Int. J. Adv. Eng. Pure Sci. 2025, 37(UYIK 2024 Special Issue): <65-76>
DOI: 10.7240/jeps.1506705

RESEARCH ARTICLE / ARAŞTIRMA MAKALESİ

Analysis of Artificial Intelligence Methods in Classifying Heart Attack Risk: Black-Box Models vs. Glass-Box Models

Kalp Krizi Riskinin Sınıflandırılmasında Yapay Zekâ Yöntemlerinin Analizi: Kara Kutu Modelleri ve Cam Kutu Modelleri

Ebru GEÇİCİ [1], Eyüp Ensar IŞIK [1], Mısra ŞİMŞİR [1], Mehmet GÜNEŞ [1]

[1] *Yildiz Technical University, Faculty of Mechanical Engineering, Industrial Engineering Department, 34349, İstanbul, Türkiye*

Abstract
Artificial Intelligence (AI) is becoming more and more involved in human life day by day. Healthcare is one of the areas where AI is widely used, such as in the diagnosis prediction, and/or classification of diseases. Techniques such as machine learning provide high-accuracy results, but many algorithms have black-box structures, where the reasoning behind the predictions is not known. Explainable AI emerges to address this by providing explanations for complex models. While interpretable ("glass-box") models are desirable, they may have lower accuracy than complex ("black-box") models. Finding the right balance is crucial, especially in critical areas such as healthcare. It is also important to provide individual explanations for the predictions. This study uses patient data to explore a model to predict heart attack risk. Therefore, we compare glass-box models (logistic regression, Naïve Bayes, decision tree, and explainable boosting) with black-box models (random forest, support vector machine, multi-layer perceptron, gradient boosting, and stochastic gradient boosting). The results show that explainable boosting achieves the highest accuracy. To delve into individual explanations on a patient basis, the explainable boosting algorithm is compared with the random forest algorithm, which gives the best results among the black-box models. Here, LIME and SHAP are used to provide interpretability of random forest algorithm. As a result, it is concluded that the random forest algorithm has differences in the importance weights of the variables compared to the explainable boosting algorithm. Both results provide valuable tools for healthcare stakeholders to choose the most appropriate model.
Keywords: Artificial Learning, Explainable Artificial Intelligence, Classification, Healthcare Industry, Heart Attack

I. INTRODUCTION

Artificial Intelligence (AI) and AI-focused applications have increased frequently in recent years. This growth continues to be seen in AI's supporting applications that assist daily life, as well as its supporting role at critical decision points. At these points, AI's only result-oriented support to the decision maker can lead to problems explaining the reasons for the decision. With current studies in AI, complex models with improved prediction accuracy are being used, making explanations even more difficult. Such problems have led to the emergence of terms such as *understandability*, *comprehensibility*, *interpretability*, *explainability*, and *transparency* in AI [1]**.** The expected starting point of these terms is the need to explain how machine learning (ML) models make their outputs or decisions, which are becoming increasingly complex and cannot be explained by themselves. With the increase in this need, the study of eXplainable Artificial Intelligence (XAI) has increased significantly in recent years [2]**.**

Corresponding Author: EYÜP ENSAR IŞIK, **Tel:** 0212 383 28 66, **E-mail:** eeisik@yildiz.edu.tr
Submitted: 28.06.2024, **Revised:** 20.08.2024, **Accepted:** 04.10.2024

XAI emphasizes the explainability of what features may result from decisions while preserving the predictive accuracy of the methods used. Therefore, decision-makers need to choose a model that ensures prediction accuracy in the field under study but is also high in interpretability of the decision made [3]. In this context, ML models can be examined under two main headings: *glass-box* and *black-box* models. While glass-box (or white-box) models include self-interpretable methods, black-box models represent methods that cannot be interpreted automatically due to the high number of layers and parameters but can be provided with local explanations using different XAI techniques [4]. Although the prediction accuracy of black-box models is relatively high through their complex structure, they require additional techniques to explain the decisions made. Rudin [5] discusses using glass-box models instead of black-box models in the first stage to avoid spending too much effort explaining the black-box models.

The increasing use of AI applications in areas that directly affect social and human life increases the importance of using the suitable model in the right area. Although the accuracy of the decision is of utmost importance, especially in health-related studies, healthcare professionals who are decision-makers need to explain the circumstances under which this decision was made [6-7]. As a result of models used in areas such as disease detection, when deciding whether an individual has a disease, an explanation needs to be given to the individual, such as what factors caused the disease. For this reason, XAI studies in the healthcare sector have increased in recent years [8].

One of the areas where AI applications are frequently used in the healthcare sector is in detecting patients' heart attack risk [9]. Many studies have been made in this field, and models that increase prediction accuracy to the highest level have been proposed. However, the number of studies examining the explainability of the proposed models is quite limited. In this study, we compare the glass-box methods and black-box methods. In predicting heart attack risk, the differences between using a complex model that is difficult to explain due to its structure and a model that may not have deficient performance but is easily explainable are being discussed.

The rest of the paper is organized as follows: The literature review section presents related studies in this field. Then, the methodology is given in section 3. Section 4 addresses the application and its results. Last, we conclude in Section 5.

II. LITERATURE REVIEW

For many years, ML algorithms have been used in a wide range of applications in various fields, such as recommendation systems, cybersecurity, image processing, industrial applications, education, and healthcare. The literature section of this study provides an overview of studies on ML applications in healthcare. ML applications in the healthcare sector can be found in many areas, from disease diagnosis to personalized treatment, drug discovery to radiology, etc. The article reviews the applications of ML and AI algorithms in healthcare [10]. It states that support vector machines (SVM), decision trees (DT), random forests (RF), and artificial neural networks (ANNs) are widely used algorithms in this field [11-15].

Heart attack/stroke is one of the most critical and focused problems in healthcare, and it causes many deaths all over the world. The latest advances in the application of ML have shown that it is possible to detect heart disease at an early stage using electrocardiograms and patient data [16]. By analyzing large amounts of patient data, ML algorithms can more accurately and quickly identify risk factors for heart attack. In contrast to traditional methods, ML algorithms can use a patient's medical history, genetic information, and lifestyle to build more complex and predictive models. This enables physicians to monitor patients more effectively, identify high-risk individuals in advance, and take the necessary preventive measures.

Sahu et al. [17] compare conventional ML algorithms (SVM, Naïve Bayes (NB), DT, RF, Logistic Regression (LR), k-nearest neighborhood (KNN)) and deep learning algorithms for using two different data sets (taken from the UC Irvine (UCI) and Kaggle repositories) to predict the heart attack and death rates related to heart attack [17]. They conclude that the one-dimensional convolutional neural network (1D-CNN) algorithm predicts heart attack and death rates with 99% accuracy, outperforming conventional methods. Rao et al. [18] attempt to predict whether a patient will have a heart disease. LR and ANN models are used to classify them. They compare the accuracy rates of the two models, and LR outperforms ANN by 90%. Mahmud et al. [19] try to predict heart failure in patients based on clinical data. They use one of the well-known data sets (taken from Kaggle repositories), which combines five different cardiac datasets, making it the most comprehensive resource available for heart disease research. RF, NB, KNN, and DT methods are used to build a combined meta-model—the results of the evaluation show that the meta-model outperforms other state-of-the-art models. The accuracy of the meta-model is 87%. Mamun and Elfouly [20] introduce a hybrid 1D-CNN model utilizing features selected by feature selection algorithms as well as a substantial data set derived from online survey data. The 1D-CNN has shown superior accuracy compared to contemporary ML algorithms and ANNs. The model's performance is compared with ANN, RF, AdaBoost, and SVM, and 1D-CNN outperforms these methods in terms of accuracy, false negative rates, and false positive rates. Ozcan and Peker [21] introduce a classification and

regression tree chart, a supervised ML method to predict whether a patient will have a heart disease. They try to explain the relationship between the input variables and the response, so they rank the features that affect heart disease by importance. The accuracy of the proposed algorithm, which is 87%, shows the reliability of the model. Yu [22] uses ML algorithms to predict the likelihood of occurrence of heart diseases in patients. The data set taken from the UCI repository is used to analyze eight different ML classifiers. LR, SVM, KNN, NB, DT, RF, gradient boosting, and AdaBoost algorithms are compared. As a result, the gradient boosting classifier achieves the highest accuracy with 95.08%.

XAI studies have increased in recent years due to the need to learn which inputs result from the decisions obtained from increasingly complex ML models [2,15]. Although the studies in this field use different terms representing similar needs, each term does not have the same meaning. Although there is no clear definition for XAI due to different terminologies, to provide a consensus, Barredo Arrieta et al. [1] define XAI as "a set of practices that produce details or reasons to make its functioning clear or understandable, given an audience.". The authors state that this definition indirectly includes causality, transferability, informativeness, fairness, and reliability, which are seen as missing in other definitions but are covered by XAI. As can be understood from this definition and the topics it should include, XAI provides explanations for AI applications that serve many purposes and enable us to understand AI models better.

AI models can be divided into two categories in terms of explainability. The former is models that can be explained independently without using additional techniques. These models can be found in the literature under “transparent models, glass-box models, intrinsic explainability, ante-hoc approaches, and inherently interpretable ML models” [4]. This study uses the term "glass-box models" for self-explanatory models. Although their explainability is at different levels, linear regression/LR, DT, KNN, rule-based methods, general additive models, and Bayesian models are considered glass-box models [1]. These models can be explained after the prediction without any post-hoc analysis. On the other hand, models that cannot be explained by themselves due to their complex structures and where the obtained predictions can only be explained by post-hoc analyses are called "black-box models" in the literature. RF, SVM, multi-layer perceptron (MLP), and ANNs are examples of black-box models. It is crucial to be able to explain these models that provide high prediction accuracy. Therefore, new techniques have been developed to explain black-box models, and they can be divided into two categories: (i) model-agnostic and (ii) model-specific. The most well-known of these techniques are SHapley Additive exPlanations (SHAP) [24] and Local Interpretable Model-Agnostic Explanations (LIME) [25], which are classified as model-agnostic.

Literature reviews conducted in recent years clearly reveal how popular XAI is. In addition to providing information about the current terminology in the field, these studies also present newly developed methods and application areas in detail [1-3], [23], [26]. Although it has applications in many fields, such as finance, education, environmental science, and agriculture, XAI stands out, especially with its uses in healthcare.

III. METHODOLOGY

This part of the study presents information on the glass-box and black-box models used in the analysis. LIME and SHAP, which allow local interpretation of black-box models, are examined. Then, performance metrics that allow the evaluation of the built models are discussed.

3.1. Artificial Learning Algorithms

The ML and ANN models used in the analyses are presented in this section. In this context, the glass-box and black-box methods are first considered, and then the methods that enable the black-box models to be explained locally are mentioned. This section also provides information about the performance metrics used to compare the models created.

ML techniques are divided into two categories: supervised learning and unsupervised learning [27]. This distinction is related to the presence or absence of the output value in the data set: (i) if the dependent variable (response), y, is present in the data set, it is called supervised learning, and (ii) if there is no dependent variable in the data set, it is called unsupervised learning. Furthermore, supervised learning is divided into prediction and classification according to the structure of the dependent variable in the data set. While regression is used to predict the dependent variable, which has a continuous structure, classification involves classifying data using the output, which has a discrete structure. Unsupervised learning algorithms, on the other hand, are preferred for purposes such as making inferences about the data or organizing the data set (such as dimensionality reduction), and the well-known applied methods in this field are clustering algorithms. Clustering, which falls under unsupervised learning, involves grouping processes by bringing together independent variables with similar characteristics [27]. In this study, the aim is to evaluate whether patients are at risk of having a heart attack. The response value has a binary structure (0: lower heart attack risk and 1: higher heart attack risk), i.e., the response is a discrete variable. In this context, the glass-box and black-box classification methods are examined below.

LR measures the statistical significance of each independent variable relative to probability. It is highly probabilistic and a powerful ML method that models binomial output [28]. The *NB* method is one of the methods based on Bayes theorem and is called the probabilistic classification method [29]. The naïve assumption is an assumption of conditional independence between each pair of features given the value of the class variable. Moreover, it is a preferred algorithm because it is easier to use and understandable and gives faster results than other complex methods [30]. In the *DT classification*, each node represents a feature, each branch illustrates a rule, and each leaf gives a result [31]. DTs have a hierarchical structure developed by dividing the data set into smaller structures. The *Explainable boosting machine (EBM) classifier* is a cyclic gradient boosting generalized additive model, and similar to the DTs, it is also a tree-based method. Furthermore, these models have as high prediction accuracy as black-box models, but their interpretation is inherently easier than black-box models.

Many DTs work together to create an RF algorithm, and then the average of all these trees is used [28]. This structure allows more consistent results to be obtained compared to DT algorithm results [27]. *SVM classification* creates hyper-planes to separate data into multiple classes [32]. Unlike other classification algorithms, it tries to maximize the distance between the created clusters. It considers the separation of points by a line or plane and the resulting distance. *MLP* is one of the well-known ANN algorithms that can indirectly detect complex nonlinear relationships between dependent and independent variables [33]. This method is inspired by the working structure of the human brain; inputs pass through layers respectively, and output is created. Each layer consists of neurons, and the values obtained from here are obtained by passing through the activation function. *Gradient boosting* is one of the ML ensemble methods that create more than one model and then combine them to produce improved results. These well-known ML models reunite several weak learners into strong learners, in which each new model is trained to minimize the loss function appropriate to the structure of the problem. At each step, the algorithm calculates the direction of improvement for the ensemble's predictions and then trains a new weak model to move in that direction. The new model's predictions are then added to the ensemble, and this process is repeated until a stopping condition is met. As in the gradient descent method, the stochastic gradient descent model tries to minimize the loss function value defined iteratively. The reason for using the concept of stochastic in its name is based on using structures such as applying mini-batches of differentiation in iterations and creating random subsets. In this way, while trying to reach the highest efficiency value in calculations, the randomness value also increases.

The local interpretation of the black-box models given above is not straightforward due to the complex structure of the models. Therefore, intermediary stages are needed to evaluate these models with respect to observations. In this context, *LIME* and *SHAP* are used in this study. These methods are visualization techniques applied to ML algorithms and are recommended to explain the model by bringing the model predictions closer to an interpretable model. To explain individual predictions, LIME creates new data points that resemble the instance of interest. These points are generated based on a statistical model learned from the features of the dataset, treating them as independent variables. Note that it is considered that the features are independent of the other and follow a normal distribution, whose parameters are inferred from the data set [34]. SHAP combines game theory concepts with local explanation techniques [24]. SHAP transforms the original input data into a more straightforward form using a specific function. In this model, using a reduced data set, the original model can be approximated with a linear function of binary variables [35].

3.2. Evaluation of the Models: Performance Metrics

The explanation of the algorithms used in the analyses provides information about the output of whether a heart attack has occurred. The output value in the data set is considered as a 0 – 1 binary structure, and classification algorithms are used to select the model. Similarly, the definition of the performance metrics used to compare algorithms should also be appropriately chosen for the output structure. In this context, the performance metrics are accuracy, precision, recall, and f1-score. Moreover, the receiver operating characteristic (ROC) curve and the area under the ROC curve value (AUC) are also reported.

A confusion matrix must be created to use the above performance metrics in classification problems. The values in the confusion matrix are used to visualize and summarize the results. In this context, the confusion matrix is created for problems with binary output values, as shown in Table 1.

Table 1. Confusion matrix for a problem which has a binary structure

		Predicted Values	
		Positive (1)	Negative (0)
Actual Values	Positive (1)	True Positive (TP)	False Negative (FN)
	Negative (0)	False Positive (FP)	True Negative (TN)

Accuracy, one of the performance metrics, shows the percentage of samples classified correctly and is mathematically expressed as $\frac{TP+TN}{TP+TN+FP+FN}$. *Precision*, another performance metrics, is used to calculate how many of the values predicted as positive are actually true positives. Its mathematical expression is presented

as $\frac{TP}{TP+FP}$. The *recall* metric, like precision, deals with positive values, but unlike the previous one, recall is a performance metric that shows how much of the operations that should be predicted as positive are predicted as positive (Mathematically, $\frac{TP}{TP+FN}$). The *f1-score* value is calculated as the harmonic average of recall and precision values, i.e., $2 \times \frac{Precision \times Recall}{Precision + Recall}$.

ROC curve is a metric that allows visual evaluation, unlike the performance metrics given above. True positive rate and false positive rate are used to obtain this curve. While the false positive rate value on the x-axis is calculated using $\frac{FP}{FP+TN}$, the true positive rate corresponds to the recall value explained above. *AUC* is obtained by calculating the area under the ROC curve. Their expressions are shown in Figure 1. In Figure 1, the dark blue line corresponds to the ROC curve, while the gray area corresponds to the AUC value.

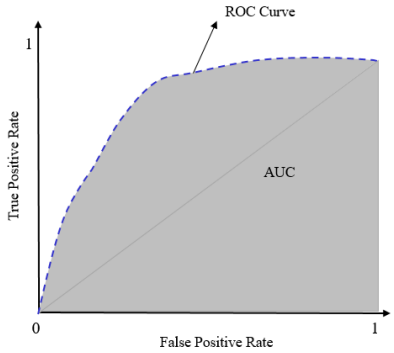


Figure 1. Representation of the ROC curve and AUC value in a graph

IV. APPLICATION AND RESULTS: PREDICTION OF HEART ATTACK RISK

In this part of the study, the methods specified are analyzed using the selected data set, and the results are presented. First, information about the data set is given, and then the application details are addressed.

4.1. Data Set

The data set "heart attack" to be used in the analysis has been shared with researchers and users as open access [36]. The data set contains information about whether the individuals whose information is included have had a heart attack. Moreover, it consists of a total of 303 observation values, 13 features, and one output value. Definitions of the features are included in Table 2.

4.2. Application and Results

This section provides information on the application and the results obtained. AI algorithms classified as glass-box and black-box in the Methodology section are used to select the model to be created to determine the risk of a heart attack. The created models are run using a computer with an Intel(R) Core(TM) i5-10210U CPU @ 1.60GHz 2.11 GHz processor and 8 GB RAM. Python programming language and existing libraries are preferred when creating the models. Accordingly, the libraries used for artificial learning are LR (*LogisticRegression*), NB (*GaussianNB*), DT classifier (*DecisionTreeClassifier*), SVM classifier (SVC), MLP (*MLPClassifier*), gradient boosting (*GradientBoostingClassifier*), stochastic gradient boosting (*SGDClassifier*) and explainable gradient boosting (*ExplainableBoostingClassifier*), whereas the libraries used to explain the black box models are LIME (*LimeTabular*) and SHAP (*shap*).

Table 2. Explanation of the features

Feature	Explanation	Structure
age (years)	The age of individuals	Integer
sex	The gender of individuals *Categories:* 0: female, 1: male	Categorical
cp	The chest pain type *Categories:* 1: typical angina, 2: atypical angina,3: non-anginal pain, 4: asymptomatic	Categorical
trestbps (mm Hg)	The resting blood pressure (on admission to the hospital)	Integer
chol (mg/dl)	The serum cholesterol level fetched via BMI sensor	Integer
fbs	The fasting blood sugar > 120 mg/dl *Categories:* 1: true, 0: false	Categorical
restecg	The resting electrocardiographic results *Categories:* 0: normal, 1: having ST-T wave abnormality, 2: definite left ventricular hypertrophy by Estes' criteria	Categorical
thalach	The maximum heart rate achieved	Integer
exang	The exercise induced angina *Categories:* 1: yes, 0: no	Categorical
oldpeak	The ST depression induced by exercise relative to rest, previous peak	Integer
slope	The slope of the peak exercise ST segment *Categories:* 1: unsloping, 2: flat, 3: downsloping	Categorical
caa	The number of major vessels (0-3) colored by flourosopy	Integer
thal	*Categories:* 3: normal; 6: fixed defect; 7: reversable defect	Categorical

According to the specified information, the data set is first analyzed, and then models are established. After the model-building phase is completed, the best model must be selected. Then, the results must be explained so that decision-makers can understand; for example, in this problem, the decision-makers are physicians. However, as mentioned, the artificial learning algorithms used in this study are presented as two pillars: (i) glass-box and (ii) black-box. Due to their structure, the models referred to as glass-box can be understood and interpreted by experts in the field who do not know ML. On the other hand, suppose a model under the black-box heading is chosen. In that case, it becomes difficult for field experts to interpret the results and interpret the data from an individual perspective. For this reason, explanatory methods are used to help explain black-box models. The models obtained afterward are the best among the established models so that the outputs can be easily interpreted. This process to be followed during the implementation phase is visualized in Figure 2.

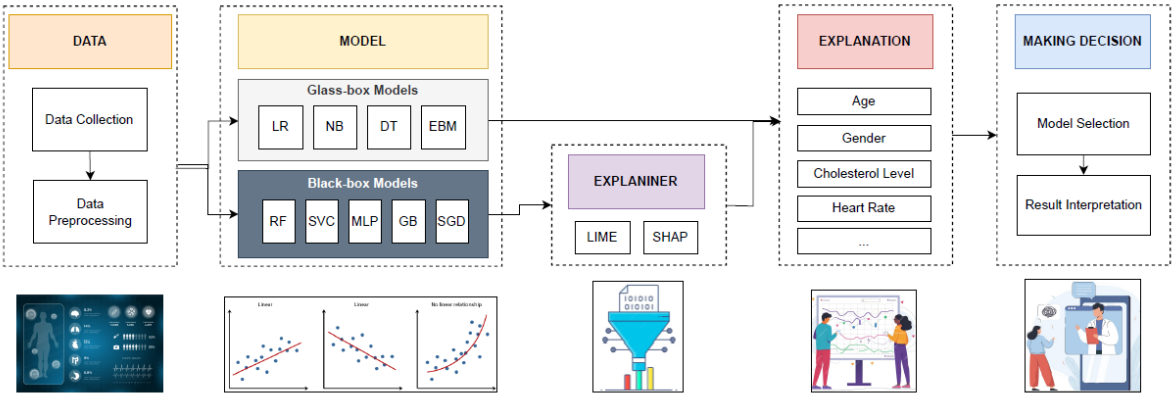


Figure 2. Flow diagram to be followed for obtaining and interpreting the results

A fine-tuned process is carried out to prepare the models for use and select the best among them. Through this process, the hyper-parameters, which are parameter values that are given externally to the model, of the models are determined. For this process, more than one value is tried for the relevant parameter in each model. The fine-tuned parameters for each model are summarized in Table 3. The established artificial learning model is included in the "Models" column in the table. Under the heading "Hyper-Parameters," hyper-parameters are differentiated in creating the established models. "Best Model Parameters" illustrates the hyper-parameters used to prepare the best model obtained due to the different hyper-parameter values run for the relevant model. Moreover, while giving the information in this column, hyper-parameter definitions are presented precisely the same as the name in the library included in the package used. Thus, it is aimed to create a clear table for users who want to use the same models. Besides, the grid search algorithm is used to fine-tune the models. After completing this process, the parameters that yield the best results are reported for each model created using the specified method. Thus, models to be compared are obtained to select the best model. The last column, "Time (sec)," presents the time to determine the best parameters among the existing ones during the fine-tuned process. No duration has been defined since existing models are used for LR, NB, and explainable boosting methods. In other models, it is seen that the method that requires the most time in the parameter definition phase is the gradient boosting method, and the method that requires the least time is the MLP method.

Table 3. Artificial learning models with hyper-parameters tuned by grid search in the inner loop

Models	Hyper-Parameters	Best Model Parameters	Time (sec)
LR	-		
NB	-		
DT Classifier	• *Criterion*: Measure the split quality and is a measure of impurity. • *Splitter*: Define the strategy which is used to choose split at each node. • *Maximum Depth*: Give the tree's maximum depth. • *Maximum Features*: The number of features to consider when looking for the best split. • *Complexity Parameter*: Used for minimal cost-complexity pruning.	*criterion: gini* *splitter: best* *max_depth: 1* *max_features: None* *ccp_alpha: 0*	20.0325
RF Classifier	• *Criterion*: Measure the split quality and is a measure of impurity. • *Number of Estimator*: The number of trees in the forest. • *Maximum Depth*: Give the tree's maximum depth. • *Weights associated with classes*	*criterion: gini* *n_estimators: 200* *max_depth: 4* *class_weight: balanced_subsample*	605.7901

Table 3. Artificial learning models with hyper-parameters tuned by grid search in the inner loop (cont.)

Models	Hyper-Parameters	Best Model Parameters	Time (sec)
SVM Classifier	• *Kernel*: The kernel type to be used in the algorithm • *Gamma:* Kernel coefficient: • *Regularization parameter*	*kernel: rbf* *gamma: auto* *C: 3.0*	1.6822
MLP Classifier	• *Activation*: The hidden layer activation function • *Solver*: The solver for weight optimization • *Alpha*: Strength of the L2 regularization term • *Learning rate* • *Maximum number of iterations*	*activation: logistic* *solver: adam* *alpha: 0.0* *learning_rate: constant* *max_iter: 300*	2703.402
Gradient Boosting Classifier	• *Loss*: Type of the loss function • *Learning Rate*: Learning rate shrinks the contribution of each tree • *Criterion*: Measure the split quality • *Maximum Depth*: Maximum depth of the individual regression estimators • *Max Features*: The number of features to consider when looking for the best split.	*loss: log_loss* *learning_rate: 0.9* *criterion: squared_error* *max_depth: 8* *max_features: sqrt*	820.8958
Stochastic Gradient Descent Classifier	• *Loss*: Loss function type • *Penalty*: Regularization term • *Alpha*: Coefficient of the regularization term • *Learning Rate*	*loss: log_loss* *penalty: l2* *alpha: 0.003* *learning_rate: optimal*	23.3565
Explainable Boosting Classifier	-		

After the fine-tuning process with grid search, the best models are compared. A run is taken using the hyper-parameters determined to compare different models. As a result of these runs, the models' performance metrics calculated using the training data are shown in Table 4. Different performance metrics are listed in this resulting table. The headings in the table provide information regarding the classification of the models operated in the first column as glass-box or black-box. The model number and the name of the model are shared in the following two columns. Furthermore, there are four subheadings under the heading "*Performance Metrics*": "*Accuracy*," "*Recall*," "*Precision*," and "*F1-Score*". The results for each model are presented for these performance metrics specified under subheadings.

When the results in Table 4 are examined, it is seen that the best results among the models called glass-box are obtained with the explainable boosting method. Black-box models are examined; on the other hand, it is observed that the best results are obtained 100% with the gradient-boosting classifier algorithm. Moreover, these results should be evaluated in terms of over-fitting. Thus, the performance metrics obtained for the test data set are considered to determine whether the models are usable. In this context, the performance metrics of the test data are listed in Table 5. Table 5 is created to resemble Table 4, where the values obtained with the training data set are reported. In addition to existing performance metrics, the AUC value is also reported. The graph of the AUC values given with the performance metrics of the test values is shown in Figure 3.

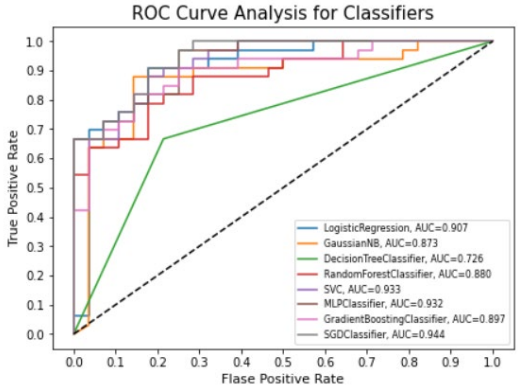


Figure 3. ROC curve graphs obtained in models for test data

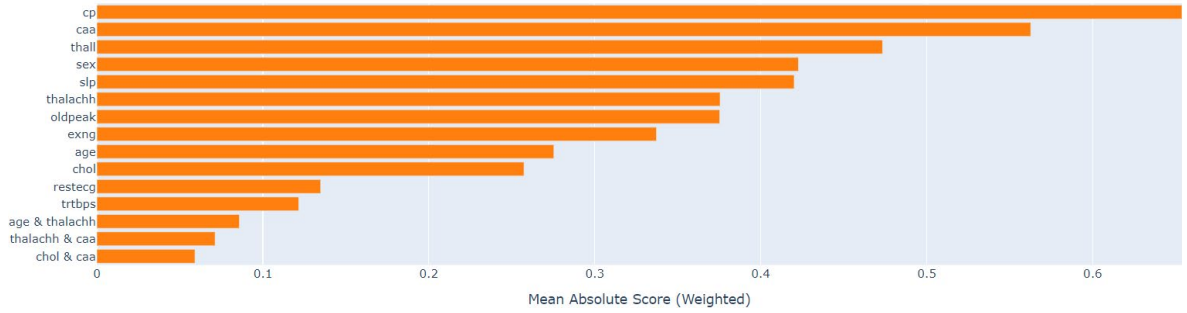
Table 4. Performance metrics for train data set

	#	Model	Performance Metrics (%)			
			Accuracy	Recall	Precision	F1-Score
Glass Box Models	1	Logistic Regression	88.8430	88.8430	88.9863	88.9146
	2	Naïve Bayes	67.7686	67.7686	77.0972	72.1325
	3	Decision Tree	77.6860	77.6860	77.6528	77.6694
	4	Explainable Boosting	***95.0413***	***95.0413***	***95.0489***	***95.0451***
Black Box Models	1	Random Forest	88.0165	88.0165	88.0128	88.0147
	2	Support Vector Machine	85.9504	85.9504	86.0250	85.9877
	3	Multi-Layer Perceptron	86.7769	86.7769	86.8602	86.8185
	4	Gradient Boosting	***100.0000***	***100.0000***	***100.0000***	***100.0000***
	5	Stochastic Gradient Descent	86.3636	86.3636	86.5579	86.4607

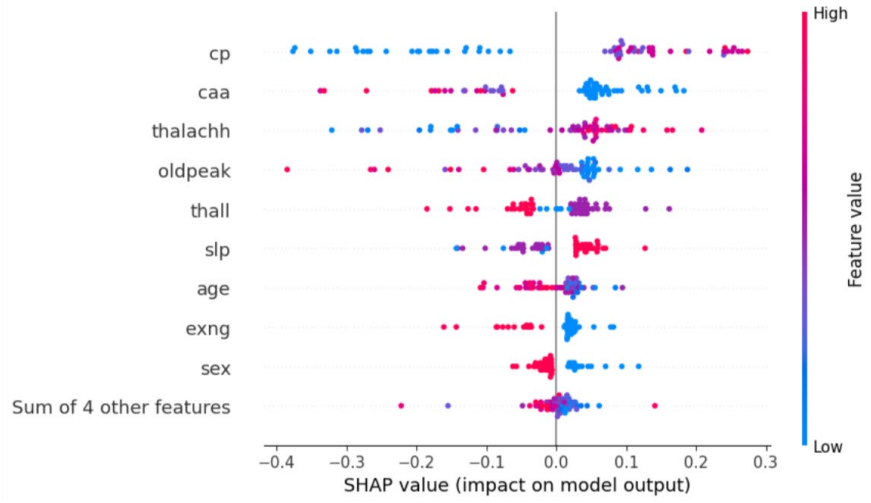
Table 5. Performance metrics for test data set

	#	Model	Performance Metrics (%)				
			Accuracy	Recall	Precision	F1-Score	AUC
Glass Box Models	1	Logistic Regression	81.9672	81.9672	83.0761	82.5179	0.907
	2	Naïve Bayes	75.4098	75.4098	81.6214	78.3927	0.873
	3	Decision Tree	72.1311	72.1311	73.1069	72.6158	0.726
	4	Explainable Boosting	***77.0492***	***77.0492***	***76.8275***	***76.9382***	***0.872***
Black Box Models	1	Random Forest	***75.4098***	***75.4098***	***77.2509***	***76.3193***	***0.880***
	2	Support Vector Machine	80.3279	80.3279	81.8386	81.0762	0.933
	3	Multi-Layer Perceptron	78.6885	78.6885	79.7530	79.2172	0.930
	4	Gradient Boosting	80.3279	80.3279	82.9217	81.6042	0.897
	5	Stochastic Gradient Descent	81.9672	81.9672	82.4220	82.1940	0.944

Suppose the performance metrics obtained for the test and train data sets are examined simultaneously. In that case, it is seen that in most of the models, better results are obtained with the train data set, and there are decreases in these metrics for the test data. Moreover, this decrease is greater in black-box models than in glass-box models. In addition, the test performance metrics for the gradient boosting algorithm, which achieved 100% success, are lower than other black-box models. This shows that over-fitting is explicitly observed for this model. For this reason, if black-box models are established and a model is selected, it would be appropriate to choose the RF method, where the change between train and test is less and good results are obtained in the training data. When the glass-box models are examined, it becomes clear that the explainable boosting algorithm is a usable model in terms of giving good results in training and good results in terms of testing.



(a)



(b)

Figure 4. Global feature importance for the selected models: (a) Explainable boosting classifier and (b) RF with SHAP

After glass-box (explainable boosting classifier) and black-box (RF) models are selected, they need to be evaluated in terms of explainability. As mentioned before, due to their structure, glass-box models are easier to interpret the results individually than black-box models. In this context, a summary of the variables and information within the scope of the variables can be obtained with the explainable boosting model. In addition, comments can be obtained for individuals. However, this process cannot be performed directly in black-box models. For this reason, this information can be obtained using methods such as LIME and SHAP. Before explaining the examples locally, we can present the weights of the variations obtained by explainable boosting, as in Figure 4(a). The graph in Figure 4(a) shows importance weights on the x-axis and features on the y-axis. According to the information obtained from this graph, the "*cp*" variable corresponding to the chest pain type is the most crucial feature in the explainable boosting classifier model. In contrast, it seems that the least important feature is the parameter formed by the "*chol&caa*" combination. Similarly, importance levels of variables can be obtained for RF using SHAP. Since different functions are used, the visualization of the results also varies. As can be seen from Figure 4(b), the most crucial variable is "*cp*." In the results obtained by evaluating RF in terms of explainability, it is seen that the variables have similarities in terms of importance.

In addition to the importance of the weights of the variables for the model, another essential feature of XAI is that it provides the opportunity for *local interpretation*. In other words, it means explaining each observation value (in this study, patients whose heart attack risk is measured are expressed). In this context, an observation value in the data set is chosen randomly, and local explanations are given based on this observation value. Since the explainable boosting classifier and LIME use the same library, namely *interpret*, the resulting graphics have a similar visual structure. SHAP, on the other hand, shows a different visuality because it comes from a different library. As mentioned before, an observation value is chosen randomly. In this context, the results of the 11th observation are shown in Figure 5, respectively.

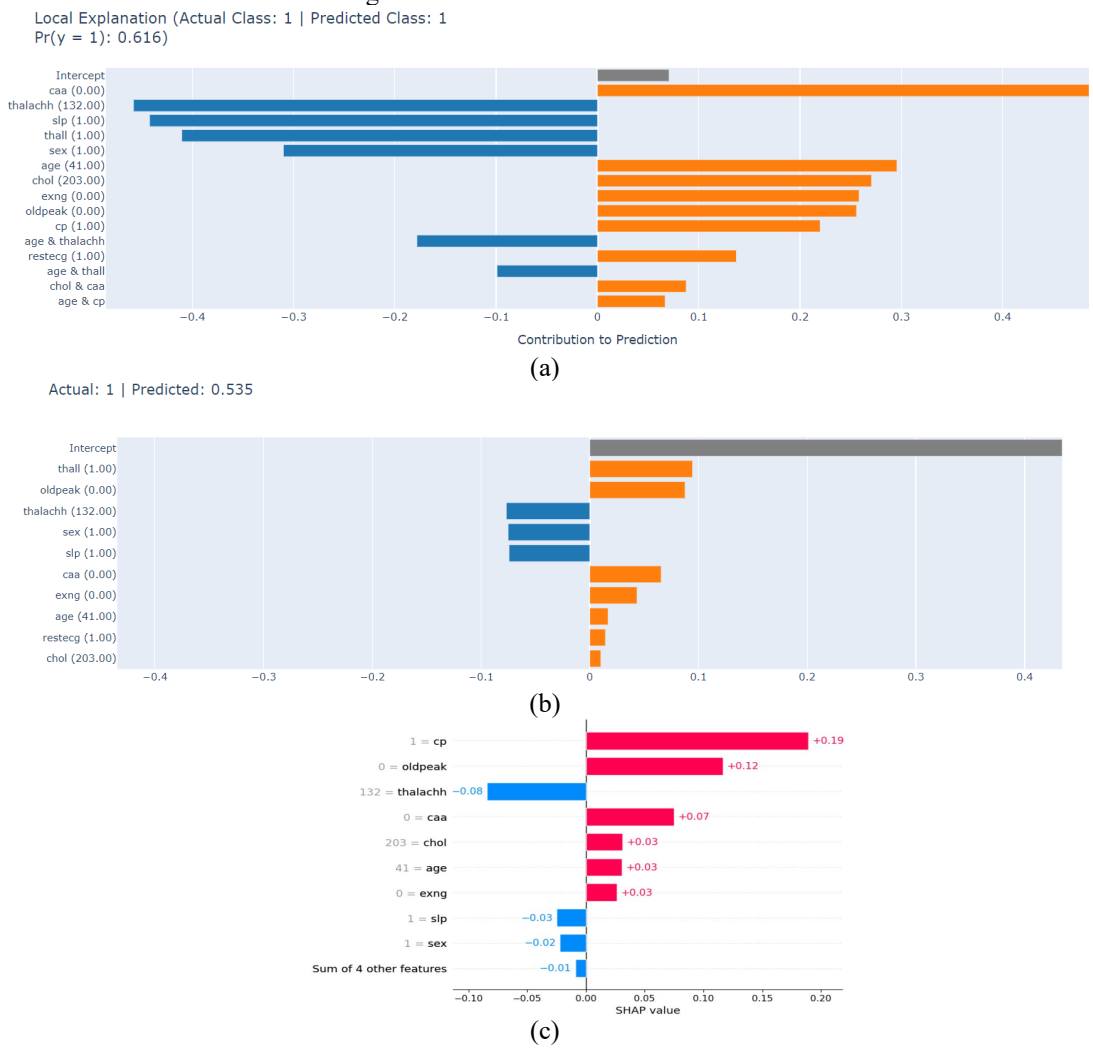


Figure 5. Local explanation of observation 11: (a) Explainable boosting classifier, (b) RF with LIME, and (c) RF with SHAP

In all three graphs in Figure 5, the x-axis shows the importance values of the variables for that patient. The y-axis shows the variables and the patient's values of those variables. For example, for this patient, the gender is shown as 1, and the age variable is expressed as 41. It can be seen from this information that it contains data from a 41-year-old male patient. In addition, the chest pain type value of this patient, expressed as "*cp = 1*", shows that the patient has typical angina. Moreover, the patient information and the observation value are estimated to have a high risk of heart attack in both models (explainable boosting classifier and RF). The importance of the variables that affect this estimate can also be obtained using the graphs in Figure 5.

The graphs in Figure 5 should be evaluated to show the effect of the variables in calculating the risk of having a heart attack after being used to obtain the patient's data. The most effective variable in classifying the risk in the explainable boosting algorithm for the patient whose information is given in Figure 5 is "caa," whereas the most effective variable for RF visualized with LIME is "thall" and the variable visualized with SHAP is "cp.". Similarly, it is seen that the influencing variables change as the effects of the variables change. Considering only the RF algorithm, using different visualization techniques for the same model also causes the variables and their effects to differ. According to the information obtained here, selecting and individually evaluating the method for problems that significantly impact human life, such as the models used to determine the risk of heart attack, is vital. In addition, presenting the results to decision-makers and physicians in this problem, with more than one explanatory model, will be effective in determining the treatments to be applied.

V. CONCLUSION

In healthcare, AI rapidly transforms how we diagnose, predict, and classify diseases. ML techniques have proven to be powerful tools, delivering impressive accuracy. However, a major hurdle lies in the complexity of some AI models. These models, often called "black-box" models, can generate highly accurate predictions but lack transparency in their reasoning process. This lack of clarity can be concerning, particularly in critical areas like healthcare decision-making, affecting human life. XAI fills this gap by providing explanations for the complex calculations performed by these models. Ideally, researchers prefer models with complete interpretability, also known as "glass-box" models. However, these models may compromise accuracy for transparency. Finding the right balance between these two aspects is crucial for XAI implementation in healthcare.

This study addresses this challenge by investigating a model specifically designed to classify heart attack risk based on patient data. LR, NB, DT, and explainable gradient boosting algorithms, called glass-box models, are developed in this context. In addition to these models, RF, SVM, MLP, gradient boosting, and stochastic gradient boosting algorithms, classified as black-box models, are established. To select the best model among the developed models, fine-tuning is done by running the models with different parameters using grid search. The results of the fine-tuning processes, whose results provide the best models of the methods within themselves, are compared. Considering all the proposed models, it is seen that the best results are obtained with the explainable gradient boosting algorithm. In addition, the best performance in black-box models is obtained as RF when considering glass-box and black-box models. The ability to explain predictions on an individual patient basis is also essential. The RF algorithm is visualized using LIME and SHAP explanation methods to compare the patient-based description of these two methods. These methods are used to unveil the inner workings of the RF model, making its predictions more interpretable. When explainable gradient boosting and RF are compared, it is seen that the importance of the features changes, and different features are considered in determining the risk of heart attack. Moreover, this study highlights another crucial point: even within the same model (RF in this case), the choice of interpretability technique (LIME vs. SHAP) can influence the perceived importance of variables. This underscores the importance of careful method selection and individual evaluation, especially when dealing with high-impact domains like healthcare, where decisions can influence life-and-death situations. Furthermore, presenting physicians with multiple interpretable models can be highly beneficial. By considering diverse perspectives on the data, physicians gain a richer understanding of the factors contributing to a patient's heart attack risk. This comprehensive view can empower them to make more informed decisions about each individual's most effective treatment course.

That is, findings offer valuable insights for healthcare professionals. By understanding the strengths and limitations of different AI models, they can make informed decisions about which tool is best suited for their specific needs, striking a crucial balance between accuracy and interpretability in the healthcare field. Future studies can expand the study regarding data set size, application area, and model. In this context, by expanding the data set, the validity of the application can be ensured, and its applicability in large data can be addressed. Apart from heart attacks, applications can be created for different cardiovascular diseases and diseases in different areas. In addition, it can be applied not only in the health field but also in areas that affect human and living life and require individual evaluation. On more extensive data sets, deep learning algorithms

can also be used, and the performance of the explanation methods can be examined. Finally, by developing hybrid models, the model's accuracy rate and interpretability rate can be improved.

REFERENCES

[1] Arrieta, A.B. et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. fusion*, vol. 58, pp. 82–115.

[2] Longo, L. et al. (2024). Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Inf. Fusion*, p. 102301.

[3] Langer, M. et al. (2021). What do we want from Explainable Artificial Intelligence (XAI)?--A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artif. Intell*., vol. 296, p. 103473.

[4] Retzlaff, C.O. et al. (2024). Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists. *Cogn. Syst. Res.*, vol. 86, p. 101243.

[5] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215.

[6] Cinà, G., Röber, T., Goedhart, R., and Birbil, I. (2022). Why we do need explainable ai for healthcare, *arXiv Prepr*. arXiv2206.15363.

[7] Wysocki, O. et al. (2023). Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. *Artif. Intell.*, vol. 316, p. 103839.

[8] Nasarian, E., Alizadehsani, R., Acharya, U.R., and Tsui, K.-L. (2024). Designing interpretable ML system to enhance trust in healthcare: A systematic review to proposed responsible clinician-AI-collaboration framework. *Inf. Fusion*, p. 102412.

[9] Riyaz, L., Butt, M.A., Zaman, M., and Ayob, O. (2022). Heart disease prediction using machine learning techniques: a quantitative review, in International Conference on Innovative Computing and Communications: *Proceedings of ICICC 2021*, Volume 3, pp. 81–94.

[10] Habehh, H. and Gohel, S. (2021). Machine learning in healthcare. *Curr. Genomics*, vol. 22, no. 4, p. 291.

[11] Liang, Z., Zhang, G., Huang, J.X., and Hu, Q. V. (2014). Deep learning for healthcare decision making with EMRs, *in 2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 556–559.

[12] Patel, M.J., Andreescu, C., Price, J.C., Edelman, K.L., Reynolds III, C.F. and Aizenstein, H.J. (2015). Machine learning approaches for integrating clinical and imaging features in late-life depression classification and response prediction. *Int. J. Geriatr. Psychiatry*, vol. 30, no. 10, pp. 1056–1067.

[13] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, vol. 542, no. 7639, pp. 115–118, doi: 10.1038/nature21056.

[14] o'Brien, A. R., Wilson, L.O.W., Burgio, G. and Bauer, D.C. (2019). Unlocking HDR-mediated nucleotide editing by identifying high-efficiency target sites using machine learning. *Sci. Rep.*, vol. 9, no. 1, p. 2788.

[15] Pan, X., et al. (2020). ToxDL: deep learning using primary structure and domain embeddings for assessing protein toxicity. *Bioinformatics*, vol. 36, no. 21, pp. 5159–5168.

[16] Ahsan, M.M. and Siddique, Z. (2022). Machine learning-based heart disease diagnosis: A systematic literature review. *Artif. Intell. Med.*, vol. 128, p. 102289.

[17] Sahu, R., Mohanty, K., Dash, S.R., Brahnam, S., and Barra, P. (2023). Prediction of Heart Attack and Death: Comparison Between 1 DCNN and Conventional ML Approaches, *in 2023 1st International Conference on Circuits, Power and Intelligent Systems (CCPIS)*, pp. 1–6.

[18] Rao, K.D., Kumar, M.S.D., Akshitha, D. and Rao, K.N. (2022). Machine Learning Based Cardiovascular Disease Prediction, *in 2022 International Conference on Computer, Power and Communications (ICCPC)*, pp. 118–122.

[19] Mahmud, I., Kabir, M.M., Mridha, M.F., Alfarhood, S., Safran, M. and Che, D. (2023). Cardiac Failure Forecasting Based on Clinical Data Using a Lightweight Machine Learning Metamodel. *Diagnostics*, vol. 13, no. 15, p. 2540.

[20] Khan Mamun, M.M.R. and Elfouly, T. (2023). Detection of Cardiovascular Disease from Clinical Parameters Using a One-Dimensional Convolutional Neural Network. *Bioengineering*, vol. 10, no. 7, p. 796.

[21] Ozcan, M. and Peker, S. (2023). A classification and regression tree algorithm for heart disease modeling and prediction. *Healthc. Anal.*, vol. 3, p. 100130.

[22] Yu, H. (2023). Analysis and Prediction of Heart Disease Based on Machine Learning Algorithms, *in In 2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP)*, pp. 1418–1423.

[23] Saeed, W. and Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Syst.,* vol. 263, p. 110273.

[24] Lundberg, S.M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.*, vol. 30.

[25] Ribeiro, M.T., Singh, S. and Guestrin, C. (2016). '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*, pp. 1135–1144.

[26] Schwalbe, G. and Finzel, B. (2023). A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Min. Knowl. Discov.*, pp. 1–59.

[27] James, G., Witten, D., Hastie, T., Tibshirani, R. and others (2013). An introduction to statistical learning, vol. 112. *Springer*.

[28] Shah, K., Patel, H., Sanghvi, D., and Shah, M. (2020). A comparative analysis of logistic regression, random forest and KNN models for the text classification. *Augment. Hum. Res.*, vol. 5, no. 1, p. 12.

[29] Aborisade, O. and Anwar, M. (2018). Classification for authorship of tweets by comparing logistic regression and naive bayes classifiers, *in 2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pp. 269–276.

[30] Stephens, C.R., Huerta, H.F. and Linares, A.R. (2018). When is the Naive Bayes approximation not so naive?. *Mach. Learn.*, vol. 107, pp. 397–441.

[31] Jadhav, S.D. and Channe, H.P. (2016). Comparative study of K-NN, naive Bayes and decision tree classification techniques. *Int. J. Sci. Res.*, vol. 5, no. 1, pp. 1842–1845.

[32] Dong, S. (2022). Virtual currency price prediction based on segmented integrated learning, *in 2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA)*, pp. 549–552.

[33] Pattanayak, S., Loha, C., Hauchhum, L., and Sailo, L. (2021). Application of MLP-ANN models for estimating the higher heating value of bamboo biomass. *Biomass Convers. Biorefinery*, vol. 11, pp. 2499–2508.

[34] Visani, G., Bagli, E., Chesani, F., Poluzzi, A. and Capuzzo, D. (2022). Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models. *J. Oper. Res. Soc.*, vol. 73, no. 1, pp. 91–101.

[35] Wang, D., Thunéll, S., Lindberg, U., Jiang, L., Trygg, J. and Tysklind, M. (2022). Towards better process management in wastewater treatment plants: Process analytics based on SHAP values for tree-based machine learning methods. *J. Environ. Manage.*, vol. 301, p. 113941.

[36] Heart Disease Prediction, dataset by informatics-edu, 2020. [Online]. Available: https://data.world/informatics-edu/heart-disease-prediction. [Accessed: 11-May-2024].