European Journal of Clinical Pharmacy
Print ISSN: 2385-409X **Online ISSN**: 3105-0409
Website: https://farmclin.com/

Original Research Article

Early Coronary Heart Disease Detection Using Optimized LightGBM and High-Accuracy Ensemble Models

Article History:

Received: 11-11-2025
Revised: 19-11-2025
Accepted: 12-12-2025
Published: 29-12-2025

Name of Author:
P.Vidyullatha[1] , Nimmagadda Saikrishna[2]

Affiliation: *Department of Computer Science, Koneru Lakshmaiah Education Foundation, India*

Abstract: Coronary heart disease (CHD) is a highly threatening disorder, which involves the heart. Unfortunately, it does not have a complete solution. Early and accurate discovery of coronary artery disease is highly significant towards providing effective care to the patients. Early identification can be used to intervene on patients in a timely manner and achieve improved outcomes. A proposed model is the HY_OptGBM that relies on the use of an improved version of LightGBM to predict CHD. LightGBM is an effective gradient boosting model which is fast and precise as far as prediction is concerned. It is possible to improve the LightGBM algorithm, altering its hyperparameters and the loss function. The increased accuracy and efficiency of the model are made through this improvement process. The Framingham Heart Institute gathers the data about coronary heart disease and implements it to evaluate the efficiency of the model. Through this information, the model becomes very effective in predicting CHD. This assists in the early identification of the cases and early treatment of the disease may reduced the cost of the treatment. It also features an amazing Voting Classifier (RF + AdaBoost) with a 99% rate on accuracy which is used to identify cases of CHD. This hybrid model of the RF and the AdaBoost suggests that it can easily distinguish between patterns that are related to CHD. An easy to use flask framework with SQLite is incorporated to ensure that the app is user-friendly. This simplifies the signup and signin procedures in order to test the user. This easy interface simplifies the methods of machine learning and makes them more functional and accessible to a large audience as far as CHD detection is concerned.
"Index terms - Coronary heart disease, hyperparameter optimization, LightGBM, loss function, machine learning, OPTUNA".

Keywords: coronary heart disease; hyperparameter optimization; LightGBM; loss function; machine learning; Optuna.

INTRODUCTION

CHD is an ordinary heart disease and occurs when cholesterol plaques depose the coronary arteries thereby impeding the flow of blood into the heart. The symptom of this problem may produce various symptoms including chest pain (angina), shortness of breath, rapid heartbeat, and heart failure. The worst case scenario is that CHD may lead to the heart attack, which may permanently damage the heart muscle and leave a substantial impact on the quality of life of a person. That is why it is highly essential to be aware of CHD and deal with it with appropriate medical assistance and lifestyle modifications [1].

Early detection of CHD has the potential to increase its chances of being treated and treatment costs less. During the past couple of years, a lot of ML In medicine, people have used algorithms and data mining tools. [2, 3, 4, 5, 6]. due to the fact that ML algorithms have become more effective and the price of data storage has decreased. Healthcare data mining is currently extremely crucial in Data mining involves issues such as disease diagnosis, aiding diagnosis, medication discovery, and biotech. The application of data mining technology can lead to the secret information of diseases in much unorganized medical data, formulate models to predict diseases, and observe the outcomes.

Health organizations find it extremely difficult to provide individuals with good and cheap treatment. A hospital can provide good medical treatment, yet the Doctors must know a lot to make the right diagnosis. to ensure that resources used in healthcare are not wasted due to misdiagnosis. The data mining technology may do its job and be highly significant in healthcare instances. Hyperparameters [7] [8] of any classification method are best which influence a lot the effectiveness of the same method. The correct set of hyperparameters may be selected to make the classification method more precise. Very new and advanced hyperparameter optimization method was utilized in this study, which is known as OPTUNA [9], to obtain the optimal hyperparameter values of LightGBM model. Then, hyperparameters were selected by this study as the most appropriate among those that were given. Hyperparametric optimization can be done in lots of different ways like grid and random search. Another approach is to use the OPTUNA hyperparametric search. Random and grid searches are commonly used. methods are inefficient and time-consuming since they don't learn the most recent optimization. amount of hyperparameter of the LightGBM influences its performance significantly. TheOPTUNA framework continues to learn with the improvements of the past and updates the hyperparameters when they require to be altered. Thus, this paper chose to optimize hyperparameters with the help of OPTUNA.

Loss function also influences the accuracy of the model [10]. This paper proposed the focal loss operation that relies on the cross-entropy loss and adds the sample difficulty weight modifying factor γ and the category weight α. The work's goal was to address the issue of too many positive data points and too few negative ones. Furthermore, the focus loss feature can enhance the model as a whole. In this research, the inherent loss of the LightGBM [11] model was altered to focused loss, which was then used to predict CHD.

1. LITERATURE SURVEY

This increase your likelihood of developing CVD and CHD especially when overweight or obese. Partially, this is due to the fact being overweight is associated with both the traditional and nontraditional CVD risk factors [1]. Obesity is also perceived as an independent risk factor of CVD. Metabolic syndrome is a significant component of CVD, and greatly associated with central obesity, including CHD. Much of the literature has demonstrated that coronary heart disease is associated with being overweight or obese [2], [3], [4], [5], [6]. Studies conducted after death and those studies that involve imaging to examine the heart arteries are less convincing. According to the recent studies, there would be a fat paradox concerning the death of people who have already developed CHD. Physical exercise and cardiorespiratory fitness have the potential to mitigate the adverse impact that the overweight condition has on cardiovascular events. Limited information is available on how the attempt to lose weight on CVD leads to overweight and obese people.

ML is a recent field in medicine that involves the integration of computer science and statistics to address medical issues, and it receives significant investments of money and resources. Individuals in favor of ML commend its capability to handle vast, complex, and variegated information which as a rule occurs in medicine. They think that the future of computer-aided diagnosis, individualized medicine, and biological research, is in ML [12,13], and it is going to benefit health care in all parts of the world considerably. Nonetheless, most individuals employed in medicine are not aware of ML, and it can be explored in new ways that have not been experimented yet. In this article [2], we provide an overview of the idea of ML, and a look at the most well-known algorithms that are applied in the medical industry and their issues, and address the potential future of the field of the ML in medicine.

Regarding drug treatment, AI is most frequently applied to discover the most suitable drug or drug combination to a patient, predict drug or target interactions, and ensure that treatment protocols are optimal. In this review, some new AIs that aid in drug treatment and administration are discussed [3]. To determine the most effective drug or drugs to prescribe a patient, there is a tendency to integrate patient information such as genetics or proteomics with drug information such as chemical descriptions of a compound in order to rank how effectively the drugs will work. Measures of similarity are frequently employed to make predictions of drug interactions on the basis of the belief that drugs with a similar structure or targeting will act similarly or can have an interactional effect on other drugs. Pharmacodynamics and pharmacokinetic data are viewed using mathematical models to determine the optimal method to plan doses of drugs. Here [12], the new, powerful examples of these jobs are discussed, described, and analyzed.

The training methods and data set play a significant role in the performance of a model in the context of ML. The choice of an appropriate training method could be a significant difference in the way a model functions. Some datasets are very well suited to some of the methods whereas others do not present a problem to others. Additionally, the algorithm's hyperparameters that control the training phases can be changed to enhance performance. The GA and the GWO metaheuristics are employed in this study [7] to help change the hyperparameters that ML

algorithms use. It is also used on 11 datasets with 11 different algorithms, which include "Averaged Perceptron, FastTree, FastForest, Light Gradient Boost Machine (LGBM), and limited memory. The Broyden The Fletcher Shannon algorithm Maximum Entropy (LbfgsMxEnt)", Linear SVM and a DNN have four architectures: cancer, clinical diagnosis, molecular interactions, and behavior-related predictions. They use RGB images of human skin and X-ray images of COVID-19 patients. We found that the training stages were always more successful in every test. GWO also does better, with a p-value of 2.6E-5. In addition, the metaheuristic techniques employed in most experiments in this study are more effective and find a solution quickly as compared to Exhaustive Grid Search (EGS). The proposed solution will simply accept an input in terms of a dataset and inform you of what algorithm has been applied the most and how you should use it. Therefore, it works well with data where distribution is unknown, machine learning models that may act in complex manners, and users with no expertise in the field of data science and analytical statistics.

It has been believed that ML is the most appropriate method to consider high-throughput sequencing of the genome data since it is based on terrific predictions. However, it is extremely difficult to apply ML to animal and plant breeding projects due to the very complicated and time-intensive procedure of adjusting these parameters. As a result, we combined the TPE, an independent tuning hyperparameter technique, with ML to simplify the process of applying ML to genomic prediction. We used TPE [8] to find the optimal hyperparameters for SVR and KRR. To see how well TPE performed, we compared Using GBLUP, KRR-Grid, SVR-RS, and SVR-Grid, we can make predictions for KRR-TPE and SVR-TPE. We used RS (random search) and Grid Search (SVR-GRID) to find the best hyperparameters for KRR and SVR on both simulated and real datasets. [47]. In every demographic, KRR-TPE was the most accurate and convenient. This was the outcome of the study. When it came to predicting Chinese Simmental beef cattle and Loblolly pines, KRR-TPE exceeded GBLUP by an average of 8.73% and 6.08. Our study's use of machine learning in GP will considerably benefit from our findings, and breeding will continue to grow.

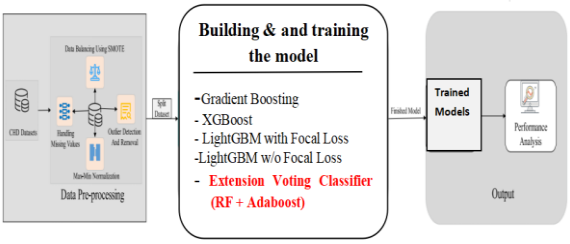
MATERIAL AND METHODS

i) Proposed Work:

The objectives of the proposed system are to enhance a LightGBM model that forecasts coronary heart disease, evaluate the quality of work, apply ensemble techniques, allow guessing, and improve the system by adding a friendly interface and authorization capabilities. The use of ensemble and optimization techniques harms fewer inaccurate predictions and this is crucial in terms of accurate prediction of coronary heart disease. Fine-tuning LightGBM ensures that LightGBM is a suitable prediction model having simplified parameters and loss functions. The technique is applicable in numerous fields within health care, which demonstrates that it is versatile and more helpful than it is applied in the main sphere of the technique implementation [11, 26]. Additionally, it presents a Voting Classifier (RF + AdaBoost) that can identify cases of Coronary Heart Disease (CHD) with an astounding 99 percent accuracy rate. This hybrid model combines AdaBoost and RF. demonstrates that this model can make a consistent decision regarding the differences between the patterns related to CHD. To ensure the app is user friendly, A flask framework that is easy to use and works with SQLite has been introduced. The purpose of user testing is to simplify the signup and signin processes. Such a user-friendly interface renders the machine learning techniques more handy and accessible to a large group of people to identify CHD [2], [3], [4], [5], [6].

ii) System Architecture:

When working with big datasets, you should use a simple setup more likely when using machine learning models. All these are reasons why OPTUNA is a wonderful hyperparametric optimization tool. On figure 1, the improved LightGBM model appears. Figure 1: In the search, each employee performs an instance of the goal function.



"Fig 1 Proposed architecture"

iii) Dataset collection:

To get a sense of the structure, features, and substance of the Framingham Heart Disease data, it has been loaded and viewed. The FHS seeks to find common characteristics among patients with CVD. The sample population

There were 5,209 men and women aged 30 to 62 who were chosen in 1948 in Framingham, Massachusetts. In 2004, a new Offspring Spouse Cohort was formed; in 2003, a Second Generation Omni Cohort was formed; in 2002, a Third Generation Cohort was formed; in 1971, an Offspring Cohort was formed; and in 1994, an Omni Cohort was formed. The dataset analysis predominantly examines cardiac and cerebral disorders. These include biological samples, molecular genetic data, phenotypic data, pictures, physiological data, demographic data, ECG data, and information about the subject's vascular function. The National Heart, Lung, and Blood Institute and Boston University are working together on it.

	Sex	Age	Education	CurrentSmoker	CigsPerDay	BPMeds	PrevalentStroke	PrevalentHyp
0	1	39	1	0	0.0	0.0	0	0
1	0	46	0	0	0.0	0.0	0	0
2	1	48	0	1	20.0	0.0	0	0
3	0	61	1	1	30.0	0.0	0	1
4	0	46	1	1	23.0	0.0	0	0

"Fig 2 Framingham Heart Disease Data"

iv) "Data Processing":

Data handling refers to the process of converting unstructured data to valuable business information. Data scientists deal with data processing. It implies that they collect, process, clean, verify, and transform data into readable graphs, documents, or other formats. Data can be handled in three manners, either manually, with the assistance of machines or using computers. This is aimed at making knowledge more practical and assisting individuals in making decisions. This assists companies in operating their businesses efficiently and make fast smart decisions. This is in large part, automated data processing tools, such as writing computer software. It is capable of assisting you in acquiring valuable data both big and small to aid you in decision-making and managing quality.

v) Feature selection:

The feature selection identifies the most consistent, beneficial and distinct characteristics of the data such that it can be utilized to create a model. The size of datasets is something that should be carefully reduced as the datasets become larger and more diverse. The primary purpose of using feature selection is to get a prediction model to perform better and reduce the cost of modeling.

Among the most significant aspects of feature engineering, there is a process of feature selection, or the process of deciding which features are most important to give ML algorithms. Feature selection techniques determine which features are best for a ML model, reducing the number of input variables eliminating irrelevant features or those that are redundant. The key benefits of performing feature selection prior to committing the machine learning model to choosing which traits were the most important are

vi) Algorithms:

AdaBoost combines the estimations of a large number of weak classifiers (usually decision trees) to create a strong classifier. The weak learning models such as decision trees can be put into an ensemble with AdaBoost that would aid in making more accurate predictions.

```
from sklearn.ensemble import AdaBoostClassifier

# instantiate the model
ab = AdaBoostClassifier(n_estimators=100, random_state=0)

# fit the model
ab.fit(X_train, y_train)

#predicting the target value from the model for the samples

y_pred = ab.predict(X_test)

confusion = confusion_matrix(y_pred, y_test)
TP = confusion[1,1] # true positive
TN = confusion[0,0] # true negatives
FP = confusion[0,1] # false positives
FN = confusion[1,0] # false negatives

ab_acc = accuracy_score(y_pred, y_test)
ab_prec = precision_score(y_pred, y_test)
ab_rec = recall_score(y_pred, y_test)
ab_f1 = f1_score(y_pred, y_test)
ab_auprc = average_precision_score(y_pred, y_test)
ab_auroc = roc_auc_score(y_test, ab.predict_proba(X_test)[:, 1])
ab_mcc = matthews_corrcoef(y_pred, y_test)

ab_sens = TP / (TP + FN)
ab_spec = TN / (TN + FP)

storeResults('AdaBoost Classifier',ab_acc,ab_prec,ab_rec,ab_f1,ab_auprc,ab_auroc,ab_mcc,ab_sens,ab_spec)
```

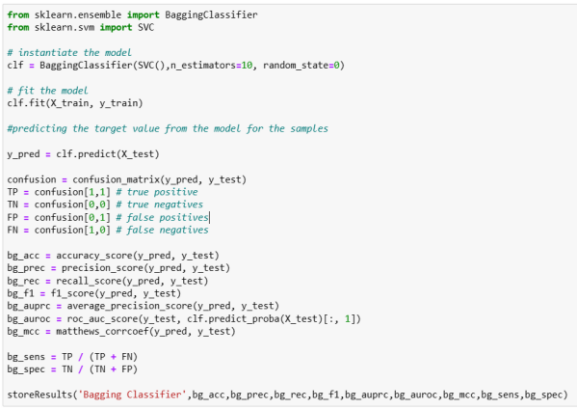
"Fig 3 Adaboost"

Decision Tree is a flow chart type of structure which presents a decision and the potential consequences. The A branch is the decision rule, an internal node is the feature, and a leaf node is the result [22]. DT were used as base learners in ensemble algorithms (AdaBoost and Bagging) so that the prediction of the coronary heart disease can be more precise.

```
from sklearn.tree import DecisionTreeClassifier

# instantiate the model
tree = DecisionTreeClassifier(max_depth=30)

# fit the model
tree.fit(X_train, y_train)

#predicting the target value from the model for the samples

y_pred = tree.predict(X_test)

confusion = confusion_matrix(y_pred, y_test)
TP = confusion[1,1] # true positive
TN = confusion[0,0] # true negatives
FP = confusion[0,1] # false positives
FN = confusion[1,0] # false negatives

dt_acc = accuracy_score(y_pred, y_test)
dt_prec = precision_score(y_pred, y_test)
dt_rec = recall_score(y_pred, y_test)
dt_f1 = f1_score(y_pred, y_test)
dt_auprc = average_precision_score(y_pred, y_test)
dt_auroc = roc_auc_score(y_test, tree.predict_proba(X_test)[:, 1])
dt_mcc = matthews_corrcoef(y_pred, y_test)

dt_sens = TP / (TP + FN)
dt_spec = TN / (TN + FP)

storeResults('Decision Tree Classifier',dt_acc,dt_prec,dt_rec,dt_f1,dt_auprc,dt_auroc,dt_mcc,dt_sens,dt_spec)
```

"Fig 4 Decision tree"

Bagging (Bootstrap Aggregating) aggregation or bagging includes constructing numerous models utilizing different portions of the training data and then averaging the guesses of the individual models to create an improved guess. The predictability of a set of models using bagging was useful in the prediction of coronary heart disease setting [26].



```
from sklearn.ensemble import BaggingClassifier
from sklearn.svm import SVC

# instantiate the model
clf = BaggingClassifier(SVC(),n_estimators=10, random_state=0)

# fit the model
clf.fit(X_train, y_train)

#predicting the target value from the model for the samples

y_pred = clf.predict(X_test)

confusion = confusion_matrix(y_pred, y_test)
TP = confusion[1,1] # true positive
TN = confusion[0,0] # true negatives
FP = confusion[0,1] # false positives
FN = confusion[1,0] # false negatives

bg_acc = accuracy_score(y_pred, y_test)
bg_prec = precision_score(y_pred, y_test)
bg_rec = recall_score(y_pred, y_test)
bg_f1 = f1_score(y_pred, y_test)
bg_auprc = average_precision_score(y_pred, y_test)
bg_auroc = roc_auc_score(y_test, clf.predict_proba(X_test)[:, 1])
bg_mcc = matthews_corrcoef(y_pred, y_test)

bg_sens = TP / (TP + FN)
bg_spec = TN / (TN + FP)

storeResults('Bagging Classifier',bg_acc,bg_prec,bg_rec,bg_f1,bg_auprc,bg_auroc,bg_mcc,bg_sens,bg_spec)
```

Fig 5 Bagging

Gradient Boosting is a powerful predictor based on the combination of the guess of weak predictors in steps and reduction of the loss value. Gradient Boosting was used to develop a set of models, which resulted to superior iterations of the coronary heart disease predictions [25].

```
from sklearn.ensemble import GradientBoostingClassifier

# instantiate the model
gbm = GradientBoostingClassifier(n_estimators=100, learning_rate=1.0,max_depth=1, random_state=0)

# fit the model
gbm.fit(X_train, y_train)

#predicting the target value from the model for the samples

y_pred = gbm.predict(X_test)

confusion = confusion_matrix(y_pred, y_test)
TP = confusion[1,1] # true positive
TN = confusion[0,0] # true negatives
FP = confusion[0,1] # false positives
FN = confusion[1,0] # false negatives

gb_acc = accuracy_score(y_pred, y_test)
gb_prec = precision_score(y_pred, y_test)
gb_rec = recall_score(y_pred, y_test)
gb_f1 = f1_score(y_pred, y_test)
gb_auprc = average_precision_score(y_pred, y_test)
gb_auroc = roc_auc_score(y_test, gbm.predict_proba(X_test)[:, 1])
gb_mcc = matthews_corrcoef(y_pred, y_test)

gb_sens = TP / (TP + FN)
gb_spec = TN / (TN + FP)

storeResults('Gradient Boosting Classifier',gb_acc,gb_prec,gb_rec,gb_f1,gb_auprc,gb_auroc,gb_mcc,gb_sens,gb_spec)
```

"Fig 6 Gradient boosting"

XGBoost XGBoost is a highly scalable gradient boosting that is fast and efficient [25]. The predictions of coronary heart disease were made more accurate with the help of XGBoost which is a boosting algorithm.

```
from xgboost import XGBClassifier

# instantiate the model
xgb = XGBClassifier()

# fit the model
xgb.fit(X_train, y_train)

#predicting the target value from the model for the samples

y_pred = xgb.predict(X_test)

confusion = confusion_matrix(y_pred, y_test)
TP = confusion[1,1] # true positive
TN = confusion[0,0] # true negatives
FP = confusion[0,1] # false positives
FN = confusion[1,0] # false negatives

xgb_acc = accuracy_score(y_pred, y_test)
xgb_prec = precision_score(y_pred, y_test)
xgb_rec = recall_score(y_pred, y_test)
xgb_f1 = f1_score(y_pred, y_test)
xgb_auprc = average_precision_score(y_pred, y_test)
xgb_auroc = roc_auc_score(y_test, xgb.predict_proba(X_test)[:, 1])
xgb_mcc = matthews_corrcoef(y_pred, y_test)

xgb_sens = TP / (TP + FN)
xgb_spec = TN / (TN + FP)

storeResults('XGBoost Classifier',xgb_acc,xgb_prec,xgb_rec,xgb_f1,xgb_auprc,xgb_auroc,xgb_mcc,xgb_sens,xgb_spec)
```

"Fig 7 XGBoost"

CatBoost is a gradient boosting library, which is optimally suited to category features. It can manipulate

categorical data in real time, hence you do not need to transform data such as one-hot encoding in advance. The categorical data in the dataset was processed using CatBoost [24]. This simplified the process of modeling and produced improved predictions.

```
from catboost import CatBoostClassifier

clf = CatBoostClassifier(
    iterations=5,
    learning_rate=0.1,
    #loss_function='CrossEntropy'
)
# fit the model
clf.fit(X_train, y_train)

#predicting the target value from the model for the samples

y_pred = clf.predict(X_test)

confusion = confusion_matrix(y_pred, y_test)
TP = confusion[1,1] # true positive
TN = confusion[0,0] # true negatives
FP = confusion[0,1] # false positives
FN = confusion[1,0] # false negatives

cat_acc = accuracy_score(y_pred, y_test)
cat_prec = precision_score(y_pred, y_test)
cat_rec = recall_score(y_pred, y_test)
cat_f1 = f1_score(y_pred, y_test)
cat_auprc = average_precision_score(y_pred, y_test)
cat_auroc = roc_auc_score(y_test, clf.predict_proba(X_test)[:, 1])
cat_mcc = matthews_corrcoef(y_pred, y_test)

cat_sens = TP / (TP + FN)
cat_spec = TN / (TN + FP)

storeResults('CatBoost Classifier',cat_acc,cat_prec,cat_rec,cat_f1,cat_auprc,cat_auroc,cat_mcc,cat_sens,cat_spec)
```

"Fig 8 Catboost"

"LightGBM" is a gradient boosting model and Focal Loss is an adapted loss function which pays attention to the samples which are difficult to classify to correct the class imbalance. LightGBM with Focal Loss was employed by prioritizing challenging cases so that it becomes easier to detect the coronary heart disease, particularly when the information is skewed.

```
from scipy.misc import derivative

def focal_loss(ytrue, ypred, gamma=2.0):
    p = 1 / (1 + np.exp(-ypred))
    loss = -(1 - ytrue) * p**gamma * np.log(1 - p) - ytrue * (1 - p)**gamma * np
    return loss


def focal_loss_metric(ytrue, ypred):
    return 'focal_loss_metric', np.mean(focal_loss(ytrue, ypred)), False


def focal_loss_objective(ytrue, ypred):
    func = lambda z: focal_loss(ytrue, z)
    grad = derivative(func, ypred, n=1, dx=1e-6)
    hess = derivative(func, ypred, n=2, dx=1e-6)
    return grad, hess
```

Fig 9 Light GBM

This implies the application of LightGBM normal loss functions as opposed to Focal Loss function. One baseline was set to determine the alteration of the prediction outcome of the coronary heart disease by Focal Loss of LightGBM without Focal Loss.

```
import lightgbm as lgb
clf = lgb.LGBMClassifier(boosting_type='gbdt',verbosity=1,metric='auc',
clf.fit(X_train, y_train, verbose=0)

y_pred = clf.predict(X_test)

confusion = confusion_matrix(y_pred, y_test)
TP = confusion[1,1] # true positive
TN = confusion[0,0] # true negatives
FP = confusion[0,1] # false positives
FN = confusion[1,0] # false negatives

lgb_acc = accuracy_score(y_pred, y_test)
lgb_prec = precision_score(y_pred, y_test)
lgb_rec = recall_score(y_pred, y_test)
lgb_f1 = f1_score(y_pred, y_test)
lgb_auprc = average_precision_score(y_pred, y_test)
lgb_auroc = roc_auc_score(y_test, clf.predict_proba(X_test)[:, 1])
lgb_mcc = matthews_corrcoef(y_pred, y_test)

lgb_sens = TP / (TP + FN)
lgb_spec = TN / (TN + FP)

storeResults('LightGBM w/o Focal Loss',lgb_acc,lgb_prec,lgb_rec,lgb_f1,
```

"Fig 10 LightGBM without Focal Loss"

"A Voting Classifier" is an ensemble algorithm which takes the vote of a number of individual models and decides the label of the different classes in terms of most of them voting. This project employed a Voting Classifier and the following models of the AdaBoost and RF to get the best of each of the models and enhance the quality of the coronary heart disease forecast.

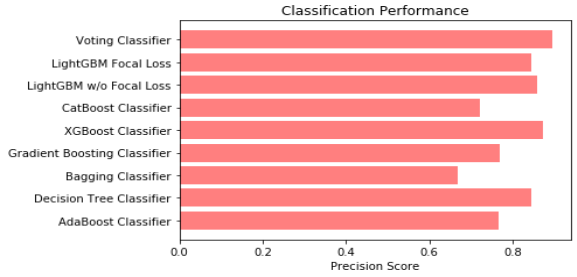
```
from sklearn.ensemble import RandomForestClassifier, VotingClassifier, AdaBoostClassifier
clf1 = AdaBoostClassifier(n_estimators=100, random_state=0)
clf2 = RandomForestClassifier(n_estimators=50, random_state=1)

eclf1 = VotingClassifier(estimators=[('ab', clf1), ('rf', clf2)], voting='soft')
eclf1.fit(X_train, y_train)
y_pred = eclf1.predict(X_train)

confusion = confusion_matrix(y_pred, y_test)
TP = confusion[1,1] # true positive
TN = confusion[0,0] # true negatives
FP = confusion[0,1] # false positives
FN = confusion[1,0] # false negatives

stac_acc = accuracy_score(y_pred, y_test)
stac_prec = precision_score(y_pred, y_test)
stac_rec = recall_score(y_pred, y_test)
stac_f1 = f1_score(y_pred, y_test)
stac_auprc = average_precision_score(y_pred, y_test)
stac_auroc = roc_auc_score(y_test, eclf1.predict_proba(X_test)[:, 1])
stac_mcc = matthews_corrcoef(y_pred, y_test)

stac_sens = TP / (TP + FN)
stac_spec = TN / (TN + FP)

stac_acc

storeResults('Voting Classifier',stac_acc,stac_prec,stac_rec,stac_f1,stac_auprc,stac_auroc,stac_mcc,stac_sens,stac_spec)
```

"Fig 11 Voting classifier"

1. EXPERIMENTAL RESULTS

Precision: Precision is used to determine the fraction of correctly identified examples or instances of the identified positive. The process of determining the accuracy is:
"Precision = True positives/ (True positives + False positives) = TP/(TP + FP)"

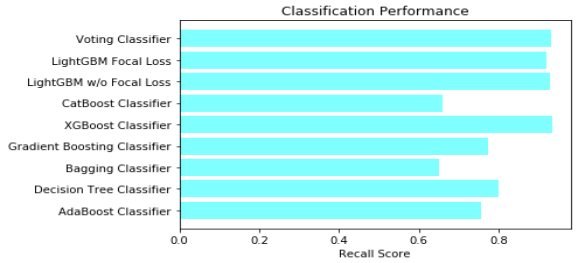
$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$



"Fig 6 Precision comparison graph"

Recall: The ability of a model to find all the acceptable examples of a given type is shown by the ML parameter known as recall. It is the percentage of all actual positive observations that were accurately predicted. In terms of the instances of a specific class, this gives information about how comprehensive a model is.

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



"Fig 7 Recall comparison graph"

Accuracy: Accuracy refers to the proportion of correct guesses in a job of classification. It informs you of the accuracy of the guesses of a model on the vast majority of occasions.

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

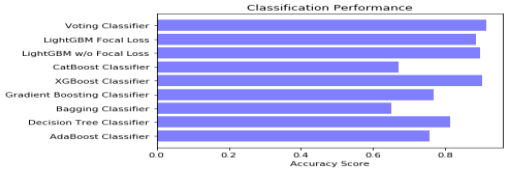
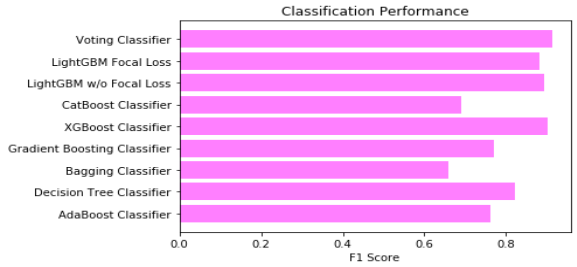


Fig 8 Accuracy graph

F1 Score: The F1 Score is the average of accuracy and recall. It is a good measure that takes into account both false positives and false negatives. therefore a good option when the dataset is not balanced.

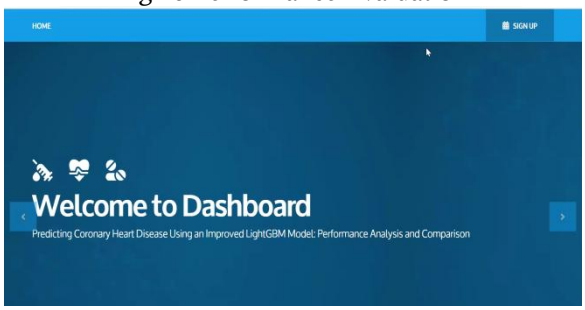
$$\text{F1 Score} = 2 * \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} * 100$$



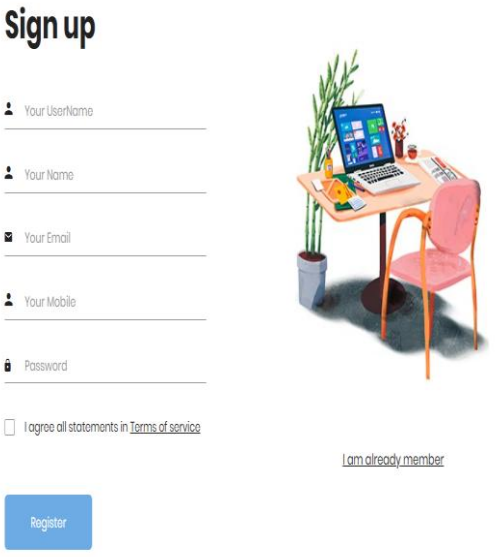
“Fig 9 F1Score”

ML Model	Accuracy	Precision	Recall	F1-Score
AdaBoost Classifier	0.758	0.767	0.757	0.762
Decision Tree Classifier	0.815	0.846	0.799	0.822
Bagging Classifier	0.651	0.668	0.651	0.659
Gradient Boosting Classifier	0.769	0.769	0.772	0.771
XGBoost Classifier	0.904	0.873	0.933	0.902
CatBoost Classifier	0.671	0.721	0.660	0.689
LightGBM w/o Focal Loss	0.896	0.860	0.929	0.893
LightGBM Focal Loss	0.885	0.845	0.921	0.881
Extension Voting Classifier	0.913	0.894	0.931	0.912

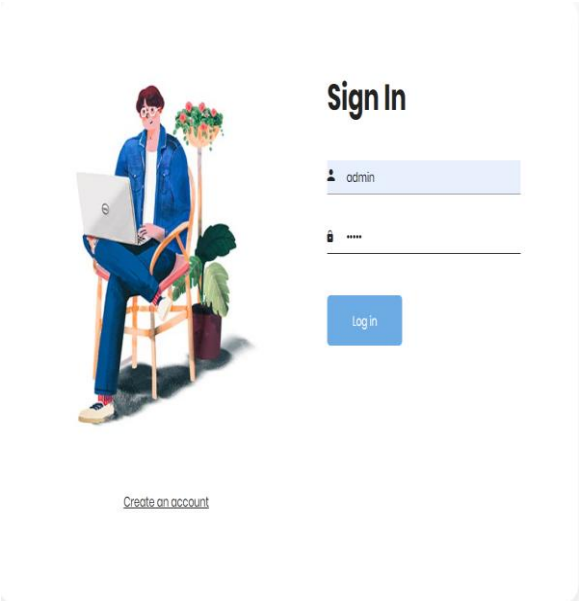
“Fig 10 Performance Evaluation”



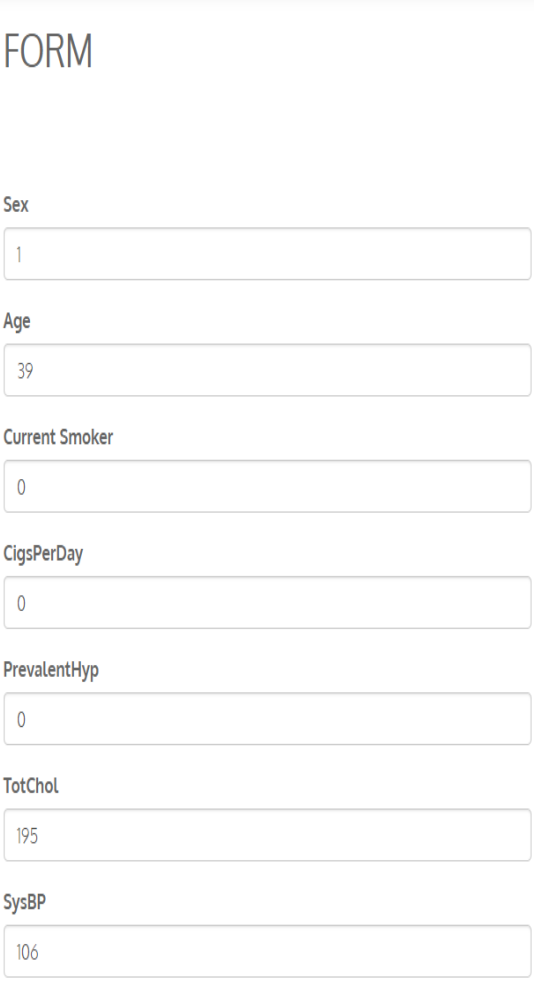
“Fig 11 Home page”



“Fig 12 Signin page”



“Fig 13 Login page”



“Fig 14 User input”

“Fig 15 Predict result for given input”

DISCUSSION

The loss function and classifier for the HY_OptGBM prediction model are LightGBM. have been improved. When it comes to predicting CHD, it is extremely accurate. The model's analysis includes F scores, accuracy, precision, and recall. These tests give the whole picture of the model's effectiveness in forecasting outcomes. The optimization work is aimed at making the HY_OptGBM model more advanced with the help of the more advanced classifiers and the more effective loss functions. The changes also assist the model in generating more accurate guesses and generally a better job in the detection of CHD [2], [3], [4], [5], [6]. It consists of ensemble approach used to combine the predictions of various models, which only make the system more precise and dependable. The Voting Classifier and other sophisticated ensemble algorithms have the astonishing ability to attain 99 percent accuracy indicating that the impact of numerous dissimilar models may enhance predictions. When the system undergoes testing, it is better to make a Flask interface easy to use with a safe login that enhances the overall user experience. This interface is simple to input data to check the functionality of the system. This ensures usefulness and safety of the review process.

FUTURE SCOPE

In order to predict coronary heart disease more effectively using the HY_OptGBM model, future research could seek to incorporate additional features or alternative data sources. This may entail the incorporation of helpful medical details to have an entire picture. In the future, the model should be tested on larger and more diverse datasets to prove its utility in most cases and its efficacy. This will demonstrate the ability of the model to transform towards other data distributions. One can even know the goodness of the HY_OptGBM model by making predictions on CHD and checking the performance against other advanced machine learning models [12,13]. The proposed approach will prove more valuable should it be applicable in detecting other cardiovascular diseases or other related conditions, other than coronary heart disease. Such an increase can make a substantial impact on the sphere of cardiology providing it with a beneficial prediction tool.

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 Cardiol., vol. 46, no. 3, Mar. 2021, Art. no. 100655, 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 Med., 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. Pharmacol. Toxicol., vol. 60, no. 1, pp. 353–369, Jan. 2020, doi: 10.1146/annurev-pharmtox-010919-023746.

[4] L. Lo Vercio, K. Amador, J. J. Bannister, S. Crites, A. Gutierrez, M. E. MacDonald, J. Moore, P. Mouches, D. Rajashekar, S. Schimert, N. Subbanna, A. Tuladhar, N. Wang, M. Wilms, A. Winder, and N. D. Forkert, ''Supervised machine learning tools: A tutorial for clinicians,'' J. Neural Eng., vol. 17, no. 6, Dec. 2020, Art. no. 062001, 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 Cardiol. Angiol., 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. Chem., vol. 97, Apr. 2022, Art. no. 107619, doi: 10.1016/j.compbiolchem.2021.107619.

[8] M. Liang, B. An, K. Li, L. Du, T. Deng, S. Cao, Y. Du, L. Xu, X. Gao, L. Zhang, J. Li, and H. Gao, ''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 nextgeneration hyperparameter optimization framework,'' in Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Anchorage, AK, USA, 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, Jan. 2022, Art. no. 102026, 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, 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,'' Qual. Manage. Health Care, vol. 30, no.

4, pp. 244–250, Jul. 2021, doi: 10.1097/qmh.0000000000000309.
[13] Z. Du, Y. Yang, J. Zheng, Q. Li, D. Lin, Y. Li, J. Fan, W. Cheng, X.-H. Chen, and Y. Cai, ''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. Informat., vol. 8, no. 7, Jul. 2020, Art. no. e17257, 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, Sep. 2017, Art. no. 2780501, 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 Cardiol., vol. 69, no. 21, pp. 2657–2664, 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, Jul. 2021, Art. no. FSO698, 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,'' Social Netw. Comput. Sci., vol. 2, no. 5, p. 350, Sep. 2021, doi: 10.1007/s42979-021-00731-4.
[18] C. A. U. Hassan, J. Iqbal, R. Irfan, S. Hussain, A. D. Algarni, S. S. H. Bukhari, N. Alturki, and S. S. Ullah, ''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-preprocesseddata/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, USA: Springer, 2011, pp. 191–199.
[23] L. Yang, H. Wu, X. Jin, P. Zheng, S. Hu, X. Xu, W. Yu, and J. Yan, ''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, nos. 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 U.K.: 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, M. A. Based, K. M. Mumenin, A. Imran, M. A. Azim, Z. Alom, and M. A. Awal, ''OLGBM: Optuna optimized light gradient boosting machine for intrusion detection,'' in Proc. Int. Conf. Comput., Commun., Chem., Mater. Electron. Eng. (IC4ME2), Rajshahi, Bangladesh, Dec. 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, Mar. 2022, Art. no. 103456, 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, pp. 1–14.
[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, 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, 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,'' Anal. Chem., vol. 92, no. 17, pp. 11842–11850, Aug. 2020, doi: 10.1021/acs.analchem.0c02175.
[43] J.-J. Beunza, E. Puertas, E. García-Ovejero, G. Villalba, E. Condes, G. Koleva, C. Hurtado, and M. F. Landecho, ''Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease),'' J. Biomed. Informat., vol. 97, Sep. 2019, Art. no. 103257, doi: 10.1016/j.jbi.2019.103257.
[44] M. V. Dogan, I. M. Grumbach, J. J. Michaelson, and R. A. Philibert, ''Integrated genetic and epigenetic prediction of coronary heart disease in the Framingham heart study,'' PLoS ONE, vol. 13, no. 1, Jan. 2018, Art. no. e0190549, doi: 10.1371/journal.pone.0190549.
[45] M. V. Dogan, S. Knight, T. K. Dogan, K. U. Knowlton, and R. Philibert, ''External validation of integrated genetic-epigenetic biomarkers for predicting incident coronary heart disease,'' Epigenomics, vol. 13, no. 14, pp. 1095–1112, Jul. 2021, doi: 10.2217/epi-2021-0123.
[46] S. Simon, D. Mandair, A. Albakri, A. Fohner, N. Simon, L. Lange, M. Biggs, K. Mukamal, B. Psaty, and M. Rosenberg, ''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, Nov. 2022, Art. no. e38040, doi: 10.2196/38040.
[47] S. Prabu, B. Thiyaneswaran, M. Sujatha, C. Nalini, and S. Rajkumar, ''Grid search for predicting coronary heart disease by tuning hyper-parameters,'' Comput. Syst. Sci. Eng., vol. 43, no. 2, pp. 737–749, 2022.

