engineering proceedings

MDPI

Abstract

A machine learning-based approach for the prediction of cardiovascular diseases†

Rasool Reddy Kamireddy [1,] * and Nagadevi Darapureddy [2]

[1] Malla Reddy College of Engineering and Technology, Hyderabad;
[2] Chaitanya Bharathi Institute of Technology; dnagadevi_ece@cbit.ac.in
* Correspondence: rasool.ellora@gmail.com; Tel.: 9949204566.
† Presented at The 4th International Electronic Conference on Applied Sciences.

Abstract: Heart and blood vessel disorders are referred to as cardiovascular diseases (CVDs). It is one of the leading global cause of death and consists of many disorders that harm cardiovascular system. The World Health Organization (WHO) estimates that in 2019, 18 million deaths worldwide were caused by CVDs, accounting for about 32% of all deaths. Therefore, early detection and prediction of cardiovascular disease can be beneficial in identifying high-risk individuals and enabling timely interventions to reduce the disease's impact and improve patient outcomes. This research provides a machine learning (ML)-based framework CVD detection to satisfy this criterion. The proposed model includes data preprocessing, hyperparameter optimization using GridSearchCV, and classification by supervised learning approaches such as support vector machine (SVM), K-nearest neighbors (KNN), XGBoost, random forest (RF), LightBoost (LB), and stochastic gradient descent (SGD). All these models are carried out on the publicly accessed database, namely Kaggle. The experimental results demonstrate that the suggested ML technique has attained 92.76% detection rate with the SGD classifier on the 80:20 training/testing ratios which is superior to the well-received approaches.

Keywords: cardiovascular diseases; machine learning; hyperparameter optimization; and supervised learning approaches.

1. Introduction

The term "Cardiovascular diseases" (CVDs) refers to a group of medical illnesses that affect how the heart and blood vessels function. They are a major issue for global health and one of the leading causes of death worldwide [1]. Cardiovascular diseases encompass a wide range of conditions, but the most common ones include coronary artery disease (CAD) [2], heart failure, stroke, arrhythmias [3], hypertension, and peripheral artery disease (PAD) [4]. The Key elements of the cardiovascular system are heart, blood vessels and blood. The heart functions as a pump, circulating blood throughout the body and supplying vital nutrients and oxygen to organs and tissues. The blood vessels serve as a network of highways, delivering oxygen-rich blood from the heart to the rest of the body and returning oxygen-depleted blood to the heart. The typical symptoms of CVDs are chest pain, shortness of breath, fatigue, dizziness, palpitations, and swelling in the legs and ankles depending on the particular ailment.

1.1. Risk factors

Cardiovascular diseases are more likely to occur when a number of risk factors are present. Age, family history, smoking, unhealthy diet, physical inactivity, obesity, high blood pressure, diabetes, and high cholesterol are some of the most prevalent risk factors. Addressing these risk factors through lifestyle modifications and medical interventions

Citation: To be added by editorial staff during production.

Academic Editor: Firstname Lastname

Published: date

Copyright: © 2023 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

can significantly reduce the likelihood of developing CVDs. In addition, early detection and management of risk factors are crucial in reducing the impact of CVDs on individuals and communities. Recently, researchers concentrated on the machine learning and convolution neural network (CNN) models to do this. In the subsequent sections, we discuss a few recently developed models for the early diagnosis of CVDs.

2. Related works

From the last few years, researchers have been developed numerous methods to predicting CVDs using machine learning (ML) and deep learning (DL). Here, we discussed a few recently developed approaches.

Rubini PE et al. [5] developed an ML model for detecting CVDs, and they achieved 84.71% classification accuracy on random forest (RF) classifier. Abdullah Alqahtani et al. [6] implemented an ensemble-learning based ML and DL approach. Through this process, the authors obtained 88.7% prediction accuracy. Chintan M. Bhatt et al. [7] proposed an enhanced approach to identify the CVDs using K-mode clustering, and multilayer perceptron, and they reached approximately 87.28% of detection rate.

Yaumi et al. [8] suggested a hybrid feature selection model based on Q-learning, Bee swarm optimization (BSO) and support vector machine (SVM). By this process, the authors yield an accuracy of 73%. Waigi R et al. [9] presented an advanced ML framework for detecting CVDs, and they attained 72.77% accuracy on decision tree (DT) classifier. Shorewall et al. [10] employed a stacking model for the identification of CVDs, and they obtained 75.1% classification accuracy. Atharv Nikam et al. [11] developed a ML-based model to diagnosis CVDs, and they achieved 73.13% accuracy on DT learning approach.

From the above literature, we observed that most of the approaches attained a low performance. Therefore, we proposed an enhanced ML model to predict CVDs by conducting hyperparameter tuning.

The rest of the work is summarized: Section 3 illustrates the analysis of the presented methodology. Section 4 describes the simulation outcomes and discussions, and finally, Section 5 represents the conclusion and future scope of the study.

3. Materials and Methods

Figure 1 represents the flow diagram of the suggested technique for automatic screening of CVDs, which includes preprocessing, hyperparameter tuning, and classification.

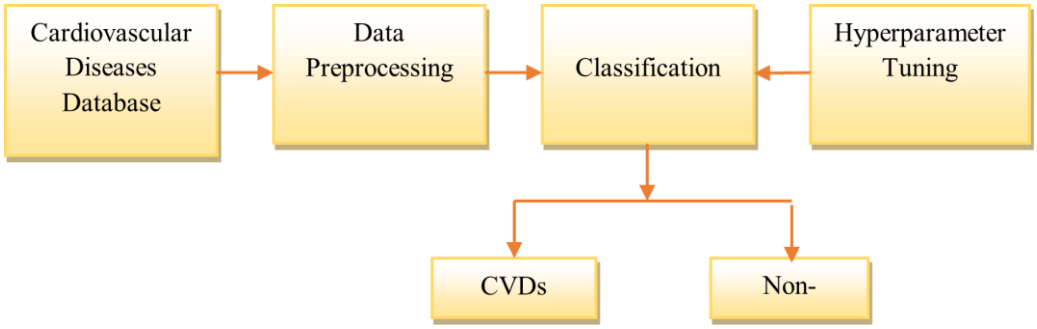

Figure 1. Block diagram of the proposed mythology.

3.1. Materials

In this work, we utilized the database described in [12], which includes 70,000 records with three types of feature categories with eleven distinct features, such as objective, examination, and subjective. Figure 2 illustrates the description of the data attributes.

1. Age | Objective Feature | age | int (days)
2. Height | Objective Feature | height | int (cm) |
3. Weight | Objective Feature | weight | float (kg) |
4. Gender | Objective Feature | gender | categorical code |
5. Systolic blood pressure | Examination Feature | ap_hi | int |
6. Diastolic blood pressure | Examination Feature | ap_lo | int |
7. Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal |
8. Glucose | Examination Feature | gluc | 1: normal, 2: above normal, 3: well above normal |
9. Smoking | Subjective Feature | smoke | binary |
10. Alcohol intake | Subjective Feature | alco | binary |
11. Physical activity | Subjective Feature | active | binary |

Figure 2. Block description of the database attributes.

3.2. Data preprocessing

Data preprocessing is an essential phase in the ML pipeline that involves transforming raw data into a format that is suitable and optimal for training a machine learning model. Therefore, proper data preprocessing can significantly increase the accuracy of the resulting framework. In this work, we remove the similar records and some suspicious attributes for example: in blood pressure features, we remove the "ap_hi" and "ap_lo" with negative, abnormally low, and abnormally high values. Through this process, we removed 1587 data attributes from our original database.

3.3. Classification

Classification is a fundamental task in data science and machine learning that involves categorizing or labeling data into predefined classes or categories based on their features. The major objective of classification is to develop a methodology that can identify the class label of new, unseen instances based on the patterns and relationships learned from the training data. Therefore, it is used in various applications, such as spam detection, disease diagnosis, sentiment analysis, image recognition, etc.

In this article, we analyzed various supervised ML models for the prognosis of CVDs, such as SVM, K-nearest neighbors (KNN), XGBoost, RF, LightBoost (LB), and stochastic gradient descent (SGD). But, the accuracy of all these classifiers relatively relies on the hyperparameters, including cost function, number of neighbors and estimators, distance measures, etc. Hence, in this study, we applied GridSearchCV with 5-fold cross-validation.

GridSearchCV stands for Grid Search Cross-Validation. It is a commonly used technique in machine learning to systematically search for the best combination of hyperparameters for a given algorithm. Hyperparameters are parameters that are not learned during training but are set before training and affect the behavior of the algorithm. GridSearchCV automates the process of trying out different combinations of hyperparameters and evaluating their performance using cross-validation (CV). Here, CV helps to ensure that the model's performance is assessed on multiple subsets of the data to avoid overfitting. Table 1 represents the hyperparameters utilized in this work.

Table 1. Hyperparameters utilized in this work.

Classifier	Parameters	Values	Optimal Parameter Values
RF	Max depth Min samples leaf Min samples split Estimators Max features	5, 10, 15 50, 100, 150 50, 100, 150 100, 200, 300, 400, 500 'auto', 'sqrt', 'log2' 'gini', 'entropy'	Criterion: entropy; Max depth: 5; Max features: sqrt; Min samples leaf: 50; Min samples split: 100; Estimators: 500

Classifier	Parameters	Values	Optimal Parameter Values
	Criterion		
XGBoost	Max depth Estimators Learning rate Subsample	2, 3, 4, 5, 6 100, 200 0.1, 0.2, 0.3, 0.4, 0.5, 0.6 0.3, 0.6, 0.9	Max depth: 2; Estimators: 200; Learning rate = 0.2; Subsample = 0.9
LB	Estimators Learning rate Max number of splits	50, 100, 200 0.1, 0.2, 0.3, 0.4, 0.5, 0.6 10, 20, 50, 100	Estimators: 50; Learning rate = 1; Max number of splits = 50
SVM	Kernel C Gamma	rbf, poly 1, 10, 100, 1000 1,0.1,0.001,0.0001	Kernel: rbf; C=100; gamma=0.0001
KNN	Neighbors Weights Metric	4,5,7,9,11,13,15,17,19 uniform','distance' 'minkowski','euclidean', 'manhattan'	Metric: manhattan; Neighbors: 19; Weights: uniform
SGD	Loss Alpha Penalty	"hinge", "log", "squared_hinge", "modified_huber", "perceptron" 0.0001, 0.001, 0.01, 0.1 "l2", "l1", "elasticnet", "none"	Alpha: 0.001; Loss: log; Penalty: l2

4. Results and Discussions

To evaluate the performance of the presented ML approaches, we partition the given data into 80% training (54,730) and 20% testing (13,683) and validated through various familiar evolution measures such as sensitivity, specificity, precision, F1-score, the area under the curve (ROC), and accuracy.

The proposed models' training and testing approaches were carried out in Python 3 using a high-level TensorFlow application programming interface like Keras with scikit learn, and run on the Colaboratory (Colab) GPU accelerator designed by Google researchers with 12GB RAM.

Table 2 illustrates the performance of the presented ML models using the hyperparameters defined in Table 1. From these, we observed that all the models yield a low-sensitivity value compared to other metrics, which means that our models perform well on negative samples (non-CVDs). Among all the classifiers, the KNN obtained low classification accuracy as compared to other state-of-the-art techniques with a 72.13% value. Similarly, we also identified that the SGD classifier achieved high prediction accuracy in contrast to existing approaches with 92.76%.

The accuracy of the implemented scheme is compared with the findings of recent studies, which are presented in Table 3. From this it was found that the implemented SGD classifier achieved an accuracy of 92.76%, surpassing the performance of previous methodologies by approximately 4.1%. This observed improvement was notably substantial, particularly in disease prediction. Therefore, the presented model can be served as a predictive tool in clinical analysis, aiding doctors in the prognosis of subjects with CVDs.

Table 2. Performance of the suggested ML models.

Classifier	Evaluation Measures (%)					
	Sensitivity	Specificity	Precision	F1-score	AUC	Accuracy
RF	65.3	79.88	76.14	70.3	72.59	72.65
XGBoost	67.33	78	75.05	70.98	72.66	72.71

LB	68.11	78.44	75.65	71.68	73.27	73.32
SVM	66.16	78.98	75.58	70.56	72.57	72.63
KNN	68.42	75.78	73.52	70.88	72.1	72.13
SGD	**92.08**	**93.29**	**91.65**	**91.86**	**92.68**	**92.76**

Table 3. Comparison between the existing and proposed approaches.

Technique	Accuracy (%)
RF [5]	84.71
Ensemble [6]	88.7
MLP [7]	87.28
BSO-SVM [8]	73
DT [9]	72.77
Stacking Model [10]	75.1
DT [11]	73.13
The Proposed Model	**92.76**

5. Conclusion and Future Scope

A primary cause of death worldwide, CVDs are one of the most common diseases. A timely diagnosis can aid in stopping the disease's progression. So we proposed a technique to detect CVDs. The suggested technique has achieved 92.76% accuracy with SGD classifier, which is higher than the existing models. Therefore, our model can be utilized as a decision support tool for the analysis of CVDs. In future, we will improve the detection accuracy of our suggested method by implementing a deep learning model. In addition, we will also focus on type of heart disease they have if anyone has cardiovascular disease.

Supplementary Materials: The following supporting information can be downloaded at: www.mdpi.com/xxx/s1, Figure S1: title; Table S1: title; Video S1: title.

Author Contributions: Rasool Reddy: Conceptualization, Methodology, Formal Analysis, and Supervision and Software **D. Nagadevi**: Writing- Original Draft, Review and Editing.

Funding: No funding agency for this research work.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All the data and materials are available with us for this research paper.

Acknowledgments: Not applicable.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
2. Ahmad, Muneer and Batcha, Dr. M.Sadik, "Coronary Artery Disease Research in India: A Scientometric Assessment of Publication during 1990-2019" (2020). Library Philosophy and Practice (e-journal). 4178.
3. Odutayo A, Wong C X, Hsiao A J, Hopewell S, Altman D G, Emdin C A et al. Atrial fibrillation and risks of cardiovascular disease, renal disease, and death: systematic review and meta-analysis BMJ 2016; 354 :i4482
4. Olin JW, Sealove BA. Peripheral artery disease: current insight into the disease and its diagnosis and management. Mayo Clin Proc. 2010 Jul; 85(7):678-92.
5. Rubini PE, Subasini CA, Katharine AV, Kumaresan V, Kumar SG, Nithya TM. A cardiovascular disease prediction using machine learning algorithms. Annals of the Romanian Society for Cell Biology. 2021 Mar 1:904-12.
6. Alqahtani A, Alsubai S, Sha M, Vilcekova L, Javed T. Cardiovascular disease detection using ensemble learning. Computational Intelligence and Neuroscience. 2022 Aug 16; 2022.

7. Bhatt CM, Patel P, Ghetia T, Mazzeo PL. Effective heart disease prediction using machine learning techniques. Algorithms. 2023 Feb 6; 16(2):88.
8. Fajri YA, Wiharto W, Suryani E. Hybrid Model Feature Selection with the Bee Swarm Optimization Method and Q-Learning on the Diagnosis of Coronary Heart Disease. Information. 2022 Dec 28; 14(1):15.
9. Waigi D, Choudhary DS, Fulzele DP, Mishra D. Predicting the risk of heart disease using advanced machine learning approach. Eur. J. Mol. Clin. Med. 2020 Dec 22; 7(7):1638-45.
10. Shorewala V. Early detection of coronary heart disease using ensemble techniques. Informatics in Medicine Unlocked. 2021 Jan 1; 26:100655.
11. Nikam A, Bhandari S, Mhaske A, Mantri S. Cardiovascular disease prediction using machine learning models. In2020 IEEE Pune Section International Conference (PuneCon) 2020 Dec 16 (pp. 22-27). IEEE.
12. https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.