Informatics and Health 1 (2024) 70–81

Contents lists available at ScienceDirect

Informatics and Health

journal homepage: www.keaipublishing.com/en/journals/informatics-and-health/

Optimizing the light gradient-boosting machine algorithm for an efficient early detection of coronary heart disease

Temidayo Oluwatosin Omotehinwa [a,*], David Opeoluwa Oyewola [b], Ervin Gubin Moung [c]

[a] Department of Mathematics and Computer Science, Federal University of Health Sciences, P.M.B. 145, Otukpo, Nigeria
[b] Department of Mathematics and Statistics, Federal University Kashere, P.M.B. 0182, Gombe, Nigeria
[c] Data Technologies and Applications (DaTA) Research Group, Faculty of Computing and Informatics, Universiti Malaysia Sabah, Kota Kinabalu 88400, Sabah, Malaysia

ARTICLE INFO

Keywords:
Clinical decision making
Coronary heart disease
Light gradient-boosting machine
Machine learning
MICE
Tree-structured Parzen estimator

ABSTRACT

Background: Coronary heart disease (CHD) remains a prominent cause of mortality globally, necessitating early and accurate detection methods. Traditional diagnostic approaches can be invasive, costly, and time-consuming, necessitating the need for more efficient alternatives. This aimed to optimize the Light Gradient-Boosting Machine (LightGBM) algorithm to enhance its performance and accuracy in the early detection of CHD, providing a reliable, cost-effective, and non-invasive diagnostic tool.

Methods: The Framingham Heart Study (FHS) dataset publicly available on Kaggle was used in this study. Multiple Imputations by Chained Equations (MICE) were applied separately to the training and testing sets to handle missing data. Borderline-SMOTE (Synthetic Minority Over-sampling Technique) was used on the training set to balance the dataset. The LightGBM algorithm was selected for its efficiency in classification tasks, and Bayesian Optimization with Tree-structured Parzen Estimator (TPE) was employed to fine-tune its hyperparameters. The optimized LightGBM model was trained and evaluated using metrics such as accuracy, precision, and AUC-ROC on the test set, with cross-validation to ensure robustness and generalizability.

Findings: The optimized LightGBM model showed significant improvement in early CHD detection. The baseline LightGBM model with dropped missing values had an accuracy of 0.8333, sensitivity of 0.1081, precision of 0.3429, F1 score of 0.1644, and AUC of 0.6875. With MICE imputation, performance improved to an accuracy of 0.9399, sensitivity of 0.6693, precision of 0.9043, F1 score of 0.7692, and AUC of 0.9457. The combined approach of Borderline-SMOTE, MICE imputation, and TPE for LightGBM achieved an accuracy of 0.9882, sensitivity of 0.9370, precision of 0.9835, F1 score of 0.9597, and AUC of 0.9963, indicating a highly effective and robust model.

Interpretation: The optimized model demonstrated outstanding performance in early CHD detection. The study's strengths include its comprehensive approach to addressing missing data and class imbalance and the fine-tuning of hyperparameters through Bayesian Optimization. However, there is a need to test with other datasets for its generalizability to be well-established. This study provides a strong framework for early CHD detection, improving clinical practice by allowing for more precise and dependable diagnostics and effective interventions.

1. Introduction

Coronary Heart Disease (CHD) is a cardiac condition that arises when the coronary arteries, responsible for delivering blood to the heart muscle, experience narrowing or blockage due to the accumulation of plaque[11,45]. This condition significantly impacts the proper functioning of the heart. Plaque is a build-up of cholesterol, fat, and various substances that can gradually accumulate along the inner walls of arteries[48]. This build-up results in decreased blood circulation to the myocardium, which can induce pain in the chest region or discomfort, or a heart attack[46]. Cardiovascular diseases including CHD are the primary factors contributing to illness and death worldwide[43]. According to this study[23], approximately 126 million people worldwide (equivalent to 1655 individuals per 100,000) are affected by coronary heart disease, accounting for roughly 1.72 % of the global population. Furthermore, CHD was responsible for 9 million fatalities globally. Several factors

* Correspondence to: Department of Mathematics and Computer Science, Faculty of Science, Federal University of Health Sciences, P.M.B. 145, Otukpo, Nigeria.
E-mail address: temidayo.omotehinwa@fuhso.edu.ng (T.O. Omotehinwa).

https://doi.org/10.1016/j.infoh.2024.06.001
Received 12 February 2024; Received in revised form 14 June 2024; Accepted 15 June 2024
Available online 2 July 2024

contribute to the risk of developing CHD. These include demographic factors such as age and gender, a family history of the disease, smoking habits, elevated blood pressure, high cholesterol levels, diabetes, obesity, and a sedentary lifestyle[21]. These risk factors can lead to the development and progression of arterial sclerosis, which is the underlying process that causes CHD[19,28]. Early detection and intervention play a critical role in preventing CHD and reducing its burden. The Framingham Heart Study, which began in 1948[5], has been instrumental in identifying the risk factors associated with CHD. However, predicting CHD risk remains a challenging task, and machine learning approaches have been used to enhance predictive accuracy.

One of the main challenges in developing predictive models for CHD is class imbalance. The Framingham dataset[20] is imbalanced, with a relatively small number of individuals who developed CHD within ten years compared to those who did not. This imbalance can lead to biased and inaccurate predictions. To address this, this study will use the Borderline-SMOTE (Borderline Synthetic Minority Over-sampling Technique) sampling method to oversample the minority class, which has been shown to improve the performance of machine learning models on imbalanced datasets[10,40,41]. Another challenge is how to improve the accuracy of the predictive model. Hyperparameter tuning is another critical aspect of developing accurate predictive models. Though it could be computationally intensive, it can enhance the predictive accuracy of machine learning models[34]. Bayesian optimization with a tree-structured Parzen estimator is an efficient and effective method for hyperparameter tuning[35,36]. This approach uses a probabilistic model to estimate the performance of a given set of hyperparameters and then searches for the optimal hyperparameters using Bayesian inference[7]. This technique has been shown to outperform other hyperparameter optimization methods, including grid search and random search[14,42].

This study is aimed at developing an effective early detection system for CHD using the Framingham dataset. The objectives are to address class imbalance using the Borderline-SMOTE sampling method, optimize the hyperparameters of the LightGBM (Light Gradient-Boosting Machine) algorithm using Bayesian optimization with a tree-structured Parzen estimator, and evaluate the performance of the developed model in predicting the 10-year risk of CHD. An effective early detection system for CHD can lead to timely interventions that can reduce the burden of this disease. Machine learning algorithms have shown promise in developing accurate predictive models for CHD[15,18,29,39,8], but further research is needed to optimize their performance. This study contributes to the development an improved predictive model predictive model for CHD, which can improve clinical decision-making and patient outcomes.

2. Related work

In recent years, many researchers have devoted considerable effort toward creating techniques for predicting cases of heart disease utilizing the aforementioned datasets. This study[8] carried out a comparative analysis of some supervised ML classifiers which include Boosted Decision Tree (BDT), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM) and Logistic Regression (LR). The classifiers were trained on three variants of the Framingham Heart Study (FHS) dataset: (i) the unprocessed FHS dataset, (ii) the FHS dataset with rows containing at least one missing value removed, and (iii) the FHS dataset where all missing values were replaced using mean imputation. The 80:20 train/test split approach was applied for all classifiers. The Random Oversampling Examples (ROSE) technique was used to address the class imbalance. In terms of accuracy, the best-performing model, the Boosted decision tree achieved an 85 % accuracy on the first variant of the dataset.

The study by[33] also focused on the development of an efficient predictive model for coronary heart disease detection. In the study, two predictive models based on the extreme gradient boost (XGBoost) and LR algorithms were modelled. The FHS dataset was used in the study and the missing values were handled by mean imputation. A 10-fold cross-validation was carried out on the train set. The XGBoost and LR models achieved accuracies of 84.46 % and 85.86 %, respectively.

This study[32], proposed an ensemble approach for CHD prediction. The Cleveland and FHS datasets were used in the study. The dataset was partitioned based on a stopping rule, and each partitioned dataset was modeled using the Classification and Regression Tree (CART). The ensemble models were constructed from the CART by modifying the Weighted Aging Classifier Ensemble (WACE). This allows the assignment of weights to each classifier based on their performance in terms of accuracy. Classifiers that achieved accuracy below a threshold were dropped from the ensemble. The proposed accuracy-based weighted aging classifier ensemble achieved an accuracy, precision, sensitivity, and f1 score of 91 %, 92 %, 90 %, and 91 % on the FHS dataset respectively.

This study[26] carried out a comparative study of the performance of the LR and RF in the detection of coronary heart disease. The FHS dataset was used in this study and all rows containing missing values were removed. The LR model achieved an accuracy of 85.04 % while the RF model achieved 84.4 % accuracy.

To prevent the progression of heart disease, Ebiaredoh-Mienye et al.,[13] applied a deep learning technique: the Sparse Autoencoder (SAE) for feature learning. The learned features were fed into a softmax classifier which is a logistic regression model generalized for classification. The softmax classifier works by taking a vector of inputs, also known as features, and computing a weighted sum of the input values. The result of this computation is then passed through a softmax function, which normalizes the output into a probability distribution over the classes. It assigns probabilities to each class and predicts the class with the highest probability. The SAE which is an unsupervised Artificial Neural Network (ANN) was trained on cervical cancer, chronic kidney disease, and the FHS dataset, and it performs dimensionality reduction. The resultant output is then passed to the softmax for supervised training. This approach achieved an accuracy, precision, recall, and F1 score of 91 %, 93 %, 90 %, and 92 % on the FHS dataset respectively. A similar study[31] to further improve the predictive accuracy of the SAE applied Particle Swarm Optimization (PSO) for the optimization of the Stacked Sparse Autoencoder (SSAE) parameters to ensure that the selected features are the most representative. The approach resulted in an improved performance in comparison to the SSAE and softmax classifier. The model achieved an accuracy of 0.973, precision of 0.948, sensitivity of 1.000, and F-measure of 0.973.

This study[38], proposed a machine learning-based technique to predict the risk of coronary heart disease based on a patient's clinical history. The study used the FHS dataset to train several classifiers, which include DT, Gradient Boosting Classifier (GBC), Naïve Bayes (NB), and LR. The p-value method was employed for feature elimination. The evaluation of the six classifiers involved the utilization of the Receiver Operating Characteristics (ROC) curve, confusion matrix, and the computation of the Area Under the Curve (AUC) value. The results showed that the GBC had the highest accuracy score of 87.61 %. The study posited that combining statistical methods such as p-value backward elimination with machine learning classifiers can improve the accuracy of the classifier and reduce the machine running time.

In the study by Masih et al.,[27], feed-forward neural networks, also known as Multilayer Perceptron (MLP), were trained on the FHS dataset in a bid to develop an early detection system for CHD. The missing values in the dataset were handled through the technique called the mean binning method. The missing values were filled with the mean of all entries of the feature with missing values. The Z-score of each attribute was used to replace its outliers. The 70:30 train/test split method was applied and the performance of the MLP model was evaluated. The model achieved an accuracy of 96.50 %, a sensitivity of 91.90 %, and a specificity of 98.28 %.

In this study, Hasan and Saleh,[17] proposed an ensemble approach that stacked SVM, DT, and XGBoost, and used LR as a meta-classifier to

combine the output of the stacked classifiers. The performance of the proposed approach was compared with other predictive models such as K-Nearest Neighbour (KNN), SVM, DT, XGBoost, and RF developed in the study. The proposed stacked ensemble approach achieved a sensitivity of 0.95, an F1 score of 0.97, an accuracy of 96.69 %, and an AUC score of 0.98 outperforming the individual models.

In a bid to improve the early detection of CHD, Correlation-based Feature Selection (CBFS) and Principal Component Analysis (PCA) were applied independently to the FHS dataset[25]. The applied CBFS and PCA resulted in five and thirteen selected features respectively. The Adaboost, NB, MLP, and Sequential Minimum Optimizer (SMO) algorithms were trained on the dataset based on both feature selection techniques, leading to eight predictive models. The CBFS+MLP model demonstrated superior performance compared to other models, achieving an accuracy score of 0.8490, a precision score of 0.805, and a recall score of 0.849.

This study[1] modelled the LR, SVM, and RF algorithms on the FHS dataset in a bid to determine an optimal model for predicting the ten-year risk of developing CHD. The median imputation was applied to address the missing values. The classifiers were trained using 10-fold cross-validation. The RF model outperformed the other traditional models and achieved an accuracy of 0.8505, an F1 score of 0.9190, and a geometric mean of 0.8840.

In a bid to ensure early detection of the risk of coronary heart disease, Yilmaz and Yağin,[49] modelled three ML algorithms: RF, LR, and SVM, on the Cleveland heart disease dataset. The RF model performed better than the LR and SVM models with an accuracy of 0.929, specificity of 0.929, F1 score of 0.928, and sensitivity of 0.928.

In this study[4], an ensemble-based approach to coronary heart disease prediction was proposed. The KNN, NB, DT, SVM, LR, RF, Stacking, and Voting classifiers were trained on a Multivariate Imputation by Chained Equations (MICE) imputed Framingham dataset. The performance of the developed models was evaluated in terms of accuracy; both the LR and stacking classifiers achieved 85.38 % on the Framingham dataset. The stacking classifier achieved an accuracy of 91.96 % on the UCI (University of California Irvine) dataset.

Devi et al.,[12], employed Framingham datasets and 14 features to increase the further the performance of predictive models for the detection of heart disease. A Feature-Selector optimization model with recursive feature elimination and the Boruto technique was utilized to choose the appropriate subset of coronary heart disease features. SMOTE was also utilized to generate synthetic examples to overcome data imbalance. For classification, Random Forest, Decision Tree, Gradient Boosting, Adaptive Boosting, and Support Vector Machine were utilised. The system's accuracy is 88 % when using the recursive feature elimination method with the random forest model. Similar research was conducted by Kigka et al.,[24], who employed machine learning to identify people at high and low risk of coronary artery disease (CAD). This study consists of five steps: data preprocessing, class imbalanced handling using the Easy Ensemble algorithm, recursive feature elimination technique implementation, gradient boosting classifier implementation, model evaluation, and fine-tuning presented hyperparameters over an internal 3-fold cross-validation. The 187 patients who previously had a CAD (Coronary Artery Disease) suspicion and underwent CTCA (Computed Tomography Coronary Angiography) in prior EVINCI (Evaluation of Integrated Cardiac Imaging for the Detection and Characterization of Ischemic Heart Disease) and ARTreat (Arterial Revascularization Treatment) clinical studies are included in this study. Imaging and non-imaging data were used to train the prediction model. This study's findings demonstrate an overall predictive accuracy of 81 %. A similar methodology was discovered in[37], where the authors combined image processing techniques with iridology, an iris analysis approach that utilizes the patterns, colours, and various attributes of the iris to gather insights about an individual's health condition. There was a total of 198 participants in this study, with 94 people having CAD and the remaining 104 volunteers not having CAD. The integral differential operator and rubber sheet approach was used to translate the iris transformation into a rectangular format. Wavelet transform, first-order statistical analysis, a Gray-Level Co-Occurrence Matrix (GLCM), and a Grey Level Run Length Matrix (GLRLM) were used to extract features in this work. The model's performance was assessed using measures such as accuracy, sensitivity, specificity, precision, score, mean, and AUC. The Support Vector Machine classifier produced a prediction accuracy of 93 % in this study.

This study[3] developed a predictive model based on the CART algorithm for the detection of CHD. In the study, two datasets - the FHS and Statlog heart datasets - were modelled. Four versions of each dataset were created by applying the SMOTE, Adaptive Synthetic (ADASYN), Borderline-SMOTE, ROSE, and safe level SMOTE to oversample the minority class. The SMOTE-based optimized model achieved 0.801, 0.79, 0.832, 0.807, and 0.68, accuracy, precision, recall, f1 score, and AUC score on the FHS dataset respectively while the ADASYN-based optimized model achieved an accuracy of 0.867, a precision of 0.75, a recall of 0.667, an F1 score of 0.65 and an Area Under the receiver operating characteristics Curve (AUC) score of 0.64 on the Statlog heart dataset.

This study[47], focused on hyperparameter optimization of the LightGBM algorithm using an advanced tuning framework called the OPTUNA developed by[2]. A 10-fold cross-validated OPTUNA-based LightGBM model was trained on the FHS dataset and its performance was compared with other models such as AdaBoost, DT, Gradient Boosting Machine (GBM), and XGBoost developed in the study. The class imbalance was addressed using the SMOTE oversampling technique and all rows with missing values were deleted. The proposed model achieved accuracy, sensitivity, specificity, F1-score, precision, the area under the receiver operating curve, and Mathews Correlation Coefficient (MCC) of 0.930, 0.897, 0.963, 0.929, 0.963, 0.978, and 0.861 respectively.

3. Methodology

3.1. Dataset

The Framingham dataset was generated as part of a continuous study focusing on cardiovascular health conducted with the participation of residents from Framingham, Massachusetts, and it is publicly accessible on the Kaggle website ([20]). This dataset is primarily employed in classification tasks to ascertain the probability of a patient's risk of developing CHD over ten years. It contains 4240 patient records and 16 distinct features, with each feature serving as an indicator of a specific risk factor. To detect the target feature, which is the 10-year risk of CHD, 15 of these input features were utilized. Table 1. provides an overview of

Table 1
An overview of the variables in the FHS dataset.

Variable	Description
male	Gender (0 = Male, 1 = Female)
age	Age in days
education	Education Level (1 to 4)
currentSmoker	Is the patient currently a smoker? 0 = No, 1 = Yes
cigsPerDay	Average daily cigarette consumption
BPMeds	Is the patient currently on blood pressure medication? 0 = No, 1 = Yes
prevalentStroke	Is the patient experiencing a stroke? 0 = No, 1 = Yes
prevalentHyp	Is the patient hypertensive? 0 = No, 1 = Yes
diabetes	Is the patient diabetic? 0 = No, 1 = Yes
totChol	Measurement of the total amount of cholesterol
sysBP	Systolic Blood Pressure
diaBP	Diastolic Blood Pressure
BMI	Body Mass Index
heartRate	Heart rate measurement
glucose	Level of glucose
TenYearCHD	The likelihood of developing heart disease in the next 10 years (target variable): 0 = No, 1 = Yes

The 'age' in days was converted to age in years before the modelling.

the various data features present in the Framingham dataset.

3.2. Handling missing values

The Framingham Heart Study dataset, like any real-world dataset, contains missing values that can affect the validity of the analysis. It is crucial to handle missing values effectively as they can distort the statistical inference and predictive modelling of the data. Fig. 1 depicts the missing data pattern in the Framingham dataset, indicating the presence of missing values for seven variables, namely heartRate, BMI, cigsPerDay, totChol, BPMeds, education, and glucose. Specifically, there are 1, 19, 29, 50, 53, 105, and 388 missing values for each variable, respectively. A total of 645 rows had missing values as revealed on the x-axis of the missing value visualization in Fig. 1.

Given that a sizeable amount of data is missing, removing the rows (List-wise deletion) that contain the missing data is not ideal as this could lead to the loss of valuable information and introduce bias. The mean/median imputation method is a simple and quick approach to handling missing data. However, one major disadvantage of mean imputation is that it cannot factor in the relationship of the missing data with other variables in the entire dataset. For example, a dataset with two columns containing the number of months and the mileage before a car failure. The average mileage of cars that failed after 40 to 60 months may not be a realistic estimate for a car that failed after only one month of use.

To handle missing values, the iterative imputer, Multivariate Imputation by Chained Equations (MICE), was employed. MICE is a powerful imputation method that can be used to handle missing values in datasets. The key advantage of using MICE is that it can capture the complex relationships between variables in the dataset and impute the missing values based on this information[6]. MICE can also preserve the distributional properties of the data, which is important for maintaining the statistical power of the analysis. According to Azur et al.,[6], MICE works by creating multiple imputations for each missing value and then averaging these imputations to obtain a single imputed value. At each iteration, the algorithm replaces the missing values with a predicted value, based on a model trained on the other variables in the dataset. The process is then repeated several times, with each iteration updating the imputed values for each variable. Each iteration results in a slightly different imputed dataset, as the predicted values for each missing value are slightly different due to the random variation in the model.

In this study, the MICE package in R was used to address the presence of missing values in the Framingham dataset. The dataset was split into 80 % train set and 20 % test set. MICE imputation was carried out separately on the train and test set. Three MICE imputation sets were generated for the train and test set and each of the imputation sets were evaluated without applying borderline-SMOTE. The results of the experiments are as presented in Table 3.

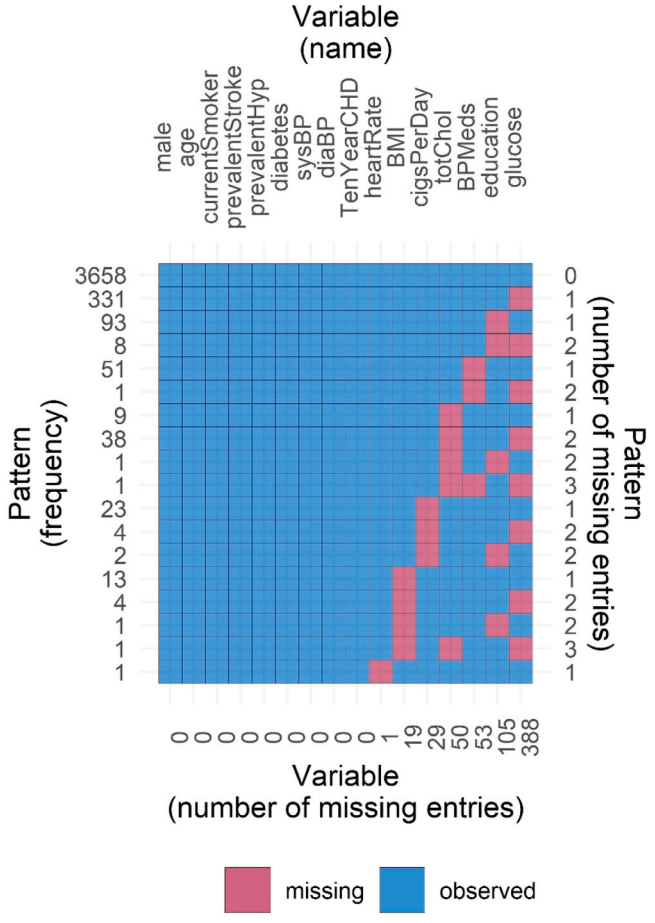

Fig. 1. Visualization of the missing and observed values in the Framingham heart study dataset.

3.3. Borderline-SMOTE algorithm

The Framingham Heart Study dataset has a class imbalance issue (Fig. 2) with the target variable TenYearCHD, where the positive class (i.e., individuals who are at risk of developing CHD in ten years) is relatively rare compared to the negative class (i.e., individuals who are not at risk of developing CHD in ten years). This class imbalance can lead to biased predictions and poor performance of classification models.

To address the class imbalance in the Framingham dataset, the Borderline-SMOTE algorithm developed by[16] was applied. This technique involves identifying borderline instances of the positive class, which are samples that are misclassified by the nearest neighbours of the same class, and then creating synthetic samples by interpolating between the borderline instances and their nearest neighbours. The Borderline-SMOTE algorithm can be summarized as follows:

1. Compute the k-nearest neighbours of each minority class instance:
   For each minority instance x_i:
   a. Compute the Euclidean distance between x_i and all other instances in the dataset.
   b. Arrange the distances in ascending order and choose the k closest neighbours that belong to both minority and majority class as x_i.
2. Identify the borderline instances:
   For each minority instance x_i:
   a. Count the number of neighbours of the same class (k_{same}) and the number of neighbours of the majority class ($k_{majority}$).
   b. If $k_{same} < k$ and ($k_{majority}$) $> k_{same}$, then x_i is a borderline instance.

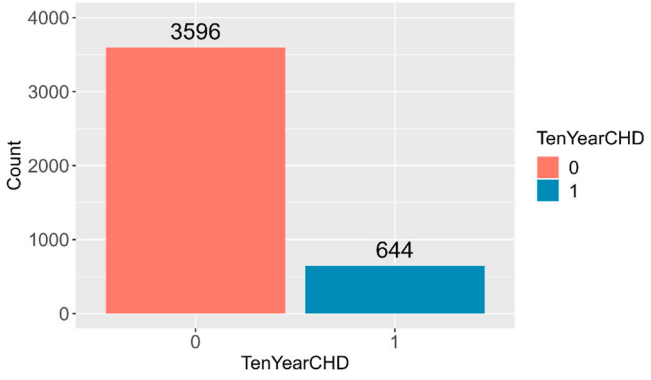

Fig. 2. Class distribution of the instances of individuals who developed CHD in ten years and those who did not, indicating the imbalance in the Framingham dataset.

3. Generate synthetic samples:
   For each borderline instance x_i:
   a. Select a random neighbour x_j from the set of k-nearest neighbours belonging to the same class.
   b. Generate a new synthetic sample x_{new} by interpolating between x_i and x_j: $x_{new} = x_i + rand(0,1) * (x_i - x_j)$
   c. Add x_{new} to the minority class instances and repeat steps 3a-c until the desired number of synthetic samples is generated.

3.4. Light gradient boosting machine (LightGBM)

The Light Gradient Boosting Machine (LightGBM) is a type of gradient-boosting machine learning algorithm that is used for supervised learning tasks such as regression, classification, and ranking. LightGBM was developed by Microsoft and is an open-source library[30]. LightGBM differs from traditional gradient boosting algorithms such as XGBoost and GBM in the way it builds trees. It uses a Gradient-based One-Sided Sampling (GOSS) technique to select only the important data points for tree construction, which reduces the training time and memory consumption. The LightGBM algorithm can be represented mathematically as follows:

Let X be the training dataset with N examples and M features, and Y be the corresponding target values. Let $f(x)$ be the function to be learned that maps from the input features to the target values.

The objective of LightGBM is to minimize the loss function $L(f)$ with respect to the function f as in Eq. (1).

$$L(f) = \sum [y_i - f(x_i)]^2 + \Omega(f) \tag{1}$$

where $\Omega(f)$ is a regularization term that controls the complexity of the learned function and prevents overfitting.

LightGBM solves this optimization problem by iteratively adding decision trees to the ensemble. At each iteration t, LightGBM constructs a decision tree $h_t(x)$ that minimizes the loss function over a subset of the training examples S_t:

$$h_t(x) = argmin_h \sum [y_i - f_{t-1}(x_i) - h(x_i)]^2 + \Omega(h) \tag{2}$$

where $f_{t-1}(x)$ is the ensemble of decision trees learned in the previous iterations.

LightGBM uses gradient boosting to optimize the loss function by iteratively adding decision trees to the ensemble. At each iteration t, LightGBM computes (Eq. (3)) the negative gradient of the loss function concerning the predicted target values of the previous ensemble.

$$g_i = -\partial L(f_{t-1}(x_i))/\partial f_{t-1}(x_i) \tag{3}$$

LightGBM then selects a subset of the training examples S_t using the Gradient-based One-Side Sampling (GOSS) technique[22]. GOSS selects examples with large gradients to preserve their importance in the learning process while under-sampling examples with small gradients reduces the computational cost and the risk of overfitting. LightGBM then constructs a decision tree on the subset S_t using a variant of the Gradient-based Decision Tree (GBDT) algorithm that grows the tree in a leaf-wise fashion. The GBDT algorithm selects the feature that results in the largest reduction in the loss function at each split and prunes the tree using a minimum gain threshold to prevent overfitting.

LightGBM continues the iterative process of adding decision trees to the ensemble until a stopping criterion is met, such as reaching the maximum number of trees or the minimum improvement in the validation set error.

Once the ensemble is trained, LightGBM can be used for prediction by computing the weighted average of the predictions of the individual decision trees as in Eq. (4).

$$f(x) = \sum_{t=1}^{T} w_t h_t(x) \tag{4}$$

where T is the number of trees in the ensemble, w_t is the weight of the t-th tree, and $h_t(x)$ is the prediction of the t-th tree. The weights are determined by the gradient boosting algorithm based on their contribution to the reduction of the loss function.

The choice of LightGBM for this study is informed by its powerful predictive power. As an ensemble algorithm, it builds trees that successively correct the predictive errors of the previous trees. The other advantage it has over ensemble algorithms such as Random Forest is that it is computationally efficient as it supports parallel computing and has a quick convergence time (due to the GOSS technique) while training with very large datasets. It has been established that LightGBM could be at least 16 times faster than CART and SVM[44], and 26 times faster than the extreme gradient boost[9].

Table 2
Hyperparameter space search configuration for the LightGBM model.

Model	Hyperparameter settings

Borderline-SMOTE + TPE + LightGBM	# Defining the search space for hyperparameters
	search_space = {
	'boosting_type': Categorical(['gbdt', 'dart', 'goss']),
	'num_leaves': Integer(2, 400),
	'max_depth': Integer(1, 750),
	'learning_rate': Real(0.01, 1.0, 'log-uniform'),
	'n_estimators': Integer(50, 1500),
	'min_child_samples': Integer(1, 100),
	'subsample': Real(0.1, 1.0, 'uniform'),
	'colsample_bytree': Real(0.1, 1.0, 'uniform'),
	'reg_alpha': Real(1e-9, 100, 'log-uniform'),
	'reg_lambda': Real(1e-9, 100, 'log-uniform'),
	'max_bin': Integer(100, 700),
	'max_delta_step': Real(0, 10, 'uniform')
	}
	# Performing hyperparameter optimization using Bayesian Search
	opt = BayesSearchCV(
	model,
	search_space,
	n_iter=100,
	cv=StratifiedKFold(n_splits=10),
	n_jobs= −1,
	scoring= 'roc_auc',
	random_state= 42
	)

3.5. Hyperparameter optimization

In this study, the hyperparameter space of the LightGBM model was searched using the Tree-Structured Parzen Estimator (TPE). The TPE is a Bayesian optimization algorithm developed by Bergstra et al.,[7]. It estimates a conditional probability distribution over the hyperparameters of a machine learning model. TPE constructs a tree of probabilistic models, where each node corresponds to a hyperparameter of the model. TPE uses a mixture of two density functions to model the distribution of the hyperparameters. One density function model the probability of good performance (For example, high accuracy), while the other models the probability of bad performance. The algorithm uses these probability distributions to sample new hyperparameters for evaluation in the next iteration. TPE seeks to maximize the expected improvement over the best-observed performance so far. The algorithm balances exploration (choosing hyperparameters that have performed well in the past) and exploitation (searching for new hyperparameters that may have a higher potential for good performance) by using a tree structure to evaluate the probability of each hyperparameter and decide which hyperparameters to sample next.

The TPE algorithm can be described mathematically as follows:

1. Let H be the hyperparameter space and let f be the objective function we want to optimize.

2. Define a prior probability density function $P(h)$ over the hyperparameter space.
3. Divide the hyperparameter space into two regions: the region of good performance, G, and the region of bad performance, B.
4. Define two conditional probability density functions, $p(h|G)$ and $p(h|B)$, that model the distribution of hyperparameters in the good and bad regions, respectively.
5. Sample a set of hyperparameters $h1$ from $p(h|G)$ and another set of hyperparameters $h2$ from $p(h|B)$.
6. Evaluate the objective function with the sampled hyperparameters and record the performance $f(h1)$ and $f(h2)$.
7. Update the regions G and B based on the recorded performance, i.e., move a hyperparameter set from B to G if it has better performance than the best hyperparameter set in G and vice versa.
8. Update the conditional probability density functions $p(h|G)$ and $p(h|B)$ using the recorded hyperparameters as a reference, i.e., fit a new density function to each region using the recorded hyperparameters as data.
9. Calculate the expected improvement over the best-observed performance. The expected improvement is given by:

$$EI(h) = E[\max(f(h') \quad - \quad f(h), 0)] \tag{5}$$

where h′ is a hyperparameter set sampled from $p(h|G)$ or $p(h|B)$.
10. Sample new hyperparameters h' from the density functions that maximize the expected improvement.
11. Repeat steps 5-10 until a stopping criterion is met.

The details of the hyperparameter search space defined are presented in Table 2.

3.6. Evaluation metrics

The performance of the proposed model was evaluated based on the confusion matrix, accuracy, recall, precision, specificity, F1 score, and the Mathews Correlation Coefficient.

3.6.1. Confusion matrix

The effectiveness of a binary classification model is tabulated in the confusion matrix. It lists the predicted and actual class labels for a group of cases so that we may assess the model's precision and mistakes. Four cells make up the confusion matrix, each of which represents a concatenation of expected and actual class labels. The cells are composed of the True Positives (TP) which represents instances where a positive outcome was accurately predicted, False Positive (FP) which are instances that are falsely identified as positive, True Negative (TN) represents instances that were accurately predicted as negative and the instances incorrectly classified as negative are referred to as False Negatives (FN).

3.6.2. Accuracy

The model's accuracy is a metric for how frequently it predicts an instance's class label accurately. It is determined as in Eq. (6).

$$\frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Recall.

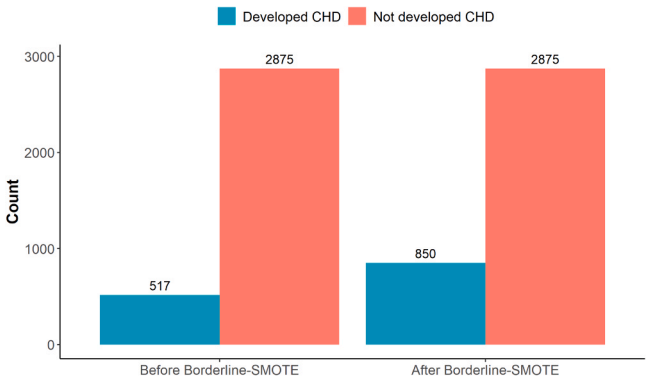

Fig. 3. Class distribution in the training set before and after oversampling of minority class with Borderline-SMOTE.

The proportion of genuine positives that were correctly identified is known as recall, often referred to as sensitivity, and it is a measurement of how well the model recognizes positive cases. It is evaluated thus:

$$\frac{TP}{TP + FN} \tag{7}$$

3.6.3. Precision

How many cases that are categorized as positive are actually positive is a measure of precision. It is characterized as the proportion of genuine positives to the total of both true and false positives. It is determined using Eq. (8).

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

3.6.4. Specificity

The proportion of true negatives that were correctly identified, sometimes referred to as the true negative rate or specificity, is an indicator of how well the model recognizes negative events. It can be computed using Eq. (9).

$$\frac{TN}{TN + FP} \tag{9}$$

3.6.5. F1 score

The accuracy of a model is assessed using the F1 score, which considers both precision and recall. The F1 score, which ranges from 0 to 1, is the harmonic mean of precision and recall; higher values denote better performance. It is determined as in Eq. (10).

$$\frac{2TP}{(2TP + FP + FN)} \tag{10}$$

3.6.6. Mathews correlation coefficient

Another evaluation metric that considers a binary classifier's true positive, true negative, false positive, and false negative rates is the Matthews correlation coefficient (MCC). MCC is a scale that goes from -1 to 1, with 0 denoting no connection between the expected and actual labels. A higher MCC value denotes improved model

Table 3
Performance evaluation result of the three (3) MICE imputed Framingham datasets.

Imputations	Confusion Matrix	Accuracy	Sensitivity	Precision	F1 score	MCC	AUC
MICE Imputation 1	$\begin{bmatrix} TN = 712 & FP = 9 \\ FN = 42 & TP = 85 \end{bmatrix}$	0.9399	0.6693	0.9043	0.7692	0.7465	0.9457
MICE Imputation 2	$\begin{bmatrix} TN = 712 & FP = 9 \\ FN = 42 & TP = 85 \end{bmatrix}$	0.9399	0.6693	0.9043	0.7692	0.7465	0.9425
MICE Imputation 3	$\begin{bmatrix} TN = 711 & FP = 10 \\ FN = 42 & TP = 85 \end{bmatrix}$	0.9387	0.6693	0.8947	0.7658	0.7415	0.9396

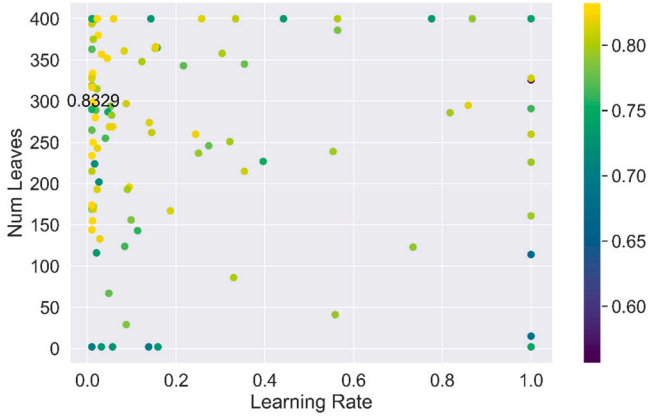

Fig. 4. The trajectory of the search process for determining optimal values of learning rate and number of leaves. The points (dots) represent the area under the curve (AUC) values obtained during the 10-fold cross-validation and hyperparameter tuning as the values of the Number of Leaves and Learning Rate change. The AUC value range from 0 to 1. The closer the AUC value is to 1 the better the discriminative power of the model.

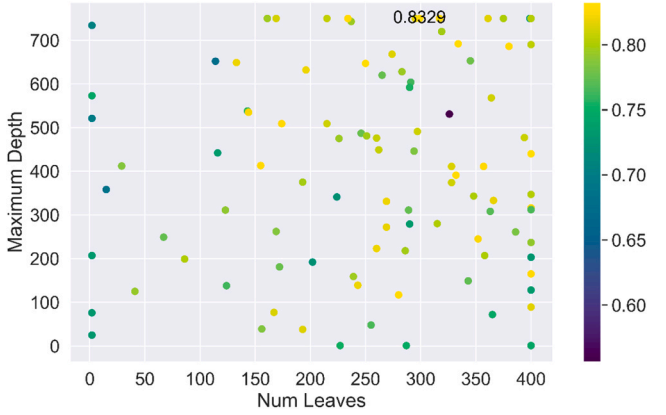

Fig. 5. The trajectory of the search process for determining optimal values of maximum depth and number of leaves. The points (dots) represent the area under the curve (AUC) values obtained during the 10-fold cross-validation and hyperparameter tuning as the values of the Number of Leaves and Maximum Depth change.

performance. The expression in Eq. (11) is used to evaluate the MCC.

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

3.7. Experimental setup

The analysis was carried out on a computer running Windows 10 Pro 64-bit operating system. The system specifications include 16 GB RAM, a 500 GB SSD, and a Corei-7–1165G7 @ 2.8 GHz processor. The algorithms were executed using Python version 3.10.9 in a Jupyter Notebook version 6.5.2. The primary libraries used in the analysis included pandas version 1.5.3, numpy version 1.23.5, matplotlib version 3.7.0, seaborn version 0.12.2, scikit-learn, imblearn version 0.0, LightGBM version 3.3.5, and scikit-optimize version 0.9.0. The data imputation for missing values was carried out on R version 4.2.2 (2022–10-31 ucrt) via Rstudio version 2022.12.0.353 using mice version 3.15.0.

A 10-fold cross-validated LightGBM tuned using Bayesian Optimization with Tree-structured Parzen Estimator (BO-TPE) was trained with 80 % (3392) of the MICE imputed Framingham dataset augmented with Borderline-SMOTE. The hyperparameter configuration of the LightGBM model is presented in Table 2.

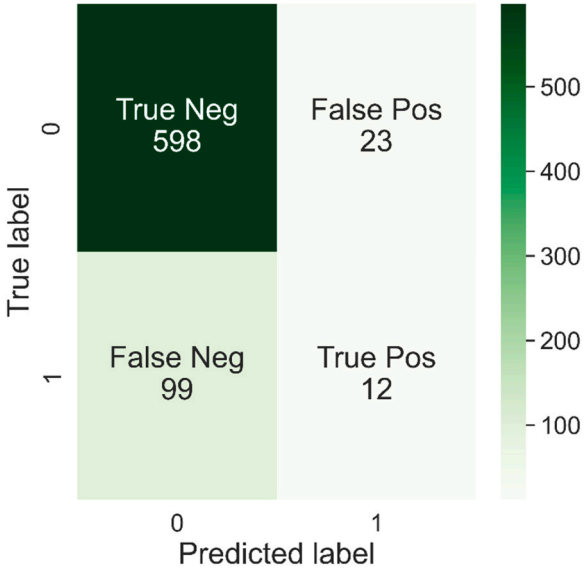

Fig. 6. The confusion matrix of the baseline LightGBM model with all rows with missing values dropped.

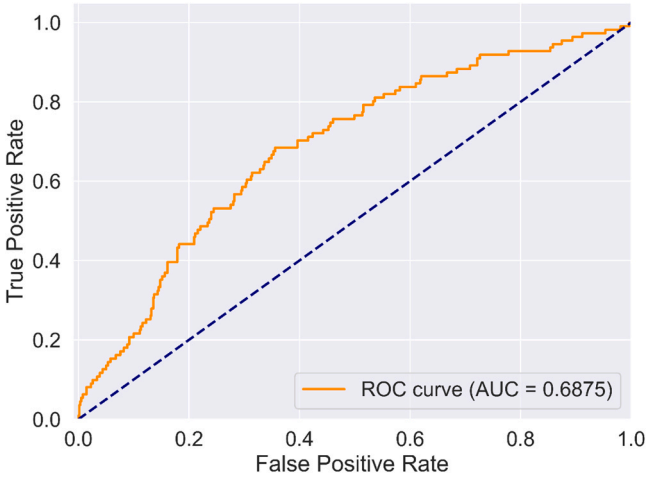

Fig. 7. The receiver operating characteristic curve of the baseline LightGBM model with missing value rows dropped.

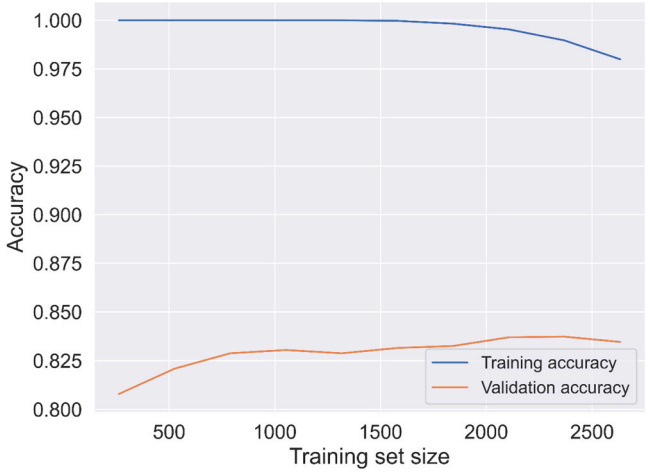

Fig. 8. The learning curve of the baseline LightGBM model with missing value rows dropped.

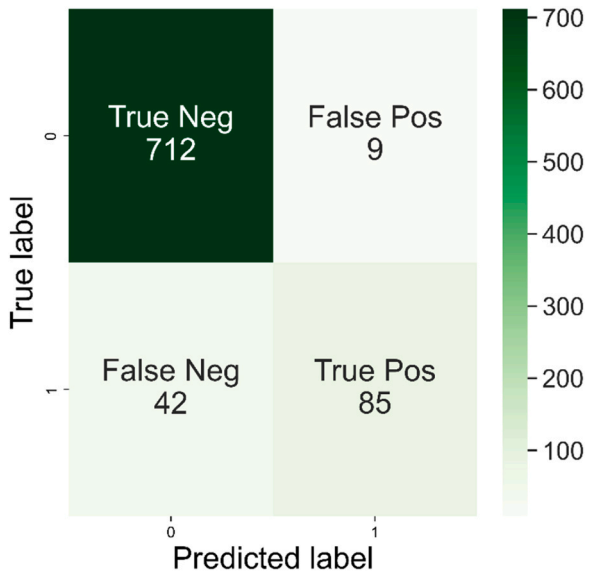

Fig. 9. The confusion matrix of the baseline LightGBM model with MICE imputation.

To assess the influence of hyperparameter tuning and borderline-SMOTE on the model performance, a 10-fold cross-validated baseline LightGBM model was built without applying hyperparameter tuning or oversampling of the minority class. This serves as a reference point to compare and evaluate the effects of these techniques on the model's predictive capabilities.

Since the dataset is imbalanced the StratifiedKFold function was applied to the cross-validation to ensures that each fold has approximately the same percentage of samples of each class as the entire dataset. The stratified sampling preserves class distribution in both sets, this ensures a representative distribution of the data and reduces the chances of bias during model evaluation. Also, setting a fixed random seed (random_state) ensures reproducibility and consistent results.

4. Results and discussion

4.1. Evaluation results of the three MICE imputation sets

Three MICE imputed train and test sets were generated. Each of the imputed sets was saved as three independent train datasets and experimented with using 10-fold cross-validation and performance evaluated on the three independent test datasets. The performance evaluation result of the three imputations is shown in Table 3.

The evaluation result of the first imputation resulted in a ROC-AUC value of 0.9457, the second imputation resulted in ROC-AUC score of 0.9425 and the third imputation set resulted in ROC-AUC score of 0.9396. From Table 3, It can be observed that the first and second imputation resulted in the same performance for all metrics evaluated except the ROC-AUC. The third imputation did not measure up to the first and second. Considering the ROC-AUC result the first imputation was selected for hyperparameter tuning and application of borderline-SMOTE.

4.2. Data oversampling

The MICE imputed train dataset (Imputation 1) represents 80 % (3392) of the entire Framingham dataset. The train set had 517 instances where individuals developed coronary heart disease in ten years and 2875 instances where coronary heart disease did not develop representing 15.24 % and 84.76 % of the training set. The dataset shows that there is an imbalance between the two classes utilized. As shown in Fig. 3, the Borderline-SMOTE algorithm was employed to address the minority class imbalance by increasing the number of samples in the "Developed CHD" class from 517 to 850. This indicates that the algorithm produced 333 synthetic samples for the minority population. It is important not to introduce too many synthetic data to avoid overfitting. As a result of using Borderline-SMOTE, the total number of cases in the training set is 3725. A 10-fold cross-validated LightGBM with hyperparameter optimization was carried out on this new training set.

4.3. Optimal hyperparameter values

The hyperparameter space of the LightGBM model was searched based on the settings in Table 2. The following hyperparameter values produced the best mean roc_auc score; boosting_type = goss, colsample_bytree = 0.9959432642084818, learning_rate = 0.013189503529031468, max_bin = 350, max_depth = 750, min_child_samples = 1, n_estimators = 1500, num_leaves = 299, reg_alpha = 1.1834004519625197e-07, reg_lambda = 1e-09, subsample = 0.3096708934167691.

The GOSS boosting type was selected as optimal for this model. This implies a faster training time. The proportion of features selected from the features in a tree was about 99.59 % as indicated by the value of the colsample_bytree. Using almost all the features ensures that complex relationship in the data are captured by the model. Aside from that, a learning rate of 0.01 indicates that the model will update its parameters by taking a step that is 0.01 times the gradient of the loss function

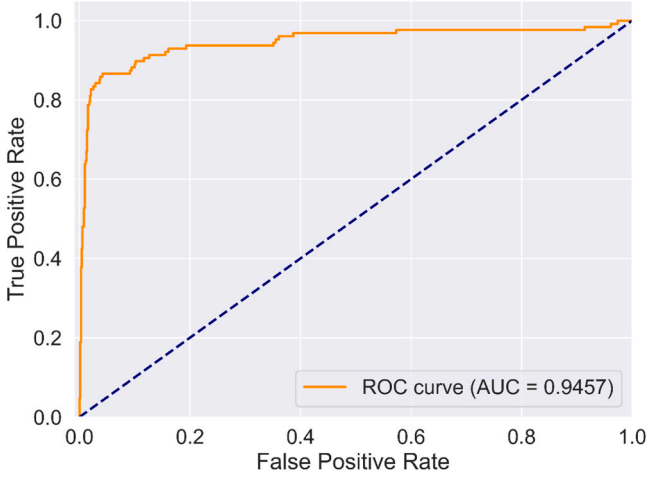

Fig. 10. The receiver operating characteristic curve of the baseline LightGBM model with MICE imputation.

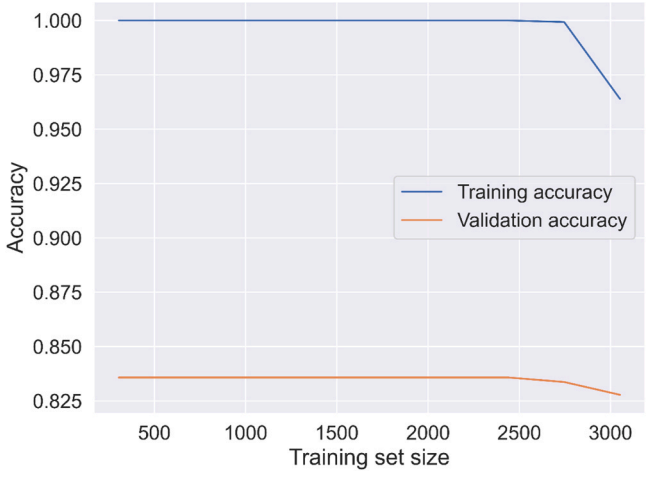

Fig. 11. The learning curve of the baseline LightGBM model with MICE imputation.

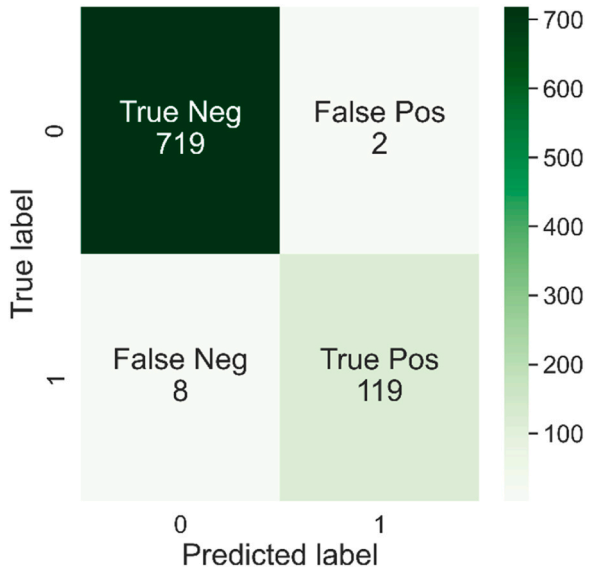

Fig. 12. The confusion matrix of the enhanced LightGBM model.

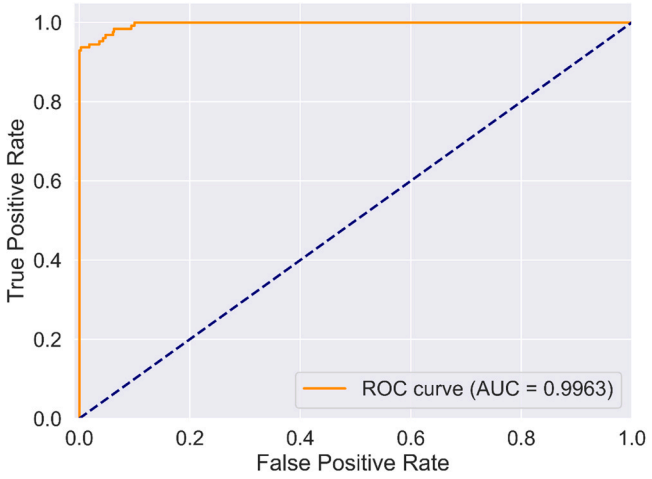

Fig. 13. The receiver operating characteristic curve of the enhanced LightGBM model.

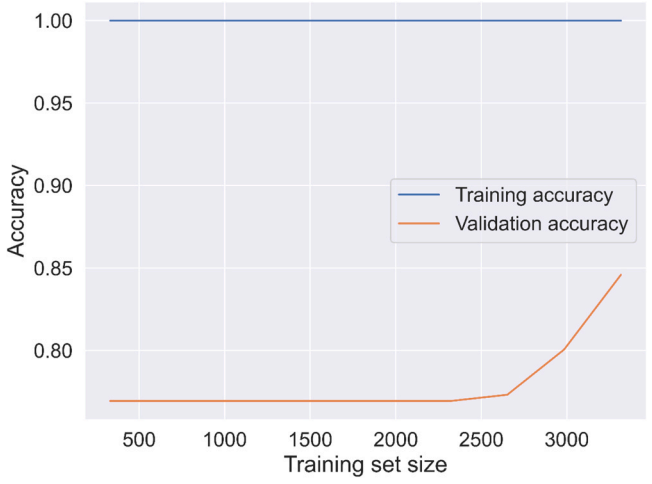

Fig. 14. The learning curve of the enhanced LightGBM model.

concerning those parameters. The learning rate in LightGBM ranges between 0 and 1 and determines how much each tree contributes to the overall prediction. A higher learning rate increases the impact of each tree but can lead to overfitting and instability. In contrast, a lower learning rate requires more iterations to converge but can improve generalization and performance. This implies that the learning rate of 0.01 controlled overfitting within the model and resulted in improved performance. A total of 1500 decision trees were created, with each tree in the boosting process having up to a maximum of 750 layers, each node was able to create a maximum of 299 leaf nodes, and each leaf node required 1 samples. A greater n_estimator value improves the model's capacity to learn complicated relationships, resulting in improved performance, but it also increases the computational cost and training time. A higher max_dept and num_leaves value enables the creation of a more complicated tree, perhaps catching more detailed patterns but risking overfitting. The value for reg_alpha and reg_lambda implies a weak regularisation, reflecting a desire for reduced feature sparsity and less sensitivity to individual data points. Cross-validation cushions the effect of these minimal penalty and ensures the model generalizes well to unseen data. A subsample value of 0.31 indicates that each tree was trained using about 31 % of the training data at random with replacement. This can help in reducing the possibility of overfitting. The search trajectory in Figs. 4 and 5 revealed the mean AUC obtained during cross-validation at the optimal hyperparameter values.

4.4. Performance evaluation

To establish the impact of the MICE imputation in addressing the missing values in the Framingham dataset, a baseline LightGBM model was trained and evaluated when all rows with missing values (NA) were dropped and the imbalance unaddressed. The confusion matrix of the baseline model without MICE imputation is presented in Fig. 6.

The result in Fig. 6 established that the model with all the missing values dropped achieved an accuracy of 0.8333, a sensitivity of 0.1081, a precision of 0.3429, an F1 score of 0.1644, and a MCC of 0.1195. The performance of the model is below average even with the accuracy being 83 %. It is important to note that accuracy is not the best evaluation metric when the class distribution is not balanced. A sensitivity of about 10 % indicates that about 90 % of individuals at risk of CHD are categorized as not having CHD.

In Fig. 7, the ROC curve of the baseline model with rows with missing values dropped is presented. With an AUC of 68.75 %, the model is unable to distinctively distinguish between those individuals who are at risk of CHD and those who are not.

The learning curve of the baseline LightGBM model with rows with missing values dropped is presented in Fig. 8. The learning curve describes the behaviour of the model as additional data are made available to the model. The training (representing the training scores for each training set size and each cross-validation fold) and test scores (representing the cross-validation scores for each training set size and each cross-validation fold) for increasing training set sizes are computed and plotted as a learning curve.

The learning curve in Fig. 8 revealed that the model was unable to sustain learning as the training set increased to about 1600 as it can be observed the accuracy was plunging downward from this point. The validation accuracy also did not improve as the data size increased and it also appears to be dovetailing as the data increases to about 2400.

The confusion matrix of the baseline LightGBM model is presented in Fig. 9. The results provide insights into the model's performance in predicting the risk of developing CHD within the next ten years. The model accurately classified 712 individuals as not having a risk of CHD (TN = 712). These individuals remained disease-free during the follow-up period, indicating the model's ability to identify low-risk individuals and potentially reduce unnecessary interventions or screenings. Nine (9) false positives (FP = 9) were detected, meaning the model incorrectly classified nine individuals as being at risk of CHD. This outcome is not

Table 4
Performance comparison of the baseline and optimized LightGBM model. MCC is the Mathews Correlation Coefficient and AUC is the area under the receiver operating characteristic curve.

Model	Accuracy	Sensitivity	Precision	F1 score	MCC	AUC
Baseline LightGBM with Nas dropped	0.8333	0.1081	0.3429	0.1644	0.1195	0.6875
Baseline LightGBM + MICE imputation	0.9399	0.6693	0.9043	0.7692	0.7465	0.9457
Borderline-SMOTE + MICE imputation TPE + LightGBM	0.9882	0.9370	0.9835	0.9597	0.9532	0.9963

Table 5
Comparison of the performance of the proposed model with previous studies on the Framingham heart study dataset.

Author & Year	Method	Missing Values Handling	Evaluation Metrics			
			Accuracy	Recall	Precision	F1 Score
Beunza et al.,[8]	ROSE + BDT	Deletion	0.850	-	-	-
Nalluri et al.,[33]	LR	Mean imputation	0.8586	-	-	-
Mienye et al.,[32]	WACE	-	0.91	0.90	0.92	0.91
Latifah et al.,[26]	LR	Deletion	0.8504	-	-	-
Ebiaredoh-Mienye et al.,[13]	Sparse Autoencoder + Softmax classifier	-	0.91	0.90	-	0.92
Saurabh Pal[38]	P-value Backward Elimination + GBC	-	0.8761	-	-	-
Masih et al.,[27]	MLP	Mean Binning	0.9650	0.9190	-	-
Hasan, Saleh[17]	Stacked Ensemble + Meta-classifier	-	0.9669	0.95	-	0.97
Kuruvilla, Balaji[25]	CBFS + MLP	-	0.8490	0.8490	0.805	-
Adel Mahmoud et al.,[1]	RF	Median imputation	0.8505	-	-	0.9190
Mienye, Sun[31]	Stacked Sparse Autoencoder + PSO	-	0.973	1.000	0.9480	0.973
Ambrews et al.,[4]	LR	MICE imputation	0.8538	-	-	-
Albert et al.,[3]	SMOTE + CART	-	0.801	0.832	0.79	0.807
Yang et al.,[47]	SMOTE + OPTUNA LightGBM	Deletion	0.930	0.897	0.963	0.929
This study.	Borderline-SMOTE + TPE + LightGBM	MICE imputation	0.9882	0.9370	0.9835	0.9597

desirable as it can create uneccessary anxiety, additional tests and interventions. Forty-Two false negatives (FN = 42) occurred, where the model misclassified individuals as low-risk who later developed CHD. This aspect is concerning as it suggests that the model failed to identify individuals who were at risk. Reducing false negatives is crucial to enhance the model's sensitivity in identifying individuals who would benefit from early interventions or closer monitoring. The model accurately identified 85 individuals as having a risk of CHD (TP = 85). These individuals indeed developed the disease during the follow-up period. Identifying high-risk individuals correctly enables timely interventions and closer monitoring to mitigate their risk.

The baseline LightGBM model's performance showed that there is room for improvement. The presence of false negatives and false positives indicates that further improvements can be made to enhance the model's ability to correctly classify individuals at risk and those not at risk of developing CHD.

From Fig. 10, it can be observed that the baseline LightGBM model achieved a 0.9457 area under the receiver operating characteristic curve which implies that the model ranked and differentiated those that are at a high risk of developing CHD (positive instances) and those that are at low risk (negative instances), with some misclassifications. This indicates that there is still room to improve accuracy and precision of the model.

From the learning curve in Fig. 11, it can be observed that the performance of the model remained consistently high as the size of the training set increased. However, there was a sharp descent as the train size increases to about 2500 this indicates that the model is not adapting well to the new data points. This relects the imbalance in class distribution. The validation performance of the model was steady without significant improvement as the instances increases and also experienced a gradual drop in accuracy. This indicates that more data could significantly improve the performance of the baseline model.

The results of the model evaluation on the separately imputed 20 % test set after separate application of MICE, Borderline-SMOTE to the train set, and hyperparameter optimization of the LightGBM model are shown in the confusion matrix in Fig. 12. The model correctly identified all but eight (8) (FN = 8) instances where coronary heart disease developed in ten years (TP = 119), two (2) instances were incorrectly detected as developed CHD in ten years (FP = 2). The model detected 719 instances where CHD was not developed in ten years correctly (TN = 719). The evaluation results revealed that the proposed method achieved 99 % accuracy, 94 % recall, 98 % precision, 96 % F1 score, 95 % Mathews Correlation Coefficient (MCC).

The model achieved an area under the receiver operating curve of 0.9963 as shown in Fig. 13. This indicates that the model was able to distinguish between the two classes to a large extent.

Fig. 14 revealed that the model performed well on the training set as the model maintain a good accuracy as the training size increases. The model exhibited some instability at the initial stage. However, the validation accuracy increased progressively as the data size went above 2100. The learning curve plot indicates that the model has potential to capture complexity and relationships in the data better particularly with increased number of instances in the positive class.

The results of the baseline LightGBM model and the enhanced model with Borderline-SMOTE, TPE, and LightGBM presented in Table 4 indicate significant performance improvements. The enhanced model achieved better evaluation scores in MCC, F1 score, accuracy, sensitivity, precision, and AUC, demonstrating its superior predictive capabilities compared to the baseline model. The enhanced model's perfect scores indicate that it accurately classifies individuals at risk of developing coronary heart disease within the next ten years.

The performance of the model proposed in this study is compared with the results obtained by other state-of-the-art methods proposed in previous studies. It can be observed from Table 5 that the proposed model outperformed other models in previous studies interms of accuracy and precision. The F1-score and recall obtained are also comparable with what exist in previous studies. For an imbalanced data, the F1-score, MCC and AUC are appropriate for evaluation. However, majority of the previous studies did not report the MCC and AUC save for[47] which reported an MCC of 0.861 and AUC of 0.978.

Upon careful analysis of Table 5, it is evident that the results of this study exhibit significantly greater potential in comparison to the existing works. Several noteworthy aspects contribute to this finding. Firstly, the proposed method adopts a comprehensive approach by combining Borderline-SMOTE, Tree-structured Parzen, and LightGBM methodologies, effectively addressing the prevalent class imbalance challenge in

heart disease classification. Secondly, the utilization of the Multiple Imputation using Chained Equations (MICE) technique for handling missing values ensures robust imputation, resulting in a reliable dataset for analysis. This is exemplified by the significant improvement in the performance of the baseline model with MICE imputation as presented in Table 4. It is important to note that the baseline model with MICE has forty-two False Negative (Fig. 9) which the enhanced model was able to reduce to 8 (Fig. 12). The improvement achieved by the enhanced model is very significant, particularly in the diagnosis of life-threatening conditions such as coronary heart disease. The findings establish this study as a promising contribution to the field, with the potential to significantly advance heart disease diagnosis and prediction.

5. Conclusion

This study aimed to develop an efficient predictive model for detecting the ten-year risk of developing coronary heart disease (CHD). The analysis was performed using the Framingham Heart Study dataset, which presented challenges of missing values and class imbalance. To address these issues, the multivariate imputation by chained equation (MICE) technique was employed for handling missing data, while the borderline-SMOTE technique was applied to oversample the minority class and mitigate class imbalance. A 10-fold cross-validated LightGBM classifier, tuned with the Tree-structured Parzen Estimator (TPE), was trained on the over-sampled and separately imputed train and test datasets. Furthermore, to assess the impact of hyperparameter tuning and borderline-SMOTE on model performance, a 10-fold cross-validated baseline LightGBM model was constructed without these techniques.

The comparison of the baseline and enhanced models revealed that the enhanced model outperformed the baseline, indicating the positive impact of oversampling and hyperparameter tuning. These findings hold significant implications as accurate identification of individuals at high risk of CHD enables timely interventions and closer monitoring, ultimately contributing to improved disease management and prevention. In terms of practical implications, the developed model has significant potential in real-world healthcare settings. However, there is room for improvement.

For generalizability, these technique needs to be validated on more heart disease datasets. Therefore, future research could consider experimenting with additional datasets, risk factors, or biomarkers.

CRediT authorship contribution statement

Temidayo Oluwatosin Omotehinwa: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **David Opeoluwa Oyewola:** Writing – review & editing, Writing – original draft, Validation, Resources, Methodology, Investigation, Formal analysis, Data curation. **Ervin Gubin Moung:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Adel Mahmoud W, Aborizka M, Ahmed Elsayed Amer F. Heart disease prediction using machine learning and data mining techniques: application of framingham dataset. *Turk J Comput Math Educ (TURCOMAT)*. 2021;12(14):4864–4870. ⟨https://turcomat.org/index.php/turkbilmat/article/view/11445⟩.
2. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. *Proc ACM SIGKDD Int Conf Knowl Discov Data Min*. 2019:2623–2631. https://doi.org/10.1145/3292500.3330701.
3. Albert AJ, Murugan R, Sripriya T. Diagnosis of heart disease using oversampling methods and decision tree classifier in cardiology. *Res Biomed Eng*. 2023;39(1):99–113. https://doi.org/10.1007/s42600-022-00253-9.
4. Ambrews AB, Gubin Moung E, Farzamnia A, Yahya F, Omatu S, Angeline L. Ensemble Based Machine Learning Model for Heart Disease Prediction. *Int Conf Commun, Inf, Electron Energy Syst, CIEES 2022 - Proc*. 2022;2022. https://doi.org/10.1109/CIEES55704.2022.9990665.
5. Andersson C, Johnson AD, Benjamin EJ, Levy D, Vasan RS. 70-year legacy of the Framingham Heart Study. *Nat Rev Cardiol*. 2019;16(11):687–698. https://doi.org/10.1038/s41569-019-0202-5.
6. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: What is it and how does it work? *Int J Methods Psychiatr Res*. 2011;20(1):40–49. https://doi.org/10.1002/mpr.329.
7. Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for Hyper-Parameter Optimization. *24th Int Conf Neural Inf Process Syst*. 2011;24:2546–2554.
8. Beunza JJ, Puertas E, García-Ovejero E, Villalba G, Condes E, Koleva G, Hurtado C, Landecho MF. *Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease)*. In: *Journal of Biomedical Informatics*. Vol. 97. Academic Press,; 2019. https://doi.org/10.1016/j.jbi.2019.103257.
9. Bhutta AA, Nisa M un, Mian AN. Lightweight real-time WiFi-based intrusion detection system using LightGBM. *Wirel Netw*. 2023. https://doi.org/10.1007/s11276-023-03516-0.
10. Chen Y, Chang R, Guo J. Effects of Data Augmentation Method Borderline-SMOTE on Emotion Recognition of EEG Signals Based on Convolutional Neural Network. *IEEE Access*. 2021;9:47491–47502. https://doi.org/10.1109/ACCESS.2021.3068316.
11. De Hert M, Detraux J, Vancampfort D. The intriguing relationship between coronary heart disease and mental disorders. *Dialog- Clin Neurosci*. 2018;20(1):31–40. https://doi.org/10.31887/dcns.2018.20.1/mdehert.
12. Devi KN, Suruthi S, Shanthi S. Coronary artery disease prediction using machine learning techniques. *8th Int Conf Adv Comput Commun Syst, ICACCS 2022*. 2022:1029–1034. https://doi.org/10.1109/ICACCS54159.2022.9785140.
13. Ebiaredoh-Mienye SA, Esenogho E, Swart TG. Integrating enhanced sparse autoencoder-based artificial neural network technique and softmax regression for medical diagnosis. *Electron (Switz)*. 2020;9(11):1–13. https://doi.org/10.3390/electronics9111963.
14. Goguelin S, Dhokia V, Flynn JM. Bayesian optimisation of part orientation in additive manufacturing. *Int J Comput Integr Manuf*. 2021;34(12):1263–1284. https://doi.org/10.1080/0951192X.2021.1972466.
15. Gonsalves AH, Thabtah F, Mohammad RMA, Singh G. *Prediction of Coronary Heart Disease Using Machine Learning: An Experimental Analysis*. ACM International Conference Proceeding Series,; 2019:51–56. https://doi.org/10.1145/3342999.3343015.
16. Han H, Wang WY, Mao BH. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. *Lect Notes Comput Sci*. 2005;3644(PART I):878–887. https://doi.org/10.1007/11538059_91.
17. Hasan OS, Saleh IA. Development of heart attack prediction model based on ensemble learning. *East-Eur J Enterp Technol*. 2021;4(2(112):26–34. https://doi.org/10.15587/1729-4061.2021.238528.
18. Hassan CA ul, Iqbal J, Irfan R, Hussain S, Algarni AD, Bukhari SSH, Alturki N, Ullah SS. Effectively Predicting the Presence of Coronary Heart Disease Using Machine Learning Classifiers. *Sensors*. 2022;22(19):7227. https://doi.org/10.3390/s22197227.
19. Jebari-Benslaiman S, Galicia-García U, Larrea-Sebal A, Olaetxea JR, Alloza I, Vandenbroeck K, Benito-Vicente A, Martín C. Pathophysiology of Atherosclerosis. *Int J Mol Sci*. 2022;23(6). https://doi.org/10.3390/ijms23063346.
20. Kaggle FHS. (n.d.). Framingham heart study dataset | Kaggle. Retrieved April 1, 2023, from ⟨https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset⟩.
21. Kannel WB, McGee D, Gordon T. A general cardiovascular risk profile: The Framingham study. *Am J Cardiol*. 1976;38(1):46–51. https://doi.org/10.1016/0002-9149(76)90061-8.
22. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Adv Neural Inf Process Syst*. 2017;30. ⟨https://github.com/Microsoft/LightGBM⟩.
23. Khan MA, Hashim MJ, Mustafa H, Baniyas MY, Al Suwaidi SKBM, AlKatheeri R, Alblooshi FMK, Almatrooshi MEAH, Alzaabi MEH, Al Darmaki RS, Lootah SNAH. Global Epidemiology of Ischemic Heart Disease: Results from the Global Burden of Disease Study. *Cureus*. 2020;12(7), e9349. https://doi.org/10.7759/cureus.9349.
24. Kigka VI, Georga E, Tsakanikas V, Kyriakidis S, Tsompou P, Siogkas P, Michalis LK, Naka KK, Neglia D, Rocchiccioli S, Pelosi G, Fotiadis DI, Sakellarios A. Machine Learning Coronary Artery Disease Prediction Based on Imaging and Non-Imaging Data. *Diagnostics*. 2022;12(6). https://doi.org/10.3390/diagnostics12061466.
25. Kuruvilla AM, Balaji N. Heart disease prediction system using Correlation Based Feature Selection with Multilayer Perceptron approach. *IOP Conf Ser: Mater Sci Eng*. 2021;1085(1), 012028. https://doi.org/10.1088/1757-899x/1085/1/012028.
26. Latifah FA, Slamet I, Sugiyanto. Comparison of heart disease classification with logistic regression algorithm and random forest algorithm. *AIP Conf Proc*. 2020;2296(1). https://doi.org/10.1063/5.0030579.
27. Masih N, Naz H, Ahuja S. Multilayer perceptron based deep neural network for early detection of coronary heart disease. *Health Technol*. 2021;11(1):127–138. https://doi.org/10.1007/s12553-020-00509-3.

28. McMahan CA, Gidding SS, McGill HC. Coronary heart disease risk factors and atherosclerosis in young people. *J Clin Lipidol*. 2008;2(3):118–126. https://doi.org/10.1016/j.jacl.2008.02.006.
29. Miao L, Guo X, Abbas HT, Qaraqe KA, Abbasi QH. Using Machine Learning to Predict the Future Development of Disease. *2020 Int Conf UK-China Emerg Technol, UCET 2020*. 2020:1–4. https://doi.org/10.1109/UCET51115.2020.9205373.
30. Microsoft. (2016). *LightGBM. Microsoft Research*. ⟨https://www.microsoft.com/en-us/research/project/lightgbm/⟩.
31. Mienye ID, Sun Y. Improved Heart Disease Prediction Using Particle Swarm Optimization Based Stacked Sparse Autoencoder. *Electronics 2021*. 2021;10(19):2347. https://doi.org/10.3390/ELECTRONICS10192347.
32. Mienye ID, Sun Y, Wang Z. An improved ensemble learning approach for the prediction of heart disease risk. *Inform Med Unlocked*. 2020;20, 100402. https://doi.org/10.1016/j.imu.2020.100402.
33. Nalluri S, Vijaya Saraswathi R, Ramasubbareddy S, Govinda K, Swetha E. Chronic Heart Disease Prediction Using Data Mining Techniques. *Adv Intell Syst Comput*. 2020;1079:903–912. https://doi.org/10.1007/978-981-15-1097-7_76.
34. Omotehinwa TO, Oyewola DO. Hyperparameter Optimization of Ensemble Models for Spam Email Detection. *Appl Sci (Switz)*. 2023;13(3):1971. https://doi.org/10.3390/APP13031971.
35. Omotehinwa TO, Oyewola DO, Dada EG. A Light Gradient-Boosting Machine algorithm with Tree-Structured Parzen Estimator for breast cancer diagnosis. *Healthc Anal*. 2023;4, 100218. https://doi.org/10.1016/J.HEALTH.2023.100218.
36. Oyewola DO, Dada EG, Omotehinwa TO, Emebo O, Oluwagbemi OO. Application of deep learning techniques and bayesian optimization with tree parzen estimator in the classification of supply chain pricing datasets of health medications. *Appl Sci (Switz)*. 2022;12(19):10166. https://doi.org/10.3390/app121910166.
37. Özbilgin F, Kurnaz Ç, Aydın E. Prediction of coronary artery disease using machine learning techniques with iris analysis. *Diagnostics*. 2023;13(6):1081. https://doi.org/10.3390/diagnostics13061081.
38. Saurabh Pal RA. Elimination and Backward Selection of Features (P-Value Technique) In Prediction of Heart Disease by Using Machine Learning Algorithms. *Turk J Comput Math Educ (TURCOMAT)*. 2021;12(6):2650–2665. https://doi.org/10.17762/turcomat.v12i6.5765.
39. Shorewala V. Early detection of coronary heart disease using ensemble techniques. *Inform Med Unlocked*. 2021;26, 100655. https://doi.org/10.1016/j.imu.2021.100655.
40. Smiti S, Soui M. Bankruptcy prediction using deep learning approach based on borderline SMOTE. *Inf Syst Front*. 2020;22(5):1067–1083. https://doi.org/10.1007/s10796-020-10031-6.
41. Sun Y, Que H, Cai Q, Zhao J, Li J, Kong Z, Wang S. Borderline SMOTE Algorithm and Feature Selection-Based Network Anomalies Detection Strategy. *Energies*. 2022;15(13):4751. https://doi.org/10.3390/en15134751.
42. Turner R, Eriksson D, McCourt M, Kiili J, Laaksonen E, Xu Z, Guyon I. Bayesian Optimization is Superior to Random Search for Machine Learning Hyperparameter Tuning: Analysis of the Black-Box Optimization Challenge 2020. *NeurIPS 2020 Competition and Demonstration Track*. *Proc Mach Learn Res*. 2021;133:3–26. ⟨https://proceedings.mlr.press/v133/turner21a.html⟩.
43. WHO. (2021). Cardiovascular diseases (CVDs). World Health Organization. ⟨https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases⟩-(cvds).
44. Xi X. The role of LightGBM model in management efficiency enhancement of listed agricultural companies. *Appl Math Nonlinear Sci*. 2023. https://doi.org/10.2478/amns.2023.2.00386.
45. Xu L, Fu T, Wang Y, Ji N. Diagnostic value of peripheral blood miR-296 combined with vascular endothelial growth factor B on the degree of coronary artery stenosis in patients with coronary heart disease. *J Clin Ultrasound*. 2023. https://doi.org/10.1002/jcu.23433.
46. Yancy CW, Jessup M, Bozkurt B, Butler J, Casey DE, Drazner MH, Fonarow GC, Geraci SA, Horwich T, Januzzi JL, Johnson MR, Kasper EK, Levy WC, Masoudi FA, McBride PE, McMurray JJV, Mitchell JE, Peterson PN, Riegel B, Wilkoff BL. 2013 ACCF/AHA guideline for the management of heart failure: A report of the American college of cardiology foundation/american heart association task force on practice guidelines. *J Am Coll Cardiol*. 2013;62(16). https://doi.org/10.1016/j.jacc.2013.05.019.
47. Yang HZ, Chen Z, Yang H, Tian M. Predicting coronary heart disease using an improved LightGBM model: Performance analysis and comparison. *IEEE Access*. 2023;11:23366–23380. https://doi.org/10.1109/ACCESS.2023.3253885.
48. Yi T, Huang S, Li D, She Y, Tan K, Wang Y. The association of coronary non-calcified plaque loading based on coronary computed tomography angiogram and adverse cardiovascular events in patients with unstable coronary heart disease-a retrospective cohort study. *J Thorac Dis*. 2022;14(9):3438–3444. https://doi.org/10.21037/jtd-22-933.
49. Yilmaz R, Yağin FH. Early detection of coronary heart disease based on machine learning methods. *Med Rec*. 2021;4(1):1–6. https://doi.org/10.37990/medr.1011924.