logic and neural networks, achieved accuracy rates exceeding 90%, offering the potential for patient-centric applications.

III. Datasets And Pre-Processing

In this study, two distinct datasets were utilized. The first dataset is the Framingham dataset, which was developed by Boston University, the National Institute of Heart (NIH), and the National Heart Lung and Blood Institute (NHLBI) to identify factors associated with cardiovascular diseases, particularly heart attacks and strokes. The second dataset used is the Heart dataset, sourced from the UCI Machine Learning Repository at the University of California, Irvine, School of Information and Computer Sciences. It comprises data from four databases, including those from institutions such as "The Hungarian Institute of Cardiology, Budapest," "University Hospital, Zurich, Switzerland," "University Hospital, Basel, Switzerland," and "V.A. Medical Center, Long Beach, and Cleveland Clinic Foundation." This dataset is available on platforms such as Kaggle.

A. Framingham Dataset

There are four types of attributes in the Framingham dataset: Demographics, behavior, Previous medical history, and Current medical condition.

Demographic Attributes:

- Sex: Categorized as either 0 or 1, with 0 representing female and 1 representing male.
- Age: The age of the patient at the time of examination.
- Education: Considered inessential, as medical issues do not occur based on someone's education level.

Behavioral:

- Current Smoker: Categorized as either 0 or 1, indicating whether a patient currently smokes or not, where 1 represents "yes" and 0 represents "no".
- Cigs Per Day: Indicates the average number of cigarettes smoked per day by individuals who smoke regularly.
- Previous Medical History-Based Information:
- Diabetes: Categorized as either 0 or 1, indicating whether a patient has diabetes (1) or not (0).
- BP Meds: Categorized as either 0 or 1, representing whether a patient is on medication for blood pressure (1) or not (0).
- Prevalent Stroke: Indicates whether the patient had a stroke previously, with 1 representing "Yes" and 0 representing "No".
- Prevalent Hyp: Indicates whether the patient was hypertensive (having abnormally high blood pressure), with 1 representing "Yes" and 0 representing "No".

Current Medical Condition-Based Information:

- Heart Rate: Heart rate of the patient.
- Tot Chol: Total cholesterol level.
- Sys BP: Systolic blood pressure.
- Dia BP: Diastolic blood pressure.
- BMI: Body Mass Index.
- Glucose: Glucose level.

Target Variable to Predict:

- 10-year risk of CHD (coronary heart disease): Binary variable where 1 represents "Yes" and 0 represents "No".

B. UCI Dataset

The dataset comprises a total of 13 decision parameters, with the target value represented by "target."

- Age (age): Age of the patient at the current time of examination.
- Sex (sex): Categorized as either 0 or 1, with 0 indicating female and 1 indicating male.
- Chest Pain (cp): Categorized into four types from 0 to 3, representing: 0 as typical angina, 1 as atypical angina, 2 as non-anginal pain, and 3 as asymptomatic.
- Resting Blood Pressure (treetops): Resting blood pressure value of the patient in mmHg (unit).

- Cholesterol (Chol): Cholesterol level of the patient in mg/dl (unit).
- Fasting Blood Sugar (fbs): Categorized as either 0 or 1, where 1 represents fbs >120 mg/dl (true) and 0 represents otherwise (false).
- Resting ECG (restecg): Categorized into three types from 0 to 2, representing: 0 as normal, 1 as having ST-T wave abnormality, and 2 as left ventricular hypertrophy.
- Max Heart Rate (thalach): Maximum heart rate achieved by any patient.
- Exercise-induced angina (exang): Categorized as either 0 or 1, where 0 represents No and 1 represents Yes.
- Oldpeak: Indicates the value of ST depression induced by exercise concerning rest (float values).
- Slope: Describes the peak of exercise during the ST segment, classified into three ranges: 0 for up-slope, 1 for flat, and 2 for down-slope.
- No. of major vessels (ca): Classified in the range 0 to 4 by coloring through fluoroscopy.
- Thalassemia (thal): Classified into three ranges from 0 to 2, where 0 represents normal, 1 represents fixed defect, and 2 represents reversible defect.
- Target: Prediction column for diagnosing heart attacks, categorized into two types: 0 indicating no possibility of a heart attack and 1 indicating possibilities of a heart attack.

C. Preprocessing

Data Pre-Processing is defined as transforming or encoding the data in such a state so that it can be easily parsed by the machines for generating accurate information. In other words, it should be transformed in such a form so that it can be easily interpreted by different algorithms producing higher accurate results. It is not necessary to have complete pure data in every dataset. There is always some missing data in every dataset in “NULL” form due to which the dataset becomes redundant and hence leads the models to predict results with poor accuracies. Hence, to overcome these poor accuracies and to attain higher and better accuracies, data pre-processing came in genre. We usually clean the tuples having missing values by either dropping those rows from the dataset or by imputing mean or median or mode values of the respective column. Framingham dataset all data are numeric but have Null value in the other hand UCI Dataset has categorical features so firstly we convert these features into numerical features and then we use in both data sets Framingham and UCI we used Feature Scaling minmax scaler to make the min equal to zero and max equal to one so Hence, in our proposed model we are using mean and median imputation approaches for imputing missing values in the data set to attain its consistency to achieve higher accuracy. Mean imputation is the way of replacing missing values (i.e., ‘NA’ or ‘NULL’) data in the dataset using that parameter. Median imputation is the way of replacing missing values (i.e., ‘NA’ or ‘NULL’) data in the dataset by a median of that parameter. Even in this mean and median imputation, there is always confusion about when we should use mean imputation and when we should use median imputation. It can be described as whenever the parameter represents a normal distribution then we can use any one of both mean and median imputation. But if the parameter represents a skewed distribution instead of a normal distribution, then the median imputation is preferred over the mean imputation.

IV. METHODOLOGY

In our proposed model, we utilize two prominent datasets in the field of cardiovascular analysis: the Framingham dataset and the UCI Heart dataset. We employ six machine learning classifiers: Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Support Vector Machine, K Nearest Neighbor, and Gradient Boosting Classifier. After identifying and addressing missing values, we preprocess the data by imputing these missing values. Additionally, we convert categorical data into numerical format and perform feature scaling. Subsequently, the dataset is divided into two subsets: training data and holdout data for testing purposes. The dataset is split in a 9:1 ratio, with 90% allocated for training the model and the remaining 10% reserved for testing the model.

A. Pseudo Code

```
1. Start
2. Importing necessary libraries
3. Reading the dataset
4. Displaying the first few rows of the dataset
5. Information about the dataset
6. Checking for missing values
7. Plotting the distribution of the target variable
8. Plotting the age distribution
9. Plotting age-wise distribution of heart disease
10. Plotting resting blood pressure level in correlation
    with heart disease positivity ratio
11. Plotting cholesterol level in correlation with heart
    disease positivity ratio
12. Plotting chest pain type distribution
13. Replacing categorical data with numerical values
14. Plotting the correlation matrix
15. Feature scaling
16. Displaying summary statistics
17. Checking the class distribution of target variable
18. Plotting the distribution of target variable after
    balancing
19. Separating features and target variable
20. Displaying the shape of feature and target
    variable
21. Importing necessary machine learning models
22. Creating a dictionary of models
23. Function to fit and score models.
24. Calling the function
25. Function to calculate cross-validation scores.
26. Calling the function
27. End
```

B. Algorithm

Algorithm for Data Exploration and Preprocessing:

1. Import necessary libraries:
   - pandas for data manipulation and analysis.
   - numpy for numerical computations.
   - matplotlib.pyplot and seaborn for data visualization.
   - warnings to ignore any warnings.
2. Read the dataset from the specified path using pd.read_csv() function.
3. Display the first few rows of the dataset using df.head().
4. Obtain information about the dataset using df.info().
5. Check for missing values in the dataset using df.isnull().sum().
6. Plot the distribution of the target variable using a bar plot.
7. Plot the age distribution:
   a. Group age into bins using pd.cut() function.
   b. Plot a bar chart and line plot to visualize the distribution of age.
8. Plot the age-wise distribution of heart disease:
   a. Group age into bins using pd.qcut() function.
   b. Calculate the sum of heart disease cases for each age group.
   c. Plot a bar chart to visualize the age-wise distribution of heart disease cases.
9. Plot the resting blood pressure level in correlation with heart disease positivity ratio:
   a. Group resting blood pressure into bins using pd.qcut() function.
   b. Calculate the mean of heart disease cases for each resting blood pressure level.
   c. Plot a bar chart to visualize the correlation.
10. Plot the cholesterol level in correlation with heart disease positivity ratio:
   a. Group cholesterol level into bins using pd.qcut() function.
   b. Calculate the sum of heart disease cases for each cholesterol level.
   c. Plot a bar chart to visualize the correlation.
11. Plot the distribution of chest pain types using a bar plot.
12. Replace categorical data with numerical values:
   a. Replace "sex" values with 1 for Male and 0 for Female.
   b. Replace "chest_pain_type" values with 0, 1, 2, and 3.
   c. Replace "fasting_blood_sugar" values with 1 for Greater than 120 mg/ml and 0 for Lower than 120 mg/ml.
   d. Replace "rest_ecg" values with 0, 1, and 2.
   e. Replace "exercise_induced_angina" values with 1 for Yes and 0 for No.
   f. Replace "slope" values with 0, 1, and 2.
   g. Replace "vessels_colored_by_flourosopy" values with 0, 1, 2, 3, and 4.
   h. Replace "thalassemia" values with 3, 6, 7, and 0.
13. Visualize the correlation matrix using a heatmap.
14. Perform feature scaling using MinMaxScaler.
15. Plot the histogram of the target variable to check for class imbalance.
16. Split the dataset into feature data (X) and target data (y).
17. Initialize machine learning models: Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, K Neighbors Classifier, and Support Vector Machine.

18. Implement model fitting and scoring function:
   a. Split the data into training and testing sets.
   b. Fit each model on the training data and calculate its accuracy score on the testing data.
   c. Plot the scores of each model using a bar plot.

19. Implement model cross-validation scoring function:
   a. Calculate cross-validation scores for each model using various performance metrics.
   b. Return a dataframe containing the mean cross-validation scores for each model.

20. Call the model_fit_and_score() function with the initialized model dictionary and dataset.
21. Call the model_cross_val_score() function with the initialized model dictionary and dataset.

V. Experimental Setup

for experimental setup various packages installed are: Scikit learn and pandas and numpy, etc over Jupyter lab and we called all these libraries from these packages:

1. pandas
2. NumPy
3. matplotlib.pyplot
4. seaborn
5. warnings
6. sklearn.preprocessing.MinMaxScaler
7. sklearn.linear_model.LogisticRegression
8. sklearn.tree.DecisionTreeClassifier
9. sklearn.ensemble.RandomForestClassifier
10. sklearn.ensemble.GradientBoostingClassifier
11. sklearn.neighbors.KNeighborsClassifier
12. sklearn.svm.SVC
13. sklearn.model_selection.train_test_split
14. sklearn.model_selection.cross_val_score

All experimental steps over each
classifier and datasets are depicted in this section of this thesis.

VI. Result And Analysis

a. Framingham Dataset Result

As per our proposed model, the highest accuracy of different classifiers over the Framingham dataset is 85.5% for the Logistic Regression classifier, 75.6% for Decision Tree Classifier, and 85.04% for Gradient Boosting Classifier, 85.07%

for Random Forest Classifier, 83.5% for KNN Classifier, and 84.7% for Support Vector Machine Classifier

1. Logistic Regression

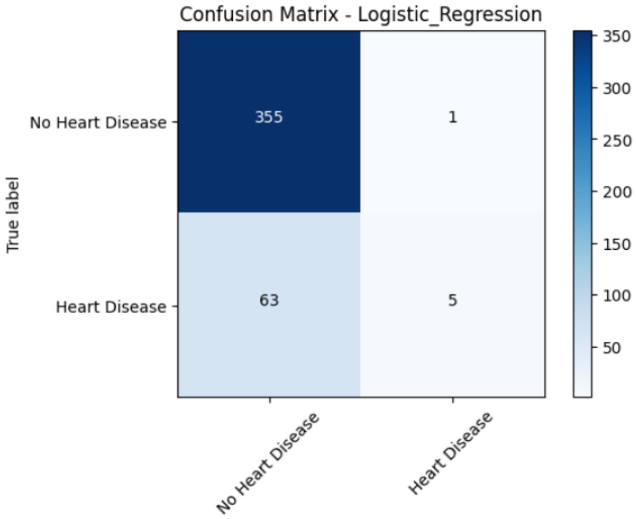

Figure 5: Confusion Matrix of Logistic Regression of Framingham Dataset

2. Decision Tree

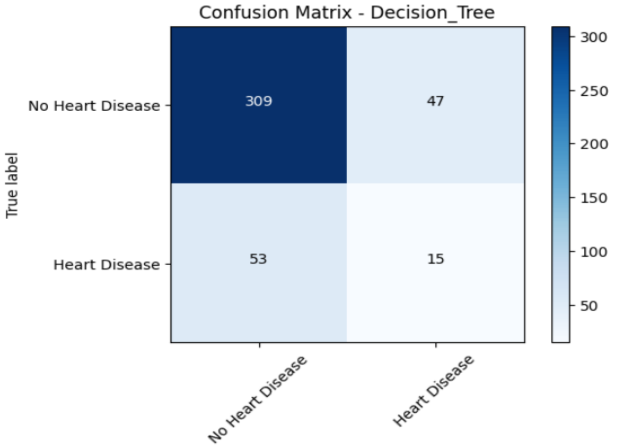

Figure: 6 Confusion Matrix of Decision Tree of Framingham Dataset

3. Gradient Boosting

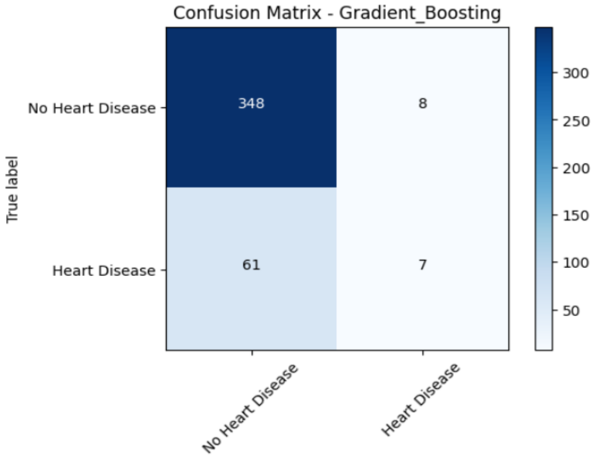

Figure 7: Confusion Matrix of Gradient Boosting of Framingham Dataset

4. Random Forest

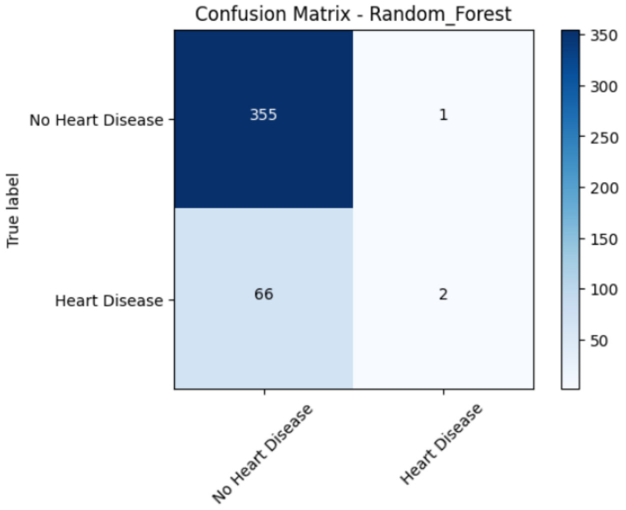

Figure 8: Confusion Matrix of Random Forest of Framingham Dataset

5. K-Nearest Neighbor

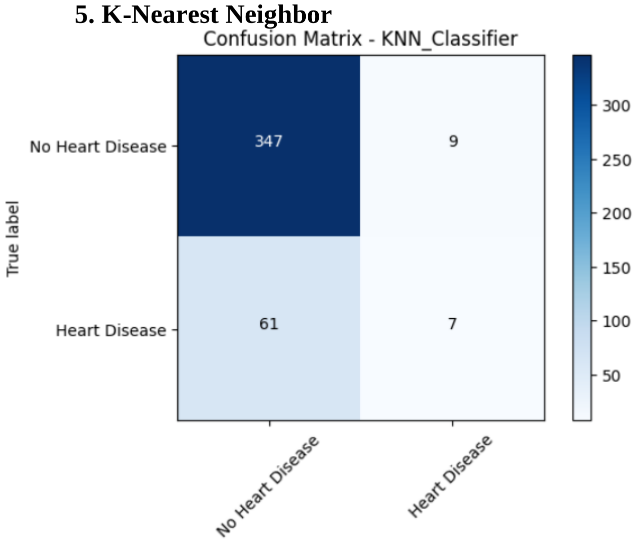

Figure 9: Confusion Matrix of K-Nearest Neighbor of Framingham Dataset

6. Support Vector Machine

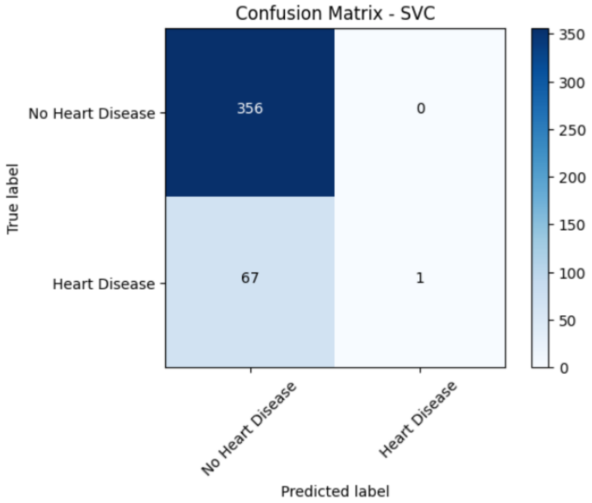

Figure 10: Confusion Matrix of Support Vector Machine of Framingham Dataset

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

b. UCI Dataset Result

While the highest accuracy of different classifiers over the UCI dataset that are two classifiers which got 99.7% for the Random Forest and Decision Tree Classifiers, 89.56% for the Support Vector Machine Classifier, 84.4% for the Logistic Regression Classifier, 84.1% for the K-Nearest Neighbor Classifier, 96.3% for Gradient Boosting Classifier.

1. Logistic Regression

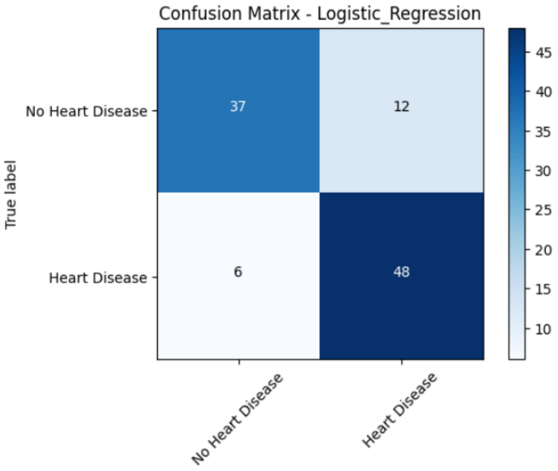

Figure 11: Confusion Matrix of Logistic Regression of UCI Dataset

2. Decision Tree

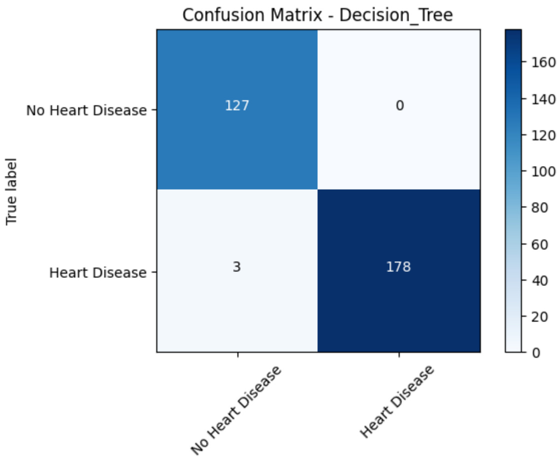

Figure 12: Confusion Matrix of Decision Tree of UCI Dataset

3. Gradient Boosting

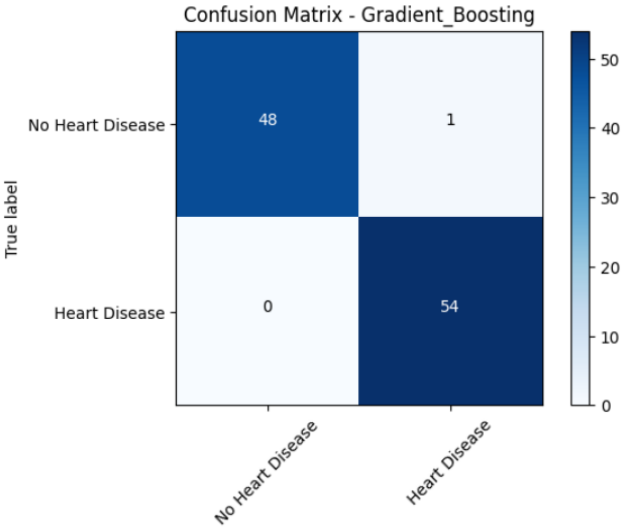

Figure 13: Confusion Matrix of Gradient Boosting of UCI Dataset

4. Random Forest

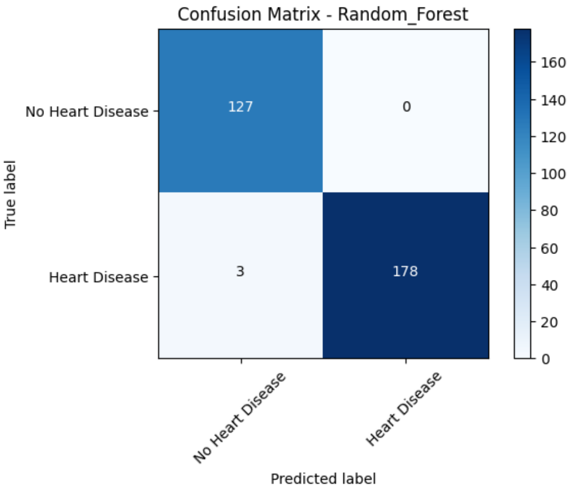

Figure 14: Confusion Matrix of Random Forest of UCI Dataset

5. K-Nearest Neighbor

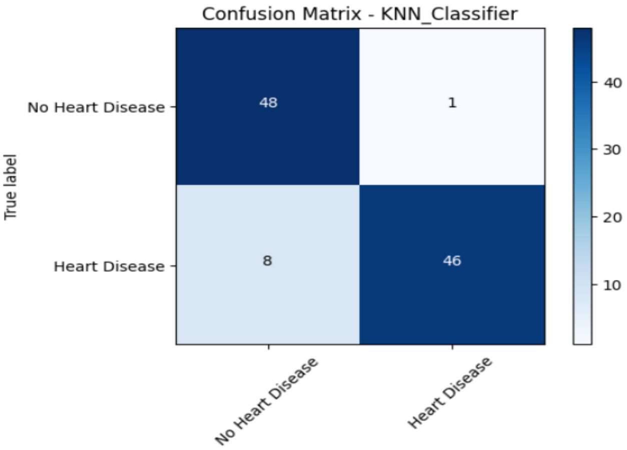

Figure 15: Confusion Matrix of K-Nearest Neighbor of UCI Dataset

6. Support Vector Machine

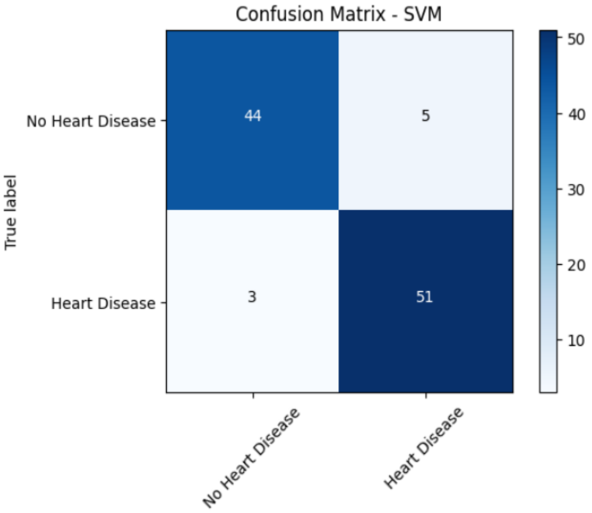

Figure 16: Confusion Matrix of Support Vector Machine of UCI Dataset

b. Inferences

i. Our model predicts the likelihood of a future heart attack using binary classification, with '0' indicating a low probability and '1' indicating a higher likelihood of experiencing a heart attack.

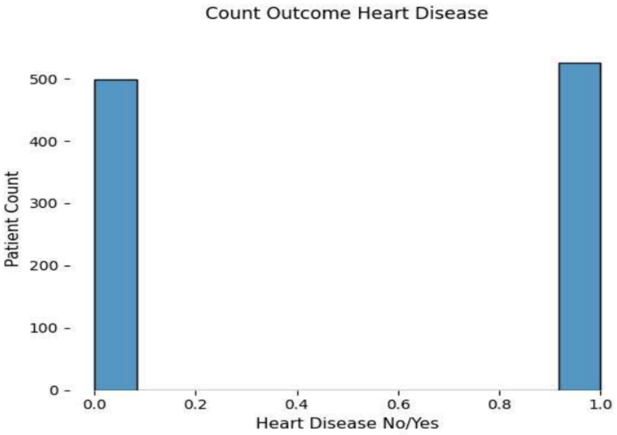

Figure 17: Possibility of heart attack as per test-set

ii. Age-wise heart disease distribution shows the distribution of age concerning the disease probability

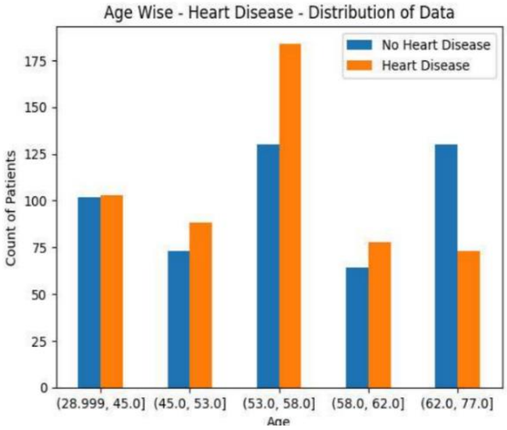

Figure 18: Possibility of heart attack with age distribution

iii. People having higher cholesterol (>200) or higher heart rate (>150) have higher probability for occurrence of heart attack.

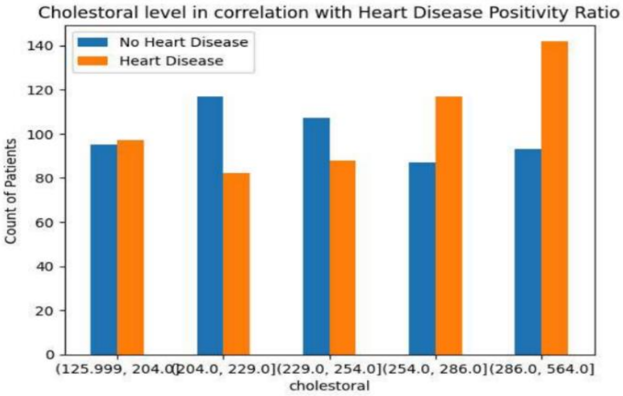

Figure 19: Graphical representation for cholesterol level w.r.t. possibility of heart attack

iv. People having regular chest pain are having higher probability for occurrence of heart attacks. Despite other types of chest pains, typical angina has lower possibilities of heart attacks.

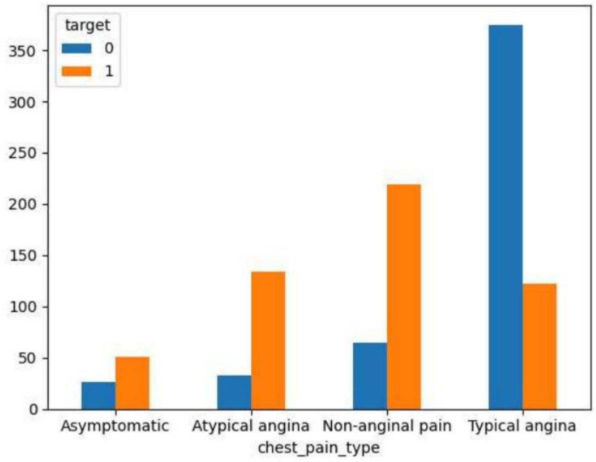

Figure 20: Chest Pain Type

Here, the chart represents various categories (0, 1, 2, 3), each corresponding to distinct types of chest pain:

0 (typical angina), 1 (atypical angina), 2 (non-anginal pain), and 3 (asymptomatic). The analysis suggests that individuals experiencing type-0 chest pain, indicative of typical angina, face a heightened risk of heart attacks compared to those with other types of chest pain. Conversely, individuals with type-2 chest pain, categorized as non-anginal pain, demonstrate a comparatively lower susceptibility to heart attacks.

VII. CONCLUSION

Various supervised machine learning classifiers, including Random Forest, Decision Tree, Gradient Boosting, Support Vector Machine, K-Nearest Neighbor, and Logistic Regression, were used in this study to develop a model for predicting myocardial infarction. Despite inconsistencies in both datasets, various feature transformers were used to improve dataset consistencies and achieve an average accuracy of 85.3% and a recall rate of 6% in the Framingham dataset based on the Logistic Regression classifier. The results showed that the Random Forest and Decision Tree classifier achieved the highest accuracy score, with predictions in binary form where 1 indicates a chance of heart attack and 0 indicates no chance. Several influential attributes were identified, including chest pain type (with typical angina being the most influential and asymptotic chest pain being the least), cholesterol level (with levels greater than 200mg/dl being more prone), increased heart rate, Thal, and age. The study concluded that premature heart attacks can be prevented in 80% of cases through a healthy diet, regular exercise, and avoiding tobacco products. Additionally, drinking more than 5 glasses of water per day was found to reduce the likelihood of developing heart attacks. Blood pressure and cholesterol levels should be checked regularly through medical checkups.

REFERENCES

1. Suraj Kumar Gupta, Aditya Shrivastava, Satya Prakash Upadhyayand Pawan Kumar Chaurasia, “A Machine Learning Approach for Heart Attack Prediction”, International Journal of Engineering and Advanced Technology (IJEAT), vol. 10, no. 6, pp. 124–134, Aug. 2021, doi: 10.35940/ijeat.F3043.0810621.
2. H. Animesh, K. M. Subrata, G. Amit, M. Arkomita, and A. Mukherje, “Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review,” Advances in Computational Sciences and Technology, vol. 10, no. 7, pp. 2137–2159, 2017. [Online]. Available: http://www.ripublication.com.
3. H. S. Buttar, T. Li, and N. Ravi, “Prevention of Cardiovascular Diseases: Role of Exercise, Dietary Interventions, Obesity and Smoking Cessation,” Experimental and Clinical Cardiology, vol. 10, no. 4, pp. 229–249, 2005.
4. I. D. Mienye, Y. Sun, and Z. Wang, “An Improved Ensemble Learning Approach for the Prediction of Heart Disease Risk,” Informatics in Medicine Unlocked, vol. 20, p. 100402, Jan. 2020. DOI: 10.1016/J.IMU.2020.100402.
5. M. Hortmann et al., “The Mitochondria-Targeting Peptide Elamipretide Diminishes Circulating HtrA2 in ST-Segment Elevation Myocardial Infarction,” European Heart Journal: Acute Cardiovascular Care, vol. 8, no. 8, pp. 695–702, 2019. DOI: 10.1177/2048872617710789.
6. Mankad R; Staff Mayoclinics, “Heart Attack,” Mayo Clinic, 2020. Available: https://www.mayoclinic.org/diseases-conditions/heart-attack/symptoms-causes/syc-20373106.
7. P. Severino et al., “Ischemic Heart Disease Pathophysiology Paradigms Overview: From Plaque Activation to Microvascular Dysfunction,” International Journal of Molecular Sciences, vol. 21, no. 21, pp. 1–30, 2020. DOI: 10.3390/ijms21218118.
8. A. Segura-Galindo, F. Javier Del Cañizo-Gómez, I. Martín-Timón, C. Sevillano-Collantes, and F. Javier Del Cañizo Gómez, “Type 2 Diabetes and Cardiovascular Disease: Have All Risk Factors the Same Strength?,” 2014. DOI: 10.4239/wjd.v5.i4.444.
9. P. B. Lockhart and Y.-P. Sun, “Diseases of the Cardiovascular System,” in Burket’s Oral Medicine, John Wiley & Sons, Ltd, 2021, pp. 505–552.
10. T. Mü nzel et al., “Reduction of Environmental Pollutants for Prevention of Cardiovascular Disease: It’s Time to Act,” European Heart Journal, 2021. DOI: 10.1093/eurheartj/ehaa745.
11. M. Ferrante et al., “Air Pollution in High-Risk Sites–Risk Analysis and Health Impact,” in Current Air Quality Issues, InTech, 2015.
12. A. W. R. N. Kandola, “Types of Heart Attack: What You Need to Know,” Medical News Today, 2018. Available: https://www.medicalnewstoday.com/articles/321699.
13. H. Yasue, Y. Mizuno, and E. Harada, “Coronary Artery Spasm-Clinical Features, Pathogenesis and Treatment,” Proceedings of the Japan Academy. Series B, Physical and Biological Sciences, vol. 95, no. 2, pp. 53–66, 2019. DOI: 10.2183/pjab.95.005.
14. G. D. Sandler, David A and Aspenson, D Erik and Johnsen, “Oklahoma Heart Institute,” Citeseer, vol. 2, no. 1, 2005.
15. R. Fass and S. R. Achem, “Noncardiac Chest Pain: Epidemiology, Natural Course and Pathogenesis,” Journal of Neurogastroenterology and Motility, vol. 17, no. 2, pp. 110–123, 2011. DOI: 10.5056/jnm.2011.17.2.110.
16. M. S. Ellulu et al., “Atherosclerotic Cardiovascular Disease: A Review of Initiators and Protective Factors,” Inflammopharmacology, vol. 24, no. 1, pp. 1–10, 2016. DOI: 10.1007/s10787-015-0255-y.
17. J. A. Perez, F. Deligianni, D. Ravi, and G.-Z. Yang, “Artificial Intelligence and Robotics,” pp. 1–56, 2018, [Online].

Available: https://arxiv.org/ftp/arxiv/papers/1803/1803.10813.pdf.

18. M. Fatima and M. Pasha, “Survey of Machine Learning Algorithms for Disease Diagnostic,” J. Intell. Learn. Syst. Appl., vol. 09, no. 01, pp. 1–16, 2017, doi: 10.4236/jilsa.2017.91001.
19. M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” vol. 349, no. 6245, 2015
20. L. Navarini et al., “Cardiovascular Risk Prediction in Ankylosing Spondylitis: From Traditional Scores to Machine Learning Assessment,” Rheumatol. Ther., vol. 7, no. 4, pp. 867–882, 2020, doi: 10.1007/s40744-020-00233-4.
21. J. Brownlee, “4 Types of Classification Tasks in Machine Learning,” Machine Learning Mastery, 2020. https://machinelearningmastery.com/types-of-classification-in-machine-learning/#:~:text=In%20machine%20learning%2C%20classification%20refers,one%20of%20the%20known%20characters.
22. T. Jiang, J. L. Gradus, and A. J. Rosellini, “Supervised Machine Learning: A Brief Primer,” Behav. Ther., vol. 51, no. 5, pp. 675–687, Sep. 2020, doi: 10.1016/J.BETH.2020.05.002.
23. I.-S. Comsa and R. Trestian, “Next-generation wireless networks meet advanced machine learning applications,” no. September, p. 17033, 2019.
24. P. Yadav, “Decision Tree in Machine Learning,” Towards Data Science, 2018. https://towardsdatascience.com/decision-tree-in-machine-learning-e380942a4c96.
25. I. H. Sarker, A. Colman, J. Han, A. I. Khan, Y. B. Abushark, and K. Salah, “BehavDT: A Behavioral Decision Tree Learning to Build User-Centric Context-Aware Predictive Model,” Mob. Networks Appl., vol. 25, no. 3, pp. 1151–1161, 2020, doi: 10.1007/s11036-019-01443-z.
26. C. Molnar, Interpretable machine learning: A Guide for Making Black Box Models Explainable. Github, 2020.
27. V. Aliyev, “Gradient Boosting Classification explained through Python,” Towards Data Science, 2020. https://towardsdatascience.com/gradient-boosting-classification-explained-through-python-60cc980eeb3d.
28. S. Peter, F. Diego, F. A. Hamprecht, and B. Nadler, “Cost efficient gradient boosting,” Adv. Neural Inf. Process. Syst., vol. 2017-Decem, no. Nips 2017, pp. 1552–1562, 2017.
29. T. Yiu, “Understanding Random Forest: How the Algorithm Works and Why it Is So Effective,” Towards Data Science, 2019. https://towardsdatascience.com/understanding-random-forest-58381e0602d2.
30. N. Donges, “A complete guide to the random forest algorithm,” Built In, 2019.
31. Z. Masetic and A. Subasi, “Congestive heart failure detection using random forest classifier,” Comput. Methods Programs Biomed., vol. 130, pp. 54–64, Jul. 2016, doi: 10.1016/J.CMPB.2016.03.020.
32. J. J. Beunza et al., “Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease),” J. Biomed. Inform., vol. 97, p. 103257, Sep. 2019, doi: 10.1016/J.JBI.2019.103257.
33. H. D. Masethe and M. A. Masethe, “Prediction of heart disease using classification algorithms,” Lect. Notes Eng. Comput. Sci., vol. 2, pp. 809–812, 2014.
34. M. Wang, X. Yao, and Y. Chen, “An Imbalanced-Data Processing Algorithm for the Prediction of Heart Attack in Stroke Patients,” IEEE Access, vol. 9, pp. 25394–25404, 2021, doi: 10.1109/ACCESS.2021.3057693.
35. P. Nag, S. Mondal, F. Ahmed, A. More, and M. Raihan, “A simple acute myocardial infarction (Heart Attack) prediction system using clinical data and data mining techniques,” 20th Int. Conf. Comput. Inf. Technol. ICCIT 2017, vol. 2018-Janua, pp. 1–6, 2018, doi: 10.1109/ICCITECHN.2017.8281809.
36. M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease Prediction by Machine Learning over Big Data from Healthcare Communities,” IEEE Access, vol. 5, pp. 8869–8879, 2017, doi: 10.1109/ACCESS.2017.2694446.
37. P. Kaur, M. Sharma, and M. Mittal, “Big Data and Machine Learning Based Secure Healthcare Framework,” Procedia Comput. Sci., vol. 132, pp. 1049–1059, Jan. 2018, doi: 10.1016/J.PROCS.2018.05.020.
38. S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, “Big data in healthcare: management, analysis and future prospects,” J. Big Data, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0217-0.
39. C. A. Alexander and L. Wang, “Big Data Analytics in Heart Attack Prediction,” J. Nurs. Care, vol.
40. Github Link for code : AhmedGad231/Employing-Machine-Learning-Techniques-for-Predicting-Heart-Attacks (github.com)

AUTHORS PROFILE

Ahmed Mohamed Yousef is a 3rd year student in the Faculty of Information Systems and Computer Science College, Sixth of October University, Artificial Intelligence Department. He has a keen interest in problem solving and machine learning, especially in the fields of data and predictive analytics and has shown great dedication towards this project.

Email: ahmedmohamedyoussef212101727@gmail.com

Mohamed Ezz El Dean Ahmed is a 3rd year student in the Faculty of Information Systems and Computer Science College, Sixth of October University, Artificial Intelligence Department. He has a keen interest in problem solving and machine learning, especially in the fields of data and predictive analytics and has shown great dedication towards this project.

Email: mohamedezzeldean212104089@gmail.com