IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Predicting coronary heart disease using an improved LightGBM model: Performance analysis and comparison

HuaZhong Yang [1], Zhongju Chen ✉, Huajian Yang [2], and Maojin Tian [2]

[1]School of Computer Science, Yangtze University, Jingzhou, Hubei 434000, China
[2]School of Mechanical Engineering, Dongguan College of Technology, Dongguan, Guangdong 523000, China
[3]School of Public Health, Zunyi Medical University, Zunyi, Guizhou 563000, China

Corresponding author: Zhongju Chen (chenzj@yangtzeu.edu.cn).

This work was supported by the Hubei Provincial Education Department, Science and Technology Research Program (B2021052).

ABSTRACT Coronary heart disease (CHD) is a dangerous condition that cannot be completely cured. Accurate detection of early coronary artery disease can assist physicians in treating patients. In this study, a prediction model called HY_OptGBM was proposed for predicting CHD by using the optimized LightGBM classifier. To optimize the LightGBM classifier, the hyperparameters of the LightGBM model were adjusted. In addition, its loss function was improved, and the model was trained using adjusted hyperparameters. In this study, the hyperparameters of the prediction model were optimized by applying the most advanced hyperparameter optimization framework (OPTUNA). The improved loss function is referred to as the focal loss (FL). In this study, a prediction model was evaluated by using CHD data from the Framingham Heart Institute. To evaluate the performance of the prediction model, various metrics, including precision, recall, F score, accuracy, MCC, sensitivity, specificity, and AUC, were used. The AUC value of the proposed model was 97.9%, which was better than that of other comparative models. The results demonstrate that the rate of early identification of CHD among the general population can be improved by utilizing the proposed method. This, in turn, could serve to mitigate the costs associated with the medical treatment of patients suffering from CHD.

INDEX TERMS coronary heart disease, hyperparameter optimization, LightGBM, loss function, machine learning, OPTUNA

I. INTRODUCTION

CHD is a prevalent cardiovascular disorder resulting from the buildup of atherosclerotic plaques in the coronary arteries, leading to a reduction in blood flow to the heart muscle. This condition presents a range of symptoms, including chest pain or angina, shortness of breath, palpitations, and heart failure. In severe cases, CHD may lead to a heart attack, which can result in permanent damage to the heart muscle and have a profound impact on an individual's quality of life. Therefore, it is imperative to recognize and manage CHD through appropriate medical intervention and lifestyle modifications. [1]. Early detection of CHD can improve the cure probability and can decrease the cost of treatment. Numerous machine learning algorithms and data mining technologies have been widely used in the medical field [2]–[6] in recent years, owing to advancements in machine learning algorithms and a significant reduction in the cost of data storage. Data mining technology has become essential for healthcare data mining, such as disease diagnosis, auxiliary diagnosis, drug mining, and biomedicine. Through data mining technology, hidden knowledge about diseases can be extracted from large quantities of unstructured medical data, disease prediction models can be developed, and results can be analyzed.

Health organizations face tremendous challenges in providing high-quality and affordable healthcare. A hospital provides quality healthcare services that require physicians to have comprehensive knowledge and a correct diagnosis for the patient to avoid wasting healthcare resources due to inaccurate diagnoses. Data mining technology can perform efficiently and can play a crucial role in clinical cases. The optimal hyperparameters [7], [8] for any classification algorithm significantly affect its performance. The accuracy of the classification algorithm can be improved by selecting the optimal set of hyperparameters. In this study, a state-of-the-art hyperparameter optimization framework (OPTUNA) [9] was employed to obtain optimal hyperparameter values for the LightGBM model. Therefore, in this study, the most suitable set of hyperparameters was determined from the available hyperparameters. Hyperparametric optimization can be accomplished by different methods, such as random and grid searches. Another method is the OPTUNA hyperparametric search. Because the number of hyperparameters in the LightGBM significantly affects its performance, conventional random and grid search methods do not learn from the previous optimization, which wastes considerable time and is inefficient. The OPTUNA framework continuously learns from previous optimizations and adjusts the hyperparameters as necessary. Therefore, OPTUNA was chosen in this paper for hyperparameter optimization.

The loss function also affects the model accuracy [10]. In this paper, the focal loss function was proposed based on the cross-entropy loss by adding the category weight α and the sample difficulty weight modulating factor γ. The aim of this study was to address the problem of unbalanced proportions of positive and negative samples. Additionally, the focal loss function can improve the overall performance of the model. In this study, the default loss function of the LightGBM [11] model was revised using the focal loss function and applied to predict CHD. The key contributions of this study are as follows:

1. This paper proposed a powerful classification model (HY_OptGBM) for predicting CHD. This model is based on optimizing the hyperparameters of LightGBM with the state-of-the-art hyperparameter optimization framework OPTUNA and the revision of the loss function of LightGBM by using the focal loss function.
2. In this study, a focal loss function was proposed. By adjusting the category weight α and sample difficulty weight modulating factor γ, the problem of unbalanced positive and negative sample proportions was addressed, and the model's overall performance was improved.
3. Data preprocessing and hyperparameter adjustment techniques were used to predict CHD.

II. RELATED WORKS

Artificial intelligence (AI) has been applied in various industries owing to rapid advancements in information technology. Numerous academics have used machine learning technology to predict and study diseases. Goldman et al. [12] proposed an improved artificial neural network (ANN) model for predicting CHD. They used the Framingham Heart Institute dataset to validate the experiment, and the results demonstrated that the ANN model had greater sensitivity and specificity in predicting outcomes than the Framingham Risk Score (FRS) (used to calculate an individual's risk of developing CHD over the next ten years based on cholesterol levels and noncholesterol factors). However, the area under the ROC curve (AUC) was lower than that for FRS. Receiver operating characteristic (ROC) curves were used to evaluate the classification performance of the machine learning models. The proposed ANN model yielded significantly better results than FRS for precision-recall measures. In 2020, Du et al. [13] predicted CHD in patients with hypertension based on electronic health record data by using machine learning technology. They divided the CHD dataset into a training set and a test set; then, they used a variety of machine learning algorithms to train the model on the training set and the test set to evaluate the model's performance and compared the results with the FRS score. The experimental results demonstrated that the highest AUC value (0.943) was obtained using the XGBoost for the test set. They compared other machine learning algorithms. The k-nearest neighbor algorithm had an AUC of 0.908, the random forest algorithm had an AUC of 0.938, and the logistic regression algorithm had an AUC of 0.865. They analyzed the relevant features and found that time-related features improved the model's performance. To predict CHD, Kim et al. [14] proposed a neural network algorithm that used feature correlation analysis (NN-FCA). In their experiment, they first selected and ranked the essential features relevant to predicting CHD and then input them into the neural network algorithm to obtain the final prediction results. The results of the experiment demonstrated that in the dataset they used (a total of 4146 patients, including 3031 with low CHD risk and 1115 with high CHD risk), their proposed neural network algorithm AUC value was 0.749 ± 0.010, which is higher than the AUC value of the FRS model (0.393 ± 0.010). Krittanawong et al. [15] predicted cardiovascular disease, primarily CHD and stroke, by using machine learning algorithms in 2020. For CHD prediction, the boosting algorithm achieved an AUC value of 0.88. For stroke prediction, the SVM algorithm yielded an AUC value of 0.92, and the CNN prediction yielded an AUC value of 0.90. These results demonstrate the promise of machine learning algorithms for predicting cardiovascular disease. In 2021, Akella and Akella [16] proposed a solution for predicting coronary artery disease (CAD). They applied six machine learning algorithms to predict CAD on the "Cleveland Dataset" to achieve a feasible clinical tool for CAD detection. They used machine learning algorithms that were over 80% accurate and neural network algorithms that were over 93% accurate. This summary presents retrospective research on the prediction of CHD using machine learning and data mining technology. Muhammad et al. [17] developed CAD prediction models by using CAD datasets from two general hospitals in Kano State-Nigeria. They applied this dataset to support vector machines, K-nearest neighbors, random forests, naïve Bayes, gradient boosting trees, and logistic regression algorithms. The performance of the model was evaluated in terms of accuracy, specificity, sensitivity, and ROC curve. For accuracy, the random forest algorithm was the best model with 92.04%; for specificity, the naïve Bayes algorithm was the best model with 92.40%; for sensitivity, the support vector machine algorithm was the best model with 87.34%; for the ROC curve, the random forest model was the best model with 92.20%. The experimental results demonstrated that the random forest algorithm was the best model in terms of accuracy and ROC curve. Hassan et al. [18] used machine learning algorithms to predict CHD by using various features to improve the accuracy of the prediction model. Among the 11 classifier algorithms used, the gradient boosting tree and the multilayer perceptron had an accuracy of 95%, and the random forest algorithm had an accuracy of 96%. The prediction results showed that using feature combinations can effectively improve the accuracy of the algorithms. Table I summarizes studies related to CHD prediction.

Some retrospective studies on machine learning and data mining techniques for predicting CHD have been conducted. The above study showed that the accuracy obtained using machine learning algorithms and neural network models to predict CHD was poor. Therefore, there is still room for improvement in the use of machine learning to predict CHD. To obtain a better prediction accuracy, this study proposes a model named HY_OptGBM, which uses a state-of-the-art hyperparameter optimization framework (OPTUNA) to optimize the hyperparameters of LightGBM and improve its loss function to predict CHD.

III. MATERIALS AND METHODS

A. DATASET

The CHD dataset, which contains 4240 records from the Framingham Heart Institute, was used to validate the model. A total of 15.188% of the records were for patients with CHD (644 cases), and 84.812% were for normal cases (3597 cases). Among the CHD patients, 53.260% were men and 46.740% were women. The attributes of the dataset are listed in Table II.

TABLE I
State-of-the-art methods for CHD. Note: Acc indicates accuracy

Year	Authors	Research Title	Method	Performance Evaluation
2017	Jae Kwon Kim et al.	Neural Network-Based Coronary Heart Disease Risk Prediction Using Feature Correlation Analysis	NN-FCA	NN-FCA model AUC is 0.749±0.010
2019	Juan-Jose Beunza et al.	Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease)	SVM+NN	NN model AUC is 0.71 and SVM model AUC is 0.75
2019	Tsatsral Marbayasgalan et al.	Reconstruction error-based deep neural networks for coronary heart disease risk prediction	AE-DNN	AE-DNN model acc is 0.8634, precision is 0.9137, recall is 0.8290, F-score is 86.91 and AUC is 0.867
2020	Zhenzhen Du et al.	Accurate Prediction of Coronary Heart Disease for Patients With Hypertension From Electronic Health Records With Big Data and Machine-Learning Methods: Model Development and Performance Evaluation	XGBoost	XGBoost model AUC is 0.943
2021	Yehong Liu et al.	Systemic immune-inflammation index predicts the severity of coronary stenosis in patients with coronary heart disease	Gensini-score	Gensini-score's sensitivity is 0.71 and specificity is 0.86
2021	Aravind Akella et al.	Machine learning algorithms for predicting coronary artery disease: efforts toward an open source solution	NN	NN model acc is 0.93 and recall is 0.93
2021	LJ Muhammad et al.	Machine Learning Predictive Models for Coronary Artery Disease	SVM+KNN+RF+NB+GB+LR	RF model acc is 0.9204, NB model specificity is 0.9240, SVM model sensitivity is 0.8734 and RF model AUC is 0.9220
2022	JoonNyung Heo et al.	Prediction of Hidden Coronary Artery Disease Using Machine Learning in Patients With Acute Ischemic Stroke	EGB+LR	EGB model AUC is 0.763 and LR model AUC is 0.714
2022	Ch Anwar Ul Hassan et al.	Effectively Predicting the Presence of Coronary Heart Disease Using Machine Learning Classifiers	KNN+RF+GBT+MLP+SVM	GBT and MLP model acc is 0.95, RF model acc is 0.96

B. DATA PREPROCESSING

Data preprocessing is necessary for machine learning algorithms and data mining technology. The performance of machine learning algorithms depends on the data structure and quality. Missing rows were removed during the data preprocessing stage [20]. Then, the outliers were processed using the statistical 3б principle, which uses the interquartile range (IQR) to detect outliers and extreme values. The IQR is a method for measuring the change in values in a dataset. For this analysis, we assumed that these outliers were due to measurement tool accuracy and other unrelated phenomena.

TABLE II
Description properties

SN	Name	Description
1	Sex	Male=1; female=0 (nominal)
2	Age	Age of the patient (continuous)
3	Education	0: High school degree, 1-4: a college degree or higher
4	CurrentSmoker	Smokes (1=yes, 0=no)
5	CigsPerDay	Average daily number of cigarettes smoked
6	BPMeds	Takes antihypertension drugs (1=yes; 0=no)
7	PrevalentStroke	Stroke (1=yes; 0=no)
8	PrevalentHyp	Suffers from high blood pressure (1=yes; 0=no)
9	Diabetes	Suffers from diabetes (1=yes; 0=no)
10	TotChol	Total cholesterol level (continuous)
11	SysBP	Systolic blood pressure (continuous)
12	DiaBP	Diastolic blood pressure (continuous)
13	BMI	Body mass index (continuous)
14	HeartRate	Heart rate (continuous)
15	Glucose	Glucose level (continuous)
16	TenYearCHD (Class)	CHD within ten years (1=yes; 0=no)

To detect outliers, the data were divided into three quartiles, namely, Q1, Q2 and Q3, where Q1 and Q3 were the boundary values, and the value of IQR is Q3-Q1. The upper boundary B$_u$ value and the lower boundary B$_l$ value were calculated using the following equations.

$$B_l = Q_1 - 1.5 * IQR \tag{1}$$

$$B_u = Q_3 + 1.5 * IQR \tag{2}$$

According to Equations (1) and (2), data larger than the B$_u$ value or smaller than the B$_l$ value were determined to be outliers. In this paper, the synthetic minority oversampling technique (SMOTE) [21] was used to reconstruct the data distribution and to balance the data. In the present study, a systematic analysis of the data was conducted to detect any abnormalities, such as abnormal values, outliers, or missing data points. This evaluation was imperative for establishing the validity and reliability of the data. Moreover, heatmap plots and correlation plots were utilized to graphically represent the interrelationships between the variables in the dataset. These graphical representations facilitated a deeper understanding of the relationships between the variables and informed the modeling process. To assess the model's generalizability and robustness, the dataset was divided into three distinct subsets: training, testing, and validation sets, with proportionate allocations of 80%, 10%, and 10%, respectively. This division of the dataset was crucial for evaluating the model's performance and ensuring that the results were robust and statistically significant. The thorough analysis and division of the data in this study laid a solid foundation for the subsequent analysis and modeling. The training set was used to train the model, the test set was used to evaluate the performance of the model, and the validation set was used to verify the hyperparameters of the model. The random parameter seed was set to 42.

C. EVALUATION METRICS

In this study, six machine learning algorithms were used for comparison with the proposed model, and a 10-fold cross-validation method was implemented to prevent algorithm overfitting. Machine learning algorithms, including the decision tree (DT) [22], random forest (RF) [23], CatBoost (CB) [24], XGBoost (XGB), AdaBoost (ADA) [25], bagging (BG) [26] and LightGBM (GBM) algorithms, were compared. In the following section, the evaluation metrics for the machine learning algorithms are described.

In this study, the sensitivity, specificity and accuracy of each algorithm were calculated. The formulas below represent the mathematical equations used for assessing the performance of the algorithms.

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

$$TPR = \frac{TP}{TP + FN} \tag{5}$$

$$FPR = \frac{FP}{FP + TN} \tag{6}$$

Among the four essential elements, TP indicates true positives, TN represents true negatives, FP represents falsepositives, and FN represents false-negatives. Sensitivity denotes the number of positive factors correctly identified by the classifier, specificity denotes the number of negative factors correctly identified by the classifier, TPR denotes the actual positive rate, and FPR denotes the false-positive rate.

In this study, precision, recall, f score, Matthew's correlation coefficient (MCC), accuracy, AUC, receiver operating characteristic (ROC) curve, and precision-recall curve (prc) metrics were used to evaluate the performance of different classifiers. Precision denotes the proportion of samples predicted by the classifier that belong to the positive class, and the closer to one the value is, the better the classification effect. Recall denotes the proportion of categories correctly predicted as positive by the classifier, and the closer to one the value is, the better the classification result. The F-score is defined as the harmonic mean of the precision and recall. The MCC is used to evaluate the effectiveness of the binary and multiclass models. Accuracy denotes the proportion of correct classifier predictions (positive and negative cases), and the closer to one the values are, the better the classification results. The AUC is defined as the area under the ROC curve. The ROC curve is the receiver operating characteristic curve, and the larger the area under the curve is, the better the classifier classification effect. The formulas below represent the mathematical equations used for assessing the performance of the algorithms.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F - Measure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{10}$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{11}$$

D. LIGHTGBM

LightGBM is an ensemble learning algorithm based on decision trees. It is more powerful and faster than the XGBoost algorithm, which requires less memory, has better accuracy, and supports parallel computing. The major features of LightGBM are gradient-based one-sided sampling (GOSS), mutually exclusive feature bundling (EFB), and differential acceleration that uses a histogram algorithm [27].

The basic concept of one-sided gradient sampling (GOSS) is to reserve samples with large gradients and randomly select samples with small gradients in proportion to their size. The fundamental idea of the exclusive feature bundling (EFB) algorithm is to bundle two nonmutually exclusive features. This binding reduces the number of features and the time complexity, thereby improving the computational effectiveness of the model.

The basic concept of the histogram algorithm [28], [29] is as follows. First, the continuous floating-point eigenvalues are discretized into k integers, and then a histogram of width k is constructed. When traversing data, statistics are accumulated in a histogram based on discrete values as indices. The continuous floating-point eigenvalues are first discretized into k integers, and then a histogram of width k is constructed. Fig. 1 shows the concept of the histogram algorithm.

The LightGBM algorithm applies parallel computing to improve the computing efficiency and is mainly divided into the feature parallel, data parallel and voting parallel algorithms. The main idea behind feature parallelism is that different machines separately search for the best split points on different sets of features, and then the best split points are synchronized among the machines. XGBoost utilizes this method of feature parallelism. However, this method has a significant drawback. It divides the data virtually, with each machine containing different data, and then uses different machines to find the optimal split points of different features, which adds complexity due to the need to communicate the split results to each machine. LightGBM does not perform vertical data partitioning; instead, it retains all training data on each machine and performs the partition locally after obtaining the best partition plan, thus reducing unnecessary communication. The detailed process is shown in Fig. 2.

Traditional data parallelism strategies mainly involve horizontally partitioning data, allowing different machines to first construct histograms locally and then perform a global merge. Finally, the optimal split points are found on the merged histogram. This data division process has a major drawback, excessive communication overhead, if point-to-point communication is used. LightGBM uses scatter-gather reduction in its data parallelism approach, distributing the task of merging histograms among different machines to reduce communication and computation. It also uses histogram differences to further reduce communication by half. The specific process is shown in Fig. 3.

Voting-based data parallelism further optimizes the communication cost in data parallelism, making the communication cost constant. When the data volume is large, using the voting parallelism approach only merges the histograms of some features to reduce the communication volume, resulting in very good acceleration effects. The specific process is shown in Fig. 4.

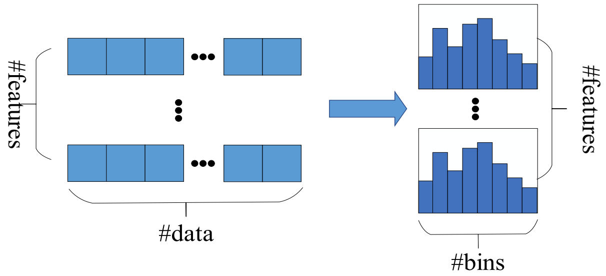


FIGURE 1. Histogram algorithm.

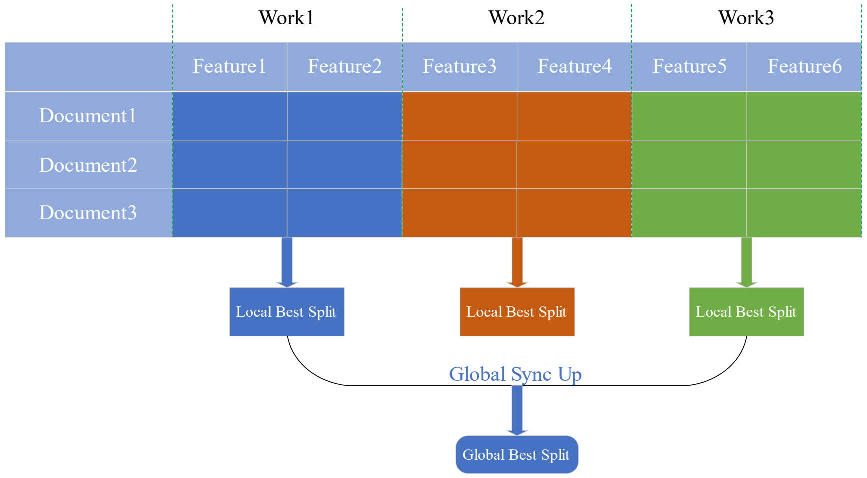


FIGURE 2. Feature/attribute parallelization diagram.

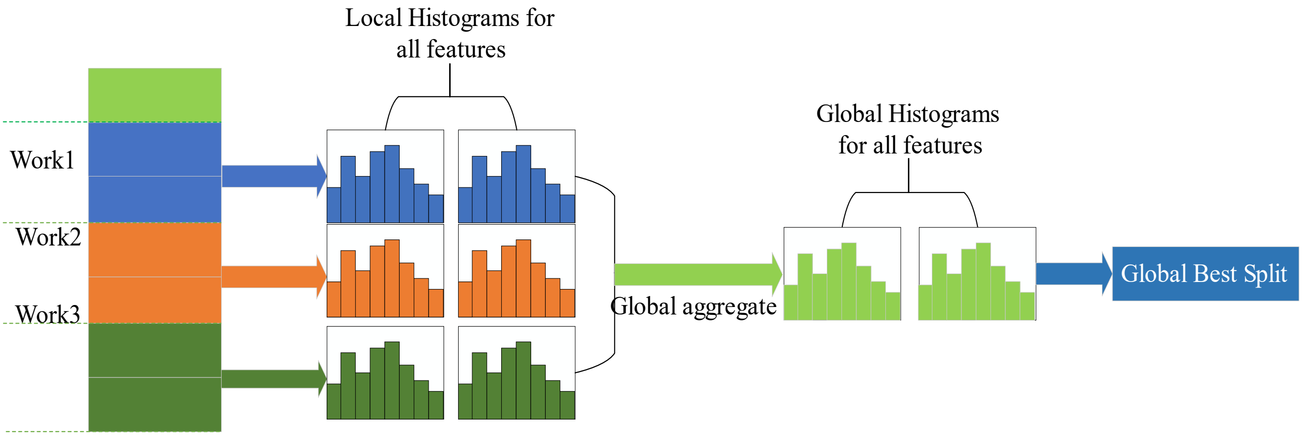

Local Histograms for all features
Work1
Work2
Work3
Global Histograms for all features
Global aggregate
Global Best Split

FIGURE 3. Data parallelism diagram.

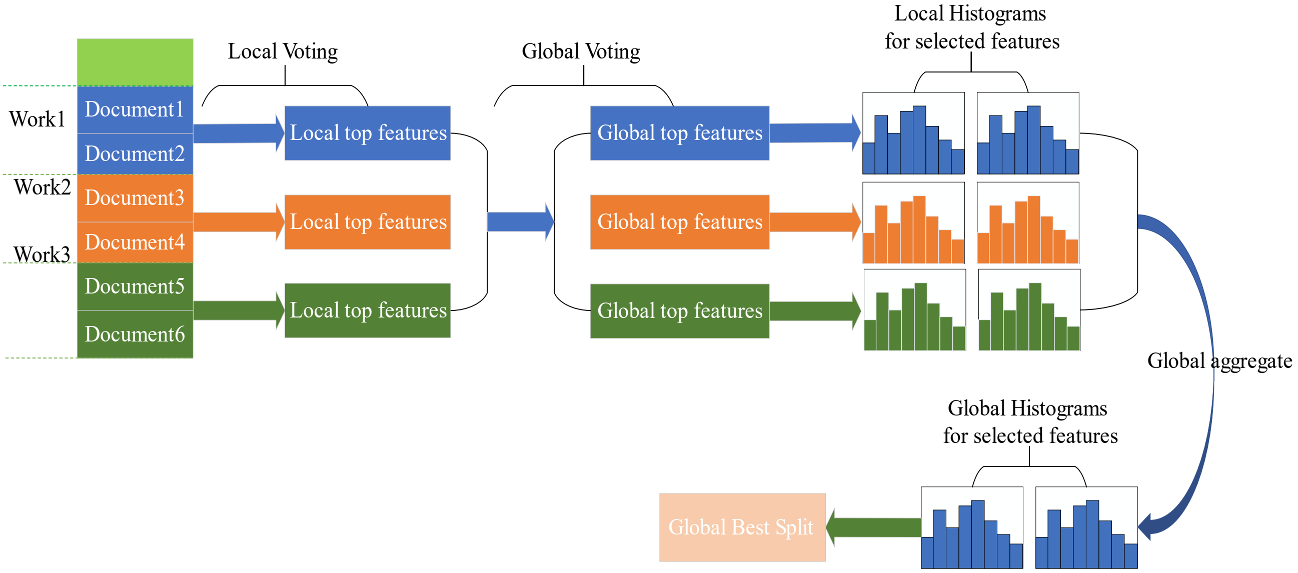


FIGURE 4. Voting-based parallel diagram.

E. FOCAL LOSS FUNCTION

In this research, a proposal for a focal loss function and its application to the default loss function of the LightGBM model are presented. The successful implementation of the focal loss function is facilitated through a four-step process that outlines the integration procedure.

STEP 1: This study proposes adding a modulating factor γ and a category weight α (α < 1) to the cross-entropy loss with a tunable focusing parameter γ ⩾ 0 and defining the focal loss function, as shown in Equation (12).

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\, log(p_t) \qquad (12)$$

STEP 2: The first-order derivative of the focal loss is calculated.

The first-order derivative of the focal loss is calculated using the chain rule, as shown in Equation (13).

$$\frac{\partial FL}{\partial z} = \frac{\partial FL}{\partial_{p_t}} \times \frac{\partial_{p_t}}{\partial_p} \times \frac{\partial_p}{\partial_z}$$

$$= \alpha_t(1-p_t)^{\gamma}\left(\frac{\gamma p_t\, log(p_t) + p_t - 1}{p_t(1-p_t)}\right) \times y \times p_t(1-p_t)$$

$$= \alpha_t y(1-p_t)^{\gamma}(\gamma p_t\, log(p_t) + p_t - 1) \qquad (13)$$

STEP 3: The second-order focal loss derivative is calculated.

In step 3, the chain rule is used to calculate the second-order derivative. The focal loss second-order derivative chain rule is given by Equation (14).

$$\frac{\partial^2 FL}{\partial z^2} = \frac{\partial}{\partial z}\left(\frac{\partial FL}{\partial z}\right) = \frac{\partial}{\partial p_t}\left(\frac{\partial FL}{\partial z}\right) \times \frac{\partial p_t}{\partial p} \times \frac{\partial p}{\partial z} \qquad (14)$$

For the convenience of calculation, this study defines the symbols u and v; therefore, according to Equation (13), the following equations are obtained.

$$\frac{\partial FL}{\partial z} = u \times v \qquad (15)$$

$$u = \alpha_t y(1-p_t)^{\gamma} \qquad (16)$$

$$v = \gamma p_t\, log(p_t) + p_t - 1 \qquad (17)$$

Therefore, the second-order derivative of the focal loss function is as follows.

$$\frac{\partial^2 FL}{\partial z^2} = \left(\frac{\partial u}{\partial p_t} \times v + u \times \frac{\partial v}{\partial p_t}\right) \times \frac{\partial p_t}{\partial p} \times \frac{\partial p}{\partial z}$$

$$= (-v\gamma\alpha_t y(1-p_t)^{\gamma-1} + u\gamma\, log(p_t) + u\gamma + u) \times y \times p_t(1-p_t) \qquad (18)$$

STEP 4: Finally, this study implements the focal loss function to revise the default loss function of LightGBM.

The above four steps are the major steps to improve the default loss function of the LightGBM model. Algorithm 1 provides a concise, step-by-step explanation of the proposed loss function. The detailed theoretical derivation of the proposed loss function can be found in the appendix.

```
Algorithm 1: Focal loss

Step 1: Define the focal loss.
Add a modulating factor $\gamma$ and a category weight $\alpha$ to the
cross-entropy loss, with tunable focusing parameter $\gamma >= 0$.
Step 2: Calculate the focal loss first-order derivative.
Step 3: Calculate the focal loss second-order derivative.
Step 4: Implement the focal loss function to override the
default loss function of LightGBM.
```

F. OPTUNA: A FLEXIBLE, EFFECTIVE AND STABILIZED HYPERPARAMETRIC OPTIMIZATION FRAMEWORK

In 2019, Takuya Akiba proposed OPTUNA, an open-source library for hyperparametric optimization. It is capable of searching for optimal hyperparameters based on the validation scores returned by the model [30]–[32]. OPTUNA implements sampling algorithms, such as independent sampling and relation sampling, and an asynchronous successive halving algorithm (ASHA) for pruning the search space. Hyperparameters have a significant impact on the accuracy of machine learning models. Hyperparameter optimization is an essential step in machine learning. In this study, the hyperparameters of the LightGBM were optimized by using OPTUNA. Next, this study describes the OPTUNA design features in three ways.

1) DEFINE-BY-RUN STYLE API

OPTUNA employs a new define-by-run style, which is very convenient for users to optimize the hyperparameters of machine learning algorithms. OPTUNA [9], [33] hyperparameter optimization is characterized by minimizing the objective function given a set of hyperparameters. OPTUNA builds the target function step-by-step by interacting with a trial object and dynamically creates the search space by using the trial object during the execution of the target function (a trial object is a special trace object of OPTUNA that uses the name and range of the hyperparameter to search for the optimal value).

2) EFFICIENT PRUNING AND SAMPLING MECHANISM

Pruning was used to define the process of unpromising wind-up trials. It is also referred to as automated early stopping [34]. This process is divided into two sections: 1) periodic monitoring of the median value of the objective function and 2) pausing the trials that do not fit the predefined probability. The Hyperopt, Spearmint, and SMAC hyperparameter optimization algorithms are not available.

The sampling method can be divided into relational sampling [35] and independent sampling [36]. Relational sampling takes advantage of the correlations between parameters. Independent sampling did not account for the correlations between the parameters. The cost-effectiveness of independent and relational sampling depends on the task and the environment. OPTUNA combines it with these two sampling algorithms, which means it can manage independent

TABLE II
Description properties

SN	Parameter	Default
1	verbosity	1
2	boosting_type	gbdt
3	metric	auc
4	objective	binary: logistic
5	num_threads	max

sampling methods such as TPE; therefore, convergence is faster and more efficient.

3) EASY SETUP

When using machine learning models, the simpler the setup is, the better, especially for large-scale training and large-scale datasets. All of the above make OPTUNA an excellent hyperparametric optimization framework. The architecture of the optimized LightGBM model is illustrated in Fig. 5. In Fig. 5, each worker performs an instance of the objective function during the search.

4) PROPOSED ALGORITHM: HY_OPTGBM

Data preprocessing has a significant impact on the performance of machine learning algorithms. First, the dataset is preprocessed, the LightGBM model hyperparameters are optimized, and the loss function is improved. Second, the data in this paper are trained utilizing the improved LightGBM. The optimization of the hyperparameters was accomplished through the use of the OPTUNA framework, and the focal loss function was employed as the improved loss function. Ten hyperparameters, including max_depth, lambda_l1, lambda_l2, num_leaves, learning_rate, n_estimators, feature_fraction, bagging_fraction, bagging_freq and min_child_samples, were chosen for parameter optimization. Although these ten hyperparameters were chosen for optimization, some of the default hyperparameters of LightGBM were retained, as shown in Table III.

Table IV shows the ten hyperparameters optimized using OPTUNA. max_depth limits the maximum depth for the tree model and is also used to control overfitting. lambda_l1 is for L1 regularization; lambda_l2 is for L2 regularization. L1 and L2 can also decrease the model complexity and avoid overfitting. num_leaves represents the maximum number of leaves in a tree. For learning_rate, we know that a higher learning rate allows the model to converge faster but decreases the accuracy. N_estimators represents the number of control decision trees. feature_fraction is used for randomly selecting a subset of features; this parameter can also be used to speed up training and address overfitting. bagging_fraction is used to obtain the percentage of training samples used to train each tree, and this parameter can also be used to speed up training and to address overfitting. bagging_freq represents the frequency for bagging, where default=0,0 means disable bagging, with k means performing bagging at every k iteration. For min_child_samples, the value depends on the number of samples in the training dataset. Fig. 6 represents the whole experimental flow chart.

IV. EXPERIMENTAL RESULTS

A. THE RESULTS FROM EXPLORATORY DATA ANALYSIS

In this study, exploratory data analysis [37] was used to understand the preprocessed dataset while removing null values and outliers, and then the proposed algorithm and other machine learning algorithms were applied to this dataset.

Fig. 7 shows a heat diagram [39] of the feature correlation. The darker the color of this diagram is, the stronger the correlation between the features, where a value less than zero denotes a negative correlation and zero denotes no correlation between two features.

Fig. 8 shows boxplot charts [38] that depict the use of the interquartile range to detect and remove outliers from the data. After filtering, there were no outliers in the dataset. Then, the dataset was prepared for further analysis.

B. RESULTS OF THE EXPERIMENT

In this study, the LightGBM model was used as the machine learning model. First, the OPTUNA framework was used to optimize its hyperparameters, and these hyperparameters were used for the LightGBM model. Second, the default loss function of the LightGBM model was improved, and finally, 10-fold cross-validation [40]–[42] was used to obtain the AUC scores of LightGBM.

Table V presents an evaluation of the performance of the algorithm by using the sensitivity, specificity, and accuracy evaluation metrics. The classification algorithms were the decision tree (DT), CatBoost (CB), XGBoost (XGB), AdaBoost (ADA), bagging (BG), LightGBM (GBM) algorithms, and the proposed HY_OptGBM classification algorithm. Table V presents the comparison of the results obtained with and without the FL function. The experimental results indicate that the proposed algorithm demonstrates superior sensitivity, specificity, and accuracy compared to other classification algorithms, with higher values for these evaluation metrics.

Table VI displays the results of the experiment conducted on the proposed algorithm in comparison with other algorithms, using the precision, recall, and F-score as performance evaluation metrics. According to Table VI, a model with improved classification performance compared to other models was introduced. The proposed method achieved a precision of 0.963, a recall of 0.897, and an F-score of 0.929.

Table VII shows an evaluation of the model based on three evaluation metrics: AUROC, AUPRC, and MCC. The AUROC of the HY_OptGBM model was 0.979, the AUPRC was 0.983, and the MCC was 0.861.

Fig. 9 shows the ROC curve, which displays the trade-off between the true positive rate (TPR) and false-positive rate (FPR) across different decision thresholds, and the area under this curve is the AUROC. In Fig. 9, the AUROC value of our proposed algorithm was 0.979, which is better than the AUROC

TABLE IV
LightGBM hyperparameters

SN	Name	Range	Tuned parameters
1	max_depth	[1~200]	117
2	lambda_l1	[1e-8~10.0]	0.0011559055
3	lambda_l2	[1e-8~10.0]	0.0131888932
4	num_leaves	[1~512]	346
5	learning_rate	[1e-8~1.0]	0.0219999851
6	n_estimators	[200~3000]	726
7	feature_fraction	[0.1~1.0]	0.5066181469
8	bagging_fraction	[0.1~1.0]	0.5600164173
9	bagging_freq	[1~10]	7
10	min_child_samples	[5~100]	2

TABLE V
Classification results of different classification algorithms.
Wo/FL indicates without FL and with OPTUNA; W/FL indicates with FL and with OPTUNA

Classifier Algorithm	Sensitivity	Specificity	Accuracy
ADA	0.763	0.777	0.769
DT	0.778	0.784	0.780
BG	0.847	0.900	0.873
GBM	0.859	0.940	0.898
XGB	0.866	0.937	0.900
CB	0.869	0.937	0.902
Proposed model (Wo/FL)	0.887	0.963	0.924
Proposed model (W/FL)	0.897	0.963	0.930

TABLE VI
Precision, Recall and F-score

Classifier Algorithm	Precision	Recall	F-score
ADA	0.785	0.763	0.773
DT	0.793	0.778	0.785
BG	0.883	0.828	0.855
GBM	0.938	0.859	0.897
XGB	0.936	0.866	0.899
CB	0.936	0.869	0.901
Proposed model (Wo/FL)	0.962	0.887	0.923
Proposed model (W/FL)	0.963	0.897	0.929

TABLE VII
AUROC, AUPRC AND MCC

Classifier Algorithm	AUROC	AUPRC	MCC
ADA	0.769	0.721	0.539
DT	0.783	0.732	0.565
BG	0.872	0.832	0.745
GBM	0.899	0.879	0.800
XGB	0.901	0.879	0.803
CB	0.902	0.880	0.806
Proposed model (Wo/FL)	0.932	0.917	0.851
Proposed model (W/FL)	0.978	0.983	0.861

values of the other algorithms.

Fig. 10 shows the precision-recall curve, demonstrating the relationship between precision and recall at varying decision thresholds. The area under the precision-recall curve (AUPRC) measures its performance, with the proposed algorithm exhibiting an AUPRC value of 0.983, which surpasses the AUPRC values of the other algorithms.

TABLE Ⅷ
The result of experimental FL function (when α ranges from 0 to 0.9 and γ ranges from one to three)

alpha	gamma	Sensitivity	Specificity	Accuracy	Precision	Recall	F score	AUROC	AUPRC	MCC
None	0	0.887	0.956	0.921	0.956	0.887	0.920	0.976	0.981	0.844
	1	0.897	0.963	0.930	0.963	0.897	0.929	0.978	0.983	0.861
	2	0.881	0.953	0.916	0.953	0.881	0.915	0.978	0.981	0.835
	3	0.868	0.960	0.913	0.958	0.868	0.911	0.976	0.981	0.830
0.1	0	0.812	0.986	0.896	0.984	0.812	0.890	0.975	0.979	0.807
	1	0.818	0.980	0.896	0.977	0.818	0.891	0.975	0.98	0.806
	2	0.815	0.983	0.896	0.981	0.815	0.890	0.975	0.979	0.806
	3	0.803	0.983	0.890	0.980	0.803	0.883	0.973	0.978	0.795
0.2	0	0.844	0.976	0.908	0.974	0.843	0.904	0.975	0.980	0.824
	1	0.863	0.976	0.917	0.975	0.863	0.915	0.976	0.980	0.842
	2	0.831	0.983	0.904	0.981	0.832	0.900	0.976	0.981	0.82
	3	0.828	0.980	0.901	0.977	0.828	0.896	0.975	0.980	0.814
0.3	0	0.856	0.960	0.906	0.958	0.856	0.904	0.974	0.979	0.818
	1	0.856	0.966	0.909	0.964	0.856	0.907	0.975	0.979	0.825
	2	0.850	0.970	0.908	0.967	0.850	0.905	0.975	0.980	0.823
	3	0.837	0.973	0.903	0.971	0.837	0.899	0.974	0.979	0.815
0.4	0	0.868	0.960	0.913	0.958	0.868	0.911	0.976	0.980	0.830
	1	0.865	0.963	0.913	0.961	0.865	0.911	0.975	0.980	0.830
	2	0.862	0.967	0.913	0.965	0.862	0.910	0.976	0.981	0.831
	3	0.856	0.966	0.909	0.964	0.856	0.907	0.975	0.98	0.825
0.5	0	0.875	0.953	0.913	0.952	0.875	0.912	0.976	0.981	0.829
	1	0.887	0.963	0.924	0.962	0.887	0.923	0.976	0.981	0.852
	2	0.884	0.963	0.922	0.962	0.884	0.921	0.977	0.982	0.849
	3	0.881	0.960	0.919	0.959	0.881	0.918	0.976	0.981	0.842
0.6	0	0.894	0.957	0.924	0.956	0.894	0.924	0.977	0.982	0.851
	1	0.887	0.950	0.917	0.949	0.887	0.917	0.976	0.981	0.837
	2	0.888	0.953	0.919	0.953	0.888	0.919	0.975	0.981	0.841
	3	0.878	0.947	0.911	0.946	0.878	0.910	0.975	0.981	0.825
0.7	0	0.896	0.927	0.911	0.928	0.897	0.912	0.975	0.98	0.823
	1	0.897	0.94	0.918	0.940	0.897	0.918	0.976	0.981	0.837
	2	0.900	0.943	0.921	0.944	0.900	0.922	0.976	0.981	0.843
	3	0.891	0.940	0.915	0.940	0.891	0.915	0.975	0.981	0.831
0.8	0	0.925	0.900	0.913	0.907	0.925	0.916	0.975	0.980	0.826
	1	0.912	0.897	0.905	0.904	0.913	0.908	0.974	0.980	0.809
	2	0.916	0.910	0.913	0.916	0.916	0.916	0.973	0.979	0.826
	3	0.919	0.900	0.909	0.907	0.918	0.913	0.973	0.979	0.819
0.9	0	0.956	0.874	0.916	0.889	0.956	0.921	0.974	0.979	0.834
	1	0.938	0.880	0.909	0.892	0.938	0.915	0.974	0.979	0.820
	2	0.950	0.860	0.906	0.878	0.950	0.913	0.973	0.978	0.816
	3	0.941	0.854	0.899	0.872	0.941	0.905	0.970	0.976	0.799

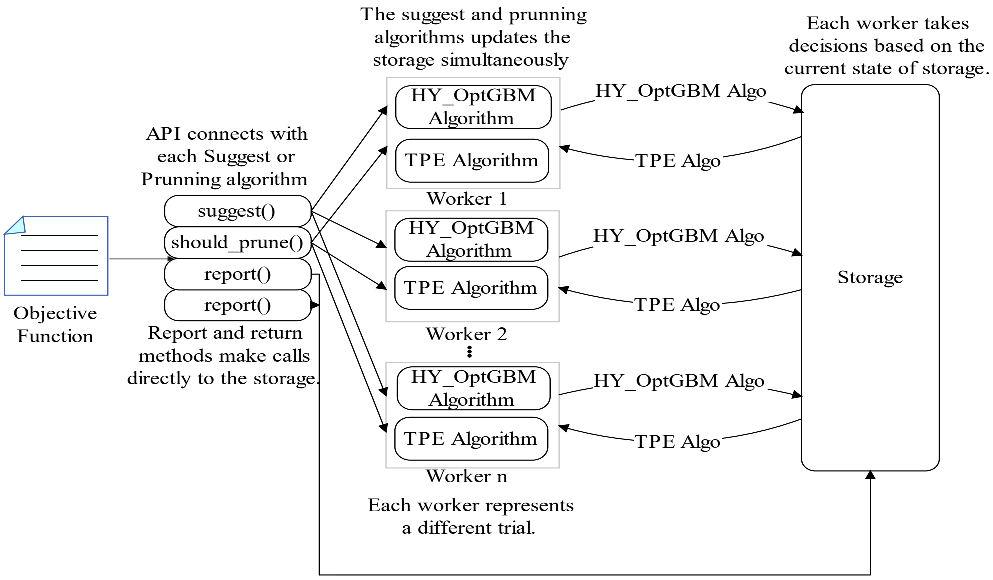

The suggest and prunning algorithms updates the storage simultaneously
Each worker takes decisions based on the current state of storage.
API connects with each Suggest or Prunning algorithm
suggest()
should_prune()
report()
report()
Objective Function
Report and return methods make calls directly to the storage.
HY_OptGBM Algorithm
TPE Algorithm
Worker 1
HY_OptGBM Algorithm
TPE Algorithm
Worker 2
HY_OptGBM Algorithm
TPE Algorithm
Worker n
HY_OptGBM Algo
TPE Algo
HY_OptGBM Algo
TPE Algo
HY_OptGBM Algo
TPE Algo
Storage
Each worker represents a different trial.

FIGURE 5. Architecture chart of the OPTUNA optimized algorithm.

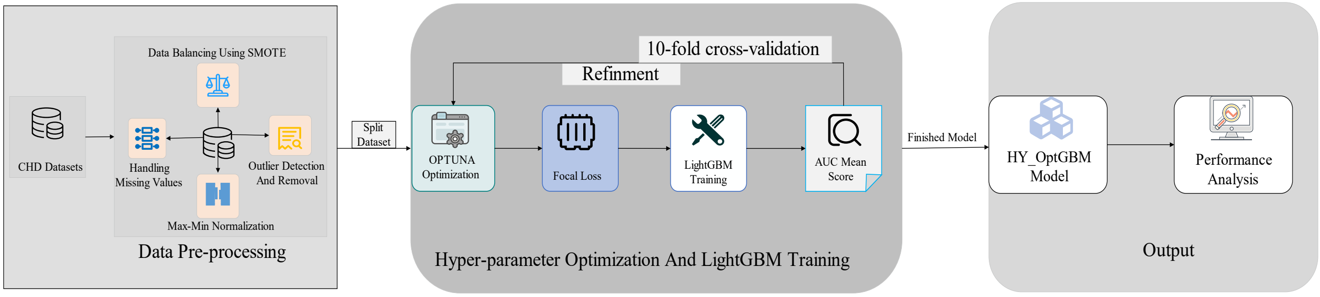
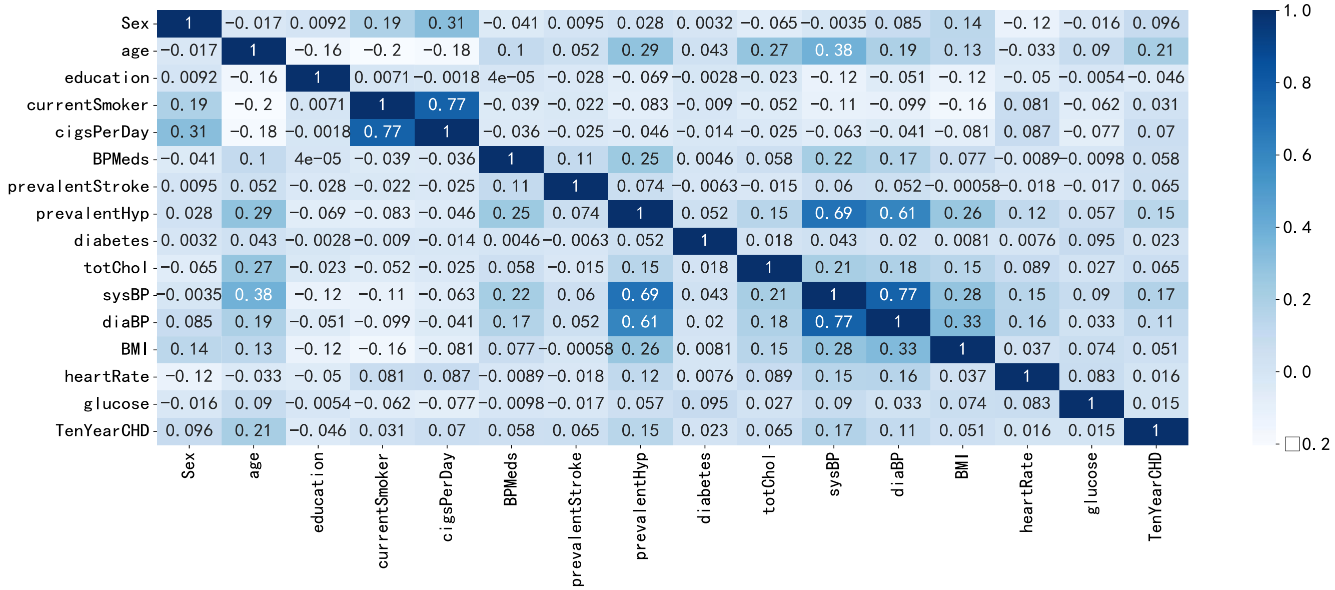


FIGURE 6. Experimental flow chart.



	Sex	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
Sex	1	−0.017	0.0092	0.19	0.31	−0.041	0.0095	0.028	0.0032	−0.065	−0.0035	0.085	0.14	−0.12	−0.016	0.096
age	−0.017	1	−0.16	−0.2	−0.18	0.1	0.052	0.29	0.043	0.27	0.38	0.19	0.13	−0.033	0.09	0.21
education	0.0092	−0.16	1	0.0071	−0.0018	4e−05	−0.028	−0.069	−0.0028	−0.023	−0.12	−0.051	−0.12	−0.05	−0.0054	−0.046
currentSmoker	0.19	−0.2	0.0071	1	0.77	−0.039	−0.022	−0.083	−0.009	−0.052	−0.11	−0.099	−0.16	0.081	−0.062	0.031
cigsPerDay	0.31	−0.18	−0.0018	0.77	1	−0.036	−0.025	−0.046	−0.014	−0.025	−0.063	−0.041	−0.081	0.087	−0.077	0.07
BPMeds	−0.041	0.1	4e−05	−0.039	−0.036	1	0.11	0.25	0.0046	0.058	0.22	0.17	0.077	−0.0089	−0.0098	0.058
prevalentStroke	0.0095	0.052	−0.028	−0.022	−0.025	0.11	1	0.074	−0.0063	−0.015	0.06	0.052	−0.00058	−0.018	−0.017	0.065
prevalentHyp	0.028	0.29	−0.069	−0.083	−0.046	0.25	0.074	1	0.052	0.15	0.69	0.61	0.26	0.12	0.057	0.15
diabetes	0.0032	0.043	−0.0028	−0.009	−0.014	0.0046	−0.0063	0.052	1	0.018	0.043	0.02	0.0081	0.0076	0.095	0.023
totChol	−0.065	0.27	−0.023	−0.052	−0.025	0.058	−0.015	0.15	0.018	1	0.21	0.18	0.15	0.089	0.027	0.065
sysBP	−0.0035	0.38	−0.12	−0.11	−0.063	0.22	0.06	0.69	0.043	0.21	1	0.77	0.28	0.15	0.09	0.17
diaBP	0.085	0.19	−0.051	−0.099	−0.041	0.17	0.052	0.61	0.02	0.18	0.77	1	0.33	0.16	0.033	0.11
BMI	0.14	0.13	−0.12	−0.16	−0.081	0.077	−0.00058	0.26	0.0081	0.15	0.28	0.33	1	0.037	0.074	0.051
heartRate	−0.12	−0.033	−0.05	0.081	0.087	−0.0089	−0.018	0.12	0.0076	0.089	0.15	0.16	0.037	1	0.083	0.016
glucose	−0.016	0.09	−0.0054	−0.062	−0.077	−0.0098	−0.017	0.057	0.095	0.027	0.09	0.033	0.074	0.083	1	0.015
TenYearCHD	0.096	0.21	−0.046	0.031	0.07	0.058	0.065	0.15	0.023	0.065	0.17	0.11	0.051	0.016	0.015	1

FIGURE 7. Heat matrix plot of the correlation of the eigenvalues.

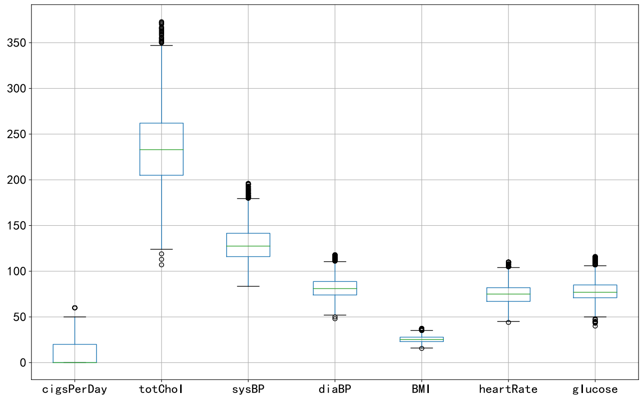


FIGURE 8. Boxplot chart.

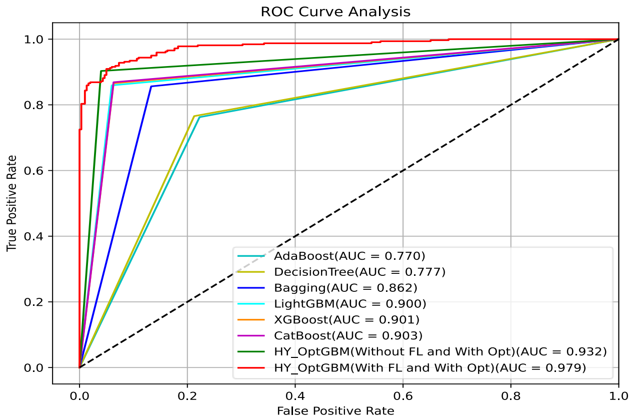


FIGURE 9. Diagram of the experimental comparison between the HY_OptGBM algorithm and other algorithms using AUROC as the evaluation metric.

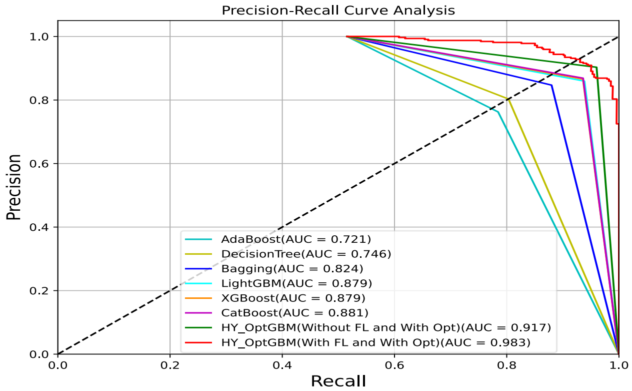


FIGURE 10. Diagram of the experimental comparison between the HY_OptGBM algorithm and other algorithms using AUPRC as the evaluation metric.

The loss function graph of the HY_OptGBM model is presented in Fig. 11. The proposed FL function can be altered to impact the outcome of the loss function through changes in the weight parameter α and the adjustment modulating factor parameter γ. The purpose of the category weight α is to let the negative sample data increase the weight. It was used to resolve the uneven proportion of positive and negative samples. The sample difficulty weight adjustment modulating factor γ was used to measure hard- and easy-to-classify samples. In the experiments, parameter α ranged from 0 to 0.9, and γ ranged from 0 to 3. The experiment showed that the optimal results were achieved when the parameter α was set to None and the parameter γ was set to 1. In Fig. 11, the loss function of HY_OptGBM converges the fastest when α=None and γ=1. At this time, the accuracy of the model was optimal. The experimental results are presented in Table Ⅷ.

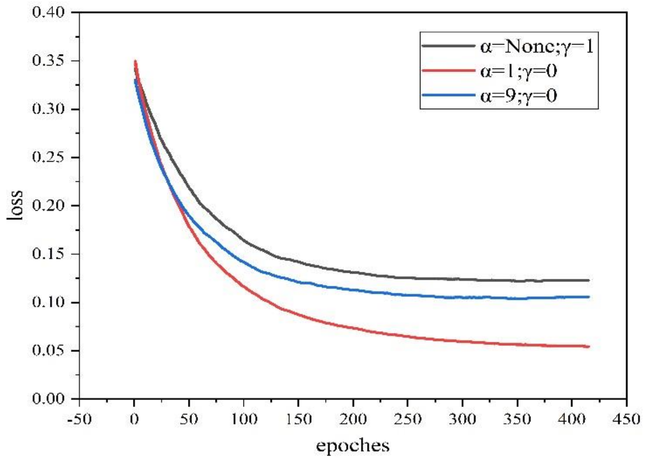


FIGURE 11. Focal loss chart.

V. CONCLUSION AND DISCUSSION

This paper proposed a CHD prediction method based on the HY_OptGBM model. Framingham Heart Institute data on CHD was selected as measurements, and the proposed method was trained using the HY_OptGBM algorithm and the comparison algorithms. Although different algorithms were used for CHD prediction in this study, the best CHD prediction was achieved by the improved LightGBM algorithm. When using data from the Framingham Heart Institute's CHD study, observing all predicted values using the HY_OptGBM algorithm yielded more successful results, which is the significance of this study. In the experiment, sensitivity, specificity, accuracy, precision, recall, F-score, AUROC, AUPRC and MCC were used as evaluation metrics. The experimental results of the DT, RF, CB, XGB, ADA, BG, GBM and HY_OptGBM algorithms were compared, and the best results were obtained using the HY_OptGBM algorithm. The sensitivity was 0.897, the specificity was 0.963, the accuracy was 0.930, the precision was 0.963, the recall was 0.897, the F-score was 0.929, the AUROC was 0.978, the AUPRC was 0.983, and the MCC was 0.861. This study proposed optimizing the hyperparameters of the LightGBM algorithm and improving its loss function (FL). The experimental results will change when changing the alpha and gramma parameters of the FL function. After the experiments were conducted, when the parameter alpha was None and gamma was 1, the accuracy, F-score, AUROC, AUPRC, MCC metrics had the best results. When alpha was 0.1 and gamma was 0, the specificity and precision had the best results. When alpha was 0.9 and gamma was 0, the sensitivity and recall had the best results. When evaluating the performance of a machine learning algorithm, usually multiple evaluation metrics are considered together, so alpha was taken as None

TABLE IX
Comparison of CHD prediction studies using the FHS dataset (2018-2022) (sen denotes sensitivity, spe denotes specificity, acc denotes accuracy, pre denotes precision, rec denotes recall, and f-se denotes f-score)

	Dataset	Method	Evaluation Metrics	Evaluation Results
Orit Goldman et al. [12]	FHS	ANN	auroc, sen, spe	ANN model has higher performance than FRS model
Juan-jose Beunza et al. [43]	FHS	SVM	auroc	auroc=0.75
Meeshanthini V Dogan et al. [44]	FHS	RF	acc, sen, spe	acc=0.78, sen=0.78, spe=0.80
Meeshanthini V Dogan et al. [45]	FHS	Epi+Gen	sen, spe	sen=0.79, spe=0.75
Steven Simon et al. [46]	FHS	LR	auroc	auroc=0.71
S. Prabu et al. [47]	FHS	GPR+KRR	rec, f-se, acc	acc=0.86, rec=0.902, f-se=0.821
proposed method	**FHS**	**HY_OptGBM**	**roc, spe, acc, pre, rec, f-se, auroc, auprc, mcc**	**acc=0.930, sen=0.897, spe=0.963, f-se=0.929, pre=0.963, rec=0.897, auroc=0.978, auprc=0.983, mcc=0.861**

and gamma was taken as 1 to obtain the final experimental results. As shown in Tables Ⅴ, Ⅵ, Ⅷ, and Fig. 9 and Fig. 10, the best results can be obtained when making predictions with the proposed method.

To compare studies in the literature with the proposed methodology, experimental studies using the Framingham CHD dataset were checked. This dataset has mostly been used to predict the probability of developing CHD within ten years. In 2021, Orit Goldman et al. [12] used ANN models to predict CHD, and the prediction results showed that the lift and gain curves of ANN models were higher than those of FRS models in terms of the highest percentile. For higher risk scores, the ANN model had higher sensitivity and specificity than the FRS model, but the ANN model had lower area under the curve (AUC) values. For the precision-recall measures, ANN models produce significantly better results than FRS models in terms of AUC values. In a 2019 study, Juan-jose Beunza et al. [43] conducted a comparative study of the dataset using machine learning methods. Decision trees, random forests, support vector machines, neural networks and logistic regression were selected for the classification study. The results of the study demonstrated that the support vector machine algorithm had the best AUC value of 0.75. Meeshanthini V Dogan et al. [44], in a 2018 study, used machine learning techniques to construct predictive CHD models. The accuracy, sensitivity and specificity obtained using the random forest classifier were 0.78, 0.75 and 0.80, respectively. In a 2021 study by Meeshanthini V Dogan et al. [45], an ensemble genetic performance genetic model for predicting 3-year coronary events was developed. This model showed a sensitivity of 0.79, a specificity of 0.75, a sensitivity of 0.15 and a specificity of 0.93 on the test set. In 2022, Steven Simon et al. [46] used logistic regression to classify and predict CHD, and AUROC values of 0.71 were obtained. In a study by S. Prabu [47] in 2021, CHD was predicted by using Gaussian process regression (GPR) and kernel ridge regression (KRR) machine learning algorithms and a hyperparametric search of the algorithm, and the final prediction results demonstrated a recall of 0.902, an F1-score of 0.821, and an accuracy of 0.86. Table Ⅸ gives details of CHD prediction studies in the past 5 years using the Framingham Heart Institute's open dataset. Using the synthetic minority oversampling technique (SMOTE) for preprocessing the dataset, optimizing the hyperparameters of the algorithm and improving its loss function in the experimental study were considered, and the success of the proposed method in predicting CHD compared with other methods was demonstrated. In contrast to other studies in the literature, the use of the most advanced algorithm in ensemble learning (LightGBM) in this study, as well as the use of the most advanced hyperparameter optimization framework (OPTUNA) for optimization of the hyperparameters of the algorithm and improvement of its loss function, led to sensitivity, specificity, accuracy, precision, recall, F score, AUROC, AUPRC and MCC enhancements, which are important for diseases such as CHD, which have lethal disease consequences. Due to the lack of similar optimized and improved prediction methods in the literature, the proposed method in this paper provides a new perspective for future CHD prediction studies.

In future studies, the Framingham Heart Institute dataset should be used to predict CHD, and multiple CHD datasets should be used to build predictive models. When experimenting with the FL function, the alpha and gamma parameters affect the study results. Thus, more accurate results can be obtained by constructing prediction models through multiple trials. The methodology proposed in this study will also be integrated in future studies. As the numbers of trials and datasets increases, it will be necessary to obtain a successful result by adjusting the default parameters presented in this paper. In addition to using a single model to predict CHD, alternatively, one may consider building a prediction model by combining multiple models.

ACKNOWLEDGMENT

The authors would like to thank IEEE Access and their respective universities for their support.

REFERENCES

[1] N. Katta, T. Loethen, C. J. Lavie, and M. A. Alpert, "Obesity and coronary heart disease: Epidemiology, pathology, and coronary artery imaging," *Current Problems Cardiology*, vol. 46, no. 3, p. 100655, Mar. 2021, doi: 10.1016/j.cpcardiol.2020.100655.

[2] G. S. Handelman, H. K. Kok, R. V. Chandra, A. H. Razavi, M. J. Lee, and H. Asadi, "eDoctor: Machine learning and the future of medicine," *J. Internal Medicine*, vol. 284, no. 6, pp. 603–619, Sep. 2018, doi: 10.1111/joim.12822.

[3] E. L. Romm and I. F. Tsigelny, "Artificial intelligence in drug treatment," *Annu. Rev. Pharmacology Toxicology,* vol. 60, no. 1, pp. 353–369, Jan. 2020, doi: 10.1146/annurev-pharmtox-010919-023746.

[4] L. Lo Vercio *et al.*, "Supervised machine learning tools: A tutorial for clinicians," *J. Neural Eng.*, vol. 17, no. 6, p. 062001, Dec. 2020, doi: 10.1088/1741-2552/abbff2.

[5] S. Rauschert, K. Raubenheimer, P. E. Melton, and R. C. Huang, "Machine learning and clinical epigenetics: A review of challenges for diagnosis and classification," *Clin. Epigenetics,* vol. 12, no. 1, p. 51, Apr. 2020, doi: 10.1186/s13148-020-00842-4.

[6] Y. Arfat, G. Mittone, R. Esposito, B. Cantalupo, G. M. De Ferrari, and M. Aldinucci, "Machine learning for cardiology," *Minerva Cardiology Angiology,* vol. 70, no. 1, pp. 75–91, Mar. 2022, doi: 10.23736/s2724-5683.21.05709-4.

[7] S. Nematzadeh, F. Kiani, M. Torkamanian-Afshar, and N. Aydin, "Tuning hyperparameters of machine learning algorithms and deep neural networks using metaheuristics: A bioinformatics study on biomedical and biological cases," *Comput. Biol. Chemistry,* vol. 97, p. 107619, Apr. 2022, doi: 10.1016/j.compbiolchem.2021.107619.

[8] M. Liang *et al.*, "Improving genomic prediction with machine learning incorporating TPE for hyperparameters optimization," *Biology,* vol. 11, no. 11, p. 1647, Nov. 2022, doi: 10.3390/biology11111647.

[9] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "OPTUNA: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Anchorage, USA: ACM, 2019, pp. 2623–2631.

[10] M. Yeung, E. Sala, C.-B. Schönlieb, and L. Rundo, "Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation," *Computerized Med. Imag. Graph.,* vol. 95, p. 102026, Jan. 2022, doi: 10.1016/j.compmedimag.2021.102026.

[11] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA: ACM, 2017, pp. 3149–3157.

[12] O. Goldman, O. Raphaeli, E. Goldman, and M. Leshno, "Improvement in the prediction of coronary heart disease risk by using artificial neural networks," *Quality Manage. Health Care*, vol. 30, no. 4, pp. 244–250, Jul. 2021, doi: 10.1097/qmh.0000000000000309.

[13] Z. Du *et al.*, "Accurate prediction of coronary heart disease for patients with hypertension from electronic health records with big data and machine-learning methods: Model development and performance evaluation," *JMIR Med. Inform.*, vol. 8, no. 7, p. e17257, Jul. 2020, doi: 10.2196/17257.

[14] J. K. Kim and S. Kang, "Neural network-based coronary heart disease risk prediction using feature correlation analysis," *J. Healthcare Eng.*, vol. 2017, p. 2780501, Sep. 2017, doi: 10.1155/2017/2780501.

[15] C. Krittanawong, H. Zhang, Z. Wang, M. Aydar, and T. Kitai, "Artificial intelligence in precision cardiovascular medicine," *J. Amer. College Cardiology,* vol. 69, no. 21, pp. 2657–2664, May. 2017, doi: 10.1016/j.jacc.2017.03.571.

[16] A. Akella and S. Akella, "Machine learning algorithms for predicting coronary artery disease: Efforts toward an open source solution," *Future Sci. OA*, vol. 7, no. 6, p. FSO698, Mar. 2021, doi: 10.2144/fsoa-2020-0206.

[17] L. J. Muhammad, I. Al-Shourbaji, A. A. Haruna, I. A. Mohammed, A. Ahmad, and M. B. Jibrin, "Machine learning predictive models for coronary artery disease," *SN Comput. Sci.*, vol. 2, no. 5, pp. 350, Mar. 2021, doi: 10.1007/s42979-021-00731-4.

[18] C. A. U. Hassan *et al.*, "Effectively predicting the presence of coronary heart disease using machine learning classifiers," *Sensors,* vol. 22, no. 19, p. 7227, Sep. 2022, doi: 10.3390/s22197227.

[19] captainozlem, "Framingham_CHD_preprocessed_data. Version 1." Accessed: May 5, 2020 [Online.] Available: https://www.kaggle.com/-datasets/captainozlem/framingham-chd-preprocessed-data/download?datasetVersionNumber=1

[20] V. Voillet, P. Besse, L. Liaubet, M. San Cristobal, and I. González, "Handling missing rows in multi-omics data integration: Multiple imputation in multiple factor analysis framework," *BMC Bioinf.*, vol. 17, no. 1, p. 402, Oct. 2016, doi: 10.1186/s12859-016-1273-5.

[21] G. Douzas and F. Bacao, "Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE," *Inf. Sci.*, vol. 501, pp. 118–135, Oct. 2019, doi: 10.1016/j.ins.2019.06.007.

[22] D. Che, Q. Liu, K. Rasheed, and X. Tao, "Decision tree and ensemble learning algorithms with their applications in bioinformatics," in *Software Tools and Algorithms for Biological Systems. Advances in Experimental Medicine and Biology*, H. Arabnia and Q. N. Tran, Eds. New York, NY: Springer, 2011, pp. 191–199.

[23] L. Yang *et al.*, "Study of cardiovascular disease prediction model based on random forest in eastern China," *Sci. Rep.*, vol. 10, no. 1, p. 5245, Mar. 2020, doi: 10.1038/s41598-020-62133-5.

[24] J. T. Hancock and T. M. Khoshgoftaar, "Catboost for big data: An interdisciplinary review," *J. Big Data,* vol. 7, no. 1, p. 94, Nov. 2020, doi: 10.1186/s40537-020-00369-8.

[25] W. Wenbo, S. Yang, and C. Guici, "Blood glucose concentration prediction based on VMD-KELM-AdaBoost," *Med. Biol. Eng. Comput.*, vol. 59, no. 11-12, pp. 2219–2235, Sep. 2021, doi: 10.1007/s11517-021-02430-x.

[26] X. Mi, F. Zou, and R. Zhu, "Bagging and deep learning in optimal individualized treatment rules," *Biometrics,* vol. 75, no. 2, pp. 674–684, Mar. 2019, doi: 10.1111/biom.12990.

[27] D. D. Rufo, T. G. Debelee, A. Ibenthal, and W. G. Negera, "Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM)," *Diagnostics,* vol. 11, no. 9, p. 1714, Sep. 2021, doi: 10.3390/diagnostics11091714.

[28] J. Feng, B. Ni, D. Xu, and S. Yan, "Histogram contextualization," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 778–788, Feb. 2012, doi: 10.1109/tip.2011.2163521.

[29] P. Łabędź, K. Skabek, P. Ozimek, and M. Nytko, "Histogram adjustment of images for improving photogrammetric reconstruction," *Sensors,* vol. 21, no. 14, p. 4654, Jul. 2021, doi: 10.3390/s21144654.

[30] L. Lin, J. Zhang, N. Zhang, J. Shi, and C. Chen, "Optimized LightGBM power fingerprint identification based on entropy features," *Entropy,* vol. 24, no. 11, p. 1558, Oct. 2022, doi: 10.3390/e24111558.

[31] O. Krivorotko, M. Sosnovskaia, I. Vashchenko, C. Kerr, and D. Lesnic, "Agent-based modeling of COVID-19 outbreaks for New York state and UK: Parameter identification algorithm," *Infectious Disease Model.*, vol. 7, no. 1, pp. 30–44, Mar. 2022, doi: 10.1016/j.idm.2021.11.004.

[32] A. Namoun, B. R. Hussein, A. Tufail, A. Alrehaili, T. A. Syed, and O. BenRhouma, "An ensemble learning based classification approach for the prediction of household solid waste generation," *Sensors,* vol. 22, no. 9, p. 3506, May. 2022, doi: 10.3390/s22093506.

[33] M. M. Arifin *et al.*, "OLGBM: OPTUNA optimized light gradient boost-ing machine for intrusion detection," in *2021 Int. Conf.Comput. Commun. Chem. Mater. Electron. Eng. (IC4ME2)*, Rajshahi, Bangladesh: IEEE, 2021, pp. 1–4.

[34] P. Srinivas and R. Katarya, "hyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using XGBoost," *Biomed. Signal Process. Control,* vol. 73, p. 103456, Mar. 2022, doi: 10.1016/j.bspc.2021.103456.

[35] D. Jensen and J. Neville, "Correlation and sampling in relational data mining," in *Proc. 33rd Symp. Interface Comput. Sci. Statist.*, 2001.

[36] S. Yan, J. M. Peck, M. Ilgu, M. Nilsen-Hamilton, and M. H. Lamm, "Sampling performance of multiple independent molecular dynamics simulations of an RNA aptamer," *ACS Omega,* vol. 5, no. 32, pp. 20187–20201, Aug. 2020, doi: 10.1021/acsomega.0c01867.

[37] M. Komorowski, D. C. Marshall, J. D. Salciccioli, and Y. Crutain, "Exploratory data analysis," in *Secondary Analysis of Electronic Health Records.* Cham: Springer International Publishing, 2016, pp. 185–203.

[38] T. R. Vetter, "Descriptive statistics: Reporting the answers to the 5 basic questions of who, what, why, when, where, and a sixth, so what?," *Anesthesia Analgesia,* vol. 125, no. 5, pp. 1797–1802, Nov. 2017, doi: 10.1213/ane.0000000000002471.

[39] B. Wang, J. J. Klemeš, P. S. Varbanov, and M. Zeng, "An extended grid diagram for heat exchanger network retrofit considering heat exchanger types," *Energies,* vol. 13, no. 10, p. 2656, May. 2020, doi: 10.3390/en13102656.

[40] M. W. Browne, "Cross-validation methods," *J. Math. Psychol.,* vol. 44, no. 1, pp. 108–132, Mar. 2000, doi: 10.1006/jmps.1999.1279.

[41] S. Parvandeh, H.-W. Yeh, M. P. Paulus, and B. A. McKinney, "Consensus features nested cross-validation," *Bioinformatics,* vol. 36,

no. 10, pp. 3093–3098, May. 2020, doi: 10.1093/bioinformatics/btaa046.

[42] S. Kucheryavskiy, S. Zhilin, O. Rodionova, and A. Pomerantsev, "Procrustes cross-validation—a bridge between cross-validation and independent validation sets," *Analytical Chemistry*, vol. 92, no. 17, pp. 11842–11850, Aug. 2020, doi: 10.1021/acs.analchem.0c02175.

[43] Beunza, Juan-Jose et al. "Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease)." *Journal of biomedical informatics*, vol. 97, p.103257,Sep. 2019, doi:10.1016/j.jbi.2019.103257.

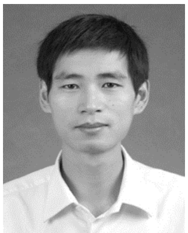
[44] Dogan, Meeshanthini V et al. "Integrated genetic and epigenetic prediction of coronary heart disease in the Framingham Heart Study." *PloS one*, vol. 13, no.1, p.e0190549, Jan. 2018, doi:10.1371/journal.pone.0190549.

[45] Dogan, Meeshanthini V et al. "External validation of integrated genetic-epigenetic biomarkers for predicting incident coronary heart disease." *Epigenomics*, vol. 13, no.14, pp.1095-1112, Jun.2021, doi:10.2217/epi-2021-0123.

[46] Simon, Steven et al. "The Impact of Time Horizon on Classification Accuracy: Application of Machine Learning to Prediction of Incident Coronary Heart Disease." *JMIR cardio*, vol. 6, no.2, p.e38040, Nov. 2022, doi:10.2196/38040.

[47] Prabu, S., et al. "Grid Search for Predicting Coronary Heart Disease by Tuning Hyper-Parameters." *Comput. Syst. Sci. Eng*, vol.43, no.2, pp.737-749, May. 2022.

HUAZHONG YANG received his B.S. degree in computer science from Yangtze University in 2021 and is now pursuing a research-based master's degree at Yangtze University. His major research interests include machine learning, data mining, computational biology, and healthcare.



ZHONGJU CHEN, associate professor, received his engineering degree in computer application and maintenance from Yangtze University in 1998. He also received his master's degree in earth exploration and information from Yangtze University in 2004. He has published ten papers in SCI and EI journals, winning one special national prize and a second for guiding university students in discipline competitions. His major research interests include computer networks, software engineering, and artificial intelligence (machine learning).

HUAJIAN YANG received his bachelor's degree in mechanical engineering from the North China University of Technology in 2021. He is pursuing a master's degree in mechanical engineering at Dongguan Institute of Technology. His major research interests include mechanical control, computer control technology, and machine learning.

MAOJIN TIAN, B.S. in preventive medicine, Binzhou Medical College, is currently studying at the School of Public Health, Zunyi Medical University, with research interests in microbiology, food preservation and computational biology.