

Asian Journal of Research in Computer Science

11(1): 17-27, 2021; Article no.AJRCOS.70004
ISSN: 2581-8260

An Improved Coronary Heart Disease Predictive System Using Random Forest

Abdulraheem Abdul[1*], Rafiu M. Isiaka[1], Ronke S. Babatunde[1] and Jumoke F. Ajao[1]

[1]Department of Computer Science, Kwara State University, Malete, Nigeria.

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJRCOS/2021/v11i130253
Editor(s):
(1) Dr. Teresa De Pilli, University of Foggia, Italy.
(2) Dr. Dariusz Jacek Jakóbczak, Koszalin University of Technology, Poland.
Reviewers:
(1) Sunday Adeola EMALEKU, Adekunle Ajasin University, Nigeria.
(2) Asia H.Al-Mashhadani, Iraq
Complete Peer review History: https://www.sdiarticle4.com/review-history/70004

Original research Article

Received 23 May 2021
Accepted 29 July 2021
Published 11 August 2021

ABSTRACT

Aims: This work aim is to develop an enhanced predictive system for Coronary Heart Disease (CHD).
Study Design: Synthetic Minority Oversampling Technique and Random Forest.
Methodology: The Framingham heart disease dataset was used, which was collected from a study in Framingham, Massachusetts, the data was cleaned, normalized, rebalanced. Classifiers such as random forest, artificial neural network, naïve bayes, logistic regression, k-nearest neighbor and support vector machine were used for classification.
Results: Random Forest outperformed other classifiers with an accuracy of 98%, a sensitivity of 99% and a precision of 95.8%. Feature selection was employed for better classification, but no significant improvement was recorded on the performance of the classifier with feature selection. Train test split also performed better that cross validation.
Conclusion: Random Forest is recommended for research in Coronary Heart Disease prediction domain.

Keywords: Coronary heart disease; machine learning; random forest; artificial neural network; K-nearest neighbor; support vector machine; and naïve bayes.

**Corresponding author: E-mail: kolabdu@gmail.com;*

1. INTRODUCTION

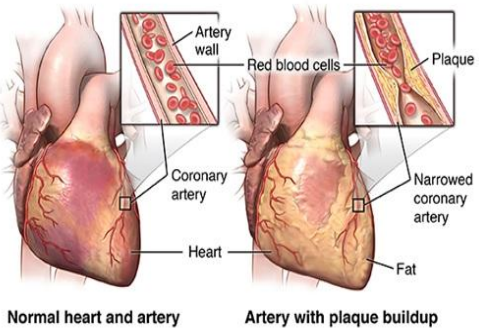


Fig. 1. Coronary-artery-disease

Cardiovascular is a compound-word in Latin, Cardio means heart while vascular means vessel. The heart is a muscular organ in the body; it is the muscular pumping machine that pumps and circulates blood throughout our body. Cardiovascular system also comprises a network of blood vessels, for example, veins, arteries, and capillaries. These blood vessels deliver blood all over the body [1]. All diseases involving this system are called Cardiovascular Diseases (CVDs). There are many diseases that can affect the cardiovascular system including stroke, heart failure, hypertensive heart disease, rheumatic heart disease, cardiomyopathy, abnormal heart rhythms, congenital heart disease, valvular heart disease, carditis, aortic neurysms, peripheral artery disease, thromboembolic disease, and venous thrombosis, [2] asserted CHD is the most prominent of the CVDs. Al-Numan ibn Bashir reported from Muhammad to have said "Verily, in the body is a piece of flesh which, if good the whole body is good and if spoilt the whole body is spoilt, truly the piece of flesh is the heart". (The Hadith, n.d)

Furthermore, world health organization (WHO) in 2017 proclaimed CVDs are the number 1 cause of death globally: more people die annually from CVDs than from any other cause. An estimated 17.9 million people died from CVDs in 2016, representing 31% of all global deaths. Of these deaths, 85% are due to CHD. Over three quarters of CVDs deaths take place in low- and middle-income countries.CVD is the leading cause of death in South Africa after HIV/AIDS. The mortality that results from CVDs in South Africa is so much that it is more than the mortality that results from all the cancers combined [3]. WHO (2017), confirmed CHD Deaths in Nigeria reached 76,410 amounting to 3.76% of total deaths in that year. Heart disease is easier to treat when detected early so, if CVDs can be predicted before patients go down with this deadly disease, measures will be put in place to avert it. [1]. States of the heart equipment are needed for the diagnose CHD according to Nhongo, Hendricks, Bradshaw, & Bail, [4], another factor worsening the effect of the disease in Nigeria is the country is far behind in the doctor patient ratio. The novel COVID-19 is another reason for tackling CHD because worst hit victims are those with underlining illness, especially CVDs, but the good news about the disease is that you don't have to compulsorily undergo clinical diagnosis before identifying the disease, since it can be predicted from indicators [5], which is why Machine Learning is used in this field to forecast and diagnose the disease.

Machine Learning (ML) is vital in the formation of medical support systems. This is as a result of the vast amount of data collected in the hospital. These information house the medical history of patients. Machine learning can be used to mine this information to uncover traces of disease in individual [6].

A lot of works have been done with ML in the domain of CHD, this study divide the works as including novel algorithm development, algorithm enhancement, hybridization, implementation and system development, comparative analysis and interface development.

A novel algorithm was developed [7] which was tested with the Alizadehsani dataset against famous classifiers like support vector machine, C4.5 and Naive Bayes, the new algorithm outperformed the famous algorithms, recording a sensitivity value of 100% and accuracy rate of 96.40%. This study has no interface where users can test for their CHD status.

Convolutional neural network was enhanced by [8], the study was conducted using randomly subsampled dataset and iterates LASSO multiple times. Majority voting system was used to identify the variables that are nonzero in major number of iterations. The enhanced algorithm outperformed multilayer perceptron. No interface was designed for this model and the accuracy could also be improved.

Merger of more than one supervised machine learning algorithm was also experimented to diagnose CHD according to the work of [9] who conducted a study using the Cleveland Heart disease dataset, on a hybridization of Naïve

Bayes and SVM which gives an accuracy of 89%.

A lot of design and implementation were done while intermarrying CHD and machine learning to improve diagnosis. Some of the works include that of [10], the study used ten thousand and thirty patients data with suspected CHD. Boosted ensemble algorithm was used. No user interface was designed

[11] also designed and implemented using the Hungarian Institute of Cardiology, Budapest and the Cleveland Clinic Foundation datasets was employed and a fuzzy expert system was build to aid interpretation of result.
Many authors did extensive comparative analysis on supervised machine learning methods used in the domain of CHD. The work of Ayatollahi et al.,[12], shows that SVM outperformed ANN on a dataset of three selected hospitals affiliated to AJA University of Medical Sciences in Iran.

[13] also compared ensemble algorithms with other classification algorithms when difficulty was express in diagnosing CHD based on the risk factor, ensemble algorithms did better than the other algorithms.

[7], compared programming expression and genetic algorithm emotional neural when it is known that angiography is the known treatment of CHD, however, it is very expensive and have very high risk. The study used the Alizadehsani dataset genetic programming expression was reported to outperform genetic algorithm emotional neural.

Some works in this domain have gone to the extent of designing interface for the implemented system to ease the use of the system for its intended end-user. Among this effort is the work of Chen et al.,[14], designed a system using the Cleveland dataset from UCI was used, with ANN as the classifier and reported accuracy of classification as 80% as well as 85% sensitivity and 70% specificity. The dataset employed is small and the accuracy could improve.

Another interface was built by [6], who designed a system using the Cleveland Heart disease dataset, random forest to rank features and used fuzzy logic for modeling, 50% correct classification accuracy was reported wish is quiet low and the dataset is also small.

[15] deployed a system which gives a yes or a no for the presence or absence of CHD respectively in the patient, when it is observed that a rampant increase in the CHD rates at juvenile ages and that it is impractical for a common man to frequently undergo costly tests like the ECG, using the Cleveland Dataset a system was designed using neural network and it gives a yes or a no for the presence or absence of CHD respectively in the patient.

2. MATERIALS AND METHODS

This work is based on an adapted heart disease prediction model as shown in Fig. 2 below. The model developed by Tarawneh and Embarak, [9]. The model is made up of three phases and each phase has sub-phases. The first phase is the preprocessing stage and has two sub-phases, namely: cleansing and feature engineering. The phase that follow is the classification phase which also has within it evaluation of classifiers, elimination of weak classifiers and hybridization of the best two classifiers. After which diagnosis is made.

2.1 Data Preprocessing Phase

The data preparation phase has sub-phases which include data cleansing, normalization, rebalancing and feature selection.

2.1.1 Data cleansing

Data cleansing is prompted because the data contain missing values and its output are supplied to the normalization sub-phase. The missing values in the data are replaced using the average of the feature that has missing values.

2.1.2 Scaling phase

The data used was scaled or normalized because of the variation in their ranges to improve the performance of our classifier. The technique employed in this task is the Standard Scalar. The result is then passed to the rebalancing sub-phase.

2.1.3 Rebalancing data

Real world datasets are characterized with imbalance class distribution; this is why this study decided to rebalance the dataset. The technique used for class rebalancing is Synthetic Minority Over-sampling Technique (SMOTE). An instance of the rebalancing function will be created and data will be passed to the initiated instance in the form of x input and y output. After which feature engineering will be carried out.

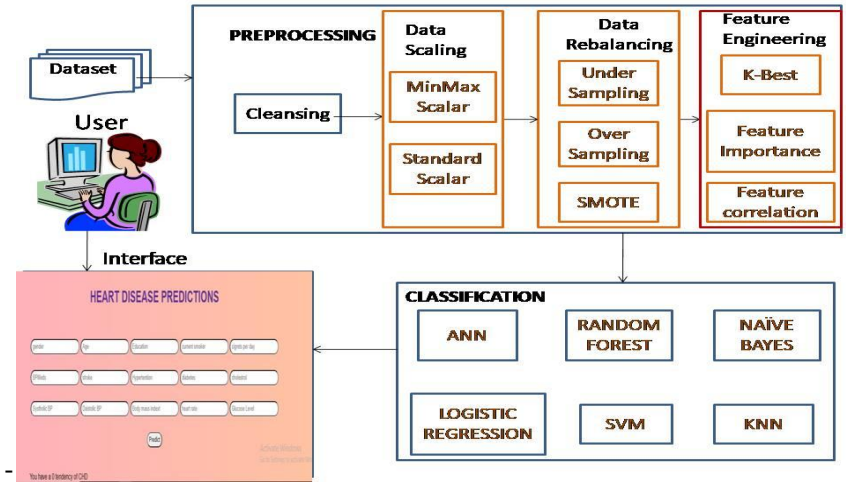


Fig. 2. Proposed heart disease predictive model

2.1.4 Feature engineering

Feature engineering is also carried out in the preprocessing stage to improve the accuracy of the classifier. Visha,[16] describes correlation as a statistical term, which refers to how close two variables are to having a linear relationship with each other; when two features have high correlation, we can drop one of the two features. The result of the entire preprocessing stage will serve as input in the classification phase.

2.2 Classification Phase

In this phase, the result of the preprocessing phase is supplied to different classifiers which will be evaluated to determine the best as experimented in this research. The classifiers as earlier mentioned are artificial neural network (ANN), support vector machine (SVM), k-nearest neighbor (KNN), Random forest, Naïve bayes and logistic regression.

2.3 Data Acquisition

This work is based on a dataset from a cardiovascular study on residents of the town of Framingham, Massachusetts. The goal of the study is to envisage the possibility of a patient having coronary heart disease (CHD). Even though there are other dataset with different feature that suit this study like the Cleveland CVD dataset which this work will use as the smaller dataset, but the choice of this particular dataset is born out the features making-up the dataset. It includes 4,240 records with fifteen (15) attributes the sixteenth been the target or the outcome which is if the patient will eventually have CHD or not. The dataset is publicly available in http//:kaggle.com.

The dataset includes the demographic risk factors, listed in Table 1.

2.4 Model Implementation

It is found while describing the dataset that there is an imbalance class distribution on the data set, 3596 account for healthy people while 644 are stricken with the disease.

There are also 645 missing values in the dataset which was filled by the mean of the attribute with missing value. These problems are taken care of in the preprocessing stage as explained in the subsequent sections.

2.4.1 Data preprocessing phase using python programming language

Python programming language was used in the implementation of this work, using the Jupyter notebook in Anaconda Environment. The data was loaded using the *pandas* python library in a tabular format because the data originally is in a comma separated value (CSV).Exploratory data analysis was carried out on the data to provide a better understanding of the data and its characteristics. Such analysis include; a check on size of the data with the *shape()*command, displaying 4240 by 16 as output. The output displayed by the *describe* command indicates that the data has missing values, with much variations in the range of values and is highly imbalanced.

These findings prompt the study to do the following:

2.4.1.1 Data cleansing

As observed in the exploratory stage stated above, the missing values in the data are replaced with the average of the feature that has missing values

2.4.1.2 Data normalization

The first feature added was the data scalar which was as a result of attribute values appearing in different ranges, such as stroke, bearing Boolean value 0 and 1, Level of education, bearing value between 1 and 4, age range having values between 32 and 70, glucose range having values between 40 and 394. This variation calls for normalization which ameliorates the disparity that could sway the learning algorithm. In the light of this, there's a need to scale the data to set all our values within the range of -1 and 1. Scaling of values of features will also reduce the computational complexity of the proposed system. The techniques employed in this task are the MinMax Scalar and the Standard Scalar. The data is passed to the MinMax Scalar function in the python library, which essentially shrinks the range such that the range is now between 0 and 1. The Standard Scalar on the other hand assumes your data is normally distributed within each feature and will scale the data such that the distribution is now centered on a standard deviation of 1.

2.4.1.3 Data rebalancing

Another feature added to the model is data rebalancing which became necessary due to class imbalance observed during the exploratory data analysis. The total number of sample is 4240and the class with CHD denoted by 1 is 644 while the class without CHD denoted by 0 is 3596. The techniques used for class rebalancing include NearMiss, SMOTE and Random Over-sampling. These work by creating an instance of the rebalancing function and data will be passed on to the initiated instance in the form of x input and y output.

2.5 Classification Phase

The preprocessed data will serve as input into the classification phase. The Sickitlearn (sklearn) is a library in python designed ML tasks and it provides a uniform interface to ML models and there is a common pattern that can be reused across different models. The data that is coming will be divided into the training and testing data using the sklearn model_selection mudule method tain_test_split. This method takes as argument the x features, the y label and the percentage of the split. The python statement for this is shown in the following lines of code:

From sklearn.model_selection import train_test_split x_train, x_test, y_train, y_test = train_test_split(x,y,test_size=0.2)

The first step in classification is to import the relevant class, in this case random forest classifier is imported from sklearn.ensembles, sklearn carefully organized into modules which will make it easy to find the class needed.

From skleran. ensemble import Random Forest Classifier

The second step in the sklearn patter is to instantiate the classifier

Clf = Random Forest Classifier

At this stage the classifier was trained, here the classifier learns the relationship between the feature and the response value. The fit method is used to train the classifier and the fit method takes two arguments.

Clf.fit (x_train, y_train).

2.6 System Evaluation

The new models performance is tested using accuracy, sensitivity and precession, these metrics are further explained below.

2.6.1 Accuracy

In addition to the fit method, there is a score method which is used to measure the accuracy of a classifier, the score method also receive two arguments the testing input features X and the testing response value Y

Clf.score (x_test, y_test).

Sklearn's model have uniform interface which means the same steps pattern can be used on a different model with ease. To use another model simply import the mlpclassifier from the nueral network module of sklearn, instantiate the classifier, fit the model with training data and score it with testing data.

2.6.2 Sensitivity

Sensitivity is the ratio of correctly positive label to the entire positive class. How sensitive is the classifier to detecting positive instances. This tells when the actual value is positive how often the prediction is correct.

Metrics.recall_score(y_test, y_pred)

2.6.3 Precision

Precision tells when a positive value is predicted, how often is the prediction correct, it is derived by

metrics.precision_score(y_test, y_pred)

3. RESULTS AND DISCUSSION

The adapted model was tested on the raw data Framingham dataset and the result obtained using the six earlier mentioned classifiers are shown in Table 2.

It can be observed that all the algorithms performed very well on the raw data after filling in missing values so that our classifier will be able to work. The best performance is produced by logistic regression with an accuracy of 85.8% while naïve bayes gave us the least performance of 83.3%, which is still good. The sensitivity of the classifiers is very poor with naïve bayes having the highest sensitivity of 17% and SVM gave a sensitivity of 16%. The precision is better with logistic regression producing a score as high as 71% and the least performance in term of precision as recorded by KNN with a score of 33.3%.

In all cases train test split performed better than cross validation, but after rebalancing the data cross validation produced better result with random forest, ANN and KNN.

The Cleveland dataset on the other hand, produce very good result too with the lowest performance recorded by SVM with an accuracy score 66.6% while random forest gave us a score as high as 98.1%. The sensitivity and precision are also very good, logistic regression and ANN paired with a score of 90.6% while the lowest recorded is from naïve bayes and random forest with a score of 84.3. The precision on the other hand is high as well with naïve bayes producing the best result with a score of 90.0% and the least score is 66.6% from SVM.

In comparing the Framingham dataset against Cleveland dataset, the accuracy are both high but the precision and recall of the Framingham dataset are very low. This result from the nature of class distribution of both dataset, the Framingham Dataset is an unbalance dataset while the Cleveland dataset is fair.

After rebalancing the dataset, both the sensitivity and precision of the Framingham dataset also increased. Random forest now scoring 99% and 95% respectively.

Table 1. The dataset includes the demographic risk factors

Feature Type	
Feature	**Meaning/ Description**
Sex	It has variables male or female with value 1 and 0 respectively
Age	The age of the patient
Educational level	This is coded 1 for high school, 2 for a high school diploma, 3 for some college or vocational school, and 4 for a college degree.
currentSmoker	This tells if the patient is a current smoker or not.
cigPerDay	This is the number of cigarettes that the patient smoked on average in one day.
bpMeds	This tells whether or not the patient was on blood pressure medication
PrevalenceofStroke	This tells whether or not the patient had previously had a stroke
prevalenceofHyp	This tells whether or not the patient was hypertensive
Diabetes	This indicate whether or not the patient had diabetes
totChol	total cholesterol level
sysBp	systolic blood pressure
diaBp	diastolic blood pressure
BMI	Body Mass Index
heartRate	Heart rate
Glucose	Glucose level

Table 2. Performance result of classifiers on raw dataset

Classification Algorithm	Train_test_split Accuracy (%)	Cross validation accuracy (%)	Recall (%)	Precision (%)
Random Forest	85.6	84.9	4.8	42.8
Artificial Neural Network	85.0	84.3	4.8	37.5
Support Vector Machine	85.6	84.7	1.6	66.6
Naïve Bayes	83.3	82.3	17.0	35.0
Logistic Regression	85.8	85.0	4.0	71.4
K-Nearest Neighbor	84.1	83.7	8.9	33.3

Table 3. Result of evaluating estimators (Train/test split and cross validation)

Classification Algorithm	Train_test_split Accuracy (%)	Cross Validation Accuracy (%)
Random Forest	85.2	84.9
Artificial Neural Network	85.0	84.3
Support Vector Machine	85.6	84.7
Naïve Bayes	83.3	82.3
Logistic Regression	85.8	85.0
K-Nearest Neighbor	84.1	83.7

Table 4. Performance result of classifiers on smaller dataset

Classification Algorithm	Train_test_split Accuracy (%)	Cross Validation Accuracy (%)	Recall (%)	Precision (%)
Random Forest	83.6	98.1	84.3	84.3
Artificial Neural Network	86.8	80.5	90.6	85.2
Support Vector Machine	70.4	66.0	87.5	66.6
Naïve Bayes	86.8	80.5	84.3	90.0
Logistic Regression	88.5	83.1	90.6	87.8
K-Nearest Neighbor	68.8	81.4	75.0	68.5

Table 5. Performance of Framingham vs. performance of cleveland dataset

Algorithms	Framingham Accuracy (%)	Cleveland Accuracy (%)
Random forest	85.2	83.6
ANN	85.0	86.8
SVM	85.6	70.4
Naïve bayes	83.3	86.8
Logistic Regression	85.8	88.5
ANN	84.1	68.8

Table 6. Testing on imbalanced and balanced dataset

Algorithms	Accuracy rebalanced (%)	Recall rebalanced (%)	Precision rebalanced (%)	Accuracy imbalanced (%)	Recall imbalanced (%)	Precision imbalanced (%)
Random forest	98	99	95.8	85.2	4.8	42.8
ANN	65.1	43.9	67.9	85	4.8	37.5
SVM	65.8	66.8	67.5	85.6	1.6	66.6
Naïve bayes	82.3	27	74.5	83.3	17	35
Logistic Regression	64	53.1	59.8	85.8	4	71.4
KNN	81.4	92.6	73.4	84.1	8.9	33.3

Table 7. Result of performance after scaling the dataset

Classification Algorithm	Train_test_split Accuracy (%) Framingham (Before scaling)	Train_test_split Accuracy (%) Framingham (After scaling)	Train_test_split Accuracy (%) Cleveland (Before scaling)	Train_test_split Accuracy (%) Cleveland (After scaling)
Random Forest	85.2	85.1	83.6	81.9
Artificial Neural Network	85.0	84.7	86.8	85.2
Support Vector Machine	85.6	85.2	70.4	86.8
Naïve Bayes	83.3	83.3	86.8	86.8
Logistic Regression	85.8	85.6	88.5	85.2
K-Nearest Neighbor	84.1	82.7	68.8	90.1

Table 8. Classifier performances after rebalancing

Classification Algorithm	Train_test_split Accuracy (%)	Cross Validation Accuracy (%)	Recall (%)	Precision (%)
Random Forest	97.2	98.0	99.0	95.8
Artificial Neural Network	62.6	65.1	43.9	67.9
Support Vector Machine	67.6	65.8	66.8	67.5
Naïve Bayes	84.0	82.3	27.0	74.5
Logistic Regression	68.2	64.0	53.1	59.8
K-Nearest Neighbor	81.0	81.4	92.6	73.4

Table 9. Result of performance of classifiers after feature engineering

Classification Algorithm	Train_test_split Accuracy (%)	Cross Validation Accuracy (%)	Recall (%)	Precision (%)
Random Forest	85.2	84.8	6.5	44.4
Artificial Neural Network	85.7	84.5	4.0	45.4
Support Vector Machine	85.6	84.8	1.6	66.6
Naïve Bayes	84.4	83.5	27.0	74.5
Logistic Regression	86.0	84.8	4.8	85.7
K-Nearest Neighbor	83.8	82.7	9.7	31.5

Both dataset were scaled to test the effect of scaling on the dataset, scaling in this sense is shown not to have a significant effect on the performance of the classifiers.

The Framingham dataset is characterized with imbalance class distribution which affects classifier performance negatively. Here is a result of the performance of classifiers after rebalancing the dataset. Random forest returned the best performance with an accuracy of 97.2% and a sensitivity of 99% and a precision of 95.8%.

After dropping the correlated features, sensitivity and precision increased why accuracy remain fairly the same.

4. CONCLUSION

Coronary heart disease which is one of the number killers of man is diagnosed in this work. Many studies that have tried to contain the disease are also reviewed. In this work, the CHD predictive system has been enhanced. The model was designed using random forest classifier which gave an accuracy of 98%, 99% recall and 95.8% precision.

This system is recommended for the poor masses who cannot afford the cost of CHD diagnosis and for those who also want to test their vulnerability to CHD. In future study, a dataset of Nigerians will be collected and used to train the classifier. The random forest can also be used to diagnose other deadly diseases as well.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Nashif S, Raihan MR, Islam MR, Imam MH. Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system. World Journal of Engineering and Technology. 2018;6(4):854-873.
2. Murray CJ, Abraham J, Ali MK, Alvarado M, Atkinson C, Baddour LM, Birbeck G. The state of US health, 1990-2010: burden of diseases, injuries, and risk factors. Jama. 2013;310(6):591-606.
3. Pillay-Van Wyk V, Laubscher R, Msemburi W, Dorrington RE, Groenewald P, Vos T, Nannan N. Second South African national burden of disease study: Data cleaning, validation and SA NBD List: South African medical research council Cape town; 2014.
4. Nhongo D, Hendricks J, Bradshaw J, Bail K. Leadership and registered nurses (RN s) working after-hours in residential aged care facilities (RACF s): A structured literature review. Journal of Clinical Nursing. 2018;27(21-22):3872-3881.
5. Grallert H, Dupuis J, Bis JC, Dehghan A, Barbalic M, Baumert J, Roberts R. Eight genetic loci associated with variation in lipoprotein-associated phospholipase A2 mass and activity and coronary heart disease: meta-analysis of genome-wide association studies from five community-based studies. European Heart Journal. 2012;33(2):238-251.
6. Mokeddem SA. A fuzzy classification model for myocardial infarction risk assessment. Applied Intelligence. 2018;48(5):1233-1250.
7. Jalali SMJ, Khosravi A, Alizadehsani R, Salaken SM, Kebria PM, Puri R, Nahavandi S. Parsimonious evolutionary-based model development for detecting artery disease. Paper presented at the 2019 IEEE International Conference on Industrial Technology (ICIT);2019.
8. Dutta A, Batabyal T, Basu M, Acton ST. An efficient convolutional neural network for coronary heart disease prediction. Expert Systems with Applications. 2020;159:113408.
9. Tarawneh M, Embarak O. Hybrid approach for heart disease prediction using data mining techniques. Paper presented at the International Conference on Emerging Internetworking, Data & Web Technologies;2019.
10. Motwani M, Dey D, Berman DS, Germano G, Achenbach S, Al-Mallah MH, Slomka PJ. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. European Heart Journal. 2017;38(7):500-507.
DOI: 10.1093/eurheartj/ehw188
11. Muthukaruppan S, Er MJ. A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease. Expert Systems with Applications. 2012;39(14):11657-11665.
12. Ayatollahi H, Gholamhosseini L, Salehi M. Predicting coronary artery disease: a

comparison between two data mining algorithms. BMC Public Health. 2019;19(1):1-9.

13. Christalin B, Jeeva C. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked. 2019;16:100203.
DOI: 10.1016/j.imu.2019.100203

14. Chen AH, Huang SY, Hong PS, Cheng CH, Lin EJ. HDPS: Heart disease prediction system. Paper presented at the 2011 computing in cardiology;2011.

15. Gavhane A, Kokkula G, Pandya I, Devadkar P. Prediction of Heart Disease Using Machine Learning;2018.

16. Vishal R. Feature selection - Correlation and P-value;2018.
Available:https://towardsdatascience.com/feature-selection-correlation-and-p-value-da8921bfb3cf

© 2021 Abdul et al.; This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Peer-review history:
The peer review history for this paper can be accessed here:
https://www.sdiarticle4.com/review-history/70004