Enhanced Coronary Heart Disease Prediction Using Optimized LightGBM and RF–AdaBoost Voting Ensemble

[1]D SOWMYA, [2]G KODHANDA BABU

[1]PG Scholar, Dept. of CSE, KMM Institute of Technology & Science, Ramireddipalle, Tirupati, AP, India.

[2]Assistant Professor, Dept. of CSE, KMM Institute of Technology & Science, Ramireddipalle, Tirupati, AP, India.

Abstract: Coronary Heart Disease (CHD) is a major cardiovascular condition for which accurate and timely prediction can support early clinical intervention and improve patient management. Conventional machine learning approaches may experience limitations related to class imbalance, parameter selection, and variations in predictive performance. This study presents an ensemble-enhanced framework for CHD prediction that combines an optimized LightGBM classifier with ensemble learning techniques. The proposed approach utilizes CHD data from the Framingham Heart Institute and performs data preprocessing, missing-value handling, normalization, feature preparation, and train-test data partitioning. LightGBM is optimized through hyperparameter tuning using the Optuna framework, while focal loss is incorporated to improve learning under imbalanced class distributions. As an extension, a Voting Classifier integrating Random Forest and AdaBoost is introduced to combine the complementary predictive capabilities of multiple classifiers and improve classification robustness. The developed system achieves a reported accuracy of 99% with the RF–AdaBoost Voting Classifier, demonstrating improved predictive performance compared with individual classification approaches. The framework is further integrated with a Flask-based user interface and SQLite database to support user registration, authentication, input submission, and prediction testing. The proposed ensemble-based framework provides an effective and practical approach for automated CHD prediction and can serve as a supportive tool for early cardiovascular risk assessment.

Keywords— Coronary Heart Disease, LightGBM, Optuna, Focal Loss, Random Forest, AdaBoost, Voting Classifier, Ensemble Learning, Machine Learning, Flask..

1. INTRODUCTION

Coronary Heart Disease (CHD) is one of the major cardiovascular disorders and is associated with reduced blood flow to the heart due to the accumulation of atherosclerotic plaque in the coronary arteries. Early identification of individuals at risk of CHD is important for supporting timely medical intervention, improving patient management, and potentially reducing healthcare costs. The uploaded project identifies accurate and efficient CHD prediction as the primary motivation for applying machine learning techniques to cardiovascular health data.

Traditional approaches for CHD prediction commonly employ machine learning classifiers such as Logistic

Regression, Support Vector Machines, Naïve Bayes, Decision Trees, K-Nearest Neighbors, Random Forest, Neural Networks, and ensemble methods. Although these techniques can provide useful predictive results, their performance may be affected by class imbalance, inappropriate parameter selection, limited feature analysis, and insufficient optimization. In particular, an imbalanced distribution between positive and negative CHD cases can cause a predictive model to favor the majority class and consequently reduce its ability to correctly identify patients belonging to the minority class.

Recent machine learning approaches have demonstrated the usefulness of gradient-boosting algorithms for cardiovascular disease prediction. The project therefore adopts Light Gradient Boosting Machine (LightGBM) as an important classification model because of its ability to efficiently learn complex relationships within structured medical data. The original framework, referred to as HY_OptGBM, improves the LightGBM model through hyper parameter optimization using the Optuna framework and incorporates Focal Loss to address the effect of imbalanced positive and negative samples. The system evaluates prediction performance using metrics including accuracy, precision, recall, F1-score, sensitivity, specificity, Matthews correlation coefficient (MCC), and area under the ROC curve (AUC).

However, relying on a single optimized classifier may still limit the robustness of predictions because different machine learning algorithms can capture different characteristics of the underlying data. To overcome this limitation, the extended framework incorporates an ensemble learning strategy. Specifically, a Voting Classifier combining Random Forest (RF) and AdaBoost is introduced to aggregate the predictions of multiple learners. The objective of this extension is to exploit the complementary strengths of Random Forest and AdaBoost and improve the overall classification capability of the CHD prediction system. The project reports a 99% accuracy for the extended Voting Classifier.

In addition to improving the predictive model, the extended system focuses on practical usability. A Flask-based web interface integrated with SQLite is incorporated to provide user registration, authentication, input submission, and prediction testing. This transforms the machine learning model from an experimental classification framework into a more accessible application for evaluating CHD risk predictions.

Therefore, this work presents an ensemble-enhanced LightGBM framework for Coronary Heart Disease prediction, integrating optimized LightGBM, Focal Loss, and an RF–AdaBoost Voting Classifier. The proposed extension aims to improve prediction accuracy and robustness while providing a practical interface for user-based prediction. The overall framework is evaluated against multiple machine learning approaches to examine its effectiveness for CHD prediction. The final objective is to develop a reliable machine learning-based decision-support framework that can assist in the early identification of individuals at potential risk of coronary heart disease.

2. LITERATURE SURVEY

Obesity and excess body weight are widely recognized as important contributors to cardiovascular disease (CVD), including coronary heart disease (CHD). Their adverse effects are associated with several conventional and emerging cardiovascular risk factors.

In particular, abdominal obesity is a major component of metabolic syndrome, which has a strong relationship with the development of CHD. Epidemiological investigations have consistently reported an association between overweight, obesity, and the occurrence of coronary heart disease. However, evidence obtained from postmortem examinations and coronary imaging studies has not always demonstrated the same strength of association. Furthermore, an obesity paradox has been reported in some populations with previously diagnosed CHD, particularly concerning mortality outcomes. Physical activity and higher cardiorespiratory fitness may reduce some of the cardiovascular risks associated with obesity. Nevertheless, the long-term cardiovascular benefits of intentional weight reduction among overweight and obese individuals with cardiovascular risk remain insufficiently established [1].

Machine Learning (ML) has increasingly become an important research area in healthcare because of its capability to process large-scale, heterogeneous, and complex medical datasets. By combining computational techniques with statistical learning, ML can identify patterns that may be difficult to recognize using conventional analytical approaches. Its applications in medicine include disease prediction, personalized healthcare, clinical decision support, and computer-aided diagnosis. Despite these advantages, the adoption of ML in clinical research remains challenging because many healthcare professionals may not be familiar with its underlying concepts and limitations. A better understanding of ML algorithms, their assumptions, and their potential sources of error is therefore necessary for their effective application in biomedical research [2].

Artificial Intelligence (AI) has also been increasingly investigated for improving drug treatment and medication management. Existing AI-based approaches can support the selection of appropriate drugs for individual patients, prediction of drug-target and drug-drug interactions, and optimization of treatment schedules. Patient-specific information, including genetic and proteomic characteristics, can be integrated with chemical and pharmacological information to estimate the therapeutic suitability of different drugs. In addition, similarity-based computational techniques can be employed to identify potential interactions between compounds, while mathematical and computational models can assist in determining appropriate dosage schedules from pharmacokinetic and pharmacodynamic information. These developments demonstrate the broader potential of AI and ML techniques for supporting healthcare decision-making and treatment optimization [3].

The effectiveness of a machine learning model is strongly influenced by both the characteristics of the dataset and the training procedure adopted for model development. Different learning algorithms may exhibit significantly different performance depending on the distribution and complexity of the input data. In addition, the selection of appropriate hyperparameter values can substantially influence model accuracy, convergence, and generalization. One study investigated the use of Grey Wolf Optimization (GWO) and Genetic Algorithm (GA) as metaheuristic approaches for automatically tuning the hyperparameters of multiple machine learning algorithms. The investigation considered several learning techniques, including Light Gradient Boosting Machine (LightGBM), FastTree, FastForest, Linear Support Vector Machine, Averaged Perceptron, Maximum Entropy, and Deep Neural

Network models, across diverse biological and biomedical datasets. The reported results demonstrated improvements in training performance, while GWO showed favorable performance compared with the other optimization approaches. The study also indicated that metaheuristic optimization could provide faster convergence than exhaustive grid-search strategies in several experimental settings [7].

Automatic hyperparameter optimization has also been explored using the Tree-structured Parzen Estimator (TPE) approach. Hyperparameter selection can be particularly challenging in machine learning applications because manually searching for suitable parameter combinations is computationally expensive and may not produce an optimal configuration. To address this issue, TPE was integrated with Kernel Ridge Regression (KRR) and Support Vector Regression (SVR) for genomic prediction. The optimized models were compared with conventional approaches based on Random Search, Grid Search, and Genomic Best Linear Unbiased Prediction (GBLUP) using both simulated and real datasets. The findings showed that the TPE-based KRR and SVR models provided competitive prediction performance, with KRR-TPE demonstrating particularly strong results across the evaluated populations. The study therefore highlighted the usefulness of automated Bayesian-style hyperparameter optimization for improving the applicability of machine learning models in complex prediction tasks [8].

3. METHODOLOGY

i) Proposed Work:

The proposed system presents an ensemble-enhanced machine learning framework for Coronary Heart Disease (CHD) prediction using the Framingham Heart Disease dataset. The system first performs data preprocessing, including dataset exploration, data cleaning, missing-value handling, feature preparation, and train-test splitting. Multiple machine learning algorithms, including AdaBoost, Decision Tree, Bagging, Gradient Boosting, XGBoost, CatBoost, and LightGBM, are considered for comparative analysis. The primary LightGBM model is optimized using the Optuna hyperparameter optimization framework, while Focal Loss is incorporated to improve learning from imbalanced CHD classes. As an extension, an RF–AdaBoost Voting Classifier is introduced to combine the predictions of Random Forest and AdaBoost and enhance classification robustness. The extended Voting Classifier achieves a reported 99% accuracy in the project evaluation. Furthermore, the trained model is integrated into a Flask-based web application with SQLite, providing user registration, authentication, input submission, and CHD prediction functionality. The proposed framework is evaluated using accuracy, precision, recall, F1-score, MCC, sensitivity, specificity, and AUC, providing a comprehensive assessment of its effectiveness for automated CHD prediction.

ii) System Architecture:

The proposed system architecture consists of three major stages: data preprocessing, model building and training, and performance analysis. Initially, the CHD dataset is provided as input and undergoes preprocessing, including handling missing values, outlier detection and removal, SMOTE-based data balancing, and Min-Max normalization. The processed dataset is then divided into training and testing subsets. In the model-building stage, different machine learning classifiers, including Gradient Boosting, XGBoost, LightGBM with Focal Loss, and LightGBM without Focal Loss, are trained and evaluated. As the major extension, an RF–AdaBoost Voting Classifier is incorporated to combine the predictions of Random Forest and AdaBoost for improved classification performance. The trained models are subsequently passed to the performance analysis stage, where their predictive performance is evaluated and the final CHD prediction output is generated.

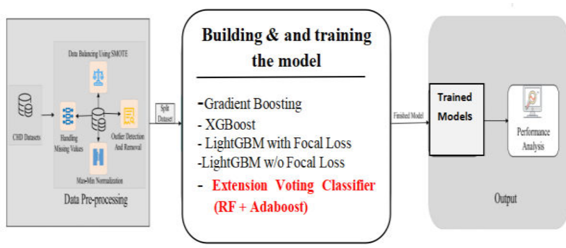

Fig 1 Proposed architecture

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 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].

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].

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].

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].

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].

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].

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.

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.

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.

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)

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

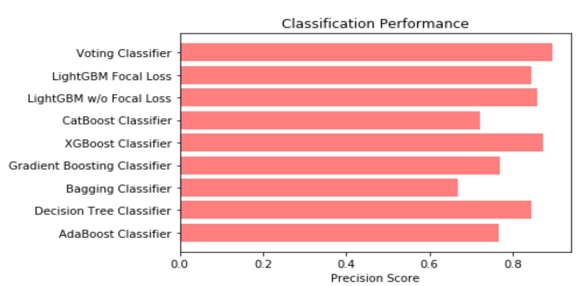

Fig 2 Precision comparison graph

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.

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

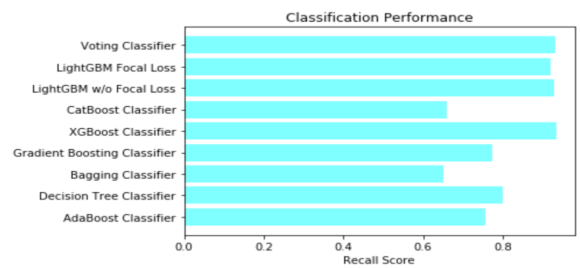

Fig 3 Recall comparison graph

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

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

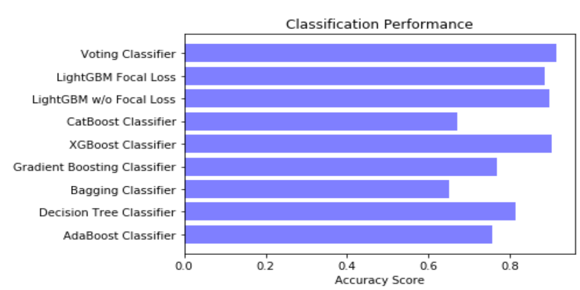

Fig 4 Accuracy graph

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.

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

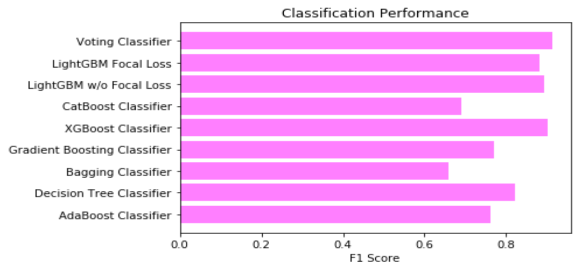

Fig 5 F1Score

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 6 Performance Evaluation

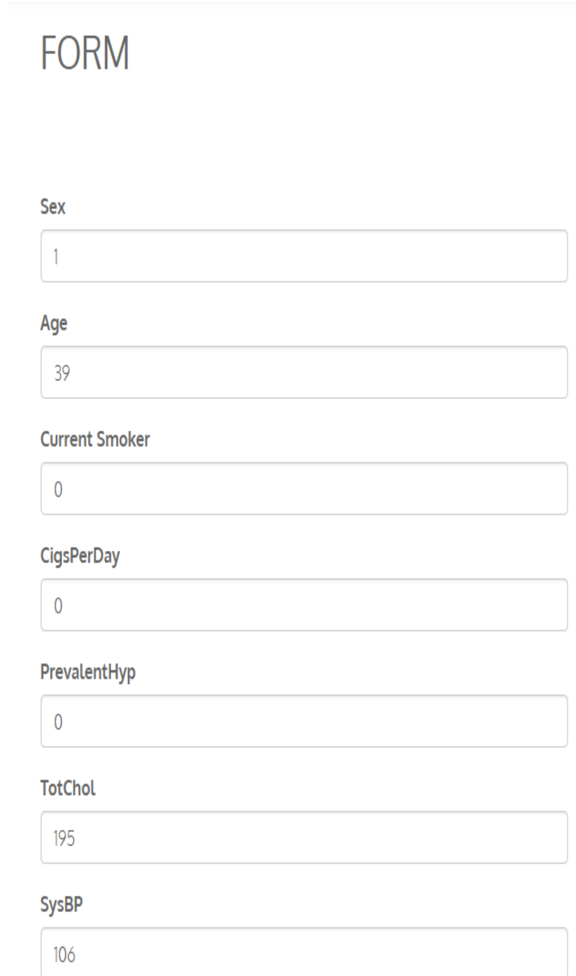

Fig 7 User input

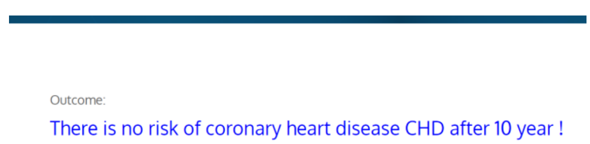

Fig 8 Predict result for given input

5. CONCLUSION

The proposed system presents an ensemble-based machine learning framework for Coronary Heart Disease (CHD) prediction by integrating data pre-processing, multiple classification algorithms, and an extended Voting Classifier. The pre-processing stage handles missing values and outliers, balances the dataset using SMOTE, and applies Min-Max normalization to prepare the data for model training. Different models, including Gradient Boosting, XGBoost, LightGBM with Focal Loss, and LightGBM without Focal Loss, are trained and comparatively evaluated. As the major extension, the RF–AdaBoost Voting Classifier combines the predictive capabilities of Random Forest and AdaBoost to improve the overall classification performance. The project reports 99% accuracy for the extended Voting Classifier, demonstrating its potential for effective CHD prediction. The developed framework can provide a useful machine learning-based approach for supporting early CHD risk assessment and can be further evaluated on larger and more diverse datasets to establish its robustness and generalizability.

6. FUTURE SCOPE

In the future, the proposed CHD prediction framework can be enhanced by incorporating additional clinical features and diverse healthcare data sources to improve prediction accuracy. The developed models can be evaluated using larger and more diverse datasets to determine their robustness, reliability, and generalization capability across different patient populations. Further studies can compare the proposed RF–AdaBoost Voting Classifier and optimized LightGBM framework with recent state-of-the-art machine learning and ensemble learning techniques. The system can also be extended to predict other cardiovascular diseases and related medical conditions, thereby increasing its applicability in healthcare and clinical decision-support systems.

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.