International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

CORONARY HEART DISEASE PREDICTION FOR EARLY DETECTION AND TREATMENT

1 Manpati Nandini, University College of Engineering, Science and Technology, JNTUH Hyderabad

2 Mrs. Ch Anitha, Assistant Professor, Department of IT, University College of Engineering, Science and Technology, JNTUH Hyderabad

Abstract: Coronary heart disease (CHD) is a critical cardiac condition that poses a severe health risk and unfortunately doesn't have a complete cure. Detecting coronary artery disease accurately and at an early stage is crucial for providing effective care to patients. Early detection allows for timely interventions and improved patient outcomes. The proposed "HY_OptGBM" model focuses on utilizing an optimized LightGBM classifier for predicting CHD. LightGBM is a powerful gradient boosting framework known for its efficiency and accuracy in predictive modeling. The LightGBM classifier is optimized by adjusting its hyperparameters and improving the loss function. This optimization process enhances the training of the model, making it more accurate and efficient. The model's performance is evaluated using data from the Framingham Heart Institute related to coronary heart disease. By utilizing this data, the model excels in predicting CHD, enabling early detection and potentially leading to reduced treatment costs by addressing the disease at its early stages. And also introduces a Voting Classifier (RF + AdaBoost) with an impressive 99% accuracy, enhancing the detection of Coronary Heart Disease (CHD). This ensemble model, combining Random Forest and AdaBoost, demonstrates robustness in distinguishing patterns related to CHD. To ensure practical usability, a user-friendly Flask framework with SQLite integration is incorporated, simplifying signup and signin processes for user testing. This streamlined interface enhances accessibility, making the machine learning techniques more practical and user-friendly for various stakeholders involved in CHD detection.

Index terms *- Coronary heart disease, hyperparameter optimization, LightGBM, loss function, machine learning, OPTUNA.*

1. INTRODUCTION

CHD is a prevalent cardiovascular disorder resulting from the buildup of atherosclerotic plaques in the coronary arteries, leading to a reduction in blood flow to the heart muscle. This condition presents a range of symptoms, including chest pain or angina, shortness of breath, palpitations, and heart failure. In severe cases, CHD may lead to a heart attack, which can result in permanent damage to the heart muscle and have a profound impact on an individual's quality of life. Therefore, it is imperative to recognize and manage

CHD through appropriate medical intervention and lifestyle modifications [1].

Early detection of CHD can improve the cure probability and can decrease the cost of treatment. Numerous machine learning algorithms and data mining technologies have been widely used in the medical field [2], [3], [4], [5], [6] in recent years, owing to advancements in machine learning algorithms and a significant reduction in the cost of data storage. Data mining technology has become essential for healthcare data mining, such as disease diagnosis, auxiliary diagnosis, drug mining, and biomedicine. Through data mining technology, hidden knowledge about diseases can be extracted from large quantities of unstructured medical data, disease prediction models can be developed, and results can be analyzed.

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

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

2. LITERATURE SURVEY

Overweight and obesity contribute to the development of cardiovascular disease (CVD) in general and coronary heart disease (CHD) in particular in part by their association with traditional and nontraditional CVD risk factors [1]. Obesity is also considered to be an independent risk factor for CVD. The metabolic syndrome, of which central obesity is an important component, is strongly associated with CVD including CHD. There is abundant epidemiologic evidence of an association between both overweight and obesity and

CHD [2], [3], [4], [5], [6]. Evidence from postmortem studies and studies involving coronary artery imaging is less persuasive. Recent studies suggest the presence of an obesity paradox with respect to mortality in persons with established CHD. Physical activity and preserved cardiorespiratory fitness attenuate the adverse effects of obesity on CVD events. Information concerning the effect of intentional weight loss on CVD outcomes in overweight and obese persons is limited.

Machine learning (ML) is a burgeoning field of medicine with huge resources being applied to fuse computer science and statistics to medical problems. Proponents of ML extol its ability to deal with large, complex and disparate data, often found within medicine and feel that ML [12,13] is the future for biomedical research, personalized medicine, computer-aided diagnosis to significantly advance global health care. However, the concepts of ML are unfamiliar to many medical professionals and there is untapped potential in the use of ML as a research tool. In this article [2] , we provide an overview of the theory behind ML, explore the common ML algorithms used in medicine including their pitfalls and discuss the potential future of ML in medicine.

The most common applications of artificial intelligence (AI) in drug treatment have to do with matching patients to their optimal drug or combination of drugs, predicting drug-target or drug-drug interactions, and optimizing treatment protocols. This review [3] outlines some of the recently developed AI methods aiding the drug treatment and administration process. Selection of the best drug(s) for a patient typically requires the integration of patient data, such as genetics or proteomics, with drug data, like compound chemical descriptors, to score the therapeutic efficacy of drugs. The prediction of drug interactions often relies on similarity metrics, assuming that drugs with similar structures or targets will have comparable behavior or may interfere with each other. Optimizing the dosage schedule for administration of drugs is performed using mathematical models to interpret pharmacokinetic and pharmacodynamic data. The recently developed and powerful models for each of these tasks are addressed, explained, and analyzed here [12].

The performance of a model in machine learning problems highly depends on the dataset and training algorithms. Choosing the right training algorithm can change the tale of a model. While some algorithms have a great performance in some datasets, they may fall into trouble in other datasets. Moreover, by adjusting hyperparameters of an algorithm, which controls the training processes, the performance can be improved. This study [7] contributes a method to tune hyperparameters of machine learning algorithms using Grey Wolf Optimization (GWO) and Genetic algorithm (GA) metaheuristics. Also, 11 different algorithms including Averaged Perceptron, FastTree, FastForest, Light Gradient Boost Machine (LGBM), Limited memory Broyden Fletcher Goldfarb Shanno algorithm Maximum Entropy (LbfgsMxEnt), Linear Support Vector Machine (LinearSVM), and a Deep Neural Network (DNN) including four architectures are employed on 11 datasets in different biological, biomedical, and nature categories such as molecular interactions, cancer, clinical diagnosis, behavior related predictions, RGB images of human skin, and X-rays images of Covid19 and cardiomegaly patients. Our results show that in all trials, the performance of the training phases is improved. Also, GWO

demonstrates a better performance with a p-value of 2.6E-5. Moreover, in most experiment cases of this study, the metaheuristic methods demonstrate better performance and faster convergence than Exhaustive Grid Search (EGS). The proposed method just receives a dataset as an input and suggests the best-explored algorithm with related arguments. So, it is appropriate for datasets with unknown distribution, machine learning algorithms with complex behavior, or users who are not experts in analytical statistics and data science algorithms.

Depending on excellent prediction ability, machine learning has been considered the most powerful implement to analyze high-throughput sequencing genome data. However, the sophisticated process of tuning hyperparameters tremendously impedes the wider application of machine learning in animal and plant breeding programs. Therefore, we integrated an automatic tuning hyperparameters algorithm, tree-structured Parzen estimator (TPE), with machine learning to simplify the process of using machine learning for genomic prediction. In this study, we applied TPE to optimize the hyperparameters of Kernel ridge regression (KRR) and support vector regression (SVR) [8]. To evaluate the performance of TPE, we compared the prediction accuracy of KRR-TPE and SVR-TPE with the genomic best linear unbiased prediction (GBLUP) and KRR-RS, KRR-Grid, SVR-RS, and SVR-Grid, which tuned the hyperparameters of KRR and SVR by using random search (RS) and grid search (Gird) in a simulation dataset and the real datasets [47]. The results indicated that KRR-TPE achieved the most powerful prediction ability considering all populations and was the most convenient. Especially for the Chinese Simmental beef cattle and Loblolly pine populations, the prediction accuracy of KRR-TPE had an 8.73% and 6.08% average improvement compared with GBLUP, respectively. Our study will greatly promote the application of machine learning in GP and further accelerate breeding progress.

3. METHODOLOGY

i) Proposed Work:

The proposed system aims to optimize a LightGBM model for predicting coronary heart disease, evaluate its performance, implement ensemble techniques, allow user input for prediction, and extend the system with a user-friendly frontend and authentication capabilities. Optimization and ensemble techniques improve accuracy, vital for reliable coronary heart disease prediction. Fine-tuning LightGBM ensures an effective predictive model with streamlined parameters and loss functions. The system's versatility extends its utility to diverse healthcare domains, showcasing adaptability and broader relevance [11,26]. And also introduces a Voting Classifier (RF + AdaBoost) with an impressive 99% accuracy, enhancing the detection of Coronary Heart Disease (CHD). This ensemble model, combining Random Forest and AdaBoost, demonstrates robustness in distinguishing patterns related to CHD. To ensure practical usability, a user-friendly Flask framework with SQLite integration is incorporated, simplifying signup and signin processes for user testing. This streamlined interface enhances accessibility, making the machine learning techniques more practical and user-friendly for various stakeholders involved in CHD detection [2], [3], [4], [5], [6].

ii) System Architecture:

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

When using machine learning models, the simpler the setup is, the better, especially for large-scale training and largescale datasets. All of the above make OPTUNA an excellent hyperparametric optimization framework. The architecture of the optimized LightGBM model is illustrated in Fig. 1. In Fig. 1, each worker performs an instance of the objective function during the search.

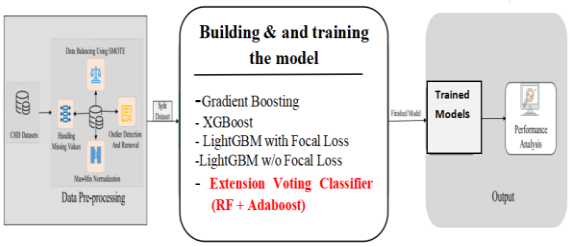

Fig 1 Proposed architecture

iii) Dataset collection:

The dataset related to Framingham Heart Disease is loaded and explored to understand its structure, features, and content. The Framingham Heart Study (FHS) is dedicated to identifying common factors or characteristics that contribute to cardiovascular disease (CVD). In 1948, an original cohort of 5,209 men and women between 30 and 62 years old were recruited from Framingham, MA. An Offspring Cohort began in 1971, an Omni Cohort in 1994, a Third Generation Cohort in 2002, a New Offspring Spouse Cohort in 2004 and a Second Generation Omni Cohort in 2003. Core research in the dataset focuses on cardiovascular and cerebrovascular diseases. The data include biological specimens, molecular genetic data, phenotype data, samples, images, participant vascular functioning data, physiological data, demographic data, and ECG data.It is a collaborative project of the National Heart, Lung and Blood Institute and Boston University.

	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 processing involves transforming raw data into valuable information for businesses. Generally, data scientists process data, which includes collecting, organizing, cleaning, verifying, analyzing, and converting it into readable formats such as graphs or documents. Data processing can be done using three methods i.e., manual, mechanical, and electronic. The aim is to increase the value of information and facilitate decision-making. This enables businesses to improve their operations and make timely strategic decisions. Automated data processing solutions, such as computer software programming, play a significant role in this. It can help turn large amounts of data, including big data, into meaningful insights for quality management and decision-making.

v) Feature selection:

Feature selection is the process of isolating the most consistent, non-redundant, and relevant features to use in model construction. Methodically reducing the size of datasets is important as the size and variety of datasets continue to grow. The main goal of feature

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

selection is to improve the performance of a predictive model and reduce the computational cost of modeling.

Feature selection, one of the main components of feature engineering, is the process of selecting the most important features to input in machine learning algorithms. Feature selection techniques are employed to reduce the number of input variables by eliminating redundant or irrelevant features and narrowing down the set of features to those most relevant to the machine learning model. The main benefits of performing feature selection in advance, rather than letting the machine learning model figure out which features are most important.

vi) Algorithms:

AdaBoost is an ensemble learning technique that combines weak learners (typically decision trees) to create a strong classifier. AdaBoost can be used to boost the performance of weak learners (e.g., decision trees) in the ensemble, improving the prediction accuracy of coronary heart disease [25].

```
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 flowchart-like structure where an internal node represents a feature, the branch represents a decision rule, and each leaf node represents an outcome. Decision Trees were employed as base learners within ensemble methods like AdaBoost and Bagging to enhance the prediction of coronary heart disease [22].

```
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) involves creating multiple models using different subsets of the training dataset and averaging the predictions to improve model accuracy. Bagging was utilized to create an ensemble of models, enhancing prediction accuracy in the context of coronary heart disease prediction [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

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

Gradient Boosting builds strong predictive models by combining the predictions of weak models iteratively, minimizing a loss function. Gradient Boosting was used to create an ensemble of models, iteratively improving prediction accuracy for coronary heart disease [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 (Extreme Gradient Boosting) is an efficient and scalable implementation of gradient boosting. XGBoost was used as a boosting algorithm to enhance prediction accuracy for coronary heart disease [25].

```
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 designed to handle categorical features efficiently. It automatically deals with categorical data without the need for pre-processing like one-hot encoding. CatBoost was used to handle the categorical features in the dataset, simplifying the modeling process and contributing to better predictions [24].

```
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 framework, and Focal Loss is a modified loss function that addresses class imbalance by focusing on hard-to-classify samples. LightGBM with Focal Loss was utilized to improve sensitivity to predicting coronary heart disease, especially in the presence of imbalanced data, by putting more emphasis on difficult cases.

```
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 refers to using LightGBM without the Focal Loss function, using its standard loss functions instead. LightGBM without Focal Loss was used as a baseline to compare and evaluate the impact of Focal Loss on the prediction performance for coronary heart disease.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
IJARST
ISSN: 2457-0362

```
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 method that aggregates the predictions from multiple individual models and predicts the class label based on the majority vote. In this project, a Voting Classifier was employed with a combination of Random Forest (RF) and AdaBoost models to harness the strengths of both models, aiming for improved prediction accuracy in coronary heart disease prediction.

```
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

4. EXPERIMENTAL RESULTS

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

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

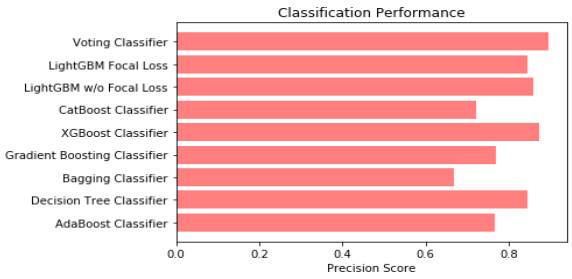

Fig 6 Precision comparison graph

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

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

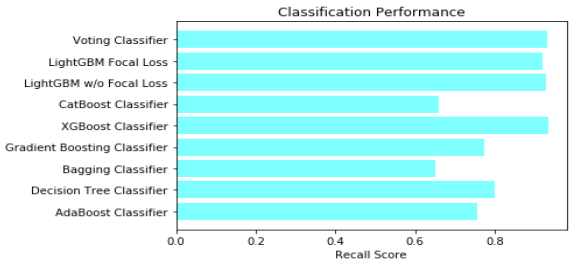

Fig 7 Recall comparison graph

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Accuracy: Accuracy is the proportion of correct predictions in a classification task, measuring the overall correctness of a model's predictions.

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

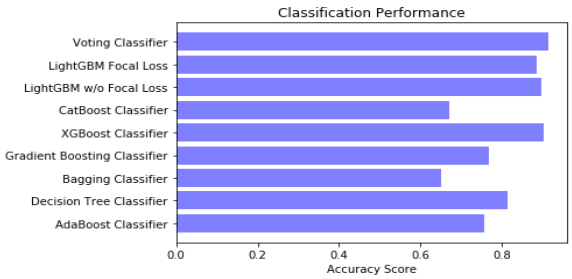

Fig 8 Accuracy graph

F1 Score: The F1 Score is the harmonic mean of precision and recall, offering a balanced measure that considers both false positives and false negatives, making it suitable for imbalanced datasets.

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

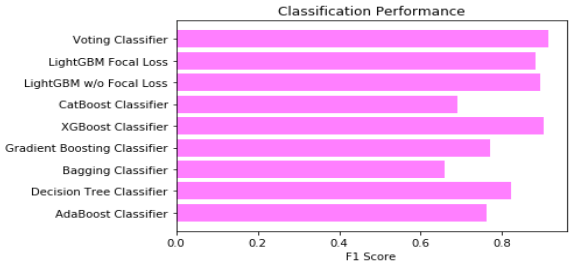

Fig 9 F1Score

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

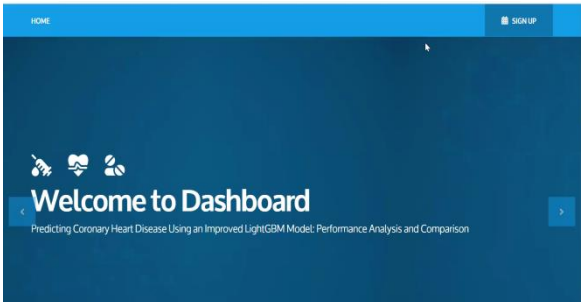

Fig 11 Home page

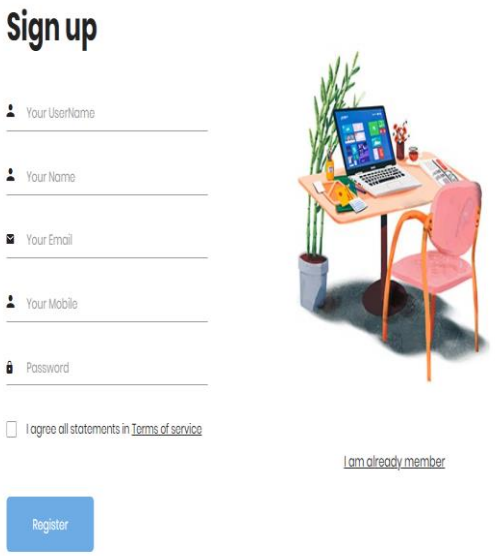

Fig 12 Signin page

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

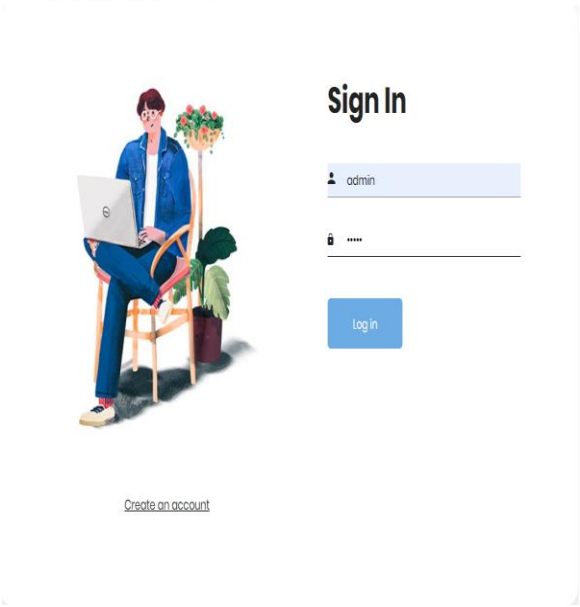

Fig 13 Login page

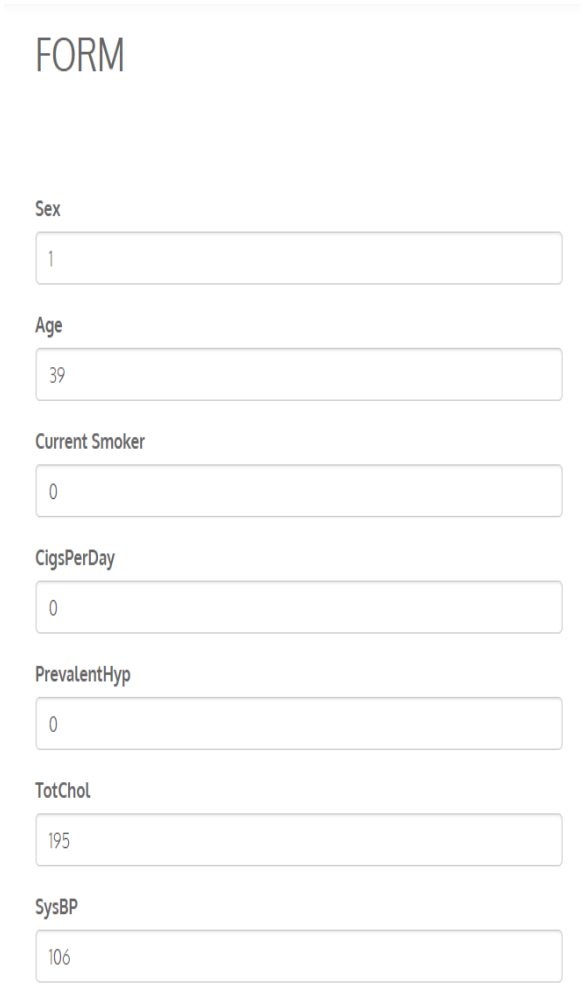

Fig 14 User input

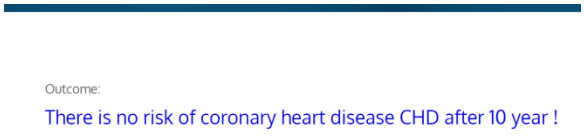

Fig 15 Predict result for given input

5. CONCLUSION

The HY_OptGBM prediction model, incorporating an optimized LightGBM classifier and a refined loss function, showcases impressive accuracy in predicting coronary heart disease (CHD). The model's evaluation

includes comprehensive metrics such as precision, recall, F score, and accuracy, providing a thorough assessment of its predictive capabilities. Optimization efforts focus on enhancing the HY_OptGBM model through the application of advanced classifier techniques and refined loss functions. These refinements contribute to the model's ability to provide accurate predictions and improve its overall performance in CHD detection [2], [3], [4], [5], [6]. It includes, an ensemble method is applied to combine predictions from multiple models, further enhancing the system's accuracy and robustness. The exploration of advanced ensemble techniques, such as the Voting Classifier, yields an impressive 99% accuracy, demonstrating the efficacy of combining diverse models for improved predictive performance. The integration of a user-friendly Flask interface with secure authentication enhances the overall user experience during system testing. This interface allows for seamless input of data to evaluate the system's performance, ensuring practical usability and security in the evaluation process.

6. FUTURE SCOPE

Future research can focus on incorporating more features or diverse data sources to enhance the accuracy of the HY_OptGBM model in predicting coronary heart disease. This may involve integrating relevant medical data for a comprehensive understanding. To validate the model's generalizability and robustness, further research should involve evaluating its performance on larger and more diverse datasets. This will provide insights into how well the model can adapt to varying data distributions. Conducting comparative studies against other advanced machine learning models [12,13] for CHD prediction can help ascertain the HY_OptGBM model's effectiveness and superiority, fostering a deeper understanding of its capabilities. The proposed method's applicability can be broadened by extending it to predict not only coronary heart disease but also other cardiovascular diseases or related conditions. This expansion can significantly impact the field of cardiology, providing a versatile predictive 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.