Early Prediction of Coronary Heart Disease using Boosting-based Voting Ensemble Learning

Subhash Mondal
Computer Science and Engineering
Central Institute of Technology Kokrajhar
Kokrajhar, India
subhash@msit.edu.in

Ranjan Maity
Computer Science and Engineering
Central Institute of Technology Kokrajhar
Kokrajhar, India
r.maity@cit.ac.in

Yash Raj Singh
Computer Science and Engineering
Meghnad Saha Institute of Technology
Kolkata, India
yash_s.cse2019@msit.edu.in

Soumadip Ghosh
Computer Science and Engineering
Institute of Engineering & Management
Kolkata, India
soumadip.ghosh@gmail.com

Amitava Nag
Computer Science and Engineering
Central Institute of Technology Kokrajhar
Kokrajhar, India
amitava.nag@cit.ac.in

***Abstract*—Coronary-Heart-Disease (CHD) risk increases daily due to the uncontrolled lifestyle of today's adult age group. The early detection of the disease can prevent unfortunate death due to heart-related complications. The Machine Learning (ML) technique is essential for the early diagnosis of CHD and for identifying its many contributing factor variables. To build the prediction model, we have used the dataset consisting of 4240 instances and 15 related features to predict the possibility of future risk of CHD in the next ten years. Initially, thirteen ML models were deployed with 10-fold cross-validation, reflecting the highest test accuracy of 91.28% for the Random Forest (RF) classifier. The models were turned further, and the boosting algorithms showed the highest accuracy of 91% and above; the Gradient Boost (GB) classifier performed better with an accuracy of 92.11%. The voting ensemble approaches using the best-performing boosting models, namely GB, HGB, XGB, CB, and LGBM, have been considered for the final prediction. The prediction results reflected an accuracy of 92.26%, an F1 score of 91.25%, a ROC-AUC score of 0.917, and the number of False Negatives (FN) values is about 6.25% of the total test dataset.**

Keywords—Machine Learning, Coronary Heart Disease, Ensemble Technique, voting classifier, Boosting classifier

I. INTRODUCTION

CHD is characterized by a blockage or the narrowing of the coronary arteries leading to causes of major death worldwide. It is usually caused by atherosclerosis, where the artery walls are stuffed with fatty deposits [1]. Vigorous exertion may result in heart attacks, chest pain, and shortness of breath. Many factors that increase the CHD risk having high blood pressure, consumption of tobacco, high cholesterol, physical inactivity, diabetes, and obesity [2]. The risk factors for CHD may combine uncontrollable elements like age, ethnicity, and family medical history with controllable factors like those influenced by a person's lifestyle. Around 17.9 million people throughout the globe die every year, which is almost 32% of the total population death [3]. Therefore, if CHD symptoms are identified early, the patient may be able to manage some of these risk factors through dietary modifications and medication, delaying the development of an extreme version of the illness.

The researchers were able to create new approaches based on machine learning and artificial intelligence thanks to ongoing technological advancements. We have analyzed the thirteen ML models' performance by considering different parameters in predicting the CHD through this study. The ML algorithms used were namely Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Gaussian NB (GNB), Decision Tree (DT), Extreme Gradient Boosting (XGB), Random Forest (RF), Bernoulli NB (BNB), CatBoost (CB), AdaBoost (AB), Hist-Gradient Boost (HGB), Gradient Boost (GB), and Light Gradient Boost (LGBM) Classifiers. The dataset comprised 4240 entries along with 15 features and one target attribute, indicating the risk of CHD in the next ten years. Further hyperparameter tuning technique, namely Randomized search CV, was applied to the top ten best-performing algorithms for better analysis of results.

This study comprised we have performed ensemble techniques like voting, including hard and soft, on the top five best-performing boosting models after hyperparameter tuning. The proposed model reflected an accuracy of 92.28% for soft voting with 10-fold cross-validation techniques. The other performance matrices like F1 score, recall, precision, and ROC-AUC score were 0.913, 86.94%, 96.01%, and 0.917, respectively. The standard deviation value of the model was 0.99, which is less than 1. The number of false-negative values was less on the model's test dataset, which was 6.25%. The detailed model results description is provided in the proposed methodology section. The boosting-based voting ensemble approach is the first to predict the CHD detection model.

The paper is organized by having a literature review in section II after the Introduction, followed by the proposal in Section III. Section IV provides the result and comparative analysis, with Section V concluding the study.

II. RELATED WORK

This section focuses on the research related to CHD prediction and detection approaches using classical ML models. We found that a few research works have already been proposed by researchers in different publication houses after extensive searches among different portals. Also, this study had not considered CHD prediction proposals on a reduced features set related work.

In [4], the authors proposed six ML algorithms like, SVM, AB, Bagging, KNN, Neural Network, and RF, to predict Acute Coronary Syndrome (ACS). Further, the thirty best features were selected with the Ranker algorithm using the Gain-Attribute-Eval parameter. Bagging performed the best

978-1-6654-9291-1/22/$31.00 ©2022 IEEE

results with an accuracy of 76.28%; the recall and precision were slightly more accurate than AdaBoost, with 0.76 and 0.75, respectively. The authors [5] proposed using several boosting algorithms belonging to ensemble techniques like CB, LGBM, XGB, AB, and GB to predict CHD. The algorithms above were applied directly to a publicly available dataset. Further, the class imbalance from the dataset was studied but not used any data balancing technique. The highest accuracy was observed in the XGB model at 87.62%, which is higher than all other boosting techniques. In [6], the authors proposed several linear ML models like the NB, KNN, Classification and Regression Trees (CART), SVM, XGB, Linear Discriminant Analysis (LDA), and LR. Further, ensemble methods were applied, like boosting and bagging, particularly to the Classification and Regression Trees (CART). The highest accuracy score of 84.96% was obtained using LR in both datasets. In [7], the authors used R-studio and RapidMiner, specific-software tools for analyzing the data, and applied some ML algorithms like DT, RF, SVM, Neural Network (NN), and LR for the prediction of CHD risk and compared results among them. Further, performance evaluation of different algorithms with the software above included considering parameters like accuracy, specificity, recall, and AUC score. It was observed that SVM in R-studio produced the highest AUC of 0.75 and an accuracy of 69%. In [8], the authors used algorithms like KNN, SVM, LR, NB, RF, and DT for the detector. To determine the correlation between the features set and the target outcome, Pearson's coefficient was considered. Also, the model was examined using ensemble techniques like stacking, bagging, and boosting. It is noted that the accuracy was slightly increased when bagging was used. The stacked ensemble model with KNN, RF, and SVM algorithm achieved a high accuracy of 75.1%. In [9], the authors proposed a predictive model of CHD using ML algorithms like KNN, SVM, DT, and RF. Further, Soft Voting Classifier (SVC) and Hard Voting Classifier (HCV) ensemble methods were used to increase the prediction model's performance; a higher accuracy of 83.2% was achieved in HVS, whereas SVC was slightly lower. In [10], ML-based prediction was presented using algorithms like GNB, BNB, and RF. The dataset is from the UCI repository, which is publicly available, consisting of 303 records and 76 columns. The model was built using GNB, BNB, and RF independently on the training dataset. For the evaluation of the model, they considered a confusion matrix, accuracy, precision, and recall values. They claimed GNB, BNB, and RF results with an accuracy of 85%, 85%, and 75%, respectively. In [11], a model for the prediction of CAD using SVM and CART was proposed. The authors concluded that both SVM and CART had almost the same accuracy of 88.33%. Still, compared to other performance matrices like sensitivity, CART performs better than SVM and the converse in the case of specificity. In [12], the LR using Stochastic Gradient Descent (SDG) and DT model was considered for the prediction of CHD. The data binarization method is used for data preprocessing. An accuracy of 77% was recorded highest for LR-SGD. In [13], a prediction model of CAD was proposed by the authors using ML algorithms like SVM, GB, DT, RF, and AB. Recursive Feature Elimination (REP) and Boruta feature (BF) selection methods were used to get the best results on a few features. The Framingham dataset was highly unbalanced, so SMOTE and random over-sampling methods were used to take care of the imbalanced data present. The RPE and RF performed best, with an accuracy of 88%.

It can be said that the accuracy of diagnosis in the studies mentioned above and many others is not remarkable and can still be improved. The significant research contributions of this paper are explicitly stated below:

(a) This study has deployed thirteen ML algorithms with 10-fold cross-validation.

(b) To the best of the authors' knowledge, this work is the first attempt at the prediction of CHD, where a boosting-based voting ensemble approach is performed.

(c) The accuracy and the RoC-AuC of the CHD diagnosis model are improved with results of an accuracy of 92.26%, and a ROC-AUC score of 0.92, respectively.

III. PROPOSED WORK

Through this section, we provided the details of used dataset acquisition, data pre-processing methods, model training that we have deployed for the prediction of CHD, and finally proposed ensemble-based voting approach. The proposed workflow diagram with a high-level model is presented in Fig. 1.

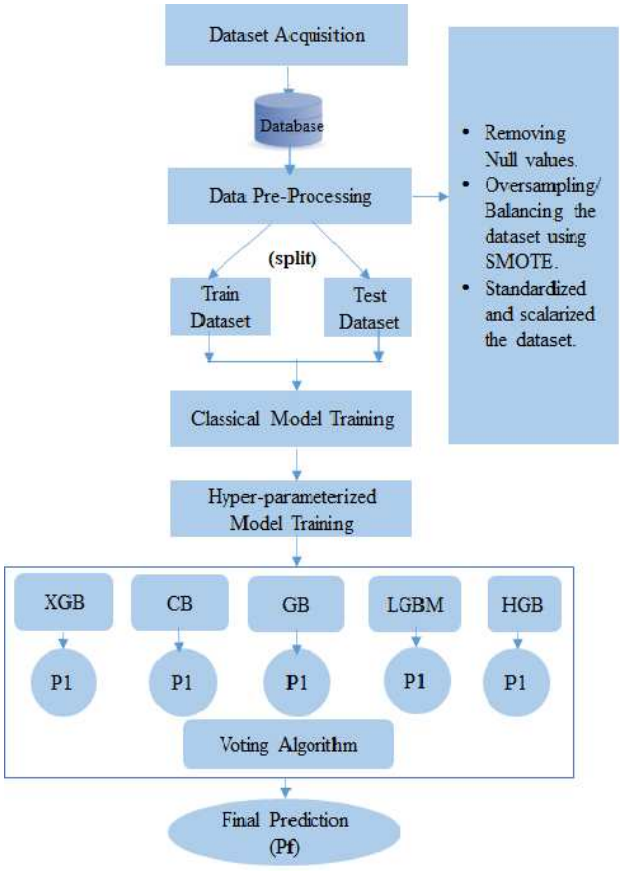

Fig. 1. The proposed workflow diagram of the prediction model

A. Dataset Acquisition & Description

We have considered the publicly accessible dataset of cardiovascular studies [14]. The dataset combined the records of 4240 patients with 15 attributes related to behavioral, demographic, and medical risk factors. The outcome attribute is the risk factor for CHD in the future ten years. The various features description, along with null values instances, are represented in Table I.

B. Data Pre-processing

One of the most crucial steps is pre-processing raw data before initializing the machine learning model. We filter out the clean data so the ML model can grasp important information from a dataset with fewer errors while predicting the results. In this study, we first handled the missing (null) attribute values using column-specific *Mean* values and scaled down the results using the *StandardScaler* library [15]. After analyzing the target variable imbalance, we fixed the whole data point by applying the oversampling SMOTE method to eliminate imbalanced data. The target column instances difference before and after the *SMOT* techniques use are presented in Table II. Although the dataset contains categorical features as they are binary and already in the numerical state, we skipped the label or one-hot encoding in feature engineering steps. After performing the data pre-processing methods, the resulting heatmap of all features, including the target attribute, is depicted in Fig. 2.

TABLE I. FEATURE DESCRIPTION WITH NULL VALUE COUNT

#	Column	# Null Values	Feature description
0	gender	0	Patients gender M (1) or F (0)
1	age	0	Patients' age, numeric values
2	education	105	Education stages, numeric valuers
3	currentsmoker	0	Smoking status, binary either o or 1
4	cigsperday	29	Count of cigarettes, numeric values
5	bpmeds	53	Either consume blood pressure medicine or not
6	prevalentstroke	0	History of stroke, either 0 or 1
7	prevalenthyp	0	Patients are hypertensive or not, numeric values.
8	diabetes	0	Diabetic patients or not, binary values
9	totchol	50	Count of cholesterol
10	sysbp	0	Lower BP level, numeric values
11	diabp	0	Upper BP level, numeric values
12	bmi	19	Body mass, values
13	heartrate	1	Heart bit-level values
14	glucose	388	Blood glucose measurements in numeric values
15	tenyearchd	0	Binary target values of ten years risk of future CHD

TABLE II. TARGEL COLUMN VALUE COUNT USING SMOTE

	Before RandomOverSampler	After RandomOverSampler
Count of Label 0	3596	3596
Count of Label 1	644	3596
Total instance Counts	4240	7192

Later we partitioned the processed data for training purposes to build the model and testing for cross-checking the model into the ratio of *0.90:0.10* respectively, here 90% of the data was used for model training with enough data the model can identify and learn the hidden patterns in data points.

C. Model Training

Initially, we developed the base model for our research study by deploying 13-ML classification algorithms on the training dataset. We analyzed the prediction model on the test dataset considering the different performance matrices like Accuracy (A), k-fold Mean accuracy (M), Precision (P), Recall (R), F1 score (F), ROC-AUC score (RA), and Cohen-kappa score (CK) to evaluate model performance and stability. We observed that the three models, namely RF, CB, and LGBM, give the best results with a mean accuracy of 90% and above. The detailed outcomes of the default hyperparameter tuning model are presented in Table III.

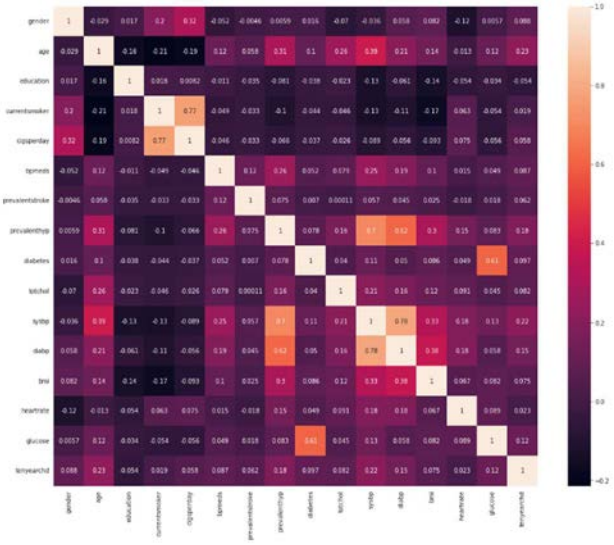

Fig. 2. Feature correlation heatmap

TABLE III. THE MODEL RESULTS WITHOUT HYPER-PARAMETER TUNING

Model Name	A	M	P	R	F	CK	RA
RF	90.69	91.29	0.94	0.87	0.90	0.81	0.907
CB	88.33	90.30	0.95	0.81	0.87	0.77	0.883
LGBM	89.17	90.03	0.97	0.81	0.88	0.78	0.892
HGB	88.33	89.85	0.94	0.82	0.88	0.77	0.883
GB	87.08	88.26	0.94	0.79	0.86	0.74	0.871
XGB	86.67	87.59	0.94	0.79	0.85	0.73	0.867
AB	82.22	84.15	0.80	0.86	0.83	0.64	0.822
DT	83.75	84.01	0.86	0.80	0.83	0.68	0.838
KNN	81.81	81.78	0.75	0.95	0.84	0.64	0.818
SVM	77.50	78.12	0.78	0.77	0.77	0.55	0.775
LR	66.81	67.65	0.66	0.71	0.68	0.34	0.668
BNB	65.69	64.23	0.67	0.63	0.65	0.31	0.657
GNB	62.64	60.94	0.76	0.37	0.50	0.25	0.626

TABLE IV. THE MODEL RESULTS WITH HYPER-PARAMETER TUNING

Model Name	A	M	P	R	F	CK	RA
GB	90.97	92.12	0.94	0.87	0.91	0.82	0.910
CB	90.69	92.10	0.95	0.86	0.90	0.81	0.907
HGB	90.97	91.80	0.94	0.88	0.91	0.82	0.910
XGB	90.97	91.42	0.98	0.84	0.90	0.82	0.910
LGBM	88.19	91.39	0.94	0.82	0.87	0.76	0.882
RF	89.44	91.07	0.93	0.85	0.89	0.79	0.894
KNN	85.14	87.25	0.79	0.95	0.86	0.70	0.851
DT	80.14	83.05	0.81	0.79	0.80	0.60	0.801

AB	79.72	80.04	0.81	0.77	0.79	0.59	0.797
LR	63.89	67.34	0.63	0.66	0.65	0.28	0.639

Afterward, we performed hyperparameter tuning on these top ten-ML models using a *Randomized Search CV* to get better results. After studying the tuned model results, we observed that boosting algorithms performed better than non-boosting type classification algorithms. The best Five-boosting algorithms that performed well out of 13 considered classification models are GB, HGB, XGB, CB, and LGBM Classifier. The result analysis for these five algorithms and also the top ten-tuned model performances are depicted in Table IV. The ROC-AUC curve with probabilistic AUC values of the tuned models is shown in Fig. 3.

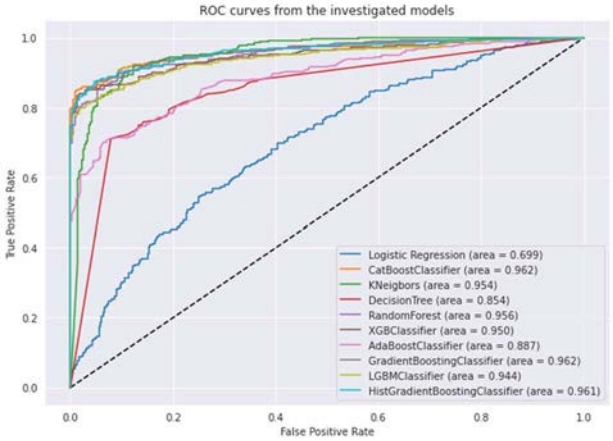

Fig. 3. Tuned model ROC-AUC curve

D. Voting Ensemble method

Voting is one of the ways via which we can tie up different performances model to get the final prediction; it is an ensemble method that works very similarly to stacking. There are two types of classifiers in voting: soft voting (SV) and hard voting (HV). Soft voting classifiers differentiate the provided data based on the probabilities of all the predictions made by different classifiers. Also, the hard voting classifier determines the supplied data based on the mode of all the predictions made by other classifiers. Since voting is much dependent on provided performance models, here we implemented our top five-best tuned boosting models for a hard and soft voting classifier with the cross-validation of Ten-fold, and we accomplished better results with the soft-voting classifier with a mean accuracy of 92.25% and ROC-AUC, F1-score of 0.917 and 0.91 respectively. Table V depicts the results of the voting method on the hyperparameter-tuned model. The confusion matrix of the voting classifiers model in Fig. 4 and Fig. 5 corresponds to soft and hard voting.

TABLE V. RESULTS OF THE VOTING ENSEMBLE MODEL

Model Name	**A**	**M**	**P**	**R**	**F**	**CK**	**RA**
SV	91.67	92.26	0.96	0.87	0.91	0.83	0.917
HV	91.11	92.09	0.96	0.86	0.91	0.82	0.911

IV. COMPARATIVE RESULT ANALYSIS

In this section, we discussed and presented the comparative analysis of results with the previously published work by keeping in mind that all literary works are done by considering the conventional ML algorithms with a selection of all the features related to CHD prediction. We have not considered the works associated with applying some feature selection methods to deploy a model for the detection of CHD. To compare the results of this study's previous works, we have considered the evaluation parameters as accuracy, recall, precision, F1 measure, Cohen-kappa score, and ROC-AUC values of the claimed model. The proposed results are far better than the literature regarding the accuracy, recall, and ROC-AUC values. A detailed analysis of the results is depicted in Table VI and Table VII.

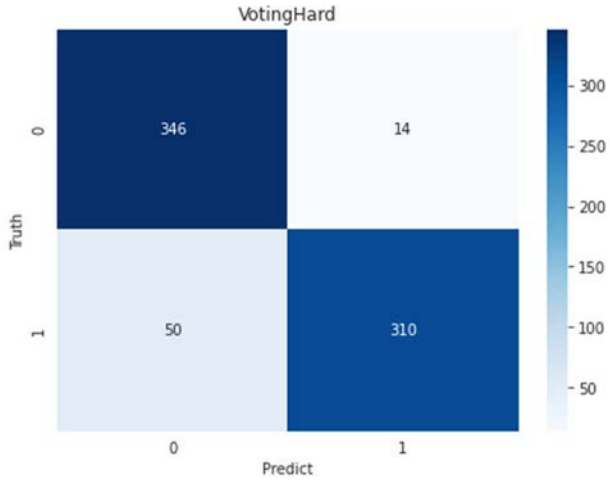

Fig. 4. Confusion matrix of the hard voting ensemble classifier

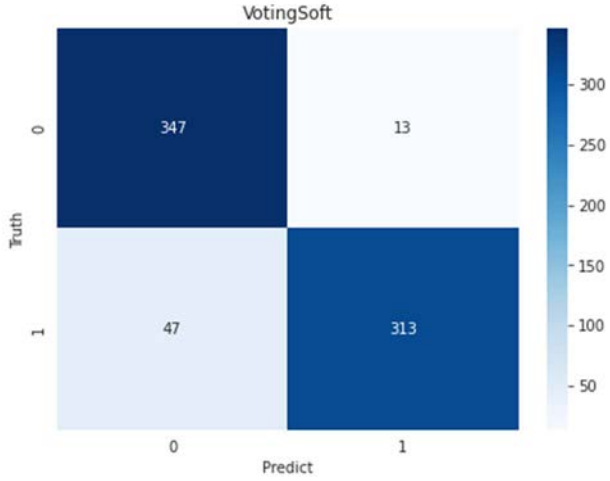

Fig. 5. Confusion matrix of the soft voting ensemble classifier

TABLE VI. LITERARY CLAIMED RESULTS AND MODEL USED

Ref#	**Algorithm(s) used**	**Claimed Results**
[4]	KNN, SVM, RF, AB, ANN, Bagging	High Accuracy of Bagging
[5]	AB, XGB, GB, CB, LGBM	High Accuracy of XGB
[6]	LR, LDA, NB, CART, SVM, KNN, MLP, DNN	High Accuracy with LR
[7]	SVM, LR, DT, RF, NN	High AUC score with SVM
[8]	KNN, SVM, RF, LR, DT, NB, MLP	Stacking (KNN, RF, SVM) accuracy
[9]	HV & SV Ensemble	High accuracy with HV classifier

[10]	GNB, BNB, RF	High Accuracy with GNB
[11]	SVM, CART	High Accuracy with SVM
[12]	DT, LR with SGD	Accuracy of DT
[13]	DT, GB, RF, SVM, AB, REP, BF	High Accuracy using RF with RPE

TABLE VII. RESULTS ANALYSIS OF THE CLAIMED MODEL USED

Ref#	A	P	R	F	RA	CK
[4]	76.28	75	76.3	74.8	0.756	
[5]	87.62	46			0.716	
[6]	84.96	71.19			0.65	
[7]	69		42		0.75	
[8]	75.1					
[9]	83.2					
[10]	85	83.78	91.81	87.32		
[11]	88.33		84			
[12]	77					
[13]	89	0.89	0.88	0.88		
Proposed	**92.26**	**0.96**	**0.87**	**0.91**	**0.92**	**0.83**

V. CONCLUSION

In this study, we have performed the prediction model of CHD in two manners. First, to develop a model using the thirteen ML classifiers among those models, we have selected the top ten best-performing models with hyperparameter tuned and observed that the boosting models performed better compared to other non-boosting models with a mean accuracy of 91% and above. In the second approach, voting ensemble methods have been deployed using the best five boosting models. This is the first attempt to use boosting classifiers as the base model for voting ensemble classifiers to develop a prediction model of CHD. We observed that the false-negative value is less than about 6.25% of the total test dataset. In the future, feature selection methods can be incorporated to eliminate the irrelevant features of the dataset and deploy the prediction model with reduced features using best-performing boosting models and intended to deploy the model using deep learning techniques and compare the results with this proposed model.

REFERENCES

[1] "Coronary heart disease," 10 March 2020. [Online]. Available: https://www.nhs.uk/conditions/coronary-heart-disease/. [Accessed 8 July 2022].

[2] "Coronary artery disease," 25 May 2022. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/coronary-artery-disease/symptoms-causes/syc-20350613. [Accessed 8 July 2022].

[3] D. T. Khan, "Cardiovascular diseases," [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1. [Accessed 8 July 2022].

[4] M. Raihan, M. . M. Islam, P. Ghosh, S. . A. Shaj, M. R. Chowdhury, S. Mondal and A. More, "A Comprehensive Analysis on Risk Prediction of Acute Coronary Syndrome using Machine Learning Approaches," in *2018 21st International Conference of Computer and Information Technology (ICCIT)*, Dhaka, Bangladesh, 2018.

[5] A.-Z. . S. B. Habib, T. Tasnim and M. M. Billah, "A study on Coronary Disease Prediction Using Boosting-based Ensemble Machine Learning Approaches," in *2019 2nd International Conference on Innovation in Engineering and Technology (ICIET)*, Dhanka, Bangladesh, 2019.

[6] N. Mangathayaru, B. P. Rani, S. M. Gajapaka, S. A. Patel, L. Bharadwaj and V. Janaki, "An Imperative Diagnostic Model for Predicting CHD using Deep Learning," in *2020 IEEE International Conference for Innovation in Technology (INOCON)*, Hyderabad, India, 2020.

[7] J. J. Beunza, E. Puertas, E. G. Ovejero, G. koleva, C. Hurtado, m. F. Landecho, G. vellalba and E. Condes, "Comparison of Machine learning Algorithms for clinical even prediction (risk of coronary heart disease)," *Journal of Biomedical Informatics,* vol. 97, p. 10325, 2019.

[8] V. Shorewala, "Early detection of coronary heart disease using ensemble techniques," *Informatics in Medicine Unlocked,* vol. 26, 2021.

[9] P. D. P. Singh, R. Bansal and S. Sharma, "Coronary Heart Disease Prediction Using Voting Classifier Ensemble Learning," in *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, Greater Noida, India, 2021.

[10] c. Bemando, E. Mranda and M. Aryuni, "Machine-Learning-Based Prediction Models of Coronary Heart Disease Using Naïve Bayes and Random Forest Algorithms," in *I2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)*, Pekan, Malasiya, 2021.

[11] M. Aryuni, E. Miranda, C. Bernando and A. Hartano, "Coronary Artery Disease Prediction Model using CART and SVM: A Comparative Study," in *2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI)*, Jakarta, Indonesia, 2021.

[12] S. Gaba, A. Aggarwal, S. Nagpal, D. Kumar and P. Singh, "A Forecast of Coronary Heart Disease using Proficient Machine Learning Algorithms," in *2021 Sixth International Conference on Image Information Processing (ICIIP)*, Shimla, India, 2021.

[13] N. Devi, S. Suruthi and S. Shanthi, "Coronary Artery Disease prediction using Machine Learning Techniques," in *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbartore, India, 2022.

[14] C. Ganteng, F. jason and W. Azariah, "Cardiovascular Study Dataset," 3 October 2020. [Online]. Available: https://www.kaggle.com/datasets/yashnaik12/heart-patients. [Accessed 8 June 2022].

[15] "sklearn.preprocessing.StandardScaler," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html. [Accessed 8 June 2022].