Predicting Cardiovascular Disease Events with Machine Learning Models

Nikhil Potluri

Delhi Public School

DOI:10.37648/ijrmst.v17i01.010

[1]*Received: 30 April 2024; Accepted: 20 June 2024; Published: 26 June 2024*

ABSTRACT

Cardiovascular diseases (CVDs) are a major cause of death worldwide, ranking among the deadliest disease. By utilizing statistical and machine learning (ML) algorithms to discover risk biomarkers, CVDs can be early detected and prevented. In this work, we use biochemical data and clinical CVD risk factors to predict CVD-related death within a 10-year follow-up period using machine learning models like Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB), Extreme Grading Boosting (XGB), and Adaptive Boosting (AdaBoost). Using the Ludwigshafen Risk and Cardiovascular Health (LURIC) study cohort, we included 2943 individuals in our analysis, of whom 484 were declared deceased from cardiovascular disease. For every model, we determined its accuracy (ACC), precision, recall, F1-score, specificity (SPE), and area under the receiver operating characteristic curve (AUC). According to the comparative analysis's results, the most dependable algorithm is logistic regression, which has an accuracy of 72.20%. In the TIMELY trial, these findings will be utilized to calculate the risk score and mortality of cardiovascular disease in patients with a 10-year risk.

INTRODUCTION

Cardiovascular disease is a dangerous condition that affects people globally. It is responsible for 17.9 million deaths annually, or 32% of all fatalities worldwide [1]. By 2030, CVD is predicted to be the cause of more than 23.6 million deaths yearly [2]. In 2016, cardiovascular disease-related mortality accounted for 1.68 million deaths in Europe, or 37.1% of all deaths [3]. Cardiovascular disease can be brought on by a number of risk factors, including blood pressure, diabetes mellitus, LDL cholesterol, irregular pulse rate, physical activity, poor diet, family history of the disease, and ethnic background. Numerous tests can be used to predict cardiovascular disease. However, early diagnosis may be challenging due to medical staff inexperience [4]. The risk of cardiovascular disease must be assessed for both primary and secondary prevention. A number of statistical risk scores are available to estimate the risk of patients who have experienced a prior CVD incident or The non-cardiovascular disease (CVD) event is categorized as follows: Atherosclerotic Cardiovascular Disease (ASCVD) [9], American College of Cardiology/American Heart Association (ACC/AHA) [9], HeartScore [10], WHO risk score [11], CoroPredict [12], QRISK3 [6], Framingham Risk Score (FRS) [7], Joint British Society risk calculator 3 (JBS3) [8].

Based on various ethnicities, the efficacy of the statistical models, risk ratings, was assessed in multiple studies [13–14]. The primary finding of these studies was that the percentage of patients who have a high or low risk of getting CVD is predicted by these scores.

When taught on appropriate medical data, machine learning algorithms represent an efficient alternative that can also be utilized for the identification of illness outcomes and events. The prediction of CVD has recently made substantial use of machine learning algorithms [15–17]. The suggested machine learning models produce accurate CVD

[1] *How to cite the article:* Potluri N. (June 2024); Predicting Cardiovascular Disease Events with Machine Learning Models; *International Journal of Research in Medical Sciences and Technology;* Vol 17, 67-75, DOI: http://doi.org/10.37648/ijrmst.v17i01.010

predictions that are greater than 90%. More precisely, in [18], the hyOPTXg model used optimized Extreme Gradient Booster and optimization approaches (min-max scaling, OPTUNA: hyper-parameter tweaking) to extract the greatest AUC value, which was equal to 0.947. In a different study [19], Extreme Gradient Boost and Gradient Boost showed the highest AUC value (0.812) among ten ML models used. Also, the results from the ML models were on par with or better than those from the Framingham and ACC/AHA risk models.

Furthermore, Pouriyeh et al. [20] employed various machine learning techniques on the Cleveland Heart Disease database, including Naıve Bayes (NB), Multilayer Perceptron (MLP), Decision Tree (DT), K-Nearest Neighbor (K-NN), Radial Basis Function (RBF), Support Vector Machine (SVM), and Single Conjunctive Rule Learner (SCRL). The bagging, boosting, and stacking techniques were used to assess the effectiveness of individual classifiers as well as the combination of these classifiers. When all classifiers are compared, SVM has the best accuracy and SCRL has the lowest. accuracy of 84.15% and 69.56%, in that order. Furthermore, anytime the bagging method is used, SVM remains the best strategy with the same accuracy %.But with 78.54%, DT is the worse model in this instance. Using bagging, SVM has also improved to 84.81%.After stacking, the MLP and SVM combination of classifiers turned out to be the most accurate, with an accuracy rating of 84.15%.

In this work, we use only basic biomarkers typically obtained in the Ludwigshafen Risk and Cardiovascular Health (LURIC) dataset [21] to predict death due to CVD after 10 years of follow-up. Specifically, we used six different machine learning techniques, and Logistic Regression showed the best results. The findings indicate that the accuracy and area under the receiver's operating characteristic curve have mean values of 72.20% and 72.97%, respectively.

SUPPLIES AND TECHNIQUES

A. The entire process

The analysis's procedure is shown in Fig. 1. The LURIC dataset, which comprises patients slated for coronary angiography, served as our starting point. The preparation of the data using feature selection and cleaning methods came next.

Afterwards, we employed machine learning methods and addressed class imbalance. We next computed how well ML algorithms performed in foretelling death from CVD. These findings are a part of the TIMELY project, which attempts to calculate the mortality and risk score for cardiovascular disease in people with a 10-year risk.

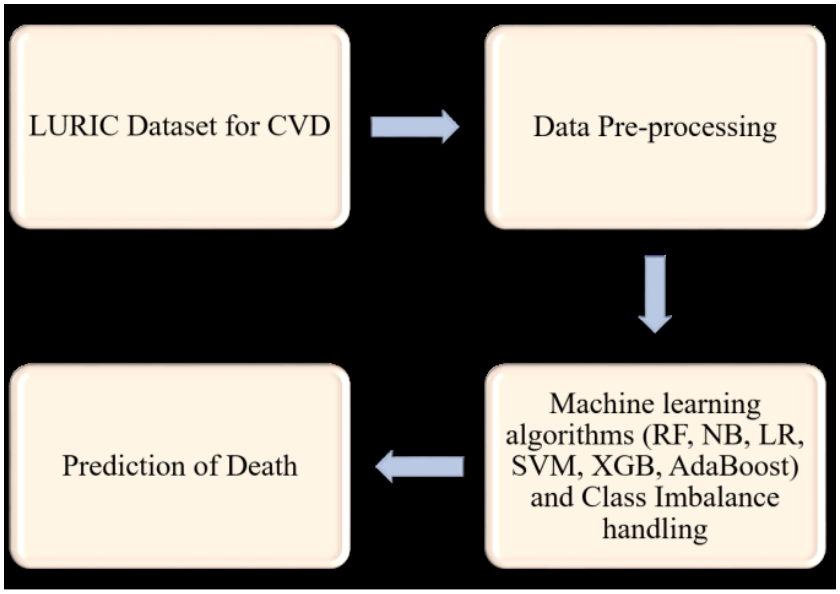

LURIC Dataset for CVD
Data Pre-processing
Machine learning algorithms (RF, NB, LR, SVM, XGB, AdaBoost) and Class Imbalance handling
Prediction of Death

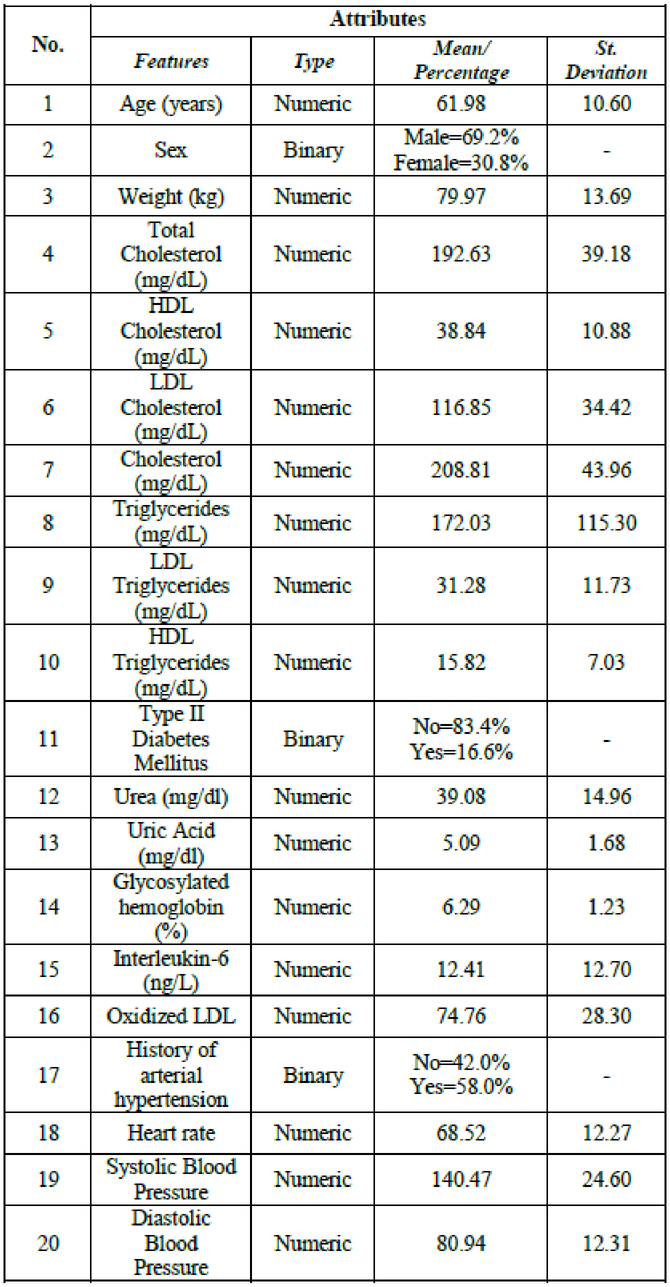
Figure 1 shows the current study's working diagram.

B. Description of the data

Ludwigshafen Risk and Cardiovascular Health (LURIC) cohort medical data were used in this study [21]. 3,023 characteristics, such as demographics, daily routines, biomarkers (inflammatory and molecular), genomics, T-cells, antibodies, and interleukins, were noted for 3,316 patients. Annotation of deaths from CVD after ten years is also included. As a result, in our study, annotation is the target and is divided into three groups. Class zero shows patients who are still living, Class 1 shows patients who have passed away from cardiovascular disease, and Class 3 shows patients who have passed away from another illness. We only included the first and zero classes, which had 484 and 2547 patients, respectively. Based on the significance of the characteristics, a feature selection technique was used to choose every feature. Table I displays these characteristics along with their mean values. The clinician has easy access to the clinical data used in this analysis.

This indicates that the results that were given line up with routine clinical practice.

Table I: The Detailed Description of the Dataset's Attributes (LURIC).



No.	Attributes			
	Features	*Type*	*Mean/ Percentage*	*St. Deviation*
1	Age (years)	Numeric	61.98	10.60

2	Sex	Binary	Male=69.2%
			Female=30.8%
				-

3	Weight (kg)	Numeric	79.97	13.69
4	Total Cholesterol (mg/dL)	Numeric	192.63	39.18
5	HDL Cholesterol (mg/dL)	Numeric	38.84	10.88
6	LDL Cholesterol (mg/dL)	Numeric	116.85	34.42
7	Cholesterol (mg/dL)	Numeric	208.81	43.96
8	Triglycerides (mg/dL)	Numeric	172.03	115.30
9	LDL Triglycerides (mg/dL)	Numeric	31.28	11.73
10	HDL Triglycerides (mg/dL)	Numeric	15.82	7.03

11	Type II Diabetes Mellitus	Binary	No=83.4%
			Yes=16.6%
				-

12	Urea (mg/dl)	Numeric	39.08	14.96
13	Uric Acid (mg/dl)	Numeric	5.09	1.68
14	Glycosylated hemoglobin (%)	Numeric	6.29	1.23
15	Interleukin-6 (ng/L)	Numeric	12.41	12.70
16	Oxidized LDL	Numeric	74.76	28.30

17	History of arterial hypertension	Binary	No=42.0%
			Yes=58.0%
				-

18	Heart rate	Numeric	68.52	12.27
19	Systolic Blood Pressure	Numeric	140.47	24.60
20	Diastolic Blood Pressure	Numeric	80.94	12.31

C. Pre-processing of data

One of the main components of our investigation has been the pre-processing of the data. In order to create a balanced dataset, a variety of strategies were used, including feature scaling, under sampling, feature selection using the Select K Best method, and cleaning using the Simple Imputer class.

We employed the down sampling strategy in our investigation. When there is a suitable sample size of data, this method is used [22]. All samples were retained in the minority class one and choosing an equal number of samples at random from the majority class zero. This process is repeated until the observations from the majority and minority classes are balanced, resulting in a new dataset with a balanced 1:1 ratio for additional analysis. Ultimately, 484 deceased patients and 484 randomly selected living patients were subjected to the analysis. Following this procedure, 40% of the dataset was split into training and test sets.

D. Machine learning algorithms under supervision

The Random Forest (RF) [23], Logistic Regression (LR) [24], Support Vector Machine (SVM) [25], Naïve Bayes (NB) [26], Extreme Gradient Boost (XGB) [27], and Adaptive Boost (AdaB) [28] are the six supervised machine learning classifiers that were selected for the current study. Most people agree that these models are suitable for predicting danger.

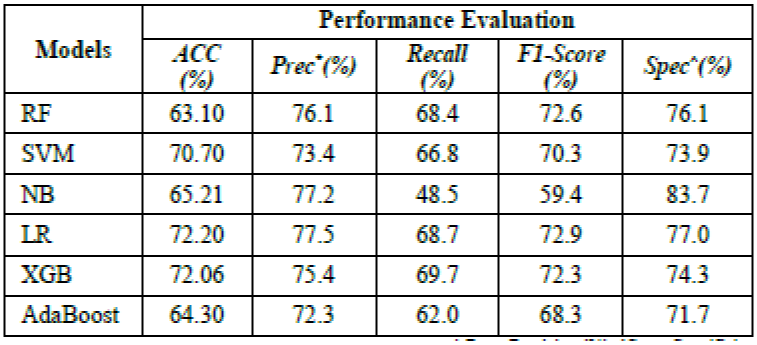
E. Metrics for performance assessments

Using conventional criteria such as Accuracy (ACC), Precision, F1-Score, Sensitivity/Recall, and Specificity, we estimated the performance in our analysis. Moreover, each classifier's performance has been compared using the Receiver Operative Characteristic Curve (ROC) and the Area under the ROC curve (AUC).

RESULTS

The Python software package was utilized to extract the results for all categorization models. We made particular use of the reliable scikit-learn 1.0.2 library, which offers effective methods for analyzing predictive data. In Table II, every measurement value is displayed and compared. The values that are displayed are the average values across ten runs.

Table II: Results of the Performance Prediction for Every Module



Models	Performance Evaluation				
	ACC (%)	*Prec*(%)*	*Recall (%)*	*F1-Score (%)*	*Spec^(%)*
RF	63.10	76.1	68.4	72.6	76.1
SVM	70.70	73.4	66.8	70.3	73.9
NB	65.21	77.2	48.5	59.4	83.7
LR	72.20	77.5	68.7	72.9	77.0
XGB	72.06	75.4	69.7	72.3	74.3
AdaBoost	64.30	72.3	62.0	68.3	71.7

Moreover, Fig. 2 displays the accuracy mean values. According to the experiment results, out of all the classifiers examined, Random Forest and Logistic Regression had the lowest accuracy at 63.10% and the best accuracy at 72.20%, respectively. For patients with a 10-year history of CVD, the best model to predict the mortality from CVD is the logistic regression classifier.

For every applied machine learning approach, the area under the receiver operating characteristic curve (AUC) has also been computed and is displayed in Fig. 3. The maximum level of accuracy is indicated by a model with an AUC value that approaches 1.

When comparing the AUC values, It is observed that the Logistic Regression model has the highest AUC value, whereas the Naïve Bayes model has the lowest AUC value.

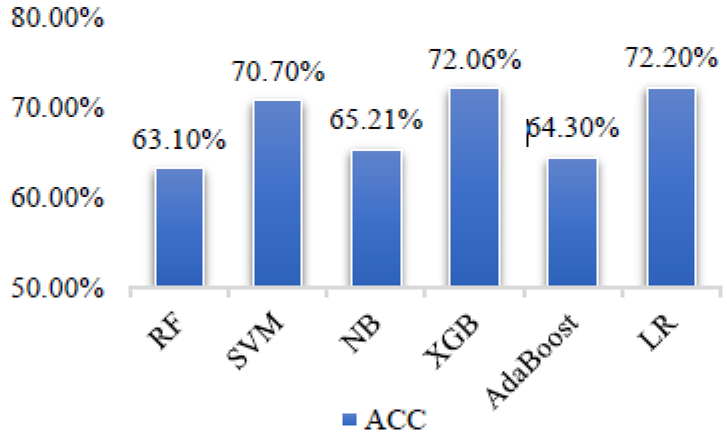


Figure 2 shows the six classifiers' average accuracy values.

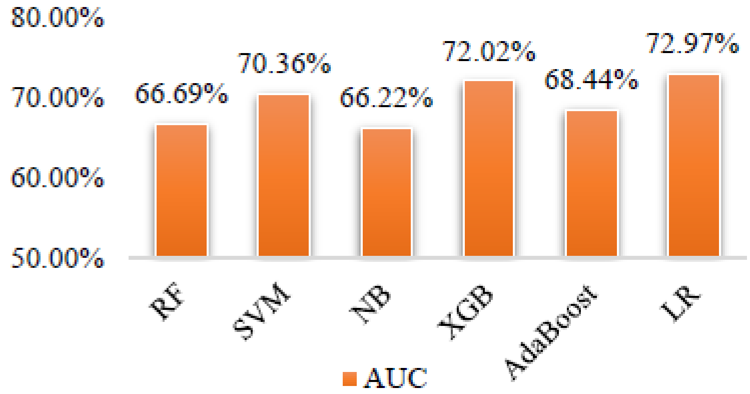


Figure 3 shows the average receiver operative characteristic curve area values for six distinct classifiers.

Additionally, a ROC curve analysis can be used to visualize the outcomes. In an illustrative run of the algorithm, we plotted the ROC curves for every classifier (Figure 4).

In addition, we calculated the receiver operating characteristic (ROC) curve to determine the models' efficiency. Based on the algorithm's indicative run, the AUC values of 0.726 and 0.706 were roughly equalized for Support Vector Machine and Logistic Regression, respectively.

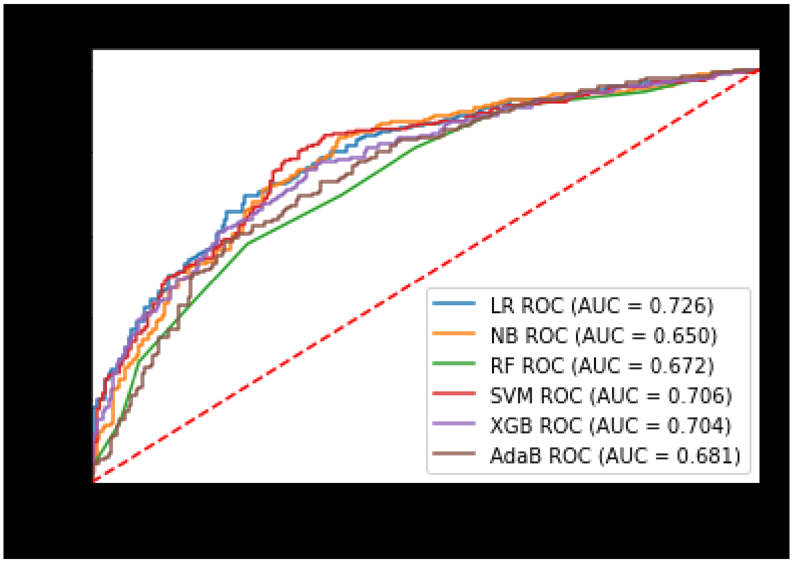


Figure 4 shows the prediction performance for each classifier in an indicative algorithm run as represented by the receiver operating characteristic curve.

CONCLUSION

keeping an eye on those with a history of cardiovascular the most crucial strategy for lowering patient mortality and preventing new cases is disease prevention. In this study, machine learning techniques were used to estimate the 10-year risk of cardiovascular mortality for patients scheduled for angiography. Prior to being compared, the six machine learning models were used and evaluated with a range of parameter values to achieve the best accuracy. Every classifier's effectiveness has been assessed. Logistic regression fared the best out of the six strategies, with the highest average accuracy (72.20%) and AUC value (72.97%). The employment of several ML models in parallel, including both contemporary (like XGB) and conventional algorithmic models (like LR), is innovative in this work. seeking to forecast the 10-year mortality from CVD by using only readily obtained biomarkers in ordinary clinical practice. Our main objective has been to use various machine learning approaches on a short dataset to forecast the mortality of CVD and to select the most effective predictive computational model by comparing them.

The population size and the lack of optimization tools were among the restrictions. Based on ten consecutive runs, the AUC and LR mean accuracy values in this research somewhat outperform those of XGB. In a subsequent analysis, we hope to incorporate 100 runs to further improve the validity and accuracy of the findings. Ultimately, creating a risk score is a future objective. Furthermore, inside the TIMELY project, the developed Coro predict score will be evaluated and compared with its calculated values in the LUC dataset.

REFERENCES

[1] World Health Organization. (2017). Cardiovascular Diseases (CVDs). [Online]. Available online: https://www.who.int/healthtopics/ cardiovasculardiseases/ (accessed on 04 January 2022).

[2] E. J. Benjamin et al., ‘‘Heart disease and stroke statistics—2019 update: A report from the American heart association,’’ Circulation, vol. 139, no. 10, pp. 56–528, Mar. 2019, doi: 10.1161/CIR.0000000000000659.

[3] Eurostat Statistics Explained, Cardiovascular diseases statistics. Available online: ://ec.europa.eu/eurostat/statisticsexplained/ index.php?title=Cardiovascular_diseases_statistics

[4] N. Garg , “Comparison of different cardiovascular risk score calculators for cardiovascular risk prediction and guideline recommended statin uses”, Indian Heart Journal, vol. 69 no. 4, pp. 458-453, Jul.-Aug. 2017, doi: 10.1016/j.ihj.2017.01.015.

[5] SCORE2 working group and ESC Cardiovascular risk collaboration, “SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe”, European Heart Journal, vol. 42, no. 25, pp. 2439–2454, Jul. 2021, doi: 10.1093/eurheartj/ehab309.

[6] S. Livingstone, “Effect of competing mortality risks on predictive performance of the QRISK3 cardiovascular risk prediction tool in older people and those with comorbidity: external validation population cohort study”, The Lancet. Healthy longevity, vol. 2, no.6, pp.352-361, Jun. 2021, doi: 10.1016/S2666-7568(21)00088-X.

[7] Y. S. Chen et al., “Identification of the Framingham Risk Score by an Entropy-Based Rule Model for Cardiovascular Disease”, Entropy, vol. 22, no.12, p. 1406, Dec. 2020, doi: 10.3390/e22121406.

[8] P. Paul et al, “Cardiovascular Risk Prediction using JBS3 Tool: A Kerala based Study”, Current medical imaging, vol.16, no. 10, pp. 1300-1322, 2020, doi: 10.2174/1573405616666200103144559.

[9] Writing Committee Members, “2020 ACC/AHA Guideline for the Management of Patients with Valvular Heart Disease: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines”, Journal of the American College of Cardiology, vol. 77, no. 4, pp. 25-197, Feb. 2021, doi: 10.1016/j.jacc.2020.11.018 .

[10] S. M. Green, “A Methodological Appraisal of the HEART Score and Its Variants”, Annals of Emergency Medicine, vol.78, no. 2, pp. 253- 266, Aug. 2021, doi: 10.1016/j.annemergmed.2021.02.007.

[11] The WHO CVD Risk Chart Working Group, “World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions”, The LANCET Global Health, vol.7, no. 10, Oct. 2019, pp. 1332-1345, doi:10.1016/S2214-109X(19)30318- 3.

[12] European Commission-CORDIS (2018). Final Report Summary - RISKYCAD (Personalized diagnostics and treatment of high risk coronary artery disease patients.), Available online: https://cordis.europa.eu/project/id/305739/reporting.

[13] Y. Wang et al., “Comparison of MESA of and Framingham risk scores in the prediction of coronary artery disease severity”, Original Articles, vol.43, no.1 pp.139-144, Dec. 2019, doi: 10.1007/s00059-019-4838-z.

[14] S. Selvarajah et al., “Comparison of the Framingham Risk Score, SCORE and WHO/ISH cardiovascular risk prediction models in an Asian population”, International Journal of Cardiology, vol.176, no.1, pp. 211-218, Sep. 2014, doi:10.1016/j.ijcard.2014.07.066.

[15] M.Amzad Hossen et al., “Supervised Machine Learning-Based Cardiovascular Disease Analysis and Prediction”, Mathematical Problems in Engineering, vol. 2021, pp.1-10, Dec. 2021, oi.org/10.1155/2021/1792201.

[16] N. Fitriyani et al., “HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System”, IEEE Access, vol. 8, pp. 133034- 133050, Jul.2020, doi:10.1109/ACCESS.2020.3010511.

[17] K.Sivaraman, V.Khanna, “Machine Learning Models for Prediction of Cardiovascular Diseases”, International Conference on Physics and Energy 2021 (ICPAE 2021), vol. 2040, 2021, doi:10.1088/1742- 6596/2040/1/012051.

[18] P. Srinivas, R. Katarya, “hyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using XGBoost”, Biomedical Signal Processing and Control, vol.73, p.103456, Mar. 2021, doi: 10.1016/j.bspc.2021.103456.

[19] J. O. Kim et al., “Machine Learning-Based Cardiovascular Disease Prediction Model: A Cohort Study on the Korean National Health Insurance Service Health Screening Database” diagnostics, vol. 11, no.6, p.943, May 2021, doi: 10.3390/diagnostics11060943.

[20] S. Pouriyeh et al., “A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease” 22nd IEEE Symposium on Computers and Communication (ISCC 2017), Jul. 2017, doi: 10.1109/ISCC.2017.8024530.

[21] B. R. Winkelmann et al., “Rationale and design of the LURIC study— a resource for functional genomics, pharmacogenomics and long-term prognosis of cardiovascular disease”, Pharmacogenomics, vol. 2, no. 1 Suppl 1, pp. 71-73, Feb. 2001, doi: 10.1517/14622416.2.1.S1.

[22] Haibo He, Yunqian Ma, Imbalanced Learning: Foundations, Algorithms, and Applications. 1st ed. Wiley-IEEE Press. 2013. 26 p.

[23] A. Chaudhary, “An improved random forest classifier for multi-class classification”, Information Processing in Agriculture, vol. 3, no. 4, pp. 215-222, Dec. 2016.

[24] Y. Yang, M. Wu, “Explainable Machine Learning for Improving Logistic Regression Models”, 2021 IEEE 19th International Conference on Industrial Informatics (INDIN), Jul. 2021, doi: 10.1109/INDIN45523.2021.9557392.

[25] S. Suthaharan. Support Vector Machine. In: Machine Learning Models and Algorithms for Big Data Classification. Integrated Series in Information Systems, vol. 36, pp. 207-235, Boston: Springer, 2016, doi:10.1007/978-1-4899-7641-3_9.

[26] D. Barrer, “Bayes' Theorem and Naive Bayes Classifier”, Encyclopedia of Bioinformatics and Computational Biology, 2019, doi:10.1016/B978-0-12-809633-8.20473-1.

[27] T. Chen, C. Guestrin, “XGBoost: A Scalable Tree Boosting System”, KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, Aug. 2016, doi: 10.1145/2939672.2939785.

[28] Y. Cao, “Advance and Prospects of AdaBoost Algorithm”, Acta Automatica Sinica, vol. 39, no. 6, Jun. 2013, pp. 745-758, doi: 10.1016/S1874-1029(13)60052-X.