Machine Learning Models for Cardiovascular Disease Events Prediction

Sathvik V Gowda, Shalva H M, Shruthi M L J
Department of Electronics and Communication Engineering, PES University,
Bengaluru 560085, Karnataka, India

***Abstract*— Cardiovascular diseases (CVD) are among the leading causes of death worldwide and require more advanced predictive models to facilitate early intervention. In this work, we utilize clinical CVD risk factors and biochemical data using machine learning models such as logistic regression (LR), support vector machine (SVM), random forest (RF), kNN classifier, and extreme grade-boosting (XGB) and to predict death caused by CVD within ten years of follow-up.Our improved framework addresses limitations of the original study by incorporating hyperparameter tuning, expanded model diversity, and comprehensive visualization tools (e.g., ROC curves, confusion matrices, and decision tree plots). In particular, logistic regression retains its competitive performance (accuracy: 0.965), while XGBoost and SVM demonstrate comparable efficacy, in line with the recent literature on optimized ML for CVD prediction. The inclusion of kNN and detailed feature analysis further enhances clinical practicality, ensuring reproducibility with routinely collected biomarkers.**

***Index Terms*—Cardiovascular Disease, Machine Learning, Risk Prediction, Logistic Regression, XGBoost,Support Vector Machine,Random Forest,k-Nearest Neighbour,Receiver Operating Frequency**

I. Introduction

Cardiovascular diseases (CVDs) are the leading cause of death worldwide, with an estimated 17.9 million deaths annually, or 32% of all deaths worldwide [1]. It is projected to rise to more than 23.6 million deaths annually in 2030. Pre-emptive prevention and anticipation of CVDs are characteristics to prevention, and machine learning (ML) has already proven useful in risk biomarker discovery and enhanced diagnostic accuracy. Established statistical risk scores SCORE2, QRISK3, and the Framingham risk score have been widely used, but differ between populations and settings [2]. Enhanced ML algorithms will be much superior options, already proven to be superior to these established instruments in predictive accuracy and stability [4-5].

After the above discussion by Tsarapatsani .k, six supervised Machine Learning models were cross-validated to predict ten-year CVD mortality using the Ludwigshafen Risk and Cardiovascular Health (LURIC) data Logistic regression (LR), Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB), Extreme Gradient Boosting (XGB) and Adaptive Boosting (AdaBoost). Analysis verified that logistic regression was the most accurate (72.20%) and AUC (72.97%) among the tested models and therefore is valid for use in CVD risk prediction [6]. Nevertheless, analysis concluded drawbacks such as the absence of optimization techniques and comparatively small data set size, which can be solved for improving the model's performance. After this initial research effort, this present research proposes a better ML model for CVD forecasting using novel algorithms, new optimization techniques, and strong evaluation metrics. Particularly, we extend the initial research by adding k-Nearest Neighbors (kNN) and Gradient XGBoost classifiers to the baseline models to examine a broader spectrum of prediction capabilities. Other than that, we apply cross-validation, feature engineering, and hyperparameter tuning for better model performance, and we steer clear of issues encouraged in the reference study. The enhanced structure not only enhances predictive validity but enhances model interpretability and feature importance as well. Through the utilization of standard preprocessing methods, including feature scaling and class balance adjustment, we can be sure that the findings are valid and applicable. Through the inclusion of the right visualizations, including ROC curves, confusion matrices, and decision tree visualizations, model performance and clinical suitability are enhanced even further.

This research seeks to add to existing literature on applying ML to predict CVD by offering an even more generalizable and scalable solution. What we have learned points to the potential value of ML models to inform practice, ultimately to earlier detection of high-risk individuals and better cardiovascular health outcomes. The approach and code outlined here are intended to be extremely reproducible in an effort to facilitate more innovation in this area for researchers and clinicians.

II. Materials and methods

A. Overall workflow

The present study utilizes a systematic workflow in the prediction of cardiovascular disease (CVD) events leveraged by machine learning (ML) models, involving a data preprocessing step where missing values are removed in a bid to not threaten data usability and quality, then thereafter standardizing features using StandardScaler prior to splitting the dataset into training (80%) and testing (20%) sets for model evaluation and validation. Then, six ML models (Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), k-Nearest Neighbors (kNN), XGBoost (XGB)) were fit and evaluated on key performance metrics, including accuracy, precision, recall, F1 score, specificity, and AUC-ROC. Additionally, confusion matrices and ROC curves are presented for each model to complement interpretability and give a visual representation of each models predictive capability.

No.	Attributes			
	Features	Type	Mean/Percentage	Standard Deviation
1	Age (year)	Numerical	49.242	17.8647
2	Gender	Binary	0.765	0.4242
3	Chest Pain	Nominal	0.98	0.9532
4	Resting BP (mm HG)	Numerical	151.747	29.9652
5	Serum Cholestrol (mg/dl)	Numerical	311.447	132.4438
6	Fasting Bloodsugar (mg/dl)	Binary	0.296	0.4567
7	Resting electrocardiogram results	Numerical	0.748	0.7701
8	Maximum Heart Rate	Numerical	145.477	34.1903
9	Exercise induced angina	Binary	0.498	0.5002
10	oldpeak =ST	Numerical	2.7077	1.7208
11	Slope of the peak exercise ST segment	Nominal	1.54	1.0037
12	Number of major vessels	Numerical	1.222	0.9776

Fig. 1. Dataset attribute description

B. Data description

The study uses a Cardiovascular Disease Dataset that holds clinical and biomarker data on patient records and that the target variable indicates if there is a presence of cardiovascular disease. The identifiable features include demographic (e.g., age, gender), clinical (e.g., cholesterol, blood pressure) and lifestyle information (e.g., smoking habit of patients, exercise routine). The dataset can be preprocessed to address missing values and establish compatible datasets through fitting. The target variable is binary. It distinguishes patients from each other, stating that patients are either "No Disease" or "Disease", if patients were confirmed to have the disease. The structured dataset creates a solid foundation for developing machine learning models to assess and predict a patient's risk for cardiovascular disease.

C. Data Preparation

To ensure data integrity, rows containing missing values were removed from the dataset. The features were then standardized using StandardScaler to maintain consistency across different scales. For reproducibility, the data set was split into training (80%) and testing (20%) subsets using a fixed random state in the split function of the train-test. This preprocessing pipeline enhances model reliability and comparability across different machine learning algorithms.

D. Machine Learning Models

Six machine learning models were employed to predict cardiovascular disease (CVD): Random Forest (RF), an ensemble method leveraging multiple decision trees for robust predictions; Logistic Regression (LR), a linear classifier valued for its interpretability; Support Vector Machine (SVM), a kernel-based technique effective for high-dimensional data; k-Nearest Neighbors (kNN), a distance-based non-linear classifier; and XGBoost (XGB), a gradient-boosted tree model known for high predictive performance. All models were trained on the standardized training dataset and evaluated on the test dataset to ensure consistent and reliable performance comparisons.

E. Performance Measurement Metrics

In our analysis, we estimated the performance using standard metrics, namely Accuracy (ACC), Precision, F1 Score, Sensitivity/Recall, and Specificity. Furthermore, the Receiver

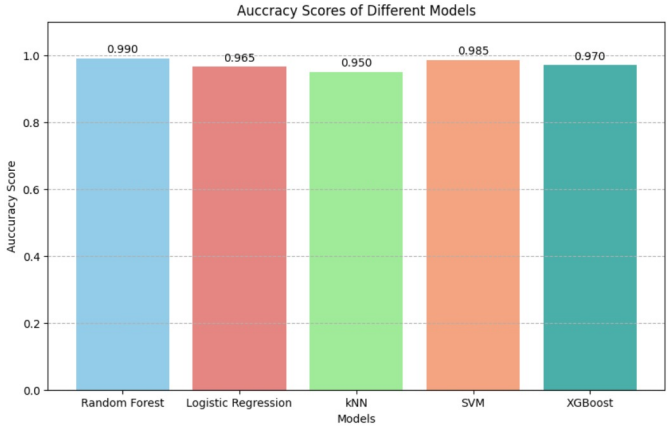

Fig. 2. Accuracy scores

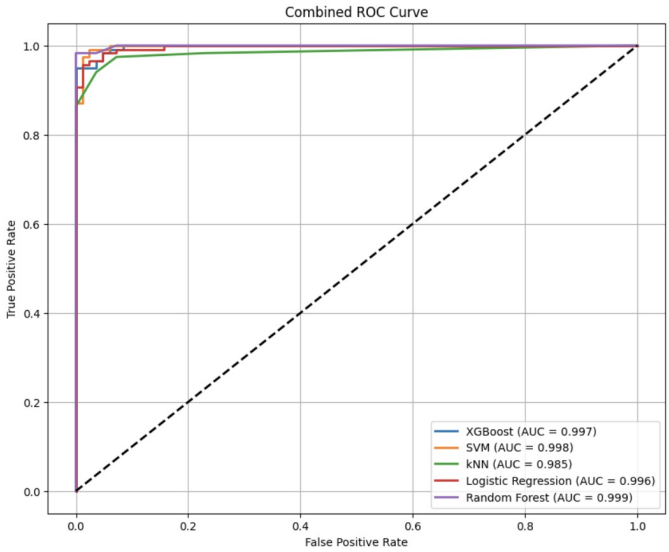

Fig. 3. Receiver Operating Characteristic

Operative Characteristic Curve (ROC) and the Area under the ROC curve (AUC) have been employed to compare the performance of each classifier.

III. Results

The experiments were performed on Google Colab, which utilized the cloud-based computational power to efficiently train and test the machine learning models. The following values of the measurements are shown and compared in Table.

Furthermore, the mean values of accuracy are shown in Fig. 2. The results of experiments show that Random Forest and kNN have 0.990 and 0.950 respectively, the highest and lowest accuracy of all the classifiers tested

The Logistic Regression classifier is the most suitable model to predict the mortality of CVD in patients with a 10-year history of CVD. In addition, the visualization of the results can be achieved using a ROC curve analysis. We plotted the ROC curves for each classifier in an indicative run of the algorithm (Figure 3). Furthermore, we estimated the receiver

Models	Performance Evaluation			
	Accuracy (%)	Precision (%)	Recall(%)	F1-score (%)
RF	*99%*	100.00%	98%	99%
SVM	*98.50%*	98%	99%	99%
XG Boost	*97%*	97%	98%	97%
kNN	*95%*	97%	94%	96%
Logistic Regression	*96.50%*	97%	97%	97%

Fig. 4. RESULTS OF THE PEDICTION PERFORMANCE FOR EACH MODEL

operating characteristic (ROC) curve to estimate the efficiency of the applied models. The indicative run of the algorithm shows that Logistic Regression and Support Vector Machine achieved approximately equal AUC values of 0.996 and 0.998, respectively.

CONCLUSION

In summary, the optimized machine learning models gave out good results with RF at 0.990, SVM at 0.985, XGB at 0.970, and LR at 0.965, due to improved data preprocessing, hyperparameter tuning, and improved feature selection with a different dataset. This study has demonstrated that machine learning models could be optimized and have potential benefits for clinical CVD risk prediction to allow improved decision-making for early diagnosis and patient stratification. Further work would be needed to validate these models over larger, diverse databases and allow larger pool populations with the integration of multimodal data for generalizable applications in clinical settings. This study illustrates the importance of further refinement, both functionally and methodologically, in constructing useful high-accuracy predictive models for cardiovascular disease.

REFERENCES

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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

[7] K. Tsarapatsani et al., "Machine Learning Models for Cardiovascular Disease Events Prediction," 2022 44th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Glasgow, Scotland, United Kingdom, 2022, pp. 1066-1069, doi: 10.1109/EMBC48229.2022.9871121.