# Evaluation of Local Interpretable Model-Agnostic Explanation and Shapley Additive Explanation for Chronic Heart Disease Detection

Tsehay Admassu Assegie*

Department of Computer Science, Injibara University, Injibara, Ethiopia

Received 15 December 2021; received in revised form 13 May 2022; accepted 14 May 2022

DOI: https://doi.org/10.46604/peti.2022.10101

## Abstract

This study aims to investigate the effectiveness of local interpretable model-agnostic explanation (LIME) and Shapley additive explanation (SHAP) approaches for chronic heart disease detection. The efficiency of LIME and SHAP are evaluated by analyzing the diagnostic results of the XGBoost model and the stability and quality of counterfactual explanations. Firstly, 1025 heart disease samples are collected from the University of California Irvine. Then, the performance of LIME and SHAP is compared by using the XGBoost model with various measures, such as consistency and proximity. Finally, Python 3.7 programming language with Jupyter Notebook integrated development environment is used for simulation. The simulation result shows that the XGBoost model achieves 99.79% accuracy, indicating that the counterfactual explanation of the XGBoost model describes the smallest changes in the feature values for changing the diagnosis outcome to the predefined output.

**Keywords** SHAP, LIME, XGBoost, counterfactual explanation

## 1. Introduction

Ensemble-based automated chronic heart disease detection systems have achieved promising performance in automated decision-making over the past few years. The smartness and maximized performance of ensemble methods in automated medical decision-making are due to the ever-growing power of computing devices and chronic heart disease datasets. Although the performance of ensemble methods is much better than simple linear methods, it is difficult to explain the decision-making process of an ensemble model to a medical expert, because of the inherent complexity in the design of the ensemble model [1]. Thus, explaining the decision-making process is crucial to increase the trust of medical experts in using ensemble models (e.g., the XGBoost model) for making significant decisions such as heart disease diagnosis; failure to accurately identify a disease will result in danger to life.

Currently, complex ensemble-based models that rely on incomprehensible inferences are not an option for existing cardiac diagnostic systems. Transparency is one of the main reasons why the adoption of ensemble-based complex models and automated diagnosis systems in the healthcare industry requires more caution than in other domains such as the e-commerce and entertainment industry [2].

Over the past few years, the use of complex models, such as deep learning and ensemble methods, has become common for disease detection [3]. Toğaçar et al. [4] proposed a support vector machine-based chronic heart disease diagnosis system. However, they found that much research effort is needed to evaluate the automated diagnosis output by using model explanation techniques [4]. They concluded that model explanation techniques such as local interpretable model-agnostic explanation (LIME)

* Corresponding author.tsehayadmassu2006@gmail.com

Tel.: +251-9-21114923

and Shapley additive explanation (SHAP) have not been adequately and formally practiced in medical applications. Interpretable ensemble-based models can explain the models' diagnostics while creating efficient and accurate diagnostics. Thus, an ensemble-based diagnosis system is critical for assisting physicians in making heart disease diagnoses [5].

Rudin et al. [6] evaluated the SHAP interaction value of an XGBoost-based gold price prediction model. The SHAP interaction value represents the feature importance score of the prediction accuracy of the XGBoost model for gold price prediction. Moreover, the SHAP summary plot is analyzed to explain the impact of features on the predictive outcome of the XGBoost model for gold price prediction. The experimental result shows that the SHAP value provides insights into the prediction process of the model, and the features that influence the outcome of the XGBoost model are identified.

Gramegna et al. [7] and Cavaliere et al. [8] proposed an explainable model for diabetes mellitus prediction. The performance of the model was evaluated using the area under the curve (AUC) (receiver operating characteristics curve (ROC)) as the performance metric and SHAP as the model explainer. The researchers employed light gradient boosting (LightGBM) for model development. Experimental results show that SHAP has the potential to explain the predicted probabilities based on the baseline LightGBM model. Furthermore, the experiment reveals that the feature with the largest average absolute SHAP value is the most important since that influences the diagnosis outcome. The SHAP value is calculated and averaged across all samples and is ranked and plotted from the highest to the lowest.

This study aims to evaluate the effectiveness of LIME and SHAP in explaining how the XGBoost model obtains the prediction outcome of chronic heart disease. To achieve the goal, a chronic heart disease dataset is employed. Diverse counterfactual explanations (DICE) are also used.

## 2. Literature Review

Machine learning (ML) is widely used to assist medical professionals in making decisions for heart disease diagnosis. However, the complexity of ensemble models (e.g., the XGBoost model) is hindering the wider applicability of automated decision-making in the medical field, where a decision is critical to a patient's life [9]. To interpret the ensemble models, researchers have conducted several studies on the model explanation techniques.

Bhagwat et al. [10] found that the use of complex models (such as ensemble and deep learning models) enabled data scientists to achieve the best possible accuracy in heart disease diagnosis. Their study was conducted using benchmark datasets such as the heart disease data repository of the University of California Irvine (UCI). However, a better-performing model is too complex to interpret and explain. Thus, providing a human-understandable explanation for such a complicated model helps domain experts trust the diagnosis outcome. Interpretable ML is an area of research devoted to interpreting diagnostics produced by complex models. Wang et al. [11] classified explainable ML problems into three categories, i.e., model interpretation, result checking, and transparent box design.

This study focuses on outcome interpretation and model explanation, and explores the answers to the following research questions:

(1) How do the explanation techniques explain the XGBoost model's diagnostic outcome?

(2) Which explanation technique is more appropriate for explaining the XGBoost model's diagnostic output to a medical expert?

(3) What is the feature weight value of the feature that alters the positive diagnosis output into the desired diagnostic output?

## 3. Methodology

This study is conducted by using a heart disease dataset obtained from the UCI repository. To develop an automated chronic heart disease diagnosis model, XGBoost is employed. For simulation, Python 3.7 programming language with Jupyter

Notebook integrated development environment is employed. To evaluate the effectiveness of LIME and SHAP for describing the XGBoost model diagnosis output, heart disease positive and heart disease negative instances are randomly selected from the testing set to simulate local or instance level explanation. The schematic diagram of the proposed model is illustrated in Fig. 1.

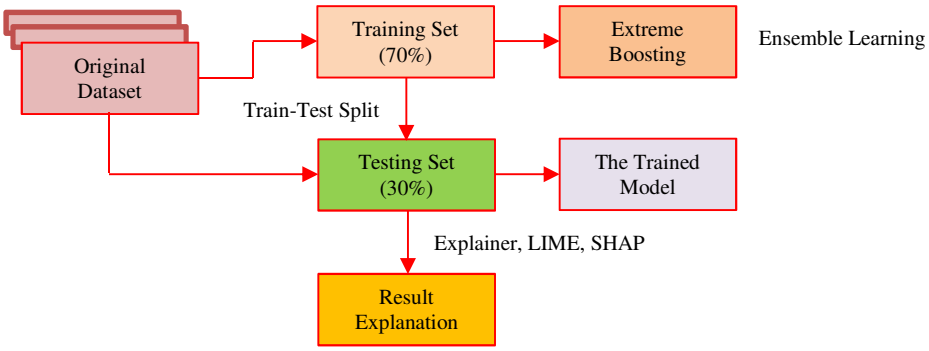


Fig. 1 Schematic diagram for model explanation

### *3.1. Model explanation problem*

The problem of ML model interpretation can be defined as finding an explanation $e$. $e$ is in a human-understandable domain by an interpretable model $g$. $E$ represents explanation. An interpretable model $g$ produces the prediction $p' = g(x)$ for a given data point $x$ and its function $c = f(x)$. The interpretable model globally mimics the prediction behavior of a complex model $G$. The explanation $E \in e$ is given by interpretable models such as $E = \alpha(c, x)$ for some training points $x$ and their prediction $\alpha$ [12].

### *3.2. LIME*

LIME is employed to describe and interpret diagnosis outcomes of the ML model at the individual level for each predicted instance in a testing set [13]. Thus, LIME is a local model explanation approach because it explains single output at a local or instance level. This study employs LIME to present the diagnosis probability of the XGBoost model at the instance level. The explanation of the XGBoost model diagnosis outcome at the local level helps healthcare experts make important decisions. Mathematically, LIME is defined as follows:

$$Explanation\ (x) = arg\ min\ g \in G\ L(f,\ g,\ \pi x) + \Omega(g) \tag{1}$$

where the model $g$ explains the instance $x$. $g$ measures the distance between the explanation and prediction of the original XGBoost model, while the XGBoost model complexity $\Omega\ (g)$ is minimized (by considering fewer features). $G$ is the family of possible explanations. The proximity $\pi x$ defines the neighborhood distance around the instance $x$. The neighborhood distance is the distance between the randomly generated feature and the original feature in the dataset.

### *3.3. SHAP*

SHAP describes the model output based on feature interaction values [14]. Feature interaction is a technique from the coalitional game theory, which is used to determine a fair pay-out distribution, based on the contributions of each player in a coalition. SHAP values are used to calculate each related feature's contribution to the predicted outcome. The SHAP values of a feature are computed based on the mean marginal difference with and without the feature in question. SHAP applies the feature interaction value to estimate the model output. The technical definition of SHAP values is "the average marginal influence of a feature value over all possible combinations". In other words, SHAP values consider all possible predictions for observation using all possible combinations between input features to each feature and the output of the XGBoost model [15].

### 3.4. Data description

In this study, the UCI heart disease dataset is used for training and testing the XGBoost model to predict if a person has heart disease or not. The dataset consists of 1025 instances of which 499 samples are healthy and 526 samples are patients. The dataset is divided into two sets: a training set and a testing set. The standard method of the train-test split is followed to divide the dataset into training and testing sets using the ratio of 70% by 30% respectively. The training set consists of 717 samples and the test set consists of 308 samples. For model training, the training set is used, and the test set is used for model testing.

The researcher presents the diagnosis outcome made by the developed model using randomly selected test samples from the testing set. Then, LIME and SHAP explanations are generated based on the diagnosis outcome of the model. Table 1 illustrates some of the data samples used in the simulation. The features of heart disease include age (age of patient), sex (gender of a patient), chest pain (cp), resting blood pressure (trestbps), cholesterol level (chol), fasting blood sugar (fbs), heart rate (thalach), angina induced by exercise (exang), oldpeak (depression), slope (slope of the heart), number of vessels (ca), thallium scan (thal), and target.

Table 1 Samples of heart disease dataset

| No. | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 293 | 58 | 1 | 0 | 128 | 259 | 0 | 0 | 130 | 1 | 3 | 1 | 2 | 3 | 0 |
| 697 | 74 | 0 | 1 | 120 | 269 | 0 | 0 | 121 | 1 | 0.2 | 2 | 1 | 2 | 1 |
| 353 | 57 | 1 | 0 | 110 | 20 | 0 | 1 | 126 | 1 | 1.5 | 1 | 0 | 1 | 1 |
| 481 | 63 | 0 | 0 | 150 | 407 | 0 | 0 | 154 | 0 | 4 | 1 | 3 | 3 | 0 |
| 823 | 54 | 1 | 0 | 140 | 239 | 0 | 1 | 160 | 0 | 1.2 | 2 | 0 | 2 | 1 |
| 462 | 52 | 1 | 0 | 118 | 186 | 0 | 0 | 190 | 0 | 0 | 1 | 0 | 1 | 1 |
| 536 | 50 | 0 | 2 | 130 | 219 | 0 | 1 | 179 | 0 | 0 | 2 | 0 | 2 | 1 |
| 438 | 47 | 1 | 2 | 130 | 253 | 0 | 1 | 179 | 0 | 0 | 2 | 0 | 2 | 1 |
| 347 | 54 | 0 | 2 | 108 | 267 | 0 | 0 | 167 | 0 | 0 | 2 | 0 | 2 | 1 |
| 148 | 59 | 1 | 0 | 164 | 176 | 1 | 0 | 90 | 0 | 1.0 | 1 | 2 | 1 | 0 |

## 4. Simulation Results and Discussion

For the simulation of LIME and SHAP, two randomly selected samples are considered. One is a healthy person or with a “heart disease negative” prediction outcome. The other is a patient or with a “heart disease positive” prediction outcome. Two simulations (case 1 and case 2) are conducted on each predicted instance by using LIME and SHAP. In the first simulation (case 1), the explanation generated by LIME for each instance is analyzed. In the second simulation (case 2), the explanation generated by SHAP for each instance is analyzed. Overall, true positive and true negative instances are used in the simulations to analyze how LIME and SHAP are used to explain the XGBoost model’s diagnosis output.

### 4.1. Case 1

In this simulation (case 1), a random sample of the patient (instance ID-1) is considered to analyze the explanation generated by LIME on the XGBoost model’s diagnosis outcome. The details of the instance employed in this simulation are shown in Table 2. The XGBoost model diagnosis probability for the patient class or target 1 is 99.35, and the probability for the healthy class or target 0 is 0.006.

As shown in Table 2, the XGBoost model detects the presence of heart disease with a probability of 1.00. Instance ID-1 has a target value of 1 (as shown in Table 2), which shows that the instance belongs to the patient class. The model correctly predicts instance ID-1 with a diagnosis probability of 1.00 for the patient class and 0.00 for the heart disease negative class. Fig. 2 summarizes the LIME explanation for the diagnosis outcome of the XGBoost model. The model’s diagnosis outcome is “patient” or “heart disease positive”. The sex feature affects the model’s output negatively while other features have a positive impact on the diagnosis outcome. As illustrated in Fig. 2, the reasons why the XGBoost model reaches a positive diagnosis

outcome are that the number of vessels (ca) is less than 0.00 and that the thallium scan is also less than 2.0. Moreover, the heart rate is greater than 165 mmHg, influencing the model's prediction as heart disease positive. The feature with the highest impact on the model's output is the number of vessels.

Table 2 Instance ID-1

| No. | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 293 | 59 | 1 | 2 | 150 | 212 | 1 | 1 | 157 | 0 | 1.6 | 2 | 0 | 2 | 1 |

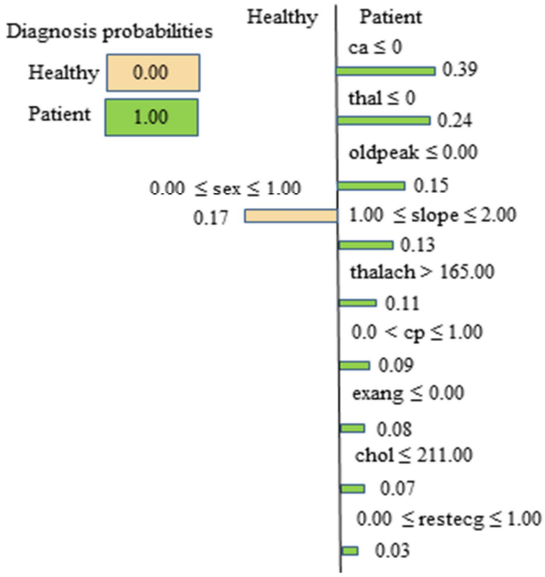


Fig. 2 Explanation generated by LIME for instance ID-1

### *4.2. Case 2*

In this simulation (case 2), two heart disease negative instances (instance ID-1 and instance ID-2) are randomly selected from the test set. Then, the diagnosis outcome of the XGBoost model is explained using LIME. Table 3 summarizes the details of the instance used in the experiment (i.e., the XGBoost model diagnosis outcome of the instance considered for the simulation).

As shown in Table 3, the XGBoost model predicts instance ID-2 as heart disease negative. The model correctly predicts instance ID-2 with a diagnosis probability of 0.99. Fig. 3 illustrates the explanation generated by LIME, which also shows that the diagnosis probability of instance ID-2 is 0.99. The features contributing to this diagnosis outcome include thallium scan (thal), oldpeak, sex, exercise-induced angina, slope, heart rate (thalach), and total resting blood pressure (trestbps) ordered by the magnitude of impact on the diagnosis outcome as illustrated in Fig. 3. In contrast, features such as the cholesterol level and the number of vessels do not affect the model output. A person with a cholesterol level ≤200.00 and vessels ≤0.00 has a lower probability of getting heart disease. Along with the LIME explanation demonstrated in Fig. 3, the influence of the features on the model output is given in Table 4 for instance ID-2.

Table 4 summarizes the contribution of heart disease features to the XGBoost model prediction outcome using instance ID-2. As illustrated in Table 4, chest pain (cp) contributes more to the model's prediction output being "heart disease positive" or "patient", as compared to other features. In Table 5, the SHAP values for instance ID-1 are demonstrated. The instance ID-1 is a positive instance considered for simulation. As demonstrated in Table 5, the features, such as heart rate (thalach) and slope, add bias to the diagnosis made by the model.

Fig. 4 illustrates the effect of heart disease features on the XGBoost model output using SHAP. As illustrated in Fig. 4, chest pain has the highest impact on the XGBoost model output. Thus, chest pain is an indicator of the presence of heart disease. Moreover, heart rate has the second highest impact on the XGBoost model. A patient with a higher heart rate has a higher risk

of getting the disease compared to a patient with a lower heart rate. The third feature influencing model's output is the number of vessels colored by X-ray. In contrast, thallium (thal), resting electrocardiogram, and slope have a lower effect on the output of the XGBoost model.

Table 3 Instance ID-2

| No. | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 49 | 60 | 1 | 0 | 117 | 230 | 1 | 1 | 160 | 1 | 14 | 2 | 2 | 3 | 0 |

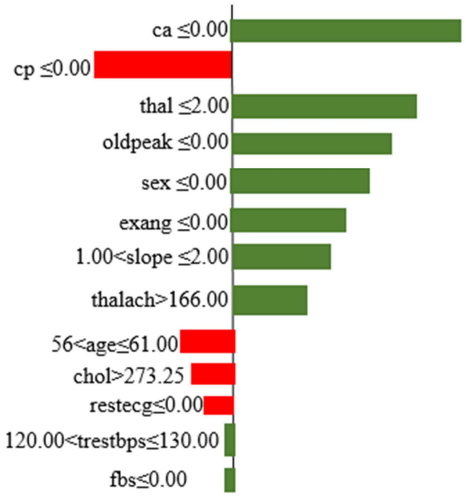


Fig. 3 Explanation generated by LIME for instance ID-1

Table 4 LIME explanation for the features of instance ID-2 (probability = 0.99)

| Feature | Value | Impact on the XGBoost model output |
|---|---|---|
| Chest pain (cp) | 2.000 | +0.14 |
| Thalassemia (thal) | 2.000 | +0.21 |
| Number of major vessels (ca) | 0.000 | +0.19 |
| Oldpeak | 0.000 | +0.09 |
| Sex | 0.000 (Female) | +0.08 |
| Age | 39 | +0.027 |
| Exercise-induced angina (exang) | 0.000 (No angina due to exercise) | +0.08 |
| Resting electrocardiogram (restecg) | 1.000 | +0.03 |
| Maximum heart rate achieved (thalach) | 133.000 | -0.03 |
| Slope | 1.000 (Upward sloping) | -0.09 |

Table 5 SHAP explanation for the features of instance ID-1 ($E[f(x)] = 0.513$)

| Feature | Value | Impact on the XGBoost model output |
|---|---|---|
| Chest pain (cp) | 1 | +0.16 |
| Number of major vessels (ca) | 0 | +0.12 |
| Thalassemia (thal) | 2 | +0.07 |
| Sex | 0 (Female) | +0.03 |
| Slope | 2 | +0.03 |
| Age | 34 | +0.02 |
| Cholesterol (chol) | 210 | +0.02 |
| Total resting blood pressure (trestbps) | 118 | +0.02 |
| Exercise-induced angina (exang) | 0 (No angina due to exercise) | +0.01 |
| Maximum heart rate (thalach) | 192 | +0.01 |
| Resting electrocardiogram (restecg) | 1 | +0.01 |
| Fasting blood sugar (fbs) | 0.000 (<120 mm/dL) | 0.0 |
| Oldpeak | 7 | 0.0 |

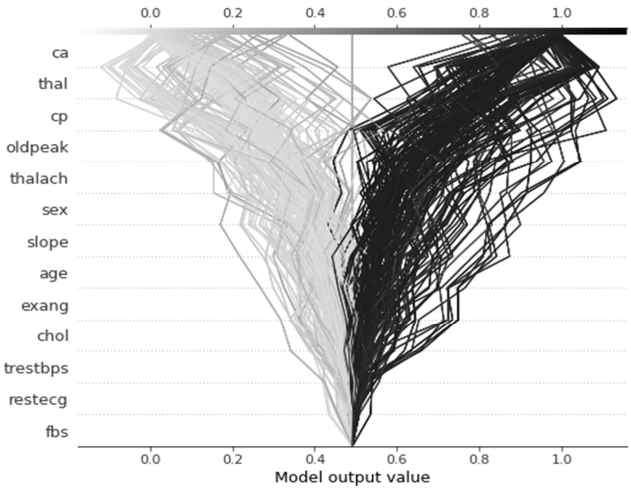


Fig. 4 Effect of heart disease features on the XGBoost model output using SHAP

### *4.3. SHAP explanation*

The SHAP value for instance ID-1 is illustrated in Fig. 5. As shown in Fig. 5, blood pressure, thallium scan, and electrocardiogram have a negative impact on the XGBoost model diagnosis outcome. In contrast, features such as the number of blood vessels, chest pain, cholesterol, age, and sex heart rate affect the XGBoost model output. The features with the greatest effect on the model's output include the number of blood vessels stained using X-ray, ca = 2 (showing that the vessels are not blocking the blood flow). The second most influencing feature for the model to make the negative diagnosis is chest pain, cp = 0 (showing that the patient is experiencing angina, e.g., symptomatic or asymptomatic angina pain).

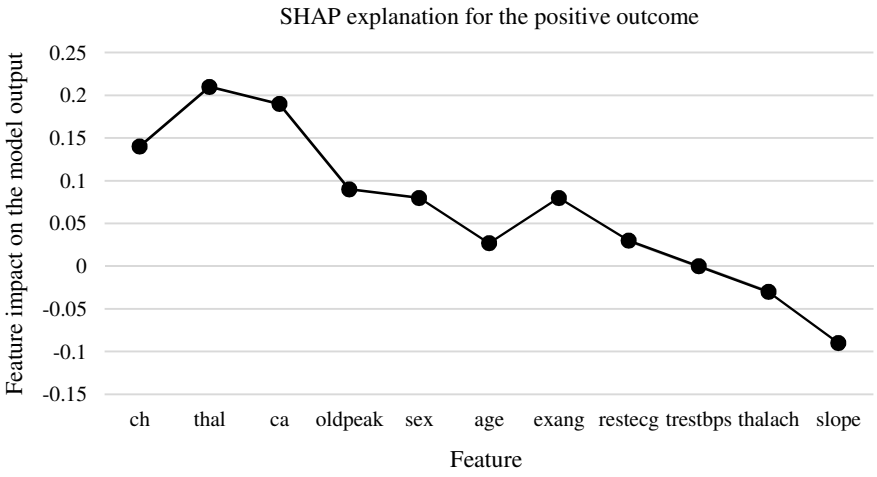


Fig. 5 SHAP explanation of the positive diagnosis outcome (instance ID-1)

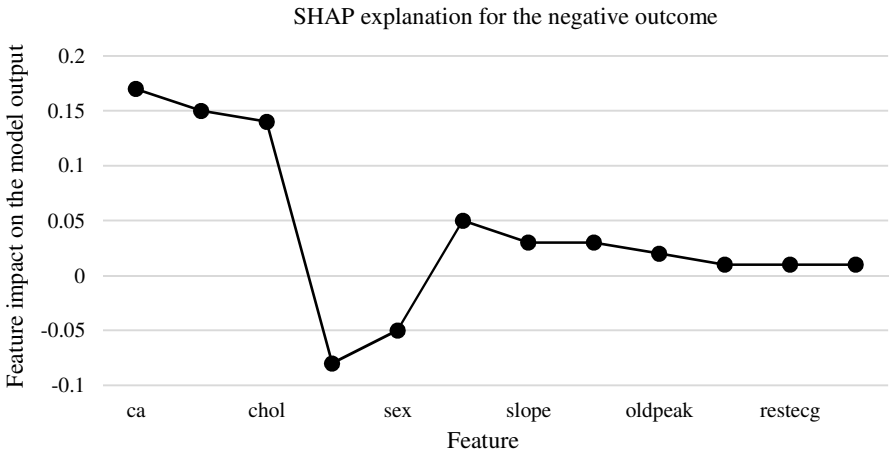


Fig. 6 SHAP explanation of the negative diagnosis outcome (instance ID-2)

In addition, as illustrated in Fig. 5, the instance has a chest pain value of 0, indicating the existence of a-typical angina. As a result, chest pain influences the model's diagnosis outcome showing "heart disease patient". Furthermore, instance ID-1 has high serum cholesterol (chol = 212 as illustrated in Fig. 5). The patient also has a thallium scan value of 2 indicating a fixed defect, and the ST segment is downward sloping (slope = 2). Moreover, the patient has a heart rate of 157 mmHg, influencing the model's positive diagnosis outcome. Thus, a counterfactual (CF) explanation for this patient should be "to reduce cholesterol level and heart rate". The change in cholesterol and heart rate could result in reduced heart disease risk.

The feature impacting the XGBoost model's diagnosis outcome is the number of blood vessels (ca = 2, showing that the blood vessels are the worst cases blocking the blood flow). The second significant feature influencing the XGBoost model's output is chest pain, cp = 2. This indicates that the patient is experiencing symptomatic or asymptomatic pain, leading the model to show that the patient is suffering from heart disease. The third influential feature is the patient's cholesterol level (chol = 212 mm/dL), which is higher than the normal cholesterol level of a healthy individual (170 mm/dL). A high cholesterol level could lead to the blockage of blood vessels. Fig. 6 illustrates the SHAP explanation for XGBoost model diagnosis outcome on instance ID-2. As demonstrated in Fig. 6, age and sex do not influence the XGBoost model diagnosis outcome. Chest pain, number of blood vessels, and cholesterol level contribute a lot to the model's diagnosis outcome.

### *4.4. Comparison of SHAP and LIME explanations*

For further comparison and analysis, the XGBoost model is tested on the heart disease positive instance using LIME and SHAP. The diagnosis outcome of the XGBoost model is explained using the SHAP value and explanation generated by LIME. Fig. 7 illustrates the effect of heart disease features on the XGBoost model output. As demonstrated in Fig. 7, the impact of the features on the XGBoost model's diagnosis outcome changes in a similar direction although the magnitude of feature impact of SHAP value is greater than the explanation generated by LIME. The XGBoost model decides that the given instance is a heart disease patient due to the higher impact of heart rate and chest pain type as illustrated in Fig. 7.

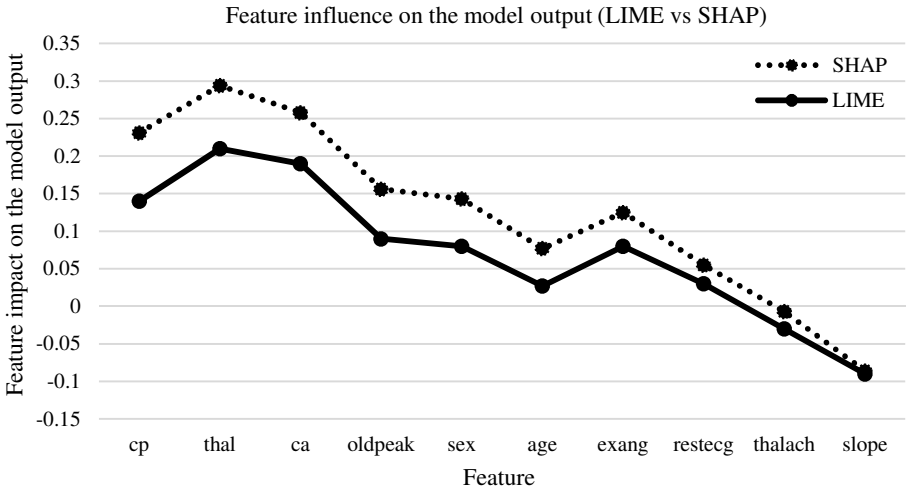


Fig. 7 Comparison of the feature impact on the XGBoost model output using SHAP and LIME

### *4.5. ROC for the XGBoost model*

ROC is a performance measure that provides a comprehensive evaluation of classification models such as XGBoost [16]. ROC considers the confusion matrices in all threshold operations and combines the confusion matrices to obtain a performance result. Fig. 8 demonstrates the ROC for the developed XGBoost model, showing that the area under the ROC (AUC) of the proposed model is scored as 0.99.

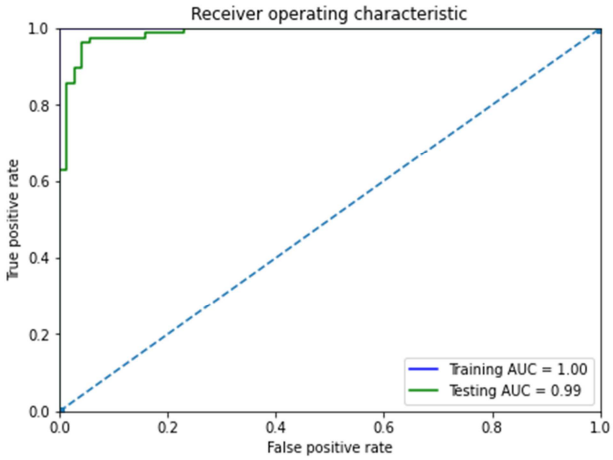


Fig. 8 ROC of the XGBoost model

### *4.6. Counterfactual explanation*

When an ML model is applied to real-world chronic heart disease data, along with the reason for its diagnosis outcome, it is also crucial to identify the variation required in feature magnitude that alters the diagnosis outcome. The CF explanation is a model-agnostic explanation procedure that provides the smallest alterations required in feature values for the diagnosis outcome to be changed to the predefined outcome [17-18]. In other words, if *X* is the feature and *Y* is an output label, the CF explanation shows the effect on *Y* due to small changes in the value *X*. Thus, CF helps determine the number of changes that need to be done in *X* to change the outcome from *Y* to *Y'* (e.g., from "heart disease patient" to "not heart disease patient"). It gives the what-if explanations for the model. CF plays a great role in showing what change is required to the feature values of a patient suffering from heart disease from being patient to non-patient. To provide the proof of CF and explore the effect of feature magnitude on the diagnosis outcome of the XGBoost model, three instances (instances ID-3, 4, and 5) are randomly taken from the heart disease dataset, as shown in Table 6.

Two heart disease positive instances and one heart disease negative instance are shown in the target column in Table 6. These instances are considered an input to the CF explanation. With the positive instances as the input, two different CFs are generated. The CFs show the minimum changes required for the feature values to change the target or label of the instance, i.e., change the target from "patient" (1) to "no disease" (0) or vice-versa. The following are the observations regarding the output to instance ID-3. In the CF output, sex, age, and type of chest pain (cp) are kept constant. Therefore, in each of the CFs, those features are unvaried. The CFs for the instance are shown in Table 7.

As shown in Table 7, the increase in oldpeak by the value of 0.5 and the change from thallium scan (thal = 2, showing the fixed defect) to the reversible defect (thal = 3) lead to a change in the intensity of heart disease. Furthermore, the reduction of heart rate (thalach) by 38 mmHg, i.e., from 165 mmHg to 127 mmHg, leads to a change in the heart disease intensity. Thus, the CF explanation is significantly important to determine the change in the values of input features to change the diagnosis outcome of the ML model.

Table 6 Instances considered for the CF explanation

| Instance ID | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 43 | 0 | 2 | 122 | 213 | 0 | 1 | 165 | 0 | 0.2 | 1 | 0 | 2 | 1 |
| 4 | 74 | 0 | 1 | 120 | 269 | 0 | 0 | 121 | 1 | 0.2 | 2 | 1 | 2 | 1 |
| 5 | 59 | 1 | 0 | 164 | 176 | 1 | 0 | 90 | 0 | 1.0 | 1 | 2 | 1 | 0 |

Table 7 CF explanation for instance ID-3

| Instance ID | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 43 | 0 | 2 | 122 | 213 | 0 | 1 | 165 | 0 | 0.2 | 1 | 0 | 2 | 1 |
| New_CFs | 43 | 0 | 2 | 122 | 213 | 0 | 1 | 127 | 0 | 2.6 | 1 | 0 | 3 | 0 |

Table 8 CF explanation for instance ID-4

| Instance ID | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 74 | 0 | 1 | 120 | 269 | 0 | 0 | 121 | 1 | 0.2 | 2 | 1 | 2 | 1 |
| New_CFs | 74 | 0 | 1 | 120 | 269 | 0 | 0 | 121 | 1 | 6.0 | 0 | 1 | 2 | 0 |

Table 9 CF explanation for instance ID-5

| Instance ID | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 59 | 1 | 0 | 164 | 176 | 1 | 0 | 90 | 0 | 1.0 | 1 | 2 | 1 | 0 |
| New_CFs | 59 | 1 | 2 | 164 | 176 | 1 | 0 | 90 | 0 | 1.0 | 1 | 0 | 1 | 1 |

As shown in Table 8, the increase in oldpeak by the value of 0.4 and the change in slope from upward sloping (slope = 2) to flat (slope = 0) leads to a change in the intensity of heart disease. Thus, changing upward sloping to flat changes the output value of the XGBoost model.

As shown in Table 9, the change in slope from flat (slope = 0) to upward sloping (slope = 2) leads to a change in the intensity of heart disease. Thus, the slope value of heart disease changes the value of the XGBoost model output from "patient" to "not patient" or "healthy class".

CF can show the changes required in the values of the input feature for changing the diagnostic outcome of the XGBoost model to the desired outcome. Hence, an explanation of the XGBoost model diagnosis output provides insights into what can be carried out to change the diagnostic outcome. For example, what can the patient do to reduce the heart disease risk? The results in Tables 7-9 show that a small change in the value of the original input feature produces the desired diagnostic outcome. As shown in Table 9, only a change in the number of vessels colored by X-ray (ca) and chest pain (cp) could produce positive diagnostic outcomes while keeping other feature values constant. This shows that the CF explanation provided above is diverse, as the changes in more than one feature, namely the number of vessels colored by X-ray (ca) and chest pain (cp), have changed the outcome of the XGBoost model.

The quality of CFs is described in terms of the proximity of the CF value to the original feature, which is the Euclidean distance between the CF-generated and the original feature value. For instance, in Table 7, changing the heart rate value from 165 mmHg to 127 mmHg based on the CF explanation changes the model output from 1 to 0. Thus, the proximity or distance between the original heart rate value and the CF-generated heart rate value is 6.164414 mmHg. The CF-generated heart rate value is feasible as reducing the patient's heart rate by a value of 6.164414 mmHg is achievable.

### *4.7. Comparison of the existing work and current study*

This section compares the existing system and the current heart disease detection model. For comparison, accuracy is considered a performance measure. The comparative result shown in Table 10 illustrates that LIME and SHAP are effective in selecting the most influential feature to simultaneously improve the model performance and make the trade-off between the model interpretability and accuracy. Overall, this study shows that LIME and SHAP are significantly important to improve the model accuracy and explain the heart disease diagnosis outcome of the ML model.

Table 10 Comparison of the existing and current systems

| Study | Publication year | Algorithm employed | Model explainer method | Accuracy achieved (%) |
|---|---|---|---|---|
| [1] | 2020 | XGBoost | SHAP | 84.98% |
| [3] | 2021 | CNN | SHAP | 99.62% |
| [6] | 2021 | XGBoost | SHAP | 99.4% |
| [9] | 2020 | SVM | Grammar based | 96.31% |
| [10] | 2021 | RF | Feature selection | 99.3% |
| [11] | 2021 | CNN | Feature selection | 99.74% |
| This study | 2022 | XGBoost | SHAP and LIME | 99.79% |

## 5. Conclusions

In this study, the XGBoost model is proposed for heart disease detection. Moreover, LIME and SHAP are used to explain the diagnosis outcome of the proposed model. The simulation shows that the proposed model has promising performance as it diagnoses heart disease with an AUC score of 0.99. Moreover, the study identifies 10 influential features closely related to the risk of heart disease with the help of SHAP and LIME. With SHAP and LIME, the diagnostic outcome of XGBoost can be explained to clinicians to better understand the reason behind the model's diagnosis outcome. The result shows that the performance of the XGBoost model improves while making the trade-off between performance and interpretability, which makes the model suitable for clinical practice. The simulation results also show that by applying LIME and SHAP, better XGBoost models with reliable results and better performance can be developed. LIME provides only local instance-based explanations while SHAP can explain the model's prediction outcome at the local and global level. In the future, the researcher plans to test the performance and explainability of the XGBoost model by using other clinical datasets. The explainability will be evaluated with the results provided by LIME, SHAP, and DICE.

## Conflicts of Interest

The author declares that there are no conflicts of interest.

## Statement of Ethical Approval

For this type of study, a statement of human rights is not required.

## Statement of Informed Consent

For this type of study, informed consent is not required.

## References

[1] Y. Meng, et al., "What Makes an Online Review More Helpful: An Interpretation Framework Using XGBoost and SHAP Values," Journal of Theoretical and Applied Electronic Commerce Research, vol. 16, no. 3, pp. 466-490, June 2021.

[2] S. M. Lundberg, et al., "From Local Explanations to Global Understanding with Explainable AI for Trees," Nature Machine Intelligence, vol. 2, pp. 56-67, January 2020.

[3] S. M. Lundberg, et al., "Explainable Machine-Learning Predictions for the Prevention of Hypoxemia During Surgery," Nature Biomedical Engineering, vol. 2, no. 10, pp. 749-760, October 2018.

[4] M. Toğaçar, et al., "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, Article no. 103128, January 2022.

[5] P. Xie, et al., "An Explainable Machine Learning Model for Predicting In-Hospital Amputation Rate of Patients with Diabetic Foot Ulcer," International Wound Journal, vol. 19, no. 4, pp. 910-918, May 2022.

[6] C. Rudin, et al., "Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges," https://arxiv.org/pdf/2103.11251.pdf, July 2021.

[7] A. Gramegna, et al., "SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk," Frontiers in Artificial Intelligence, vol. 4, Article no. 752558, 2021.

[8] F. Cavaliere, et al., "Parkinson's Disease Diagnosis: Towards Grammar-Based Explainable Artificial Intelligence," IEEE Symposium on Computers and Communications, pp. 1-6, July 2020.

[9] S. J. Sushma, et al., "An Improved Feature Selection Approach for Chronic Heart Disease Detection," Bulletin of Electrical Engineering and Informatics, vol. 10, no. 6, pp. 3501-3506, December 2021.

[10] R. Bhagwat, et al., "A Framework for Crop Disease Detection Using Feature Fusion Method," International Journal of Engineering and Technology Innovation, vol. 11, no. 3, pp. 216-228, June 2021.

[11] K. Wang, et al., "Interpretable Prediction of 3-Year All-Cause Mortality in Patients with Heart Failure Caused by Coronary Heart Disease Based on Machine Learning and SHAP," Computers in Biology and Medicine, vol. 137, Article no. 104313, October 2021.

[12] T. Suresh, et al., "A Hybrid Approach to Medical Decision-Making: Diagnosis of Heart Disease with Machine-Learning Model," International Journal of Electrical and Computer Engineering, vol. 12, no. 2, pp. 1831-1838, April 2022.
[13] R. Bharti, et al., "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning," Computational Intelligence and Neuroscience, vol. 2021, Article no. 8387680, 2021.
[14] S. Rana, et al., "Slow Learner Prediction Using Multi-Variate Naïve Bayes Classification Algorithm," International Journal of Engineering and Technology Innovation, vol. 7, no. 1, pp. 11-23, January 2017.
[15] G. Vilone, et al., "Notions of Explainability and Evaluation Approaches for Explainable Artificial Intelligence," Information Fusion, vol. 76, pp. 89-106, December 2021.
[16] M. Toğaçar, et al., "Intelligent Skin Cancer Detection Applying Autoencoder, MobileNetV2 and Spiking Neural Networks Mesut," Chaos, Solitons, and Fractals, vol. 144, Article no. 110714, March 2021.
[17] A. F. Anderson, "Case Study: NHTSA's Denial of Dr Raghavan's Petition to Investigate Sudden Acceleration in Toyota Vehicles Fitted with Electronic Throttles," IEEE Access, vol. 4, pp. 1417-1433, 2016.
[18] T. A. Assegie, "Evaluation of the Shapley Additive Explanation (SHAP) Technique for Ensemble Learning Methods," Proceedings of Engineering and Technology Innovation, vol. 21, pp. 20-26, April 2022.