Heart Attack Prediction using Machine Learning and XAI

by

Mumtahina Ahsan
20266025

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
M.Engg. in Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University
September 2022

© 2022. BRAC University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is my own original work while completing degree at BRAC University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. I have acknowledged all main sources of help.

Student's Full Name & Signature:

Mumtahina Ahsan

20266025

Approval

The project titled "Heart Attack Prediction using Machine Learning and XAI" submitted by

Mumtahina Ahsan (20266025)

Of Summer, 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of M.Engg. in Computer Science and Engineering on September 18, 2022.

Examining Committee:

Supervisor:
(Member)

Md. Golam Rabiul Alam, PhD

Professor
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

Amitabha Chakrabarty, PhD

Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Sadia Hamid Kazi

Chairperson and Associate Professor
Department of Computer Science and Engineering
BRAC University

Abstract

Predictive analytics has received a lot of attention in recent years due to advances in supporting technology, particularly in the areas of big data and machine learning. In recent years, the uses of disease prediction has been seen in the healthcare area. Among so many predictions, this project will show the prediction of a heart attack. Heart disease, often known as cardiovascular disease, refers to a variety of illnesses that affect the heart and has become the leading cause of mortality worldwide in recent decades. It links a slew of risk factors for heart disease with a pressing need for precise, dependable, and practical methods for making an early diagnosis and managing the condition. In the healthcare industry, data mining is a typical methodology for analyzing large amounts of data. Because predicting cardiac illness is a difficult undertaking. It is necessary to automate the process in order to avoid the risks connected with it and to inform the patient well in advance. Heart diseases can be determined using data mining techniques such as XGBOOST, Logistic Regression, Stochastic Gradient Descent, Support Vector Classifier, Kneighborsclassifier, and Naive Bayes. With this project, I have shown that among all the above machine learning models, XGBOOST outperforms other techniques in terms of predicting heart attacks. As a result, this paper conducts a comparative study of the performance of several machine learning algorithms. For any type of prediction features of the dataset plays a very important role. Features can give positive or negative impact on the final prediction. The features importance can be visualized by the XAI methods. This paper also takes an approach to interpret the explainability of the model's prediction. By using the XAI method SHAP and LIME with the help of the concept of black box, this research conducts the KNN algorithms prediction.

Keywords: Heart Attack; ML (Machine Learning); XAI (Explainable Artificial Intelligence); SHAP (SHapley Additive exPlanations); Shapley Value; LIME (Local Interpretable Model-Agnostic Explanations); Black-Box; XGBoost; KNN

Dedication (Optional)

This thesis is dedicated to all the individuals surviving and battling with heart disease. I hope this work may provide a significant impact to the healthcare area for predicting the heart disease accurately.

Acknowledgement

At first, I would like to thank Almighty ALLAH for helping me and making me able to complete my project work at the time successfully.

I would also like to show my gratitude to my supervisor Dr. Golam Rabiul Alam Sir for allowing me to do my project under his supervision. Sir encouraged me to do my work on my thinking capability. I am so grateful to Sir for his guidance.

I also thanked my friends and my well-wishers for helping me.

And finally, I should show my acknowledgment to my respected parents and my beloved sister for believing me and always stand by me.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

ECG	Electrocardiogram
FN	False Negative
FP	False Positive
$MIMIC$	Medical Information Mart for Intensive Care
TN	True Negative
TP	True Positive
UCI	University of California Irvine

Chapter 1

Introduction

Damage to the heart muscle due to inadequate blood flow to a part of the heart is called a heart attack [1]. Given that heart attacks are one of the primary causes of unexpected mortality, particularly in developing nations, heart attack prediction is crucial. Cardiologists can forecast heart attacks using standard clinical techniques like electrocardiography and blood testing, but they can also use computer-aided diagnosis systems that employ machine learning techniques.

1.1 Motivation

Heart diseases cover a wide range of conditions that affect the heart. Some types of cardiac disorders include coronary heart disease, cardiomyopathy, and cardiovascular disease. The term ”cardiovascular disease” refers to a broad spectrum of disorders that have an impact on the heart, blood arteries, and how the body pumps and circulates blood. Cardiovascular disease (CVD) results in severe illness, disability, and death [2]. Coronary heart disease is brought on by the narrowing of the coronary arteries, which reduces the amount of blood and oxygen delivered to the heart (CHD). The CHD includes myocardial infarctions, also referred to as heart attacks, and angina pectoris, or chest pain. A heart attack is caused by the sudden blockage of a coronary artery, which is typically brought on by a blood clot. When the blood supply to the heart muscles is insufficient, chest torments emerge [2].

According to the World Health Organization, Cardiovascular Diseases (CVD) accounts for 31% of world deaths, with heart attacks and stroke accounting for 80% of these [3]. Heart disease is estimated to kill around 23.6 million people worldwide by 2030. Heart disease is caused by a variety of variables, including personal and professional behaviors, as well as hereditary predisposition. Predisposing factors for heart disease include smoking, excessive alcohol and caffeine use, stress, and physical inactivity, as well as physiological variables such as obesity, hypertension, high blood cholesterol, and pre-existing cardiac diseases. The ability to diagnose the cardiac disease quickly, accurately, and accurately plays a critical role in adopting preventative actions to avoid death.

In healthcare, it is a very difficult task to enrich the efficiency of prediction and make it more accurate. Machine learning algorithms are one of the most beneficial techniques for addressing these difficulties. However, additional issues develop when it comes to heart diseases, which might be classified as complex systems.

Due to the potential of AI (Artificial Intelligence) to improve efficiency, consistency, and accuracy in decision making, it is widely spread out in many application areas. Although the AI algorithms seem to be very powerful in predicting and becoming more and more popular, most of these models are considered as black box to the user, opaque and difficult to understand their inner workings. Applications of artificial intelligence systems are ever-expanding, using XAI (eXplainable Artificial Intelligence) technology explains why machine learning models making certain predictions is just as important as the accuracy of the predictive, as it ensures trust and transparency in the decision-making model. In recent days, even experts struggle to explain complex AI models, such as ensemble models or deep learning models. There is a great gap between accuracy and interpretability. In the presence of this challenge, XAI is proposed and utilized to make AI systems more transparent. There is a large gap between accuracy and interpretability. In the presence of this challenge, XAI is proposed and utilized to make AI systems more transparent [4].

Nowadays, there are many interpretable patterns, most of which belong to local and agnostic models. The agnostic models are visualization methods, knowledge extraction, influence methods, and example-based explanations [4]. The representation of visualization methods is usually adopted, in order to understand ML models. The method of knowledge extraction in the agnostic model is popular at present [5]. By using the SHAP method, the Shapley value has been used to calculate the importance of the features in the ML model.

Nowadays AI and ML techniques have become the most significant elements in healthcare for making decisions and predicting accuracy. However, most widely used ML models are opaque. In the health sector, sometimes the binary “yes” or “no” answer is not sufficient. Sometimes questions like “how” or “where” something occurred are more influential [6].

Local Interpretable Model-Agnostic Explanations (LIME) faithfully explains the predictions of any classifier by fitting a local linear regression model around the prediction. The sampling operation for LIME is a homogeneous random distribution remembrance, simple but flawed, omitted correlation between features [7].

1.2 Contribution

Machine Learning is a very effective testing method that is based on training and testing. It is a branch of Artificial Intelligence (AI), which is a large area of learning in which machines mimic human abilities. Machine learning is a subset of AI. Machine learning systems, on the other hand, are taught how to interpret and use data, hence the combination of the two technologies is known as Machine Intelligence [8]. The gradient boosting algorithm was created with the goal of having a very high predictive power. However, because the technique requires just one decision tree to be generated at a time in order to reduce the errors of all prior trees in the model, its use has been limited. As a result, even the smallest models required a significant amount of time to train. Then came a new algorithm called eXtreme

Gradient Boosting (XGBoost) which changed the way gradient boosting was done. In XGBoost, individual trees are created using multiple cores, and data is organized in order to minimize the lookup times.

The goal of this research is to discover the most effective machine learning system for detecting cardiac problems. For predicting heart disease, this study compares the accuracy scores of the XGBOOST, Logistic Regression, Stochastic Gradient Descent, Support Vector Classifier, Kneighborsclassifier, and Naive Bayes algorithms. Amongst all the algorithms that we tested in our experiment, XGBoost outperformed all the algorithms with an accuracy of 86.89%.

SHAP (SHapley Additive exPlanations) is a mathematical method that explains if the prediction of a machine learning algorithm is fairly distributed among the individual feature. Shapley values tell us the average contribution of a feature or not. SHAP is generally a local explanation technique. So that means, it aims to explain individual predictions of black box models.

The contribution of using the SHAP method applied in this paper shows, the importance of individual features of agnostic models. As SHAP method usually preform best on the regression model, in this paper I have applied SHAP method to the KNN machine learning algorithm.

In healthcare, it is however, desirable to obtain an instance-specific explanation. This allows for more individualized decision-making, thus providing the patient with more personalized care [9].

LIME is a recently developed framework that can be used with any BlackBox classification model in order to obtain an explanation for one specific instance. It works by providing a local explanation for the classification and the minimum number of features that contribute to the maximum likelihood of the outcome for a given class of observations [10].

The contribution of using the LIME method applied in this paper describes, how the KNN algorithm predicted the heart attack accuracy.

1.3 Organization of the Report

The report is structured in the following manner - In Chapter 2, I shall try to explain the concepts related to the subject matter in detail and also a walkthrough of the previous studies. In Chapter 3, I shall briefly explain the methodology and the framework architecture along with the dataset preparation. Afterward, in Chapter 4, the implementation and final outcome of the results will be discussed. Finally, I'll wrap up the conclusion and some limitations along with future scopes of this research in Chapter 5.

Chapter 2

Literature Review

In this chapter, I'll discuss some of the related works that have direct or indirect impacts on the paper.

2.1 Existing Works related to the Heart Attack Prediction

Using the UCI Machine Learning dataset, a lot of work has been done to predict cardiac disease. Various data mining approaches have been used to achieve different levels of accuracy, which are detailed below.

In [11], the authors talked about various different ML algorithms that can be used for the classification of heart disease. The accuracy of the Decision Tree, KNN, and K-Means algorithms that can be utilized for classification was compared in this study. The accuracy obtained by Decision Tree was found to be the highest, and it was concluded that it may be made more efficient by combining multiple methodologies and parameter adjustments. In paper [12], the authors presented a system that used data mining techniques in conjunction with the MapReduce algorithm. According to this article, the accuracy obtained for the 45 instances of the testing set was higher than that obtained using standard fuzzy artificial neural networks. Because of the employment of a dynamic schema and a linear model, the accuracy of the method was increased.

Authors of [13] have designed an ML model comparing five different algorithms. When compared to Matlab and the Weka tool, the Rapid Miner tool produced greater accuracy. The accuracy of Decision Tree, Logistic Regression, Random Forest, Naive Bayes, and SVM classification algorithms was compared in this study. The most accurate algorithm was the decision tree algorithm.

Paper [14] proposes a system for disease prediction that employs NB (Nave Bayesian) approaches for dataset classification and the AES (Advanced Encryption Standard) algorithm for safe data communication.

In [15], a survey was done with several classification algorithms to be used for predicting heart diseases. Naive Bayes, KNN (K-Nearest Neighbour), Decision tree, and Neural network were employed as classification techniques, and the accuracy of the classifiers was evaluated for a variety of variables.

The prediction of heart disease using Naive Bayes classification and SVM(Support Vector Machine) was done by the authors in [16]. The performance measures employed in the analysis are Mean Absolute Error, Sum of Squared Error, and Root Mean Squared Error, and it has been determined that SVM outperforms Naive Bayes in terms of accuracy.

The initial approach of the paper [2] was to design and deploy an efficient model for extracting patterns, which are significant to heart attacks, from the heart disease data warehouses. According to their approach, they aimed to utilize the frequent pattern mining and clustering data mining techniques.

[17] proposed heart disease prediction by using the aid of neural networks.

Paper [18] proposed a novel heuristic for efficient computation of sparse kernel in SUPANOVA. It is used for benchmarking the Boston Housing Market Dataset and socially significant issues. The problem of improving heart disease detection Population using new non-invasive methods magnetic measurement of heart activity field produced by the human heart 83.7% of predictions. The result is correct and therefore better than the result obtained by SVM and equivalent kernels.

A bagging algorithm and J48 decision tree algorithm were used by [19] for the heart attack prediction. Their research showed that the bagging algorithm provides better results than the decision tree algorithm.

Furthermore, to predict heart attacks [1] used classification algorithms such as rule based decision trees, naive Bayes, and ANN (Artificial Neural Network). In their experiments, they used the One Dependency Augmented Naive Bayes Classifier (ODANB) and Naive Creedal Classifier 2 (NCC2) for data preprocessing.

Much more research related to this study, such as: using naiıve Bayes, decision trees, and neural net classifiers for stroke crisis; a K-means clustering algorithm with the decision tree for heart attack prediction; ECG signals for heart attack prediction are proposed by [20], [21], [22] respectively.

The main idea behind the proposed system after reviewing the above papers was to create a heart disease prediction system based on the inputs. This study analyzed the classification algorithms namely XGBOOST, Logistic Regression, Stochastic Gradient Descent, Support Vector Classifier, Kneighbors classifier, and Naive Bayes based on their Accuracy, Precision, Recall, and f-measure scores and identified the best classification algorithm which can be used in the heart disease prediction.

2.2 Existing Works related to SHAP

There are some related papers that explained how SHAP provides a visual representation of feature importance. Some of those works are discussed below.

Paper [23] used the SHAP algorithm for explaining the internal layers of a DNN (deep neural network) classifier to discriminate between normal and adversarial inputs.

The authors of [24] used the MIMIC-III clinical database to predict extubation failure by applying the LightGBM (Light Gradient Boosting Machine). They performed the analysis of feature importance and the visualization of important features by applying the SHAP method.

On the other hand, paper [25] clarified various limitations of the SHAP method. This work aimed at the constraints of Shapley value for feature selection.

For the first time, paper [4] deployed the application of SHAP in the DRL (deep reinforcement learning) model of the power system. However, this work is unique in the power system field, for presenting the first use of the SHAP method. The probability of SHAP value gives explanations for emergency control based on DRL in the power system.

A dynamic pricing mechanism of cross-node on-demand bandwidth based on Shapley value has been proposed by [26]. This paper used the Shapley value which has the interpretation ability to assign a specific predicted important value to each feature.

In many recent works, the implementation and the use of SHAP have been seen. [27] proposed a SHAP method based on the Shapley value method. SHAP method can not only be used for global explanations, but also for local explanations, which can be adopted to explain individual predictions.

Utilization of SHAP in various applications has been seen in recent years. [28] discovered the importance of biomarkers. This is helpful for accurate diagnosis and prediction of certain cancer types, which uses gradient enhancement trees and the SHAP method.

In addition, paper [29] proved that the SHAP library is not only actually built for feature performance, but can also use for Dimensionality Reduction with an extra parameter of confidence.

Three methods of SHAP (TreeSHAP, KernelSHAP, and DeepSHAP) are employed in the paper [30], to explain the predictions of classifiers that are applied in the RFC (Random Forest Classifier), XGBoost, and the Sequential model.

Research presented by [31], used the Deep Feed Forward (DFF) and Random Forest (RF) models to classify their data set. Also, they used the SHAP technique to explain the importance and influence of the characteristics of the data set.

In [32], researchers have proposed a method to detect valuable warnings about network security threats by using SHAP to determine the selected important features based on the SHAP plots.

Furthermore, [33] indicates that KernelSHAP is used to interpret the expectations of the RNN models and the average calculation time is less than ten minutes. Therefore, the explanations are analyzed at different scales relevant clinical outcomes are expected.

The main objective of this study is to assess the performance of multiple ML models to detect the prediction of heart attack and as well as to analyze the importance of individual features for heart attack prediction using SHAP.

2.3 Existing Works related to LIME

LIME is a technique, that attempts to comprehend the model by understanding how the prediction of any model works. Some works related to this topic are represented below.

The authors of [34], proposed the Local Interpretable Model-agnostic Explanation (LIME), which explains the predictions of any classifier faithfully by fitting a linear regression model locally around the prediction.

Authors in [35], determined the Detection of COVID-19 findings by the local interpretable model-agnostic explanations method of types-based activations extracted from CNNs. Their proposed approach was based on pre-trained CNN models. By using the LIME method they aimed to perform a successful classification of the activation sets extracted for each type.

Also, the paper [36] explained the use of LIME in precision medicine and discussed the importance of interpretability to understand the contribution of important features in decision-making.

[37] expanded the LIME method by using the concept of decision rules. In the same vein, Leave-One-Covariate-Out (LOCO) and Local Rule-based Explanation (LORE) are other popular techniques for generating local explanation models that offer local variable importance measures make-03.

An approach proposed by [38] explained that, instead of perturbing the data points, partition the dataset using k-means around an instance. As the default implementation of k-means picks a centroid randomly, it makes this approach non-deterministic. A similar approach was proposed in [39] by using a supervised tree-based approach for partitioning the dataset.

Lack of "stability" is one of the issues of the existing locally interpretable models. It was identified by [40] as "explanation level uncertainty". The authors of this paper showed that explanations generated by different locally interpretable models have an amount of uncertainty associated with them due to the simplification of the black box model.

To deal with this issue, [41] proposed a deterministic version of LIME, named DLIME. Their work successfully proved that DLIME determined stability and faithfulness compared to LIME.

In another paper, a method to decompose the predictions of a classifier on the individual contribution of each feature was proposed by [42]. The working process of this paper was based on the computation of the difference between original predictions and predictions made by eliminating a set of features.

In order to obtain an explanation for one specific instance [9] explained LIME framework can be used with any black box classification. In healthcare, it is very beneficial for making more individualized decisions, thus providing the patient with more personalized care.

Another approach of [43] was said while making the prediction to make a global explanation by listing what features are generally more important.

Furthermore, authors of the paper [6] offered the Deterministic Local Interpretable Model-Agnostic Explanations (DLIME) framework, which uses Hierarchical Clustering (HC) to partition the dataset into different groups.

The paper [7], proposed a Modified Perturbed Sampling method for LIME known as (MPS-LIME), which takes into full account the correlation between features. In terms of understandability, fidelity, and efficiency their offered method MPS-LIME showed significant performance over LIME.

By considering all the previous works, this paper will explain the prediction of the heart attack accuracy of the machine learning model by using LIME.

Chapter 3

Proposed Methodology

In this section, the approaches of this paper will be shown step by step.

3.1 Dataset Collection

The dataset that I have used here in order to train the model has been extracted from the Cleveland database. The names and social security numbers of the patients were recently removed from the database, and replaced with dummy values. One file has been "processed", that one containing the Cleveland database. All four unprocessed files also exist in this directory. In particular, the Cleveland database is the only one that has been used by Machine Learning researchers to this date.

3.2 Dataset Description

The Cleveland database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. There are 13 features in this dataset that directly increases the possibility of having a heart attack. The "output" field refers to the presence of heart disease in the patient. It is integer-valued "0" is no/less chance of heart attack and "1" is more chance of heart attack. The description of the attributes included in this dataset are as follows:

- age : Age of the patient
- Sex : 1 = male; 0 = female
- exng: exercise induced angina (1 = yes; 0 = no)
- oldpeak: ST depression induced by exercise relative to rest
- slp: the slope of the peak exercise ST segment (2 = upsloping; 1 = flat; 0 = downsloping)
- thall: 2 = normal; 1 = fixed defect; 3 = reversible defect.
- caa: number of major vessels (0-3)
- cp : 4 types of chest pain
- Value 0: typical angina

- Value 1: atypical angina
- Value 2: non-anginal pain
- Value 3: asymptomatic
- trtbps : resting blood pressure (in mm Hg)
- chol : cholesterol in mg/dl fetched via BMI sensor
- fbs : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- restecg : resting electrocardiographic results
- Value 0: normal
- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
- thalachh : maximum heart rate achieved
- output: target : 0 = less chance of heart attack 1 = more chance of heart attack

3.3 Data Preprocessing

In this paper, I have splitted the dataset and used 20% data for testing and 80% data for training from the whole dataset to perform the accuracy prediction of multiple machine learning algorithms.

3.4 Features of Heart Disease Dataset

Table 3.1 shows the list of all the features, that are present in the Heart Disease dataset and used to implement the models mentioned in this paper.

No	Feature Name	No	Feature Name
1	age	8	caa
2	sex	9	trtbps
3	exng	10	chol
4	oldpeak	11	fbs
5	slp	12	restecg
6	thal	13	thalachh
7	caa		

Table 3.1: List of features

3.5 System Architecture for Predicting A Heart Attack by using Multiple ML Algorithms

The main idea behind the proposed system architecture was to create a heart disease prediction system based on the inputs as shown in Fig 3.1. For predicting the accuracy of heart attack, this study analyzed the classification algorithms namely XGBoost, Logistic Regression, Stochastic Gradient Descent, Support Vector Classifier, KNeighbors Classifier and Naive Bayes based on their Accuracy, Precision, Recall and F-measure scores. And after that, identified the best classification algorithm which can be used in heart disease prediction.

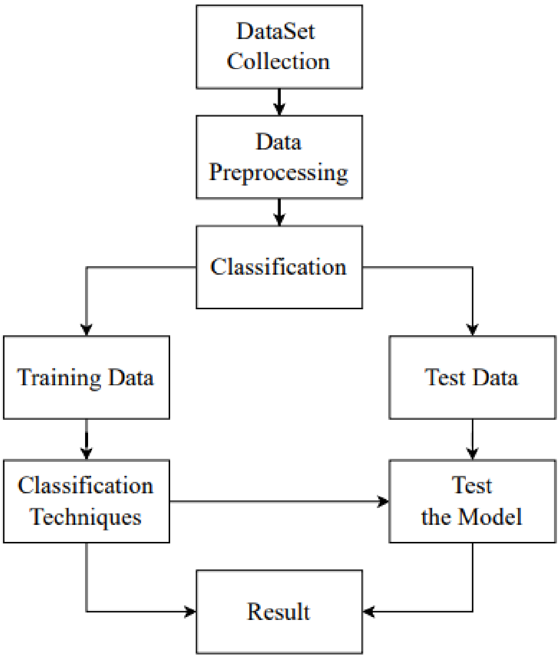


Figure 3.1: Process Flow for Predicting Heart Attack

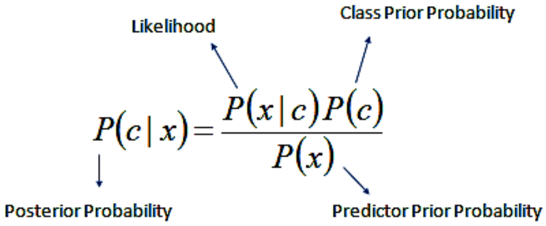
3.5.1 Naive Bayes Classifier

The attributes mentioned in Table 3.1 are provided as input to different algorithms such as the Naive Bayes Classifier. To give a brief introduction of the Naive Bayes algorithm we can say that it is a collection of classification algorithms based on the Bayes Theorem and it is the most popular and simple machine learning classification algorithm as stated in [3]. We cannot call it a single algorithm because it is a family of algorithms where all of them share a common principle. Bayes theorem provides a way of calculating posterior probability $P(c|x)$ from $P(c)$, $P(x)$, and $P(x|c)$ as shown in the equation below:

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times ... \times P(x_n|c) \times P(c) \tag{3.1}$$

Here,

- $P(c|x)$ is the posterior probability of class (c, target) to the given predictor (x, attributes).



- $P(c)$ is the prior probability of class
- $P(x)$ is the prior probability of the predictor
- $P(x|c)$ is the likelihood which is the probability of the predictor given class.

There are three types of Naive Bayes model and we have used the Gaussian Naive Bayes Classifier. The Gaussian Naive Bayes is used in classification and it assumes that features follow a normal distribution.

3.5.2 XGBoost

On classification, regression, and predictive modeling challenges, Extreme Gradient Boosting (XGBoost) has been seen to dominate structured or tabular datasets [44]. XGBoost is a scalable and improved version of the gradient boosting algorithm designed for efficiency, computational speed, and model performance. XGBoost combines a convex loss function (based on the difference between the anticipated and target outputs) with a penalty term for model complexity to minimize a regularized (L1 and L2) objective function (in other words, the regression tree functions). The process is repeated iteratively, with new trees being added that anticipate the residuals or errors of previous trees, which are then combined with previous trees to make the final prediction. It's called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models.
$Fm(x)$ should be a function that minimizes the loss function or MSE (mean squared error), in this case:

$$F_m(X) = F_{m-1}(X) + \alpha_m \times h_m(X, r_{m-1}) \tag{3.2}$$

In equation 3.2, α_i, and r_i are the regularization parameters and residuals computed with the i^th tree respectfully, and h_i is a function that is trained to predict residuals, r_i using X for the i^th tree. To compute α_i, we use the residuals computed, r_i compute the following:

$$\arg\ min_\alpha = \sum_{i=1}^{m} L(Y_i, F_{i-1}(X_i) + \alpha h_i(X_i, r_{i-1}) \tag{3.3}$$

Here, $L(Y, F(X))$ is a differentiable loss function.

3.5.3 Logistic Regression

The classification algorithm logistic regression is mostly used for binary classification problems. Instead of fitting a straight line or hyperplane, the logistic regression algorithm squeezes the output of a linear equation between 0 and 1 using the logistic function. Because there are 13 independent variables, logistic regression is a great choice for categorization. The central mathematical concept that underlies logistic regression is the logit—the natural logarithm of an odds ratio [45]. The simplest form of the logistic regression model can be written as:

$$Logit(Y) = ln\frac{\pi}{\pi - 1} = \beta_0 + \beta_1 \times X \tag{3.4}$$

Here, π is the probability of occurring the outcome Y and $\pi/(1 - \pi)$ is the odds of success the ratio of the probability of occurring the outcome Y and the probability of not occurring the outcome Y. β_0 and β_1 are called intercept and slope (regression coefficient) respectively. The logistic model can be extended for more than one predictor:

$$Logit(Y) = ln\frac{\pi}{\pi - 1} = \beta_0 + \beta_1 \times X_1 + ... + \beta_p X_p \tag{3.5}$$

3.5.4 Stochastic Gradient Descent

In simple words, gradient refers to the slope or slant of a surface. So, descending a slope to reach the lowest point on that surface is known as gradient descent. There are three types of Gradient Descent and one of them is the stochastic gradient descent. The term "stochastic" refers to a system or process that is linked to random probability. As a result, instead of selecting the entire data set for each iteration in Stochastic Gradient Descent (SGD), a few samples are chosen at random. As stated in [46], SGD performs a parameter update for each training example $x^{(i)}$ and label $y^{(i)}$:

$$\theta = \theta - \eta.\nabla_\theta \times J(\theta; x^{(i)}; y^{(i)}). \tag{3.6}$$

Because only one sample from the dataset is picked at random for each iteration in SGD, the path taken by the algorithm to reach the minima is usually noisier than the path taken by a traditional Gradient Descent method.

3.5.5 Support Vector Classifier

Another simple approach that any machine learning expert should know about is the support vector machine (SVM). Many people prefer the support vector machine because it produces great accuracy while using fewer computing resources. SVM is a type of algorithm that may be used for both regression and classification. However, it is extensively used in classification objectives. In the SVM algorithm, each data point is plotted as a point in n-dimensional space (where n is the number of features), with the value of each feature being the value of a certain coordinate in the SVM algorithm. Then the classification is accomplished by locating the hyper-plane that clearly distinguishes the two classes.

3.5.6 KNeighbors Classifier

K-means clustering is one of the clustering techniques used to cluster datasets based on nearest-neighbor. Here the data is clustered in k clusters based on a similarity between them. We can also fill missing values of data using k-means [20]. The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It's easy to implement and understand but has a major drawback of becoming significantly slower as the size of that data in use grows. KNN works by finding the distances between a query and all the examples in the data, selecting the specified number of examples (K) closest to the query, then voting for the most frequent label (in the case of classification) or averages the labels (in the case of regression).

3.6 Working Process for Explaining Prediction of Heart Attack by using SHAP

In the methodology recommend in this thesis for explaining the use of SHAP, first I have trained all the machine learning models based on the existing dataset. Then, calculate the prediction accuracy of all these models, and after that picked the model which gives the highest accuracy. At last, then would do the SHAP analysis on that machine learning model using the SHAP library. After the analysis is done, the Shapley value of each and every feature is obtained and returned as a list in python. This working process showed in Fig 3.2.

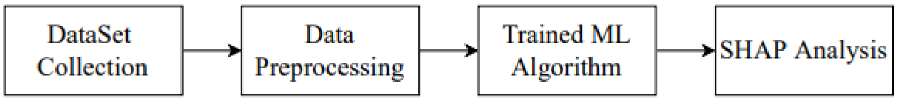


Figure 3.2: Process Flow for Explaining Heart Attack Prediction using SHAP

3.6.1 XAI Method

XAI which stands for Explainable AI helps in justifying why a model has made a particular choice. It makes the decision transparent by providing explanations that can be understood by non-experts. There are many XAI frameworks available. Amongst them, this thesis has used SHAP in the project.

SHAP (Shapley Additive exPlanations) proposed by Lundberg and Lee, is used to interpret the output of the model. For an individual prediction, Shapley values tell us then how this prediction is fairly distributed among the individual inputs. The XAI technique SHAP simply makes use of these Shapley values. SHAP is generally a local explanation technique. So that means, it aims to explain individual predictions of black box models. However, it is also possible to get valid global explanations by aggregating these individual predictions.

3.6.2 SHAP Method

SHAP is based on the concepts of cooperative game theory which, can be used to explain the predictions of any machine learning algorithm. the key idea behind the cooperative game theory is that there is a game in which players are cooperating to obtain an outcome or a game. The players of the game form coalitions to achieve the gain. Some players may contribute more to the coalition than others. And Shapley values are a fair way to attribute the total gain to the players based on their marginal contributions to a specific coalition. In the context of explaining machine learning explanations, the model output is the coalition game, the prediction is the gain or the payout, and the input features are the players that cooperate with each other to help the model arrive at a prediction.

In SHAP, the contribution of each feature (ϕ_i is the contribution of feature i) on the model output $v(N)$ is allocated based on their marginal contribution. Shapley value can be defined as the average marginal contribution of a feature value over all possible coalitions. The below formula is used to derive the Shapley value for a feature: To get the Shapley value for feature i, marginal contribution of the feature

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(N-|S|-1)!}{N!}(v(S \cup \{i\}) - v(S))$$

$[v(S \cup i) – v(S)]$ is calculated over all the subsets S. The Shapley value approach used in this study is a type of agnostic model of feature importance method. The relevance value of each feature is determined using the Shapley value approach. The feature importance method assesses each input variable's or feature's contribution to a complicated machine learning model's functionality. The rate of model error following feature substitution is calculated to determine the significance of features. Model error rises when crucial features are replaced. At present, the Shapley value method has been used in a number of cases to calculate the importance of features in the ML model, so as to get an intuitive explanation [4].

In this paper, I have showed the SHAP method on the KNN ML algorithm. Being a regression model, it is very easy to implement SHAP on this algorithm.

3.7 Working Process for Explaining Prediction of Heart Attack by using LIME

The working process to use the LIME method of this thesis will address the explanation of the prediction of a heart attack. LIME is a kind of method that not only visualize the prediction but also gives an explanation of the individual prediction. Fig 3.3 indicates the working flow of LIME in this paper.

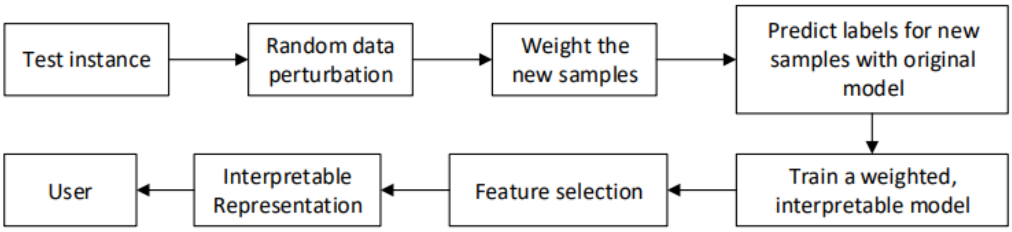

Test instance
Random data perturbation
Weight the new samples
Predict labels for new samples with original model
Train a weighted, interpretable model
Feature selection
Interpretable Representation
User

Figure 3.3: Process Flow of LIME [41]

3.7.1 LIME Framework

The ability of XAI is crucial in the healthcare industry. Machine learning and deep learning models were formerly thought of as "black boxes" that accepted some input and made a decision to produce an output, but it was unclear from which parameters these judgments were made. The necessity for Explainability in AI has increased, as a result of the growing use of AI in our daily lives. And also the decision-making capabilities of AI in situations like autonomous vehicles and cancer prediction software.

To deal with this issue the concept of black-box is very prominent. The usage of black-box in machine learning makes the machine learning models more transparent. LIME is one of the promising frameworks of the XAI method which explains how any machine learning model works.

LIME stands for Local Interpretable Model-agnostic Explanations. The basic idea of the LIME is that we just have to zoom into the local area of the individual prediction. LIME can easily create a simple explanation that makes sense in that local region. This way we do not have to worry about the rest of the model, and still get a valid explanation of why the prediction was made for an instance. Any black box model can be explained by creating such a local approximation. The complex models are complete black boxes and the internals are hidden for LIME. So, it's just based on the inputs and outputs of the model. To explain individual predictions LIME focuses on training local surrogate models Instead of training a global surrogate model [47].

Mathematically, local surrogate models with interpretability constraints can be expressed as follows [47] :

$$\text{explanation}(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

The explanation model for instance x is the model g (e.g. linear regression model) that minimizes loss L (e.g. mean squared error), which measures how close the explanation is to the prediction of the original model f (e.g. an xgboost model), while the model complexity $\Omega(g)$ is kept low (e.g. prefer fewer features). G is the family of possible explanations, for example, all possible linear regression models. The proximity measure x defines how large the neighborhood around instance x is that we consider for the explanation. In practice, LIME only optimizes the loss part. The user has to determine the complexity, e.g. by selecting the maximum number of features that the linear regression model may use. Reference.

This paper conducts the LIME method on the KNN ML algorithm.

Chapter 4

Implementation & Result Analysis

This section of the thesis will describe the findings and analyzing the results of the research.

4.1 Result Analysis of Heart Attack Prediction Using Multiple ML Algorithms

In this paper, the chance of a heart attack is predicted using different models. On Google Colab, which operates under Ubuntu 64 bits and is composed of a single core hyper threaded Intel Xeon processor@2.3 GHz and 13 GB of RAM, I performed all the computations. As Python programming has open source packages, I used those to simulate our code and experiments.

This work has used confusion matrices such as accuracy, sensitivity, specificity, and F1-score for the XGBoost Classifier algorithm. Accuracy is the percentage of total subjects classified correctly. Sensitivity is the proportion of those who do have the disease who test positive. Specificity is the proportion of those who do not have the disease who test negative. Sensitivity can also be identified as Recall. Precision is the number of subjects correctly identified as positive out of the total subjects identified as positive. F1-Score is a harmonic mean of precision and recall [48].

$$Accuracy = \frac{\sum TP + \sum TN}{\sum TP + \sum TN + \sum FP + \sum FN} \tag{4.1}$$

$$Sensitivity = Recall = \frac{\sum TP}{\sum TP + +\sum FN} \tag{4.2}$$

$$Specificity = \frac{\sum TN}{\sum TN + +\sum FP} \tag{4.3}$$

$$Precision = \frac{\sum TP}{\sum TP + +\sum FP} \tag{4.4}$$

$$F_1 = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{4.5}$$

Here, TP and FP denote the number of correctly and wrongly classified subjects having heart disease, respectively. Similarly, TN and FN denote the number of correctly and wrongly classified subjects not having heart disease, respectively [48].

The paper is shown the confusion matrix which is contained the summary of prediction results of all instances for the XGBoost Classifier of the dataset used for both testings as well as training in Fig 4.1 (a) and (b) respectively. The paper is also shown the performance of the XGBoost Classifier algorithm in Fig 4.2.

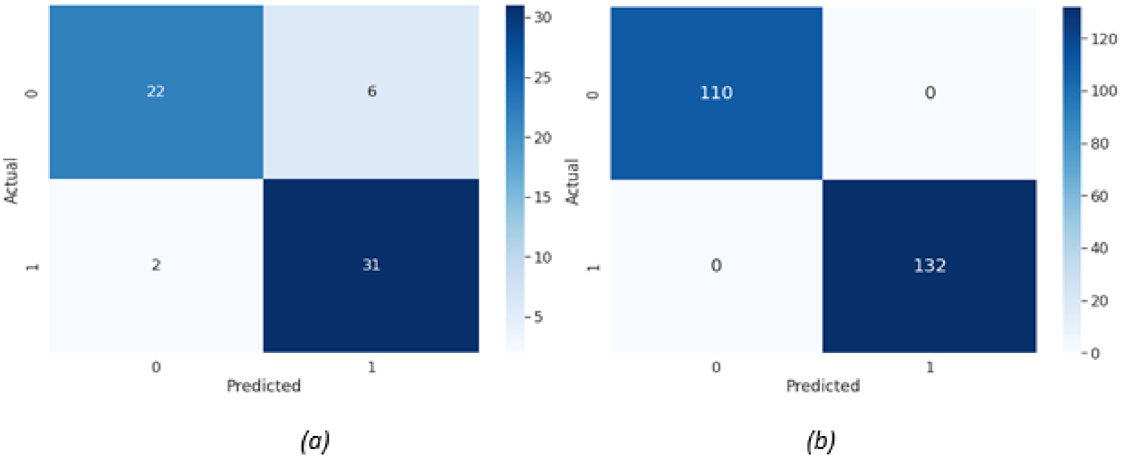
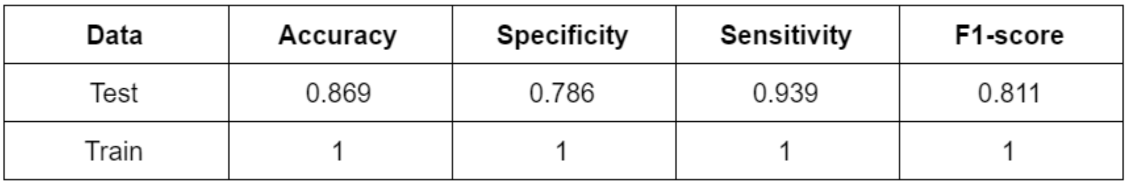


Figure 4.1: (a) Confusion Matrix of Test data prediction result; (b) Confusion Matrix of Train data prediction result



Data	Accuracy	Specificity	Sensitivity	F1-score
Test	0.869	0.786	0.939	0.811
Train	1	1	1	1

Figure 4.2: Performance of XGBoost Classifier

4.1.1 Accuracy Comparison of All Models

In the paper, the heart attack rate is predicted for some different models. We have used Logistic Regression, Stochastic Gradient Descent, Support Vector Classifier, Naive Bayes Classifier, XGBoost Classifier, and KNeighbors Classifier algorithms to find the accuracy of a heart attack. We have shown the accuracy percentage of different algorithms in Fig 4.3. Here, XGBoost Classifier gives the best performance with 86.885% accuracy among all other algorithms.

4.1.2 Feature Engineering

As a lot of features can affect the accuracy of the algorithm, that's why working with the features is very important. Here are a few reasons why some may want to work with some selected features. Choosing fewer features may help to train faster. Sometimes picking the most important features may give effective results for some algorithms. But we have used all the features from the dataset to predict the accuracy.

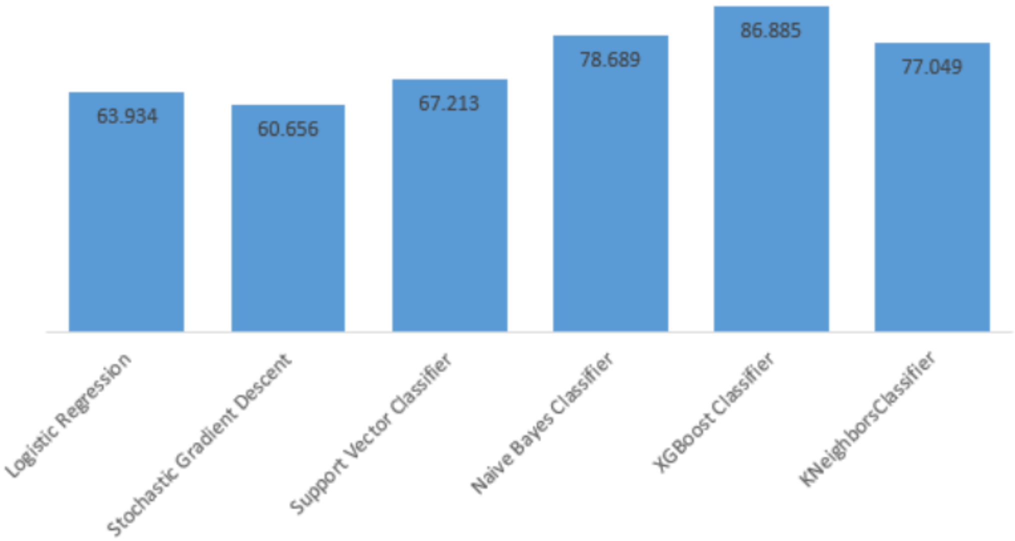


Figure 4.3: Accuracy of Different Algorithms

4.1.3 Feature Importance

A heart attack may vary depending on various circumstances. In machine learning, these circumstances are defined as features. The features that have the biggest impact on predictions are called feature importance. In some cases, a few attributes may decrease the accuracy level of a model. So, it is important to work with the correct attributes. Though we have used all the features, feature importance from the dataset is shown in Fig 4.4. As shown in Fig 4.4, the feature cp which is chest pain has the most impact on the chances of having a heart attack.

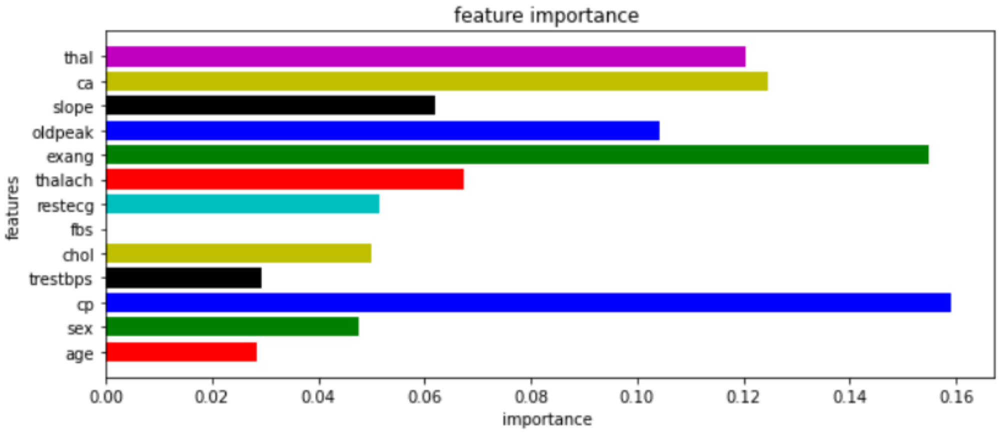


Figure 4.4: Important Features of The Dataset

4.1.4 Feature Analysis

All the features have value to count. By analyzing all the value counts we predict the rate of a heart attack. In the paper, Fig 4.5 represents the more and less chances of heart attack from value counts. It means the dataset has 54.46% of ‘1’ which predicts more chances of heart attack and 45.54% of ‘0’ which predicts fewer chances of a heart attack. All the features have value to count. By analyzing all the value counts we predict the rate of a heart attack.

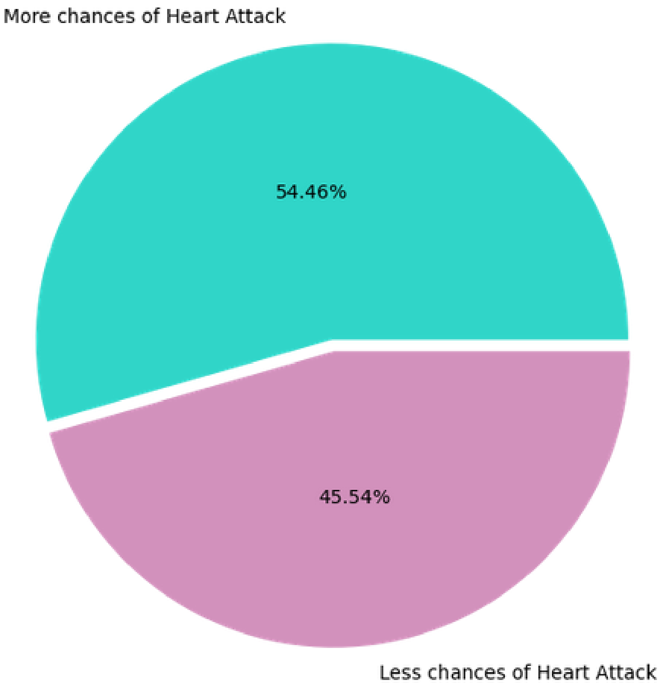


Figure 4.5: Value Counts of The Features

4.2 Result Analysis of Heart Attack Prediction Using SHAP

In this paper, the prediction of heart attacks by the KNN machine learning algorithm has been explained by the SHAP XAI method.

4.2.1 Feature Importance

The features that have the biggest impact on predictions is called feature importance. In some cases, few attributes may decrease the accuracy level of a model. So, it is important to work with the correct attributes. Though we have used all the features, feature importance from the dataset is shown in Fig 4.6. As shown in Fig 4.6, the feature "thalach" has the most impact on the prediction of heart attack.

4.2.2 Partial Dependence Plots

To understand a feature's importance in a model it is necessary to understand both how changing that feature impacts the model's output, and the distribution of that feature's values. To visualize this for the KNN algorithm, in this project I build a partial dependence plot and show the distribution of feature values as a histogram on the x-axis.
In Fig 4.7(a), the gray horizontal line in the plot represents the expected value of the model when applied to the test data. The vertical gray line represents the average

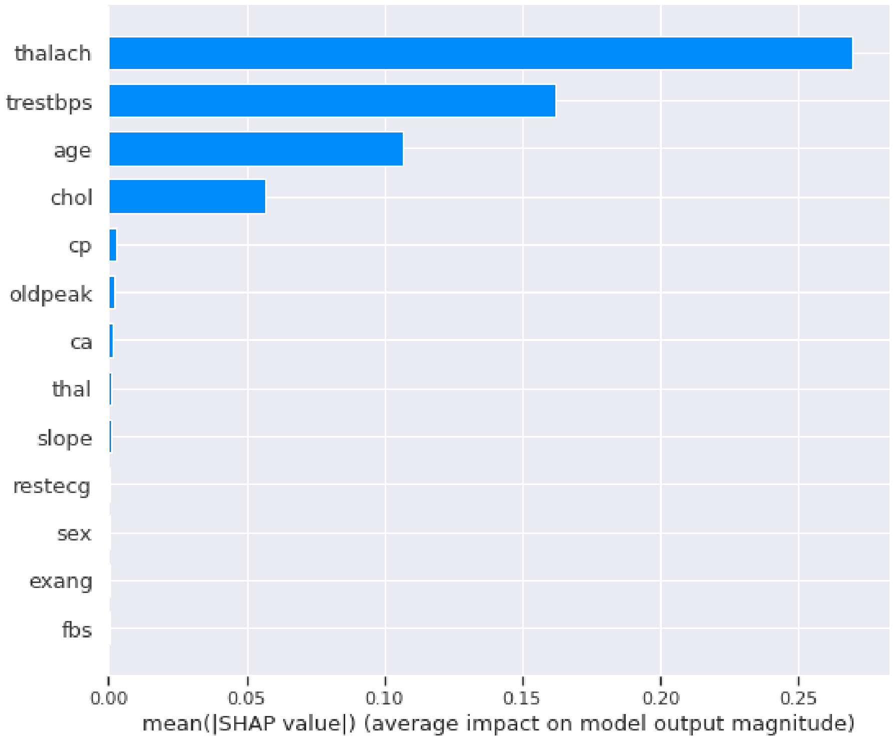


Figure 4.6: Value Counts of The Features

value of the "thalach" feature. The blue partial dependence plot line (which is the average value of the model output when I fix the"thalach" feature to a given value) always passes through the intersection of the two gray expected value lines. So, we can consider this intersection point as the "center" of the partial dependence plot with respect to the data distribution. This study has shown a partial dependence plot for "trestbps" feature also in Fig 4.7(b).

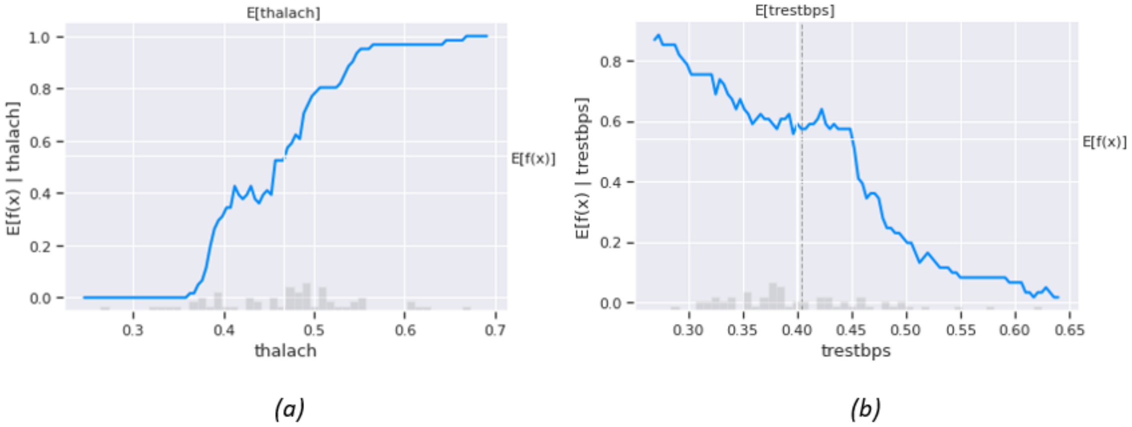
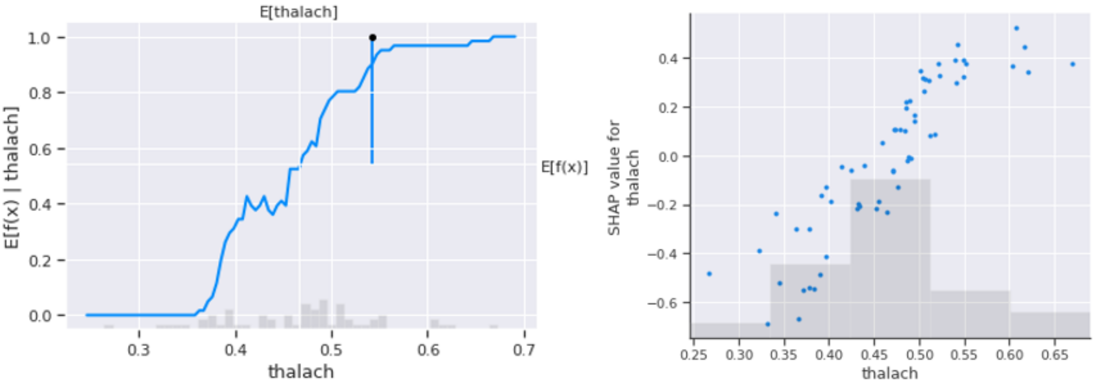


Figure 4.7: Partial Dependence Plots

4.2.3 SHAP values from Partial Dependence Plots (PDP)

Shap values measure the contribution of each feature in a ML model. When we are explaining a prediction , the SHAP value for a specific feature is just the difference

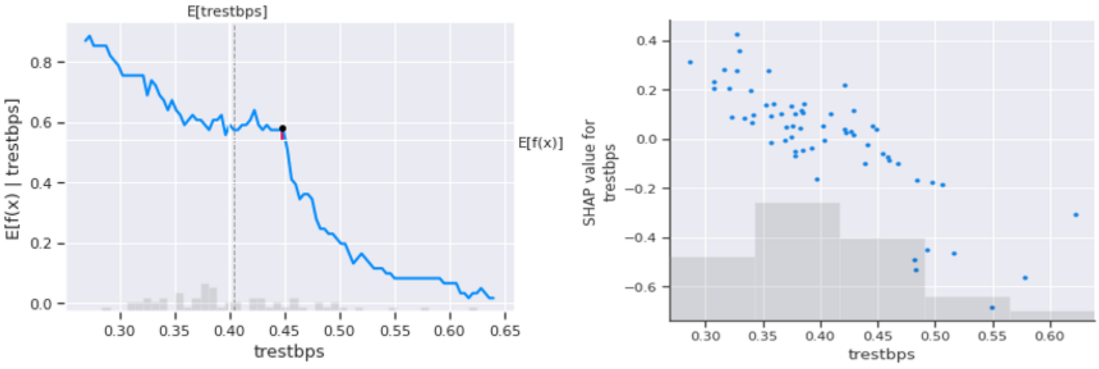
between the expected model prediction (output) and the partial dependence plot at the feature's value.
Fig 4.8 shows the SHAP values for "thalach" feature. The black dot on the figure indicates the predicted value. Now the difference between average of the model output and the partial dependence plot line (expected value of the model) is the SHAP value which is shown in blue curve line.



(a) *(b)*

Figure 4.8: SHAP Values for thalach Feature

If we plot the SHAP value for a specific feature across a whole dataset, we will exactly trace out a version of the partial dependence plot for that feature. Like as, if I plot all the SHAP values from the dataset, then the partial dependence plot for "thalach" feature will look like Fig 4.8. In the paper, I have showed just 1st 100 rows from the dataset in this figure.
This paper also shows the SHAP values for partial dependence plot for "trestbps" feature shown in Fig 4.9.



(a) *(b)*

Figure 4.9: SHAP Values for trestbps Feature

4.2.4 SHAP Waterfall Plot

For machine learning models SHAP values of all the input features shows the difference between expected model output and the predicted model output. Waterfall plot is a visual representation of it. Fig 4.10 shows the SHAP Waterfall plot. It shows the path of how SHAP values were added to the expected output (E[f(x)] = 0.541) to come to the current model output (f(x) = 0). The gray values before the feature names show the value of each feature. Blue color features give more impact towards the model output and red color gives less impact. Here, "thalach" feature gives a huge impact by increasing the rate of getting heart attack. Also "trestbps", "chol" had a good impact.

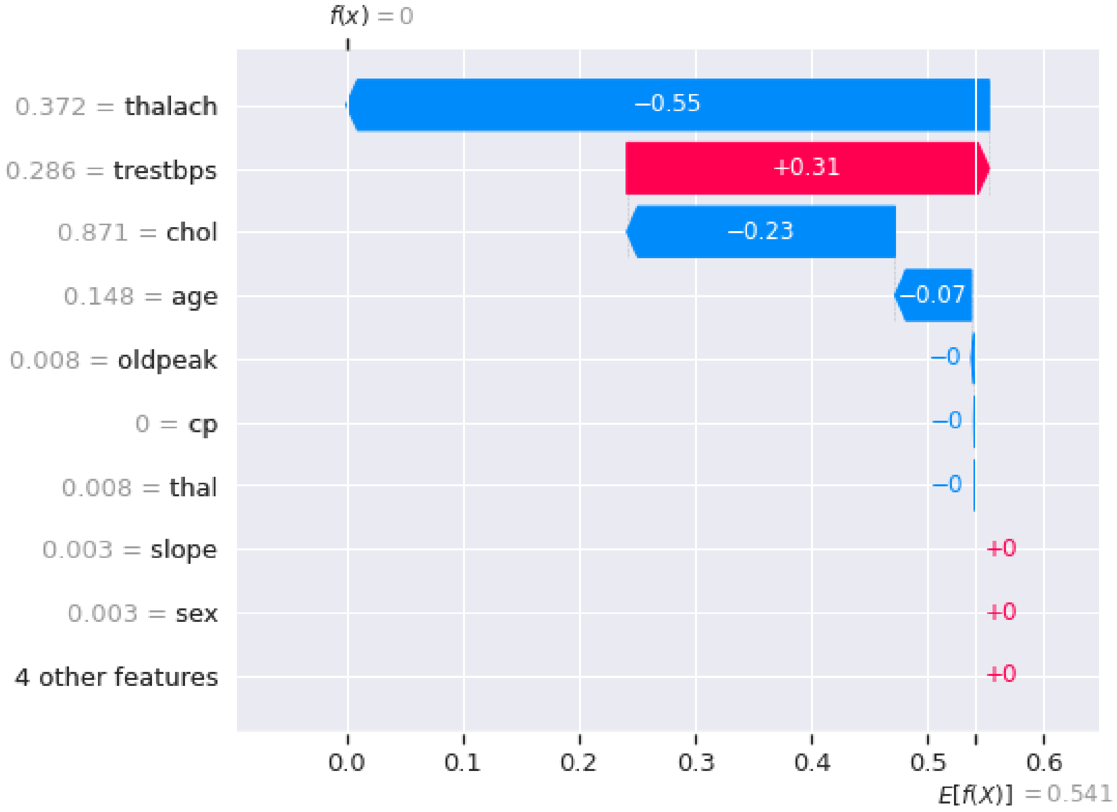


Figure 4.10: SHAP Waterfall Plot

So, the SHAP waterfall plot visually interpret why the current output is 0 whereas, the expected output was 0.541.

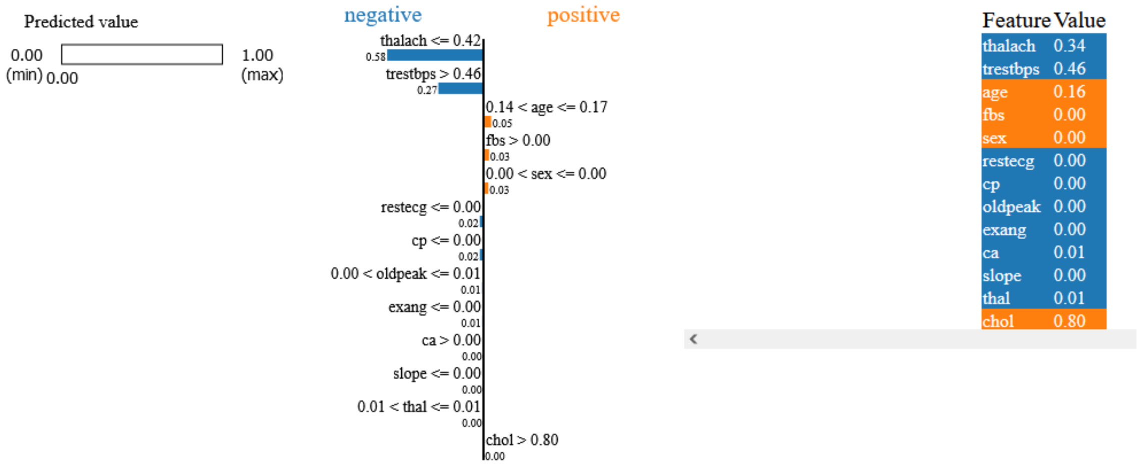
4.3 Result Analysis of the Heart Attack Prediction Using LIME

This section of the paper will explain the heart attack prediction through KNN Machine Learning algorithm using LIME.

4.3.1 Feature Importance

Which features give more contribution to the prediction, LIME method can give the visual representation of it. Using LIME is so beneficial, because it can provide the feature importance by using 2 different methods.

By using show_in_notebook() method Fig 4.11 shows the volume of the features impact. The figure consists 3 types of representation- progress bar, bar chart, and table. In the figure, the progress bar indicates the range of the which value varies and actual prediction; the bar chart shows the features of their weights positively and negatively to prediction; and the table represents the features importance by showing the actual feature values. Here, orange color indicates the positive contribution and blue color indicates the negative contribution toward the prediction [49].



Feature	Value
thalach	0.34
trestbps	0.46
age	0.16
fbs	0.00
sex	0.00
restecg	0.00
cp	0.00
oldpeak	0.00
exang	0.00
ca	0.01
slope	0.00
thal	0.01
chol	0.80

Figure 4.11: Feature Importance Using LIME

Another method is as_pyplot_figure() shown in Fig 4.12. This method is used for visualizing the bar chart of the features importance.

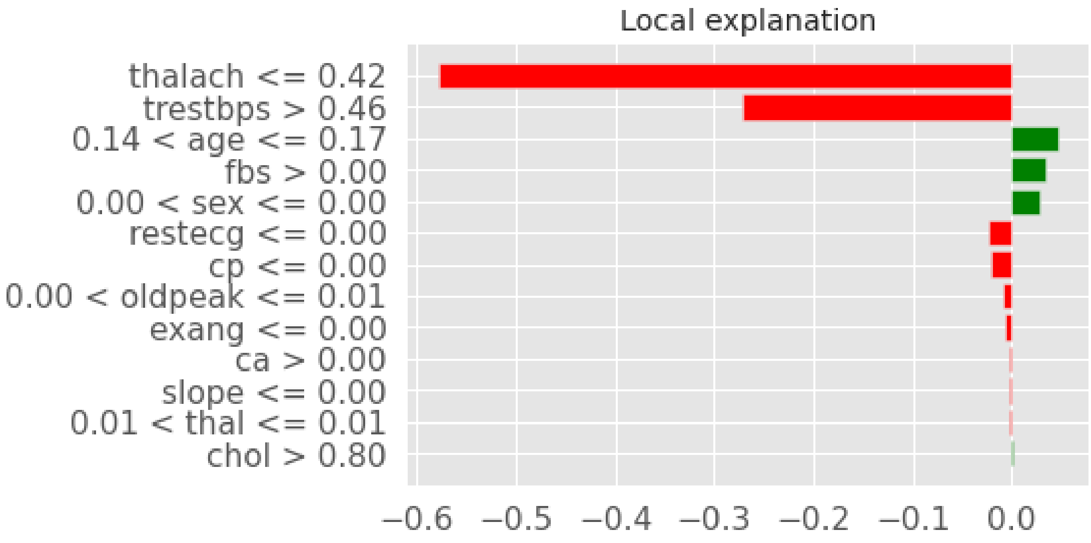
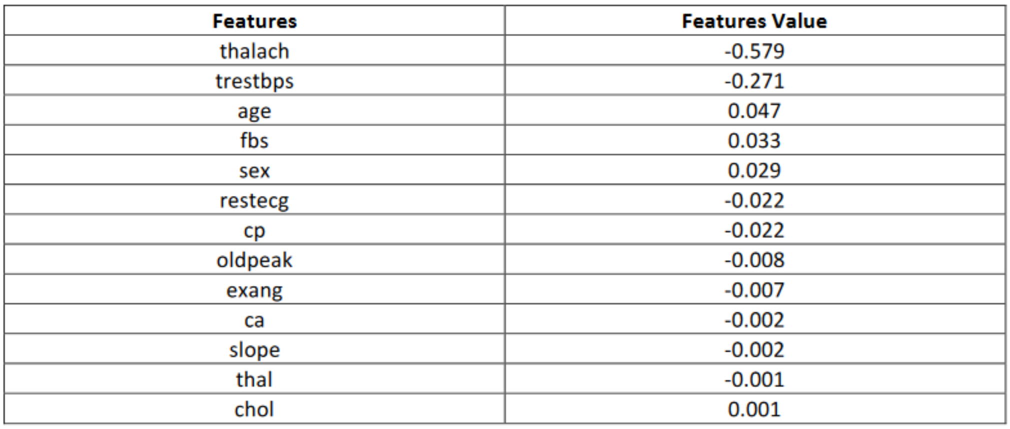


Figure 4.12: Bar Chart of Feature Importance Using LIME

4.3.2 Retrieve Features Importance

LIME XAI method has the advantage of retrieve the features importance showed in Fig 4.13. Here the first value of tuple is condition and the second value is the feature value based on condition.



Features	Features Value
thalach	-0.579
trestbps	-0.271
age	0.047
fbs	0.033
sex	0.029
restecg	-0.022
cp	-0.022
oldpeak	-0.008
exang	-0.007
ca	-0.002
slope	-0.002
thal	-0.001
chol	0.001

Figure 4.13: Retrieving Features Using LIME

As we know, KNN algorithm is a regression model. But this algorithm is also defined as a classifier. So if we want to retrieve the features importance for classifier task LIME will allow us to do so. Fig 4.14 shows that, it returns a dictionary where the key is each class of task and value is a list of feature index and their contribution in predicting that class.

4.3.3 Local and Global Prediction Value

Finally, Table 4.1 shows the local prediction and global prediction using LIME. And Table 4.2 shows the actual prediction and the predicted prediction that is done by the KNN algorithm using LIME method. Here, we can see that, our predicted prediction is exactly equal to the actual prediction, which is 0. Moreover, the local prediction and the global prediction are also identical.

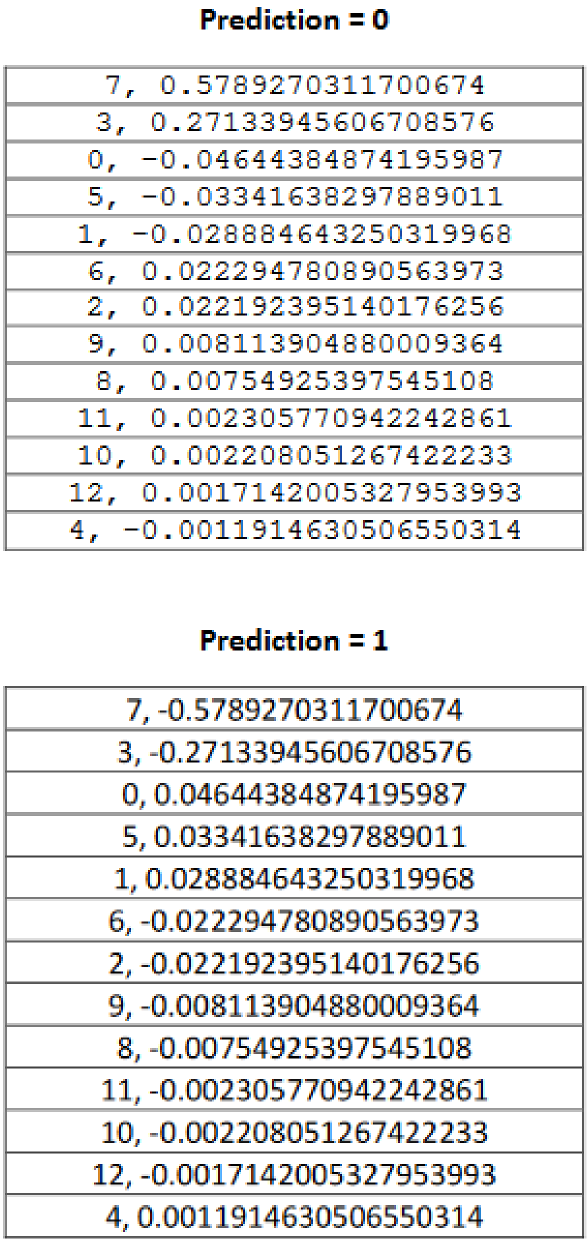
So, this thesis exhibits that the LIME method worked really good on the KNN machine learning algorithm.

Local Prediction	Global Prediction
-0.06819538	0

Table 4.1: Local & Global Prediction of LIME

Actual Prediction	Predicted Prediction
0	0

Table 4.2: Actual & Predicted Prediction of LIME



Prediction = 0

7, 0.5789270311700674
3, 0.27133945606708576
0, -0.04644384874195987
5, -0.03341638297889011
1, -0.028884643250319968
6, 0.022294780890563973
2, 0.022192395140176256
9, 0.008113904880009364
8, 0.00754925397545108
11, 0.002305770942242861
10, 0.002208051267422233
12, 0.0017142005327953993
4, -0.0011914630506550314

Prediction = 1

7, -0.5789270311700674
3, -0.27133945606708576
0, 0.04644384874195987
5, 0.03341638297889011
1, 0.028884643250319968
6, -0.022294780890563973
2, -0.022192395140176256
9, -0.008113904880009364
8, -0.00754925397545108
11, -0.002305770942242861
10, -0.002208051267422233
12, -0.0017142005327953993
4, 0.0011914630506550314

Figure 4.14: Retrieving Features Using LIME

Chapter 5

Conclusion

Now-a-days heart attack rate is increasing in Bangladesh. That's why, I wanted to develop a paper which will provide the chances of heart attack for different machine learning algorithms. In this paper, I have tried to show the comparison of multiple machine learning algorithms. The paper has predicted heart attack rate for various features. This thesis also showed the features importance. The paper also gives the prediction accuracy of these algorithms. Based on accuracy results XGBoost Classifier gives significant predictions with 86.885% accuracy.

In machine learning, it is very important to find out the models complexity, explainability and prediction performance. But there are also many complex machine learning algorithms, then it will be very difficult to deploy the model explainability. The concept of black box deals with this issue. The XAI method SHAP and LIME are very optimal for model explainability. SHAP gives the visual representation of the models prediction. This thesis used the SHAP XAI method on KNN machine learning algorithm to analyse the features importance. Moreover, SHAP offers an insightful means to interpret the results from the KNN machine learning algorithm.

After that, the paper used the LIME XAI method to explain how the black box model works on the KNN machine learning algorithm. LIME gives the explanation of the prediction of the KNN model with a faithful manner.

Overall the paper will give knowledge about heart attack prediction of various machine learning algorithms. After that, this thesis also exhibited the SHAP and LIME XAI methods to explain the accuracy of KNN.

5.1 Limitations and Future Work

It is very challenging to implement the ML algorithms in healthcare field for it's limited availability of dataset. The paper proved that the XGBoost algorithm gives the best accuracy. So the SHAP and LIME methods should be implemented on the XGBoost algorithm. To do so, a huge and populated dataset must be needed. Furthermore, the dataset needed to be preprocessed. All of these above reasons limited the papers significance.

In recent years, the ML algorithms along with AI has been shown great impact on healthcare area. This thesis is a little approach to it. As the paper requires some limitations, I want to move forward the implementation as future approach. Implementation of SHAP and LIME on XGBoost algorithm can be targeted as future work. Moreover, in future I would like to work on a huge dataset to see how accurately the model predicts. This study also showed SHAP values from Partial Dependence Plots only for 1st 100 rows. I will extend this work and try to get SHAP values for the whole dataset.

Despite having some limitations this thesis performed all the predictions and provides prominent future works. So, I dream to take the work forward and enrich further.

Bibliography

[1] S. Konda, B. Rani, and D. Govardhan, "Applications of data mining techniques in healthcare and prediction of heart attacks," *International Journal on Computer Science and Engineering*, vol. 2, pp. 250–255, Jan. 2010.

[2] S. B. Patil and Y. S. Kumaraswamy, "Extraction of significant patterns from heart disease warehouses for heart attack prediction," 2009.

[3] I. Rish, "An empirical study of the naïve bayes classifier," *IJCAI 2001 Work Empir Methods Artif Intell*, vol. 3, Jan. 1, 2001.

[4] K. Zhang, J. Zhang, P.-D. Xu, T. Gao, and D. W. Gao, "Explainable ai in deep reinforcement learning models for power system emergency control," *IEEE Transactions on Computational Social Systems*, vol. 9, no. 2, pp. 419–427, 2022. DOI: 10.1109/TCSS.2021.3096824.

[5] O. Bastani, C. Kim, and H. Bastani, "Interpreting blackbox models via model extraction," *arXiv preprint arXiv:1705.08504*, 2017.

[6] M. R. Zafar and N. M. Khan, "Dlime: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems," *arXiv preprint arXiv:1906.10263*, 2019.

[7] S. Shi, X. Zhang, and W. Fan, "A modified perturbed sampling method for local interpretable model-agnostic explanation," *arXiv preprint arXiv:2002.07434*, 2020.

[8] A. Singh and R. R. Kumar, "Heart disease prediction using machine learning algorithms," *2020 International Conference on Electrical and Electronics Engineering (ICE3)*, pp. 452–457, 2020.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," New York, NY, USA: Association for Computing Machinery, 2016, ISBN: 9781450342322.

[10] N. B. Kumarakulasinghe, T. Blomberg, J. Liu, A. S. Leao, and P. Papapetrou, "Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models," in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, IEEE, 2020, pp. 7–12.

[11] A. Golande and T. Pavan Kumar, "Heart disease prediction using effective machine learning techniques," *International Journal of Recent Technology and Engineering (IJRTE), ISSN*, pp. 2277–3878, 2019.

[12] T. Nagamani, S. Logeswari, and B. Gomathy, "Heart disease prediction using data mining with mapreduce algorithm," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 3, pp. 137–140, 2019.

[13] F. S. Alotaibi, "Implementation of machine learning model to predict heart failure disease," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 6, 2019. DOI: 10.14569/IJACSA.2019.0100637. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2019.0100637.

[14] A. N. Repaka, S. D. Ravikanti, and R. G. Franklin, "Design and implementing heart disease prediction using naives bayesian," *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 292–297, 2019.

[15] J. Thomas and R. Princy, "Human heart disease prediction system using data mining techniques," *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, pp. 1–5, 2016.

[16] N. M. Lutimath, C. Chethan, and B. Pol, "Prediction of heart disease using machine learning," *Int. J. Recent Technol. Eng*, vol. 8, pp. 474–477, 2019.

[17] A. Kumar, "Decision support system for heart disease diagnosis using neural network," Mar. 2021.

[18] B. Szymanski, L. Han, M. Embrechts, A. Ross, K. Sternickel, and L. Zhu, "Using efficient supanova kernel for heart disease diagnosis," in *proc. ANNIE*, vol. 6, 2006, pp. 305–310.

[19] M. C. Tu, D. Shin, and D. Shin, "Effective diagnosis of heart disease through bagging approach," in *2009 2nd International Conference on Biomedical Engineering and Informatics*, 2009, pp. 1–4. DOI: 10.1109/BMEI.2009.5301650.

[20] M. Shouman, T. Turner, and R. Stocker, "Integrating decision tree and k-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients," in *Proceedings of the International Conference on Data Science (ICDATA)*, The Steering Committee of The World Congress in Computer Science, Computer …, 2012, p. 1.

[21] A. Sudha, P. Gayathri, and N. Jaisankar, "Effective analysis and predictive model of stroke disease using classification methods," *International Journal of Computer Applications*, vol. 43, no. 14, pp. 26–31, 2012.

[22] P. Kora and S. R. Kalva, "Improved bat algorithm for the detection of myocardial infarction," *SpringerPlus*, vol. 4, no. 1, pp. 1–18, 2015.

[23] G. Fidel, R. Bitton, and A. Shabtai, "When explainability meets adversarial learning: Detecting adversarial examples using shap signatures," in *2020 international joint conference on neural networks (IJCNN)*, IEEE, 2020, pp. 1–8.

[24] T. Chen, J. Xu, H. Ying, *et al.*, "Prediction of extubation failure for intensive care unit patients using light gradient boosting machine," *IEEE Access*, vol. 7, pp. 150 960–150 968, 2019.

[25] D. Fryer, I. Strümke, and H. Nguyen, "Shapley values for feature selection: The good, the bad, and the axioms," *IEEE Access*, vol. 9, pp. 144 352–144 360, 2021.

[26] W. Shi, C. Wu, and Z. Li, "A shapley-value mechanism for bandwidth on demand between datacenters," *IEEE Transactions on Cloud Computing*, vol. 6, no. 1, pp. 19–32, 2015.

[27] S. M. Lundberg, G. Erion, H. Chen, *et al.*, "From local explanations to global understanding with explainable ai for trees," *Nature machine intelligence*, vol. 2, no. 1, pp. 56–67, 2020.

[28] G. R. Vásquez-Morales, S. M. Martinez-Monterrubio, P. Moreno-Ger, and J. A. Recio-Garcia, "Explainable prediction of chronic renal disease in the colombian population using neural networks and case-based reasoning," *Ieee Access*, vol. 7, pp. 152 900–152 910, 2019.

[29] C. S. Kumar, M. N. S. Choudary, V. B. Bommineni, G. Tarun, and T. Anjali, "Dimensionality reduction based on shap analysis: A simple and trustworthy approach," in *2020 International Conference on Communication and Signal Processing (ICCSP)*, IEEE, 2020, pp. 558–560.

[30] R. Alenezi and S. A. Ludwig, "Explainability of cybersecurity threats data using shap," in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, IEEE, 2021, pp. 01–10.

[31] M. Sarhan, S. Layeghy, and M. Portmann, "Evaluating standard feature sets towards increased generalisability and explainability of ml-based network intrusion detection," *arXiv preprint arXiv:2104.07183*, 2021.

[32] H. Kim, Y. Lee, E. Lee, and T. Lee, "Cost-effective valuable data detection based on the reliability of artificial intelligence," *IEEE Access*, vol. 9, pp. 108 959–108 974, 2021.

[33] L. V. Ho, M. Aczon, D. Ledbetter, and R. Wetzel, "Interpreting a recurrent neural network's predictions of icu mortality risk," *Journal of Biomedical Informatics*, vol. 114, p. 103 672, 2021.

[34] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[35] M. Toğaçar, N. Muzoğlu, B. Ergen, B. S. B. Yarman, and A. M. Halefoğlu, "Detection of covid-19 findings by the local interpretable model-agnostic explanations method of types-based activations extracted from cnns," *Biomedical Signal Processing and Control*, vol. 71, p. 103 128, 2022.

[36] G. J. Katuwal and R. Chen, "Machine learning model interpretability for precision medicine," *arXiv preprint arXiv:1610.09045*, 2016.

[37] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[38] P. Hall, N. Gill, M. Kurka, and W. Phan, "Machine learning interpretability with h2o driverless ai," *H2O. ai*, 2017.

[39] L. Hu, J. Chen, V. N. Nair, and A. Sudjianto, "Locally interpretable models and effects based on supervised partitioning (lime-sup)," *arXiv preprint arXiv:1806.00663*, 2018.

[40] A. Gosiewska and P. Biecek, "Ibreakdown: Uncertainty of model explanations for non-additive predictive models," *arXiv preprint arXiv:1903.11420*, 2019.

[41] M. R. Zafar and N. Khan, “Deterministic local interpretable model-agnostic explanations for stable explainability,” *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 525–541, 2021.

[42] M. Robnik-Šikonja and I. Kononenko, “Explaining classifications for individual instances,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 5, pp. 589–600, 2008.

[43] M. A. Ahmad, C. Eckert, and A. Teredesai, “Interpretable machine learning in healthcare,” in *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, 2018, pp. 559–560.

[44] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” Mar. 2016.

[45] J. Peng, K. Lee, and G. Ingersoll, “An introduction to logistic regression analysis and reporting,” *Journal of Educational Research - J EDUC RES*, vol. 96, pp. 3–14, Sep. 2002. DOI: 10.1080/00220670209598786.

[46] S. Ruder, “An overview of gradient descent optimization algorithms,” *ArXiv*, vol. abs/1609.04747, 2016.

[47] C. Molnar, *Interpretable machine learning a guide for making black box models explainable*, https://christophm.github.io/interpretable-ml-book/lime.html, An optional note, Sep. 2022.

[48] Y. Song, X. Jiao, Y. Qiao, *et al.*, “Prediction of double-high biochemical indicators based on lightgbm and xgboost,” in *Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science*, 2019, pp. 189–193.

[49] S. Solanki, *How to use lime to interpret predictions of ml models?* https://coderzcolumn.com/tu learning/how-to-use-lime-to-understand-sklearn-models-predictions, An optional note, Aug. 2022.