A Comparative Study of Machine Learning Algorithms to Detect Cardiovascular Disease with Feature Selection Method

Chapter · August 2022

DOI: 10.1007/978-981-19-2347-0_45

Badhan Chandra Das
Florida International University
16 PUBLICATIONS **33** CITATIONS
SEE PROFILE

Suman Saha
Bangabandhu Sheikh Mujibur Rahman Digital University, Bangladesh
13 PUBLICATIONS **59** CITATIONS
SEE PROFILE

Al Amin Biswas
Bangabandhu Sheikh Mujibur Rahman University, Kishoreganj
36 PUBLICATIONS **462** CITATIONS
SEE PROFILE

Partha Chakraborty
Comilla University
61 PUBLICATIONS **656** CITATIONS
SEE PROFILE

A Comparative Study of Machine Learning Algorithms to Detect Cardiovascular Disease with Feature Selection Method

Md. Jubier Ali, Badhan Chandra Das, Suman Saha, Al Amin Biswas, and Partha Chakraborty

Abstract Heart disease is considered one of the calamitous diseases which eventually leads to the death of a human, if not diagnosed earlier. Manually, detecting heart disease needs doing several tests. By analyzing the result of tests, it can be assured whether the patient got heart disease or not. It is time consuming and costly to predict heart disease in this conventional way. This paper describes different machine learning (ML) algorithms to predict heart disease incorporating a Cardiovascular Disease dataset. Although many studies have been conducted in this field, the performance of prediction still needs to be improved. In this paper, we have focused to find the best features of the dataset by feature selection method and applied six machine learning algorithms to the dataset in three steps. Among these ML algorithms, the random forest algorithm gives the highest accuracy which is 72.59%, with our best possible feature setup. The proposed system will help the medical sector to predict heart disease more accurately and quickly.

Keywords Heart disease · ML algorithms · Feature selection

1 Introduction

Also known as cardiovascular disease, heart disease is a condition that affects the heart and circulatory system. If there is an abnormality in the heart, that is considered

Md. Jubier Ali (✉) · B. Chandra Das
Bangladesh University of Business and Technology, Dhaka, Bangladesh
e-mail: shondhi1997@gmail.com

S. Saha
Bangabandhu Sheikh Mujibur Rahman Digital University, Kaliakair, Bangladesh

A. A. Biswas
Daffodil International University, Dhaka, Bangladesh

P. Chakraborty
Comilla University, Cumilla, Bangladesh
e-mail: partha.chak@cou.ac.bd

V. Skala et al. (eds.), *Machine Intelligence and Data Science Applications*, Lecture Notes on Data Engineering and Communications Technologies 132,
https://doi.org/10.1007/978-981-19-2347-0_45

heart disease. There are many reasons for heart disease. Every year a significant number of people die due to this fatal disease. If any abnormality in the heart can be detected at an early stage, a person can take the necessary steps to be safe from any devastating effects of it. Thus, the death rate can be controlled. So, it is highly needed to identify heart disease at the initial stage in order to be safe and conscious. Though many types of tests are needed to detect heart disease manually, that is the only conventional way to detect it, which is time consuming and costly. Therefore, an automatic heart disease detection system can ease some sort of difficulties when it comes to time and money.

In this paper, we will be using ML algorithms to predict the heart disease of a person by analyzing the attributes of a cardiovascular dataset. There are several works with the problem, but their accuracy is not so high to the Cardiovascular Disease dataset. In this research, by using the feature selection approach, we will find the different experimental configurations of features. We have used six ML algorithms (Naïve Bayes, Logistic Regression, K-nearest neighbor, Support Vector Machine (SVM), Random Forest, and Decision Tree) in order to detect the heart disease and find out the performance measurement of each ML algorithms. The feature selection method has been used to find out the best features of the dataset based on the scores provided by it. Then, we applied six ML algorithms on the different sets of features separately. Using the feature selection strategy, we have found out our important features. Three different feature sets have been set through the outcomes from the feature selection method.

The following are the main contributions of this paper:

- Using the feature selection approach, we have selected the different experimental configurations for features.
- We have applied six ML algorithms and measured the performance of different experimental configurations for each set of features.
- An extensive experiment has been performed on a cardiovascular dataset and compared the performance of applied ML algorithms.

The remaining parts of this paper is constructed as follows. The related works are described in Sect. 2. Section 3 gives the description of Cardiovascular Disease dataset is included. Section 4 consists of the methodology. Section 5 contains the ML algorithm's evaluation and comparison of the results. In Sect. 6, we have discussed the outcome of the result of the algorithms for all feature configurations. At last, we have concluded the paper in Sect. 7.

2 Related Works

2.1 *Heart Disease Prediction System and Application*

Gavhane et al. [1] discussed how to anticipate heart disease at the initial stage. They proposed to develop an application that can predict heart disease by taking some symptoms as their features. They found neural network has given the most accurate result. Palaniappan et al. [2] proposed an Intelligent Heart Disease Prediction system (IHDPS) by developing a Web site for the IHDPS with the best data mining technique. They used Naïve Bayes, decision tree, and neural network techniques. Shah et al. [3] proposed a system that can easily predict the early diagnosis of heart disease. A dataset from the database which is named as Cleveland database of heart disease patients repository. They considered only 14 important attributes among 73 attributes. Among ML algorithms, KNN gave the highest accuracy. Rajdhan et al. [4] have proposed a system to predict the chances of heart disease and classify the patient's risks. Among some ML algorithms, random forest yielded the highest accuracy which is 90.16%. They also compared the results of different ML algorithms on the dataset. Repaka et al. [5] proposed a system for disease prediction that uses Nave Bayesian (NB) techniques for dataset categorization and the advanced encryption standard (AES) algorithm for safe data transit. Mohan et al. [6] proposed a hybrid technique, dubbed a hybrid random forest with a linear model by (HRFLM). Kelwade et al. [7] have devised a system that uses a radial basis function neural network (RBFN) to predict eight heart arrhythmias. The linear and nonlinear aspects of each arrhythmia's heart rate time series are detected. Jabbar et al. [8] discussed how Hidden Naïve Bayes (HNB) can be used to forecast heart disease. They showed how HNB performed on the heart disease dataset. They also described the algorithm and the performance of the algorithm in different parameters.

2.2 *Comparative Analysis of Machine Learning Algorithms to Predict Heart Disease*

Bhatla et al. [9] have executed a comparative analysis of different types of data mining classifiers and various techniques. Sowmiya et al. [10] cited a comparative study on heart disease dataset to predict the heart disease diagnosis by different data mining techniques. They used some medical symptoms or attributes to predict heart disease. They compared the result of data mining techniques or algorithms. Thomas et al. [11] have shown a comparative study to predict heart cancer using different data mining techniques. They have tried to find the risk rate of heart disease by using these algorithms. Singh et al. [12] have discussed a systematic review of different types of machine learning techniques to predict heart diseases.

Table 1 Features of dataset

Features	Data type	Description
id	Numerical	Anonymous ID of patients
age	Numerical	Age of patients in days
gender	Categorical	Gender of patients. (women or men)
height	Numerical	Patient's height in cm
weight	Numerical	Patient's weight in kg
ap_hi	Numerical	Systolic blood pressure
ap_lo	Numerical	Blood pressure in the diastole
cholesterol	Categorical	Cholesterol status of a patient (normal, above normal, or well above normal)
gluc	Categorical	Glucose status of patient (normal, above normal, or well above normal)
smoke	Categorical	whether patient smokes (1) or not (0)
alco	Categorical	whether patient drinks alcohol (1) or not (0)
active	Categorical	Physically active (1) or not (0)
cardio	Categorical	The result of a candidate value 0: no heart disease value 1: heart disease

2.3 Feature Selection to Predict Heart Disease

Using data science, Bashir et al. [13] proposed a cardiac disease prognosis. Their study focused on feature selection algorithms and strategies for various heart disease prediction datasets.

3 Dataset

We collected the dataset from Kaggle[1]. The dataset called Cardiovascular Disease dataset is formed on a CSV file with 70,000 instances and 12 features. In this dataset, data attributes are the important features that can help us to predict heart disease or cardiovascular disease. The value of the target column is dependent on 12 column features which are discussed in Table 1. There are 12 features such as id, age, gender, height, weight, ap_hi, ap_lo, cholesterol, gluc, smoke, alco, and active. The target column is cardio, which represents whether a patient is affected by heart disease or not. In the dataset, there is a total of 70,000 instances.

Table 2 shows some of the instances of the dataset. The target column "cardio" has two values, they are—0 (no heart disease) and 1(heart disease). It is working as a dependent variable which depends on the other 12 independent variables at its left.

[1] https://www.kaggle.com/sulianova/cardiovascular-disease-dataset.

Table 2 Instances of dataset

ID	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	17474	1	156	56.0	100	60	1	1	0	0	0	0
8	21914	1	151	67.0	120	80	2	2	0	0	0	0
9	22113	1	157	93.0	130	80	3	1	0	0	1	0

4 Methodology

The proposed method of this task has been initiated by the collection of the Cardiovascular Disease dataset. In the next phase, data pre-processing is performed. Then, the feature selection method is used by us, which is the most significant part of our research. Then, we applied the following six ML algorithms, decision tree, logistic regression, SVM, Naive Bayes, random forest, and KNN. Finally, we compared results and discussed the outcomes of those machine learning models. Figure 1 shows the overall procedure of our work.

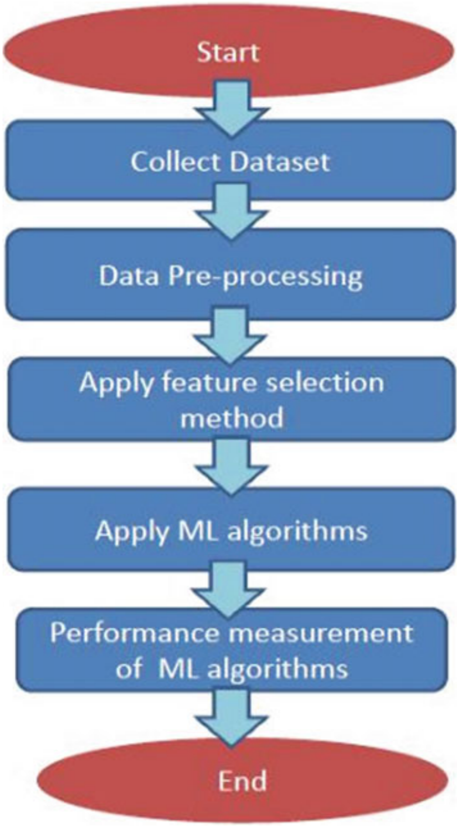
Start
Collect Dataset
Data Pre-processing
Apply feature selection method
Apply ML algorithms
Performance measurement of ML algorithms
End

Fig. 1 Methodology of detecting cardiovascular disease with feature selection method

4.1 Dataset Collection

We collected the dataset from Kaggle. The name of our dataset is the Cardiovascular Disease dataset. In the dataset, there are 70,000 instances. These instances help us to identify whether a patient has heart disease or not.

4.2 Data Pre-processing

Data pre-processing is provided in a comprehensible manner by transforming raw data into an understandable context for a specific goal. We performed several pre-processing steps like data cleaning (null and incomplete instances removal, redundancy elimination), outlier removal, and data transformation on the dataset.

4.3 Feature Selection

Finding a subset of input features that are most correlated with the target feature can be done by the feature selection method. After the data has been pre-processed, we applied the feature selection approach to find the importance of features. There were 12 features columns and a target column. The dataset was subjected to feature selection, and the score values for each feature were obtained. The ways of selecting features is described as follows.

ANOVA F-statistic. ANOVA views an analysis of variance and parametric statistical hypothesis test which determines the means of more than two instances are belong to the same trading or not. It is also depicted as F-test and it belongs to a type of ANOVA that appraises the proportion of variance values, like whether the variance from the different samples or not and elucidated or not. An ANOVA f-test is a form of F-statistic that uses the ANOVA approach. We utilized *scikit-learn* machine library to get the feature scores of all the independent attributes. *f_classif()* function includes an implementation of ANOVA f-test which can be used in a feature selection technique, like—using *SelectKBest* class to find the top *k* most admissible features (highest values). *f_classif()* function incorporates an implementation of ANOVA f-test which can be used in a feature selection technique, like—using *SelectKBest* class to find the top *k* most admissible features (highest values).

Dividing the features into different sets. For our work, we have used the ANOVA F-statistic feature selection method because there are numerical and categorical values in the dataset. We used *f_classif()* function for our work. We used SelectKBest to find the best features. We divided the features into three parts based on the feature scores ranged in three categories and sort these features in descending order. First, we have selected three features for the first part. In the second part, we have included the first

Table 3 Features with feature selection score value

Features	Features selection score
age	4209.008
cholesterol	3599.361
weight	2388.777
gluc	562.773
ap_lo	303.629
ap_hi	208.339
active	89.091
smoke	16.79
height	8.197
gender	4.604
alco	3.761
id	1.010

three features and added more five features with them. In the third part, we have included all 12 features for the experiment. We termed these parts to feature sets. For every feature set, we have applied ML algorithms and measured the performance of ML algorithms. Table 3 shows the score value of every feature in descending order.

4.4 Machine Learning Models for Different Feature Sets

In this phase, we have applied six machine learning algorithms on the pre-processed Cardiovascular Disease dataset for all of the features sets constructed by the values of feature selection method.

Figure 2 shows our working process. First, we performed pre-processing of the data. After data pre-processing, we have applied the feature selection approach and feature importance, with the highest score value, firstly we considered three features, then eight features, and at last we considered 12 features. We divided the dataset into train and test data phases. We applied ML algorithms to the dataset. Then, we measured the performance of every ML algorithms through the widely known evaluation metrics. We do our experiment using *Scikit-learn* library of Python in order to apply ML algorithms to our dataset. Following ML algorithms are applied to our pre-processed data.

Decision Tree: It is also known as decision support, uses a tree-like graph to represent various outcomes, such as chance event outcomes, resource codes. It is a conditional control statement-based representation of an algorithm. The values of the parameters are as follows. Criterion = ‘Gini’, Splitter = ‘Best’, Minimum samples split = 2.

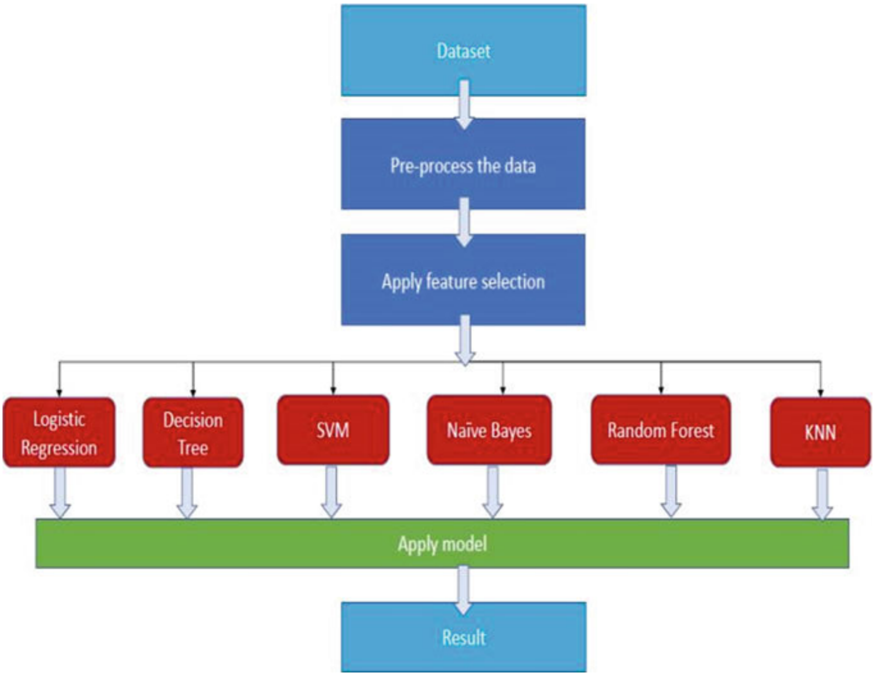

Fig. 2 The process of how ML algorithms applied to the dataset

Logistic Regression: Logistic regression evaluates the parameters of the logistic model in regression analysis. Here, we used Penalty = ’l2’, Tolerance for stopping criteria = 0.0001, Maximum iteration = 10.

Support Vector Machine (SVM): SVM is precise and generates better results in terms of classification. Kernel type = ‘rbf’, Degree of the polynomial kernel function = 3, gamma = ‘scale’, Tolerance for stopping criteria = 0.001.

Naïve Bayes: The probability of given dataset is calculated by Nave Bayes classifiers when they do classification. Each property in a set of data is treated as though it were unrelated to the others. The output class of a particular instance is the high probability class. Prior probabilities = None, Portion of the largest variance = 1e-09.

Random Forest: It is a well-known model for classification as well as regression analysis. Multiple trees are built, and the majority voting is used to classify the records. No. of estimators = 100, Criterion = ‘Gini’, Splitting the smallest samples = 2, Minimum samples leaf = 1.

***K*-nearest neighbor (K-NN)**: The resemblance is assumed by KNN algorithm, between the new data and puts the new data into the category that is most similar to the available categories. No of neighbors = 5, Weights = ‘Uniform’, Metric = ‘Minkowski’.

5 Experimental Results

The analysis of the acquired results is presented in this section. Applying the feature selection method to the dataset, we found the score value of every feature.

5.1 Performance Measurement

We compared the results of ML algorithms to the Cardiovascular Disease dataset and all the outcomes of all six algorithms for all of our features configurations. We constructed a confusion matrix for the performance measurement and then demonstrate the results in Sect. 5. For the confusion matrix, there are four terms described as follows as per our dataset and experiment.

True positive: Heart disease is correctly predicted by the model.

True negative: The model indicates that the patient does not have heart disease, and the patient does not have heart disease in reality.

False positive: Despite the fact that the patient does not have heart disease, the model predicts that he or she has it.

False negative: The model predicts that the patient does not have heart disease, yet the patient does.

The mathematical formulas and the notations are as follows [14].

Accuracy (Acc) describes the correct classification from all samples.

$$\text{Acc} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \tag{1}$$

Sensitivity (Sen) refers to a test's capacity to appropriately identify patients who have cardiac disease. It is also called recall.

$$\text{Sen} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{2}$$

Specificity (Spec) describes the ability of a test to correctly identify people without the heart disease.

$$\text{Spe} = \frac{\text{TN}}{(\text{TN} + \text{FP})} \tag{3}$$

Precision (Prec) describes the ratio of truly positive samples among all the samples that model predicted positive.

Table 4 Performance measurement of ML algorithms for top three features

Algorithms	Sensitivity (%)	Specificity (%)	Precision (%)	Accuracy (%)
Logistic regression	*84.8*	14.9	50.61	50.34
Support vector machine (SVM)	59.5	60.52	60.52	*60.01*
Random forest	57.34	57.2	57.95	57.27
Decision tree	53.86	58.69	57.29	56.24
Naïve Bayes	45.58	*69.03*	*69.03*	54.9
K-nearest neighbor (KNN)	55.7	59.6	58.65	57.62

$$\text{Prec} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{4}$$

5.2 Performance of ML Algorithms for Top Three Features

There is a total of 70,000 instances in the dataset. We put 49,000 instances for training and 21,000 for testing. Using the feature selection method, first, we considered the top three features: age, cholesterol, and weight. Then applied all six ML algorithms for this feature set. The performance of ML algorithms through the mentioned evaluation metrics has shown in Table 4.

Table 4 shows the performance measurement of ML algorithms to the three features of the dataset. These three features give a higher score value of feature selection than other features. For sensitivity, logistic regression gives the highest value, which is 84.8%. Naive Bayes gives the highest specificity value, which is 69.03%. In terms of precision, Naive Bayes gives the highest precision score value which is 69.03%. Again, SVM gives the highest accuracy, which is 60.01%.

5.3 Performance of ML Algorithms for Top Eight Features

After taking the top three features, we included have more features with the previous set. Thus, we got top eight features in total (age, weight, ap_hi, ap_lo, cholesterol, gluc, smoke, and active). Then again, we applied six ML algorithms to the new feature set. The performance measurement of ML algorithms has shown in Table 5 for the top eight features.

Table 5 shows the performance measurement of ML algorithms to the eight features of the dataset. For sensitivity, random forest gives the highest value, which is 69.49%. Naive Bayes gives the highest specificity value, which is 90.23%. For precision, Naive Bayes gives the highest precision score value which is 73.92%. In terms of accuracy, random forest gives the highest accuracy which is 69.41%.

Table 5 Performance measurement of ML algorithms for top eight features

Algorithms	Sensitivity (%)	Specificity (%)	Precision (%)	Accuracy (%)
Logistic regression	62.28	48.52	55.45	55.5
Support vector machine (SVM)	54.98	66.11	62.63	60.47
Random forest	*69.49*	69.33	69.97	*69.41*
Decision tree	62.8	64.25	64.37	63.51
Naïve Bayes	26.92	*90.23*	*73.92*	58.13
K-nearest neighbor (KNN)	66.24	70.4	69.71	68.29

Table 6 Performance Measurement of ML algorithms for 12 features

Algorithms	Sensitivity (%)	Specificity (%)	Precision (%)	Accuracy (%)
Logistic regression	65.49	73.82	72.01	69.6
Support vector machine (SVM)	65.42	54.29	59.55	59.93
Random forest	*69.48*	75.99	74.85	*72.69*
Decision tree	63.62	63.42	64.14	63.52
Naïve Bayes	11.74	*96.66*	*78.32*	53.6
K-nearest neighbor (KNN)	54.74	56.38	56.35	55.5

5.4 Performance of ML Algorithms for 12 Features

After taking the top eight features, we considered all 12 features. Then applied six ML algorithms to the top 12 features. The performance measurement of ML algorithms has shown in Table 6 for the top 12 features.

Table 6 shows the performance measurement of ML algorithms to the 12 features of the dataset. For sensitivity, random forest gives the highest value, which is 69.48%. Naive Bayes gives the highest specificity value, which is 96.66%. For precision parameter, Naive Bayes gives the highest precision score value which is 78.32%. In terms of accuracy parameter, random forest gives the highest accuracy which is 72.69%.

Figure 3 depicts the graphical representation of the accuracy of all six Machine learning algorithms for different experimental configurations of features. From this figure, it can be easily diagnosed that SVM gives the maximum accuracy for three features. Random forest gives the highest accuracy for 8 and 12 features, respectively.

6 Discussion

Using feature selection score value, we considered 3, 8, and 12 feature sets respectively for this work and then applied ML algorithms. We discussed all performance measurements of ML algorithms for, respectively, 3, 8, and 12 features. It is clear

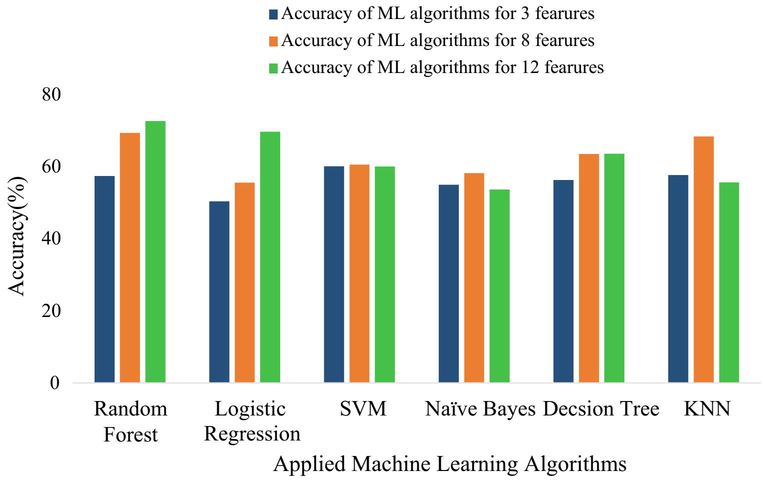

Fig. 3 Accuracy of ML algorithms to 3, 8 and 12 features

from the resulting numbers in terms of the evaluation measurements that when we considered three features, the performance of ML algorithms is quite low. The accuracy is not that much good. For eight features, ML algorithms gave a little better performance than the previous set of features. When we consider all the features. Now, the ML algorithms gives the highest performance. For precision and specificity parameters, Naive Bayes gives the highest score. Random forest gives the highest score in sensitivity and accuracy parameters. It can be inferred that incorporating more features yields better performance in terms of evaluation metrics.

7 Conclusion and Future Scope

The work represents the comparative study of different supervised ML algorithms to the Cardiovascular Disease dataset. The overall aim is to define various ML algorithms useful to effective heart disease prognosis. By employing the feature selection method, we took different experimental setups of features and applied ML algorithms to them. Only 12 critical features are considered in this study. We compared the result of ML algorithms, respectively, for different experimental configurations for different feature sets. In the future, we are planning to expand this research incorporating new contexts, like, how the heart condition of a person changes as he or gets older. We are about to build up a recommendation system to what extent a person should get aware of his or her heart diseases based on the deciding factors of heart disease.

References

1. Gavhane A, Kokkula G, Pandya I, Devadkar K (2018) Prediction of heart disease using machine learning. In 2018 Second international conference on electronics, communication and aerospace technology (ICECA), pp 1275–1278
2. Palaniappan S, Awang R (2008) Intelligent heart disease prediction system using data mining techniques. In 2008 IEEE/ACS international conference on computer systems and applications, pp 108–115
3. Shah D, Patel S, Bharti SK (2020) Heart disease prediction using machine learning techniques. SN Comput Sci 1(6):1–6
4. Rajdhan A, Agarwal A, Sai M, Ravi D, Ghuli P (2020) Heart disease prediction using machine learning. Int J Res Technol 9(04):659–662
5. Repaka AN, Ravikanti SD, Franklin RG (2019) Design and implementing heart disease prediction using naives Bayesian. In 2019 3rd International conference on trends in electronics and informatics (ICOEI), pp 292–297
6. Mohan S, Thirumalai C, Srivastava G (2019) Effective heart disease prediction using hybrid machine learning techniques. IEEE access 7:81542–81554
7. Kelwade JP, Salankar SS Radial basis function neural network for prediction of cardiac arrhythmias based on heart rate time series. In 2016 IEEE first international conference on control, measurement and instrumentation (CMI), pp 454–458
8. Jabbar MA, Samreen S (2016) Heart disease prediction system based on hidden naïve bayes classifier. In 2016 International conference on circuits, controls, communications and computing (I4C), pp 1–5
9. Bhatla N, Jyoti K (2012) An analysis of heart disease prediction using different data mining techniques. Int J Eng 1(8):1–4
10. Sowmiya C, Sumitra P (2017) Analytical study of heart disease diagnosis using classification techniques. In 2017 IEEE international conference on intelligent techniques in control, optimization and signal processing (INCOS), pp 1–5
11. Thomas J, Princy RT (2016) Human heart disease prediction system using data mining techniques. In 2016 International conference on circuit, power and computing technologies (ICCPCT), pp 1–5
12. Singh D, Samagh JS (2020) A comprehensive review of heart disease prediction using machine learning. J Crit Rev 7(12):281–285
13. Bashir S, Khan ZS, Khan FH, Anjum A, Bashir K (2019) Improving heart disease prediction using feature selection approaches. In 2019 16th international Bhurban conference on applied sciences and technology (IBCAST), pp 619–623
14. Zulfiker MS, Kabir N, Biswas AA, Nazneen T, Uddin MS (2021) An in-depth analysis of machine learning approaches to predict depression. Curr Res Behav Sci 2:100044