International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-10 Issue-6, August 2021

A Machine Learning Approach for Heart Attack Prediction

Suraj Kumar Gupta, Aditya Shrivastava, S. P. Upadhyay, Pawan Kumar Chaurasia

Abstract: ***A heart attack also known as cardiac arrest, diversify various conditions impacting the heart and became one of the chief-reason for death worldwide over the last few decades. Approximately, 31% of total deaths globally are due to CVDs. It constitutes the pinnacle of chronic processes which involve complex interactions between risk factors which can and cannot be improved. Most of the instances or cases of cardiovascular diseases can be allocated to revisable risk factors where most of the instances are considered preventable. ML became the enhancing approach for the evolution of predictive models in health care industries and was decided to test various algorithms to check what extent their prediction scores estimate or ameliorate upon the results acquired. Researchers deploy various machine learning and data mining techniques over a set of enormous data of cardiovascular patients to attain the prediction for heart attacks before their occurrence for helping healthcare industries and professionals. This research comprises various Supervised ML classifiers like, Gradient Boosting, Decision Tree, Random Forest and Logistic Regression that have been used to deploy a model for Myocardial Infarction prediction. It uses the existing datasets from the Framingham database and others from the database of the UCI Heart repository. This research intends to ideate the prediction for probabilities of occurrence of a heart attack in the patients. These classifiers have been deployed in pipeline approach of machine learning to attain the prediction using both ways i.e., without optimizations and feature transformations as well as vice-versa. The results impersonate that the Gradient Boosting classifier is achieving the highest accuracy score in such a way that prediction used by our model is of binary form in where 1 means a chance of heart attack and 0 means no chance. Some of the most influential attributes are chest pain type among which the typical angina is the most influential and asymptotic chest pain is least, cholesterol level in which the level greater than 200mg/dl are more prone, increased heart rate, thal, and age. It is concluded that premature heart attack is preventable in 80% of the total cases just by using a healthy diet along with regular exercises and not using tobacco products also the person who drinks more than 5 glasses of water daily are less likely to develop attacks.***

The medical checkup of Blood-pressure level, cholesterol level and heart rate on daily basis along with meditation can help you prevent the major heart attacks.

Keywords: ***Cardiovascular-Disease, Framingham Model, Gradient Boosting, Machine Learning, Mayo-cardinal Infarction, UCI Model.***

Manuscript received on August 04, 2021.
Revised Manuscript received on August 20, 2021.
Manuscript published on August 30, 2021.
* Correspondence Author
Suraj Kumar Gupta, Student, Pursuing B.Tech., Department of Computer Science Engineering, Mahatma Gandhi Central University, Motihari (Bihar), India.
Aditya Shrivastava, Student, Pursuing B.Tech., Department of Computer Science and Engineering, Mahatma Gandhi Central University, Motihari (Bihar), India.
Satya Prakash Upadhyay, Registrar, Central University Gujarat, India.
Pawan Kumar Chaurasia*, Associate Professor, Department of Computer Science and Information Technology, Mahatma Gandhi Central University, Motihari (Bihar), India.

I. INTRODUCTION

Heart, the main organ of the human body used for pumping blood into the whole body through the vessels of the circulatory system. In the circulatory system, the most crucial role is played by the heart [1]. The circulatory system is the most important part of our body as it is responsible for the transport of blood carrying food, oxygen, water, minerals, and other important substance important for our body throughout our entire body. If the working of the heart is disrupted due to any circumstance and it does not function properly then it may cause serious health issues including death.

Cardiovascular is a term that is used to refer to the pathologies altering or affecting the function or structure of the heart or blood vessel having the most common type of cardiovascular disease as coronary artery disease. The prevalence of the most common cardiovascular diseases (CVDs) represents the pinnacle of incurable processes which involve complex interactions between risk factors which can and cannot be improved. Most of the instances or cases of cardiovascular diseases can be allocated to improvable risk factors where most of the instances are considered preventable.

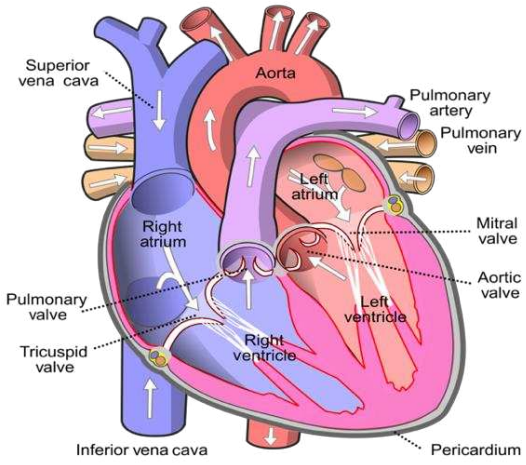

Figure 1:Human Heart Image

Retrieval Number: 100.1/ijeat.F30430810621
DOI:10.35940/ijeat.F3043.0810621
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication

Traditional efforts for preventing cardiovascular diseases have focused on modifiable behaviors governed by individuals [2]. The use of physical inactivity, tobacco, a poor diet, and obesity majorly contribute to known risk factors for cardiovascular diseases, such as mellitus, diabetes, hypertension, and the development of atherosclerosis. As per the current scenario approximately 17.9 million people every year lose their lives due to cardiovascular diseases (CVDs), apparent mainly as attacks and strokes. The above scenario of death is 31% of total deaths globally [3]. Heart-diseases or cardiovascular diseases (CVD) are the type of diseases that involve the heart and blood vessels.

A. Coronary artery disease

In coronary arteries, blockage formation due to fat, cholesterol also called plaques. These plaques on being damaged release platelets which cause blood to clot. They accumulate around the plaque and hence restrict the blood flow, which eventually damages the heart muscle [4] [5]. This damage may become more substantial if the blood flow to any segment of the heart is completely blocked [6]. Coronary artery disease is the tapering or blockage of the coronary arteries which is usually caused by atherosclerosis which builds up plaques in the artery-blocking blood flow. The most serious complication of coronary artery disease is myocardial infarction which is best known as a heart attack [7]. This disease at its primary stage is a preventable condition but preventing its quick and early treatment, it can lead to serious health problems and in the worst case even death [8].

Environmental agents also have an important role in the growth and ferocity of cardiovascular diseases although the individual often has very little control over them. The cardiovascular system are very much prone to a variety of different environmental agents which includes solvents, tobacco, pesticides, smoke, and other inhaled or ingested pollutants, as well as extremes in noise and temperature [9]. Exposure to environmental pollutants occurs via three general routes: inhalation, ingestion, and absorption. Condemnation to environmental agents mainly occurs via three general routes which are inhalation, ingestion, and absorption. In 2014, the World Health Organization proclaimed air pollution as the deadliest single environmental health risk. In 2012, more than 3.5 million deaths in people who were aged >60 years were recognized to outdoor air pollution out of which 80% of these deaths were the consequences of cardiovascular diseases [10].

B. Heart attack

Heart Attack consequences from one of these CV diseases. STEMI, an attack, which occurs due to atherosclerosis, restricting blood flow to a wide area of the heart. This led to continuous damage to the heart muscle due to which the functioning of the heart is completely stopped and may cause death. This attack is severe and needs rapid attention. NSTEMI, an attack that occurs due to partial blockage of coronary arteries, restricting blood flow severely. Although being less dangerous, it led to permanent damage of a restricted part of the heart which is not receiving the blood flow. Coronary artery spasm, the silent heart attacks. Sometimes due to the contraction of arteries connected to the heart, blood flow to major parts of the heart is restricted which causes coronary artery spasm. Being less severe with respect to other attacks, it never causes any permanent damage to the heart [11][12].

- **Symptoms**

Major symptoms of the occurrence of heart attack are tightfistedness or affliction majorly in the chest, neck, back, and arms, tiredness, dizziness, abnormal heartbeat, and consternation. Risk factors including unchangeable factors like age, sex, family background, and changeable factors like smoking, high cholesterol, high blood pressure, fatness, deficiency of proper diet as well as exercise, and a huge amount of stress. Arterial reclamation to medication, ECG, and bypass surgery is the most used treatment methods in case of heart attacks [13][14].

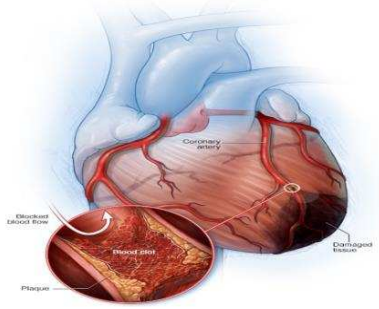

Figure 2: Blockage in Coronary Artery [5]

- **Factors causing Heart Attacks**

The comestible factors that assist or protect against the growth of coronary heart diseases ensuing to the arrival of atherosclerosis. The primary cause or initiator of atherosclerosis is unrevealed, despite one conjecture is now acquiring attention that the primary event is free radical damage to cholesterol in circulating low density lipoproteins and it is possible to acknowledge a seven dietary factor having two promoter and five protective feature for developing coronary heart disease [15]. Myocardial infarction, most serious complication of coronary heart disease, represents an amalgamation of having two distinct effects of dietary factors. There are evidences that lowering serum homocysteine containing folic acid reduces the risk of cardiovascular diseases is largely observational. Drug Treatment for preventing the cardiovascular disease events and heart strokes has been limited to single risk factors which targets the small amount of patients' population with values in the appendage of the risk factor distribution, and to reducing the risk factors to average population values [16].

C. Artificial Intelligence

Artificial intelligence can be termed as the proficiency of a computer program to mimic the human brain by developing intelligence. It gives computers the ability to think which makes them more intelligent. Alan Mathison Turing first proposed the concept of artificial intelligence more than sixty years ago. According to him if a human cannot distinguish between the human and computer based on the answers or solutions or responses given by a machine then the machine is called as "intelligent".

Retrieval Number: 100.1/ijeat.F30430810621
DOI:10.35940/ijeat.F3043.0810621
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication

Its importance's can be felt by witnessing the last two decades which showed the advances of the artificial intelligence and robotics and the future can be expected to be more speculated. The term artificial intelligence was first coined in 1956 [17]. AI is based on the principle that the machine tries to mimic and execute tasks which are most simple to those that are even more complex by using human intelligence.

- **Machine Learning**

Machine Learning can be called as a subdivision of Artificial Intelligence in which the program learns through data or input [18]. It primarily focused on two principles i.e., in what way a computer system be developed so that it automatically improves the results through its previous experiences and the fundamental statistical computational-information-theoretic laws those regulating all learning systems, including humans, computers and organizations [19]. It was designed to solve a specific problem by developing an intelligent system that is able to solve that problem without being categorically programmed for it. The learning of the program is done by deriving knowledge from a large amount of data which is useful in making predictions [20].

D. Classification

In Machine Learning, Classification is a function which uses ML algorithms to learn for assigning class label to cases from problem set. Classification is described as predictive modeling of any problem with a labelled class to be predicted for a given dataset. For example, to determine whether a human is male or female, to determine whether a mail is spam or not, to classify any given handwritten character i.e., whether it is known or unknown character [21]. Supervised learning algorithms are the most basic machine learning algorithm which is trained on labeled data. This algorithm is extremely robust when used in the right circumstances. It usually performs the prediction tasks since the goal to achieve is to either foresee or classify the target from a certain outcome of interest i.e., availability or unavailability of the heart problems [22][23].

- **Decision Tree**

A **Decision tree** is a non-parametric supervised machine learning algorithm used for solving classification and regression problems which uses decision support tool that uses a flowchart-like graph structures in making decision. In decision tree, each internal node refers a test on a feature taken via the pathways along which the branches describing the classification rules to attain the final result (i.e., class label) defined by each leaf node. The decision tree looks like, learning by splitting the dataset into subsets based on value test which is carried on an attribute. This process is continued on every extracted subset recursively which is handled until the subset at a node have the same value of the target variable, or adds no value to the prediction when splitting [24][25].

- **Logistic Regression**

Logistic Regression is a parametric supervised machine learning algorithm which is a linear model used for solving classification problem which predicts outcome of a categorical dependent variable from the set of predictor or independent variables. It is a very simple method and much more efficient for binary and linear classification problem where response variable tends to be binary having continuous explanatory variables despite having regression on its name. This model can be generalized further for multi-class classification problems. It models the liability for classification problems with two possible outcomes for target. It uses logistic equations for finding the results between 0 and 1 [26]. Logistic function (or, sigmoid function) is defined as:

$$logistic\ (\eta) = \frac{1}{1+\exp(-\eta)}$$

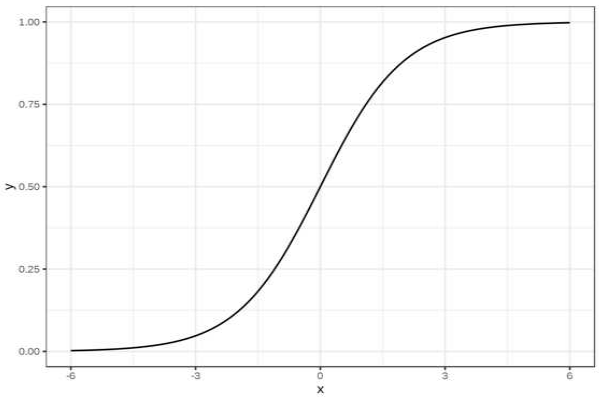

Figure 3: Classification of target value (y) w.r.t feature (x) [26]

- **Gradient Boosting Classifier**

Gradient Boosting Classifier is one of the Ensemble learning techniques. The different types of gradient boosting classifier algorithms have been increased with several interesting proposals naming its different types. Here, boosting means integrating multiple weak learning algorithms with lower accuracy to a stronger model predicting with higher accuracy. It follows mistakes done by predecessor's algorithms for maintain the upcoming algorithms to attain higher accuracy. In Gradient Boosting, each predictor tries to overcome the mistakes done by its predecessor's algorithms. It always fits a new predictor for overcoming the errors done by its predecessor's algorithms [27][28].

- **Random Forest**

Random Forrest is a supervised learning algorithm, as its name look alike, it is an integration of large number of individual decision trees created during training time which functions as Ensemble learning (can be termed as nearest neighbor predictor) used to solve classification as well as regression problem. In the crowd of such decision trees, the individual tree whose class prediction is upvoted most or the class having the largest mode value becomes the prediction of model [29]. In other words, initially it develops many individual decision trees and then integrates them to obtain a higher accuracy and better prediction. It finds a natural balance between the two extreme problems of high variance along with high bias by averaging them [30][31].

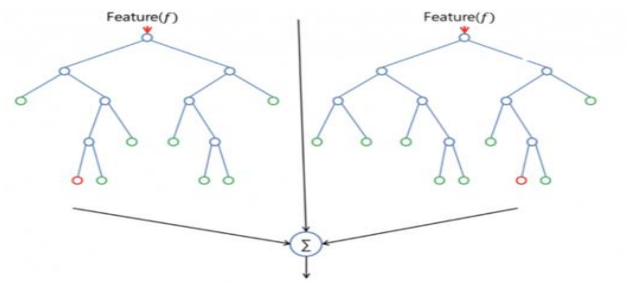

Figure 4: Basic Random Forest Model [30]

E. Problem

According to the WHO (World Health Organization) the 32% of all global deaths were due to cardiovascular diseases out of which, approximately 85% deaths were due to heart attacks. These surveys also represents that approximately 48% of US population have any kind of heart disease and around every 19 deaths one person is due to heart problems. According to the National Crime Records Bureau (NCRB) in India, the percentage of deaths due to heart attacks is increased by approximately 53% in last five years and approximately 33% of deaths in India were due to heart attacks and other problems. Although in many cases the patient may be saved by doctor even after minor or major attacks, but also in most of heart attack cases patient gets died due to unavailability of medical treatment under minimal time due to the incapability of patient to inform anyone during the attack period and also for their family members as being unknown from this kind of emergency condition to any of their relatives. And due to which, this kind of sudden attacks may lead to the death of patient even before consulting to any doctor. So, we cannot prefer consulting doctors after having attacks as the perfect solution for precaution from occurrence of heart attacks.

F. Approaches

Hence, to overcome and reduce the chances of death, AI and ML comes in role to predict the chances of upcoming heart attacks according to the health conditions as well as previous medical conditions of the patient. Although there are numerous ml-models have been already developed but most of the models are prepared for heart disease prediction system instead of being specialized for heart attacks only. Many other models have been deployed using various algorithms and over various datasets along which many are simpler whereas many are complex. In this project we applied various ML-algorithms over two datasets to take a survey and to achieve a better accuracy to deploy a model for predicting the upcoming heart attack scenario earlier than the actual attack time to prevent the patient from going to a critical condition which can lead him/her to death.

G. Objective

For the both datasets, the risk factors provided in it associated with heart attacks for the patients to find out whether they have any risk of cardiovascular problems in upcoming years or not.

Based on the datasets provided:

- To predict the probability of suffering of any patient from heart attacks in the upcoming future**.**
- Identifying the major factors influencing heart attacks.
- Identifying cholesterol level with higher possibilities of heart attacks.
- Identifying chest pain types having higher possibilities of heart attacks.
- Recommendations for preventing / reducing chances of getting a heart attacks.

II. RELATED WORKS

As per initial prospects, ML can enhance the evolution of predictive models in health care industries, it was decided to test various algorithms to check what extent their prediction scores estimate or ameliorate upon the results acquired in the authentic Framingham model. This model is one of the most important cardiac arrest risk prediction tables from the point of view of clinical practice. Several CV-risk prediction algorithms have been proposed till now [32]. The performance of various algorithms that have been used for calculating CV-risk can still be considered as a problem. Particularly, different scores like, Framingham Score, Systematic Coronary Risk Evaluation tend to underrate cardiac risk in patients [20]. Comparative study has been done for CV risk prediction using some well-known ML algorithms like k-nearest neighbor, support vector machine, classification, gradient boosting, logistic regression and regression tree and random forest[3]. Among the comparison between various features selection ML algorithms, for heart attack prediction Naive Bayes, SVM, and KNN the most optimistic classifiers.

Another approach used for the cardiac infarction risk prediction involves partitioning the dataset in a random way and deploying various old data mining approaches like J48, REPTREE, Naïve Bayes, Bayes Net, and CART. The deployed model was able to answer more complex queries in cases of cardiac infarction prediction [33]. The research was carried out in February 2021 to make a model which uses an algorithm by leveraging random under sampling, clustering, and oversampling techniques which were better known as under sampling-clustering-oversampling algorithm (shortly, UCO algorithm). The difference in this algorithm was that the data that was used for training on machine learning algorithms were nearly very balanced data. This algorithm was very good in extracting features which were then tested on different classifiers giving the best predicting performance with an accuracy of 70.29%, precision of 70.05%, 1-Recall of 75.59%, 0-Recall of 63.95% of the random forest [34]. There was research carried out for predicting the heart attack early with taking an account of chest pain with 24 other attributes. The data mining algorithms used were decision tree and random forest classifier to analyze the heart attack datasets. The deep classification was done by a decision tree algorithm where the random forest was used for the classification of targets [35].

There was another technique used for the prediction of heart disease risk which involves partitioning the dataset randomly using mean based splitting approach into smaller subsets and then using homogeneous ensemble created using various classification and regression trees models using an accuracy based weighted aging classifier ensemble.

Retrieval Number: 100.1/ijeat.F30430810621
DOI:10.35940/ijeat.F3043.0810621
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication

There were two datasets Cleveland and Framingham which received classification accuracies of 93% and 91% respectively [36]. A research was carried out during July 2020 to make a model which predict the risk of cardiovascular disease using two different techniques. The support vector machine(SVM) was first trained and tuned perfectly for its parameters and then on training the SVM model for 1000 times , the average accuracy attained for the model to predict the cardio-vascular disease accurately was up to 96.5% with its average recall rate 89.8% while the recall rate using K-nearest neighbours reaches to 92.9% [37].

With growing big data in biomedical community and healthcare industries, we have enormous amount of data whose accurate analysis can be used for earlier detection of several cardiovascular disease or strokes. There was proposed algorithm which used latent factor model to rebuild mislaid data. This new algorithm is best known as a convolutional neural network (CNN) which in this case is based on multimodal disease risk prediction algorithm which gives a prediction accuracy of 94.8% [38]. It was observed that there are many models which predict the cardio arrest problems but they have not given a significant or proper amount of weightage in data privacy as the data used in training, testing and validation of model are private data of patients. Different techniques like masking encryption, dynamic data encryption, granular access control, activity monitoring, and end point validation have been incorporated in dealing with the personal data and the algorithm used is explained [32].

Acute myocardial infarction, also known as heart attack, is the deadliest cardiovascular related disease that patients face. Big data which is well known for its vast and important use in different areas of medical field can be used against these chronic diseases for protection and prediction and for their treatment. The national and international databases of various heart patients were examined thoroughly with minute details to identify various studies which then conducted about big data analytics in healthcare, myocardial infarction prevention and prediction and for that a total of 31 studies of different datasets were studied and assessed [39]. While the prediction of myocardial infarction was being carried out by various researchers and there were various approaches depolyed by various researchers, the desired accuracy of the model was not achieved and there were certain experiments being carried outlooking for optimal solution. One of the approaches was using machine learning algorithm along with feature selection algorithm together for the prediction of myocardial infarction.

In this research area various ML approaches along with optimum parameters as well as different feature selection techniques were deployed and the results were noted. The Statlog (Heart) dataset was used during this survey to attain the experimental results for the best algorithm which was SVM used with the linear kernel giving highest accuracy of 84.21% which when used by the relief method which is a feature selection algorithm [40]. A different kind of unsupervised classification algorithm, called as Fuzzy C Means Classifier, was used to predict the heart attack using patients Early's medical report, where 13 attributes were used in classification stage for the FCM (Fuzzy C Means Classifier) for the better prediction of cardiac infarction. FCM is an unsupervised machine learning algorithm which is a different type of clustering in which a single data point may exist in two or more clusters containing every chance of that data point whether to include it or not during analysing its dependencies. It was designed for the physicians for better diagnosis of heart attacks in an efficient way. The FCM classifier was later tested on real data of approximately 270 patients in diagnosing their heart attack giving an accuracy of 92%, which was a mere high accuracy [41].

Neuro-fuzzy is a hybrid intelligent system that applies two humans like reasoning techniques of fuzzy logic with the learning and combining structure of neural networks in which fuzzy logic system is used to illustrate the knowledge in explainable manner and using the neural network for optimizing the parameters. The above system was testing data of some patients in which the accuracy of the mentioned model was above 90% and was deployed so that it can be usable by patient himself [42].

III. DATASETS AND PRE-PROCESSING

Dataset, generally said to be as a collection of data organised for specific purposes. Here in this research, we have used two different datasets. The first one dataset we used is Framingham dataset, which is published by collaboration of Boston University, National Institute of Heart (NIH), National Heart Lung and Blood Institute (NHLBI). This dataset is completely dedicated for identifying factors for identifying cardiovascular diseases, especially attacks and strokes. And the second one dataset we used is Heart dataset from UCI Machine Learning Repository (University of California, Irvine, School of Information and Computer Sciences). This dataset is a combination of four different databases from "Hungarian Institute of Cardiology, Budapest", "University Hospital, Zurich, Switzerland", "University Hospital, Basel, Switzerland", "V.A. Medical Center, Long Beach and Cleveland Clinic Foundation"[43]. Both the above datasets are also available on Kaggle website.

A. Framingham Dataset

Attributes of the Framingham dataset are classified in four types like Demographics, Behavioral, Previous medical history-based information, Current medical condition-based information.

Demographic Attributes:

- Sex: Categorised as either 0 or 1 such as 0 = female and 1 = male
- Age: Age of the patient at the current time of examination
- Education: It is an inessential data, because any medical issue doesn't occur as per someone's education level

Behavioral:

- Current Smoker: Categorised as either 0 or 1 depending upon whether a patient smoke currently or not i.e., 1 = yes and 0 = no

- Cigs Per Day: Depends upon if someone is smoking regularly, then the average no. of cigarettes being smoked by him.

Previous medical history-based information:

- Diabetes: Categorised as either 0 or 1 depending upon whether a patient had diabetes or not, 1 means Yes & 0 means No.
- BP Meds: Categorised as either 0 or 1 depending upon whether a patient is based on medication for blood pressure or not, 1 means having blood pressure medication and 0 means not having blood pressure medication.
- Prevalent Stroke: Categorised as either 0 or 1 depending upon whether the patient had a stroke previously or not, where 1 represents YES & 0represents NO.
- Prevalent Hyp: Categorized as either 0 or 1 depending upon whether the patient was hypertensive (i.e., having abnormal high blood pressure), 1 means Yes & 0 means No.

Current medical condition-based information:

- Heart Rate: heart rate
- 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 - (binary: 1 means "Yes" & 0 means "No")

B. UCI Dataset

The dataset consists of total 13 decision parameters. The target value is represented by "target".

- Age (age): Age of the patient at the current time of examination
- Sex (sex): Categorised as either 0 or 1 such as 0 = female and 1 = male
- Chest Pain (cp): Categorised into four types from 0 to 3 defining: 0 as typical angina, 1 as atypical angina, 2 as non-anginal pain, and 3 as asymptotic.
- Resting Blood Pressure (trestbps): Resting blood pressure value of patient in mmHg (unit)
- Cholesterol (chol): Cholesterol of patient in mg/dl (unit).
- Fasting Blood Sugar (fbs): Categorized as either 0 or 1 such as 1 = if fbs >120 mg/dl (true) else 0 (false).
- Resting ECG (restecg): Categorized into three types from 0 to 2 defining: 0 = normal, 1 = having ST-T wave abnormality, 2 = left ventricular hypertrophy.
- Max Heart Rate (thalach): Maximum heart rate achieved by any patient.
- Exercise induced angina (exang): Categorized as either 0 or 1 such as 0 = No and 1 = Yes
- oldpeak: Displays the value of ST depression of any patient induced by exercise w.r.t. rest (float values).
- Slope: It describes the peak of exercise during ST segment, classified in three ranges, 0 = up-slope, 1 = flat, 2 = down-slope.
- No. of major vessels (ca): It is classified in range 0 to 4 by coloring through fluoroscopy.
- Thalassemia (thal): It is classified into three ranges from 1 to 3, where 1 = normal, 2 = fixed defect, 3 = reversible defect
- Target: This is the prediction column for diagnosis of heart attacks. It is classified into two types 0 and 1, where 0 means no possibility of heart attack and 1 means possibilities of heart attack.

C. Preprocessing

Data Pre-Processing is defined as transforming or encoding the data in such a state so that is can be easily parsed by the machines for generating accurate information. In other words, it should be transformed in such a form so that is can be easily interpreted by different algorithms with producing higher accurate results. It is not necessary to be complete pure data in each and every dataset. There is always some missing data in each and 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 tuples from dataset or by imputing mean or median values of respective column or some other hyperparametric optimization to attain the imputable values for replacing those missing values. Since the both datasets used in our model consist numeric data only.

Hence, in our proposed model we are using mean and median imputation approaches for imputing missing values in data set for attaining its consistency to achieve higher accuracy. Mean imputation is the way of replacing missing values (i.e., 'NA' or 'NULL') data in dataset by mean of that parameter. And median imputation is the way of replacing missing values (i.e., 'NA' or 'NULL') data in dataset by median of that parameter. Even in this mean and median imputation there is always a confusion that 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 skewed distribution instead of normal distribution, then the median imputation is preferred over mean imputation.

IV. METHODOLOGY

In our proposed model we are using two major datasets of the field of Cardiovascular analysis. The first one dataset is Framingham dataset and the other one is UCI heart dataset. We are using four ML classifiers which are Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and Gradient Boosting Classifier. Here, we are using iterative approach for deploying our model over both the datasets. Both the first datasets are loaded iteratively and checked for the missing values and further pre-processed by imputation of the missing values.

Then dataset is split into two parts, one is training data and other is holdout data for testing purpose. The ratio of splitting of dataset into training and holdout part is 9:1 i.e., 90% of dataset will be used for training purpose of model whereas remaining 10% will be used for testing purpose of model. not number text heads-the template will do that for you.

A. Work flow of Proposed Model

After splitting and pre-processing, we are iteratively deploying each classifier over the dataset for their respective predictions and accuracies attained. For each classifier, we are generating four pipelines by optimizing and enhancing some features from previous pipeline. Pipeline in machine learning is an approach to code and for automating its workflow to build the model. Out of total sixteen pipelines, each classifier has four pipelines in order, first without any optimization then, second with hyperparameters optimization (i.e., marked as **"hpo-1"**) then, third with previous hyperparameters optimization and feature engineering (i.e., marked as **"hpo-1 + fe"**) and at last, with previous hyperparameters optimization, feature engineering and final updating again with hyperparameters optimization (i.e., marked as **"hpo-1 + fe + hpo-2"**). After these four iterations on pipelines of each classifier, the best pipeline of each classifier is selected. And then the final four pipelines, best one pipeline of each classifier are analyzed together as per their performance and accuracies. And then the final best of all sixteen pipelines is selected to deploy the final model for further predictions in future. The complete workflow designed for this study as flowchart.

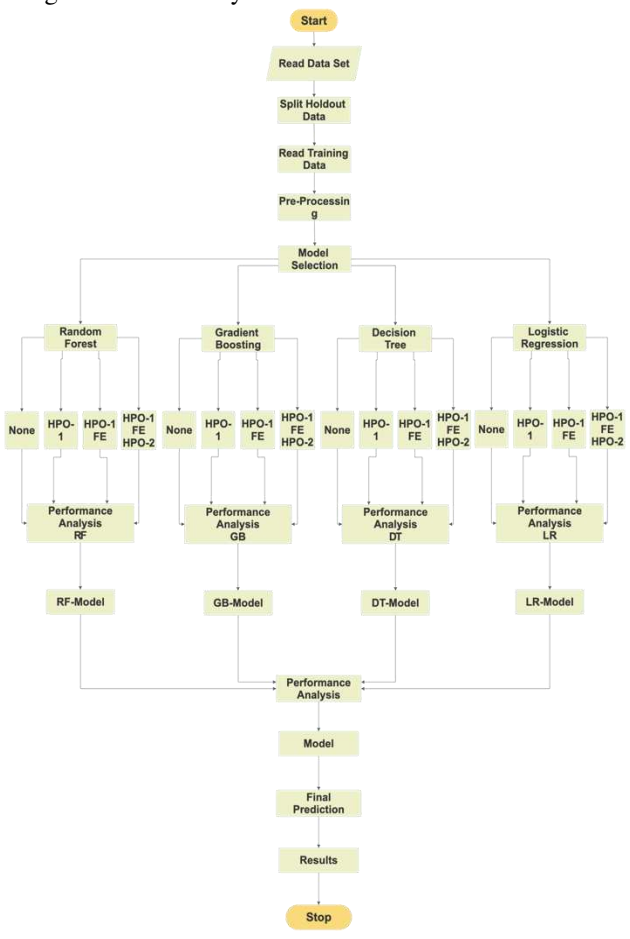

Figure 5: Flowchart for Proposed Model

B. Algorithm

Algorithm for the proposed model as per flowchart.

Set t to no of datasets

Set data[t]

Set clf[ ]

Set enhance[ ]

Set x = 0

Repeat while x < t:

Repeat for i in clf:

Repeat for j in enhance:

Deploy clf[i] with enhance[j] over data[x]

Set best_enhance[i] = max(clf_accuracy)

Set x = x+1

Set best_clf = max(best_enhance)

Set td to test data

Use best_clf to predict td

Return result

Here initially, we have taken some inputs t no datasets, data[] list of datasets, clf[] list of classifiers and enhance[] a list of enhancements to be added in every forth iteration. Thereafter some variables declaration are done for work flow of model such as x=0 for approaching each datasets in list data[] in while loop. Variables clf_accuracy is the accuracy prediction for each iteration enhance over a classifier and best_enhance[] is dict of enhancements over which the highest accuracy for each classifier predicted and last best_clf is max of values of best_enhance for prediction of best classifier and enhancements for model deployment.

V. EXPERIMENTAL SETUP

For experimental setup various packages installed are: ibm_watson_machine_learning, autoai-libs, scikit-learn, etc over IBM cloud Watson Studio. Thereafter, pipeline generation is done and parameters are configured using get_params() function of pipeline optimizer. The summary () method is used to list trained pipelines and evaluation metrics information in the form of a Pandas Data Frame. After all these setup and comparison of pipelines in the project, and after completion of pre-processing we have deployed Scikit Learn ML Pipeline model for each classifier to obtain its best accuracy after applying all approaches i.e., without optimization or feature engineering and with hyperparameter optimization, feature engineering or both continuously followed by hyperparameter optimization. Since optimizing a model is always been the toughest challenge in ML implementation. Hyperparameter optimization is the approach for learning algorithm by selecting the set of optimal hyperparameters. All experimental steps over each classifier and datasets are depicted in this section of this thesis.

A. Feature Engineering

Various feature transformers are customized and applied to features during feature engineering phase in model. Feature transformers used in this model for customizing in feature transformation are square root, round, principal component analysis, univariate feature selection, product, and sum. Iteratively each and all steps are deployed over both datasets. Square Root is a number which when multiplied by itself gives another number. It is a transformer which returns a homogeneous result over variances for different independent features. Round transformer returns the value after being rounded up for matching the records or clustering them. PCA (Principal Component Analysis) is an approach, which is used in most of the ML algorithms for reducing the dimension of large datasets where required operations are costly into smaller ones which still, being reduced, contains most of the information of the large dataset. Univariate Feature Selection is a unique method of feature selection where the selection of any particular feature is based on mathematical univariate statistical tests in which we compare the dependency of each feature to the label by checking its significant relationship with each other. It's well known as Analysis of Variance (ANOVA). Product returns the result after multiplying selected features as value for new feature. For ex. Let say two features x and y then take product of (x * y = z) then z will be the value for new feature generated by this product transformer. And Sum returns the result after adding selected features as value for new feature. For ex. Let say two features x and y then take addition of (x + y = m) then, m will be the value for new feature generated by this sum transformer.

VI. RESULT AND ANALYSIS

As per our proposed model, the highest accuracy of different classifiers over Framingham dataset are 85.5% for Gradient Boosting classifier with enhancements "HPO-1, FE, HPO-2", 84.7% for Random Forest classifier with enhancements "HPO-1, FE", 77.1% for Decision Tree classifier with enhancements "HPO-1, FE", and last 67.7% for Logistic Regression classifier with enhancements "HPO-1, FE, HPO-2". Similarly, the highest accuracy of different classifiers over UCI Heart dataset are 85.6% for Gradient Boosting classifier with enhancements "HPO-1, FE, HPO-2", 85.6% for Random Forest classifier with enhancements "HPO-1, FE, HPO-2", 80.2% for Decision Tree classifier with enhancements "HPO-1, FE, HPO-2", and last 84.5% for Logistic Regression classifier with enhancements "HPO-1, FE, HPO-2". The ranking of pipelines with having same accuracies are done with the help of time elapsed in its execution. However, for both dataset, Gradient Boosting classifier with all enhancements i.e., hyperparameter optimization and feature engineering (HPO-1 + FE + HPO-2) is attaining the most accurate results with highest accuracy. Hence after deployment of all classifiers proposed in this research, the final model is deployed with Gradient Boosting classifier with all enhancements.

A. Framingham Model Accuracies

Since, in this proposed research we are using four classifiers as per our proposed model. Hence after completion of every four pipelines for each of these classifiers, the best pipeline is selected for further studies. Best of all four pipelines for each classifier over Framingham dataset are described below.

Figure 6: Random Forest Classifier

Figure 7: Gradient Boosting

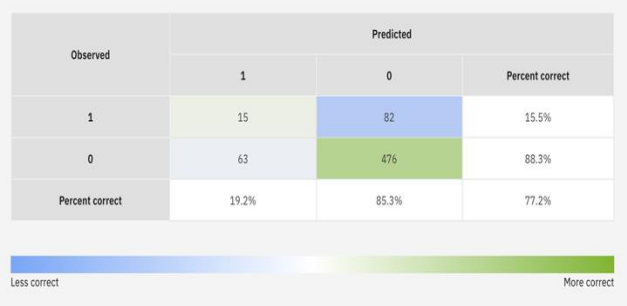

Figure 8: Decision Tree

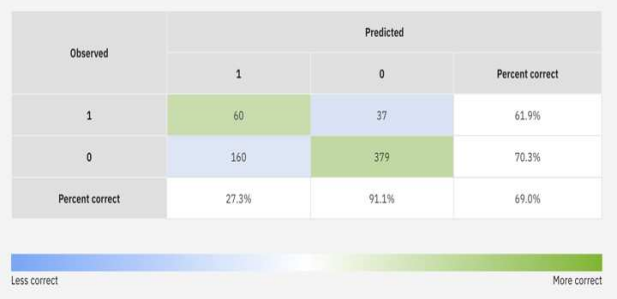

Figure 9: Logistic Regression

B. UCI Model Accuracies

Similarly as of Framingham dataset, here also in UCI heart dataset after completion of every four pipelines for each of these classifiers, the best pipeline is selected for further studies. Best of all four pipelines for each classifiers over Framingham dataset are described below in subtopics.

Retrieval Number: 100.1/ijeat.F30430810621
DOI:10.35940/ijeat.F3043.0810621
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication

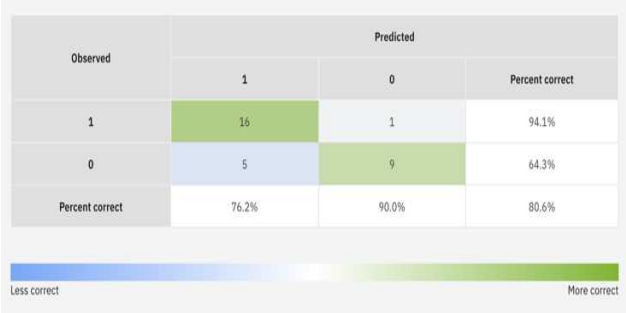

Figure 10: Logistic Regression over UCI

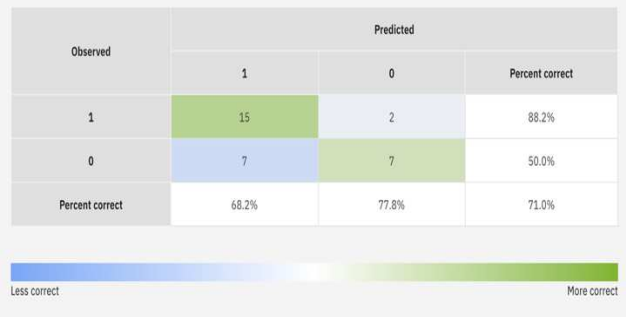

Figure 11: Gradient Boosting Classifier over UCI

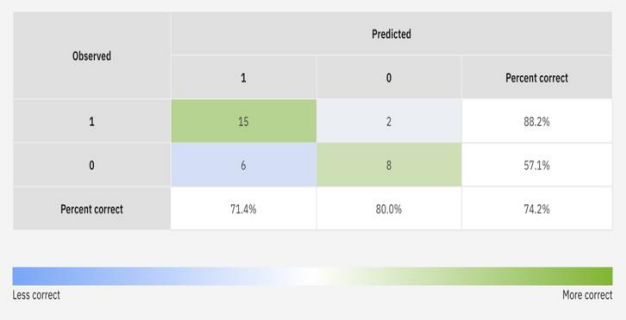

Figure 12: Random Forest Classifier over UCI

Figure 13: Decision Tree over UCI

C. Inferences

i. Heart attack in future is predicted by our model using 0 or 1 where, 0 means no possibilities and 1 means possibility of having heart attacks.

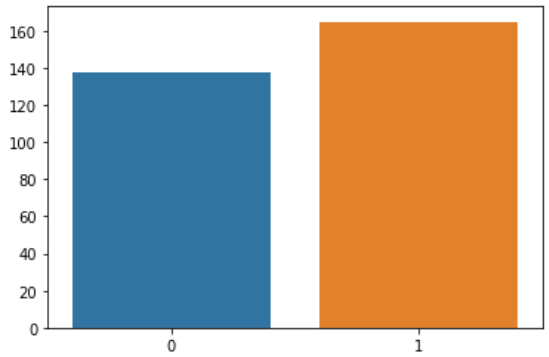

Figure 14: Possibility of heart attack as per test-set

Here, as per this graph plot in our study after prediction done over test set of near about 300 test-cases it is found that approximately 160 test-cases exhibit the occurrence of heart attacks while approximately 140 test-cases don't exhibit the occurrence of heart attack in forthcoming future.

ii. Major factors influencing heart attacks are higher cholesterol level, higher heart rate, chest pains and blood pressure.

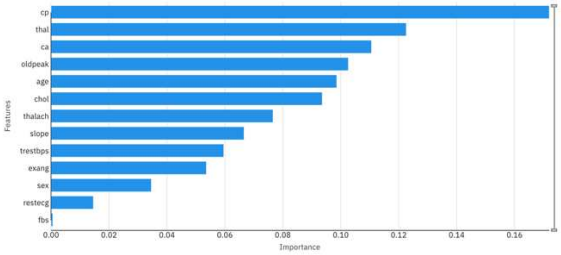

Figure 15: Major factors influencing heart attacks

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

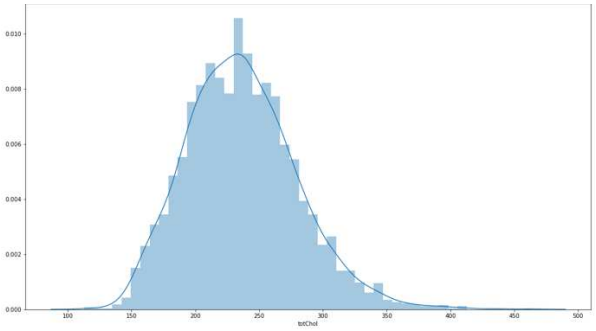

Figure 16: Graphical representation for cholesterol level w.r.t. possibility of heart attack

As per the above graphical representation for cholesterol level (totChol level) in this study, i.e., cholesterol range vs heart attack possibilities (as X-Y graph), it is found that the possibility of occurrence of heart attack due to cholesterol is approaching the peek from range about >200mg/dl while the highest scenario is at cholesterol level 250mg/dl which underlie the most prone range of occurrence of heart attack(i.e. >200mg/dl – 280mg/dl).

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

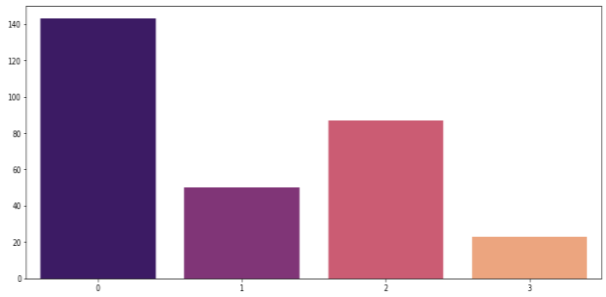

Figure 17: Chest pain types prone to heart attacks

Here, different types i.e. 0, 1, 2, 3 are as categorised into four types from 0 to 3 defining: 0 = typical angina, 1 = atypical angina, 2 = non-anginal pain, and 3 = asymptotic. As per this graph-plot is is found that people having chest-pain of type-0 i.e., typical angina are more prone to heart attacks with respect to others where as people having chest pain of type-2 i.e., non-anginal pain are mild prone to heart attacks.

v. For prevention one should always go for medical check of their cholesterol level, heart rate, blood pressure level on regular basis and also regular meditation is to be done by any cardiac patient. And for reducing the chances for future occurrence of heart attacks one must consult with specialist and take required medications regularly along with all preventions.

VII. CONCLUSION

In this research, various Supervised ML classifiers namely, Random Forest, Decision Tree, Gradient Boosting, and Logistic Regression have been used to deploy a model for Myocardial Infarction prediction. In spite of having many inconsistencies in both the datasets, various feature transformers have been used to improve the consistencies of datasets and to attain an average accuracy equal to 85.5% and a recall rate of 82% in Framingham dataset based on Gradient Boosting classifier, and more enhancing the recall rate reaches to 89.1% when training of Gradient Boosting classifier over UCI Heart dataset. After these all approach our final model is deployed for cardiac arrest prediction using Gradient Boosting Classifier. We can additionally diversify this research integrating semi-supervised and deep learning techniques.

The results impersonate that the Gradient Boosting classifier is achieving the highest accuracy score in such a way that prediction used by our model is of binary form in where 1 means a chance of heart attack and 0 means no chance. Some of the most influential attributes are chest pain type among which the typical angina is the most influential and asymptotic chest pain is least, cholesterol level in which the level greater than 200mg/dl are more prone, increased heart rate, thal, and age. It is concluded that premature heart attack is preventable in 80% of the total cases just by using a healthy diet along with regular exercises and not using tobacco products also the person who drinks more than 5 glasses of water daily are less likely to develop attacks. The medical checkup of Blood-pressure level, cholesterol level and heart rate on daily basis along with meditation can help you prevent the major heart attacks.

REFERENCES

1. H. Animesh, K. M. Subrata, G. Amit, M. Arkomita, and A. Mukherje, “Heart Disease Diagnosis and Prediction Using Machine LearAnimesh, Hazra et al. 2017. ‘Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review.’ Advances in Computational Sciences and Technology 10(7): 2137–59. ht,” *Adv. Comput. Sci. Technol.*, vol. 10, no. 7, pp. 2137–2159, 2017, [Online]. Available: http://www.ripublication.com.
2. H. S. Buttar, T. Li, and N. Ravi, “Prevention of cardiovascular diseases: Role of exercise, dietary interventions, obesity and smoking cessation,” *Exp. Clin. Cardiol.*, vol. 10, no. 4, pp. 229–249, 2005.
3. I. D. Mienye, Y. Sun, and Z. Wang, “An improved ensemble learning approach for the prediction of heart disease risk,” *Informatics Med. Unlocked*, vol. 20, p. 100402, Jan. 2020, doi: 10.1016/J.IMU.2020.100402.
4. M. Hortmann *et al.*, “The mitochondria-targeting peptide elamipretide diminishes circulating HtrA2 in ST-segment elevation myocardial infarction,” *Eur. Hear. J. Acute Cardiovasc. Care*, vol. 8, no. 8, pp. 695–702, 2019, doi: 10.1177/2048872617710789.
5. Mankad R; Staff Mayoclinics, “Heart attack,” *Mayoclinic*, 2020. https://www.mayoclinic.org/diseases-conditions/heart-attack/symptoms-causes/syc-20373106.
6. P. Severino *et al.*, “Ischemic heart disease pathophysiology paradigms overview: From plaque activation to microvascular dysfunction,” *Int. J. Mol. Sci.*, vol. 21, no. 21, pp. 1–30, 2020, doi: 10.3390/ijms21218118.
7. 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.
8. 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.
9. T. Mü nzel, M. R. Miller, M. Sørensen, J. Lelieveld, A. Daiber, and S. Rajagopalan, “Reduction of environmental pollutants for prevention of cardiovascular disease: it’s time to act,” doi: 10.1093/eurheartj/ehaa745.
10. M. Ferrante *et al.*, “Air Pollution in High-Risk Sites–Risk Analysis and Health Impact,” in *Current Air Quality Issues*, InTech, 2015.
11. A. W. R. N. Kandola, “Types of heart attack: What you need to know,” *Medical News Today*, 2018. https://www.medicalnewstoday.com/articles/321699.
12. H. Yasue, Y. Mizuno, and E. Harada, “Coronary artery spasm-Clinical features, pathogenesis and treatment-,” *Proc. Japan Acad. Ser. B Phys. Biol. Sci.*, vol. 95, no. 2, pp. 53–66, 2019, doi: 10.2183/pjab.95.005.
13. G. D. Sandler, David A and Aspenson, D Erik and Johnsen, “Oklahoma Heart Institute,” *Citeseer*, vol. 2, no. 1, 2005.
14. R. Fass and S. R. Achem, “Noncardiac chest pain: Epidemiology, natural course and pathogenesis,” *J. Neurogastroenterol. Motil.*, vol. 17, no. 2, pp. 110–123, 2011, doi: 10.5056/jnm.2011.17.2.110.
15. M. S. Ellulu, I. Patimah, H. Khaza’ai, A. Rahmat, Y. Abed, and F. Ali, “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.
16. R. Hajar, “Risk factors for coronary artery disease: Historical perspectives,” *Hear. Views*, vol. 18, no. 3, p. 109, 2017, doi: 10.4103/heartviews.heartviews_106_17.
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 machine learning%2C classification refers,one of the known characters.
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. 06, no. 02, 2017, doi: 10.4172/2167-1168.1000393.
40. K. Raza, "Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule," in *U-Healthcare Monitoring Systems*, Elsevier, 2019, pp. 179–196.
41. R. Chitra, "Heart Attack Prediction System Using Fuzzy C Means Classifier," *IOSR J. Comput. Eng.*, vol. 14, no. 2, pp. 23–31, 2013, doi: 10.9790/0661-1422331.
42. O. Opeyemi and E. O. Justice, "Development of Neuro-fuzzy System for Early Prediction of Heart Attack," *Int. J. Inf. Technol. Comput. Sci.*, vol. 4, no. 9, pp. 22–28, Aug. 2012, doi: 10.5815/ijitcs.2012.09.03.
43. A. Janosi, S. William, M. Pfisterer, and R. Detrano, "UCI Machine Learning Repository." 1998.

AUTHORS PROFILE

Suraj Kumar Gupta, is a final year student pursuing B.Tech. in Computer Science Engineering at Mahatma Gandhi Central University Bihar. He has keen interest in problem solving and machine learning specially in fields of data and predictive analytics and has shown great dedication towards this project. He has interned with Pantech Prolabs Pvt. Ltd.

Aditya Shrivastava, is a final year student pursuing B.Tech. in Computer Science and Engineering at Mahatma Gandhi Central University Bihar. He has keen interest in problem solving and machine learning specially in fields of natural language processing and computer vision. He has previously interned with Remark Skill Pvt. Ltd.

Satya Prakash Upadhyay, is working as a Registrar in Central University Gujarat. He has more than 20 years administrative, 10 years teaching and research experience. He has published various research paper and books in reputed journal. His area of interest is Data Mining and Cloud Computing.

Pawan Kumar Chaurasia, is currently working as Associate Professor in the Department of Computer Science and Information Technology, Mahatma Gandhi Central University, Motihari, Bihar. He has 15 years teaching and research experience. His area of interest Software Engineering, Machine Learning, Data Mining and Blockchain.