Journal of Cardiovascular Disease Research, 12 (1)
ISSN: 0975-3583, 0976-2833
10.31838/jcdr.2021.12.01.05

ORIGINAL ARTICLE

Machine learning techniques for heart disease prediction

D. Hemalatha[1], S. Poorani[2]

[1]Assistant Professor(s), Department of Computer Technology-UG
[2]Kongu Engineering College, Perundurai

*Corresponding author: D. Hemalatha, Assistant Professor, Department of Computer Technology-UG, Kongu Engineering College, Perundurai. Email: hema.arun2011@gmail.com

Received: 9 July 2020; Accepted: 2 January 2021; Published: 14 April 2021

Abstract

Nowadays, cardiovascular deaths and diseases have increased at a fast rate worldwide. The early prediction of this disease is necessary to prevent the deaths. Detection of heart disease requires more experience and good knowledge about heart problems. To detect heart disease in medical science massive quantity of information is collected and accumulated as databases. All the accumulated information could not be worthwhile. Mostly, the datasets are assorted and dispersed in nature. So there is a need for extracting predictive information about heart diseases. Machine learning, an evolving technique that can be used to develop automated systems/frameworks to analyze the data in different domains. It performs well in health care also. There are many techniques available for analyzing and predicting the heart disease, this paper aims to develop predictive models to analyze the datasets relevant to heart disease based on random forest, SVM, J.48, Bayesian prediction and MLP.

Keywords: MLP, Machine Learning, Classification, SVM, Bayesian, Decision Tree, Random forest.

Introduction

According to WHO, in recent years 17.9 million deaths occur worldwide.[1] The coronary disease is a fatal and dangerous disease because if patient ignores its earlier symptoms, which seems to be a warning signs, it gives no time to patient for recovery and eventually may lead to death on spot. Several classification algorithms have been implemented to develop the predictive models for heart disease with various databases like ceveland and stalog(UCI library).[2] The algorithms like Logistic regression and BPNN were used to predict the heart disease using Cleveland dataset.[3] MLPSNM,[4] Decision Tree, k-NearestNeighbor (KNN) and Naive Bayes[5] have also been applied in the prediction of heart disease. In recent days, there are number of hybrid methods are proposed for heart disease prediction. The hybrid method can have two stages of processes. In the first stage the subset of features can be selected to reduce complexity. The classifier is trained in the second stage.[6] The roughsets have also been implemented in heart disease prediction.[7] A prediction model based on ANN and decision tree has also been developed for heartdisease prediction.[8] We use five different classifiers to predict the heart disease with ceveland dataset and finally we analyses the performance of each other.

Materials and Methods

Figure 1 illustrates the process done by the proposed system.

Dataset and Feature Selection

We use the Cleveland dataset from UCI. Totally, it has 76 attributes and 303 subjects. But according to the database reference, only the following

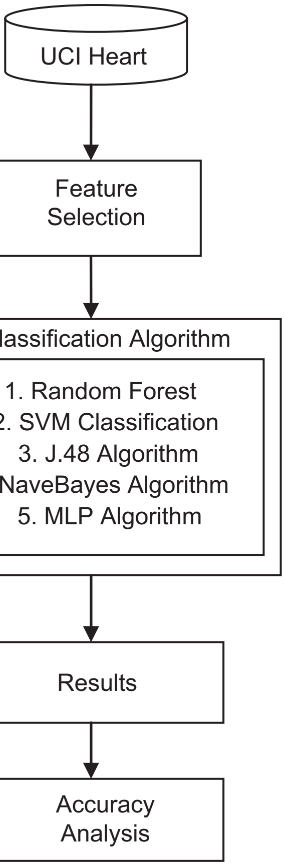

Figure 1 Flow chart of proposed system

Table 1 Attribute Description

S.No.	Attribute Name	Description
1	age	patient's age
2	sex	male/female
3	cp	type of chest-pain
4	trestbps	Blood-pressure value in rest
5	chol	cholesterol level of the patient
6	Fbs	Blood-sugar in fasting
7	Restecg	Electrocardiographic result during rest
8	Thalach	Maximum heart-rate
9	thalrest	Heart-rate in rest
10	Exang	Angina which is induced by exercise(y/n)
11	oldpeak	ST depression
12	Slope	ST-segment's slope
13	ca	Count of main vessels
14	Thal	7 = reversible; defect ; 6 = fixed-defect; 3 = normal
15	Num	analysis of heart-disease

14 attributes were taken for classification. The following table expresses the attributes and their descriptions.

Classification

A classification algorithm discovers associations among the values of the predictors and the values of the target. Different classification algorithms use different techniques for finding relationships. The list of classifiers applied in the anticipated model are as follows.

- Decision Tree
- MLP
- Nave Bayes
- Random Forest
- SVM

Decision tree do the learning process based on the class labels of train-data. The structure of decision tree is in the form of flowchart. The attributes are denoted by non-leaf node. The label of a class is denoted by leaf node. It does not need any knowledge about particular domain. It can be applied for multidimensional data. Our data has 14 attributes, so this can be a suitable method to train and test our data in an efficient way.

We use MLP with one hidden layer and it takes long time for training and suitable for data with noises and it can handle the new data which has not been trained yet. By using SVM the train-data is mapped into non-linear space an N number of hyperplanes are generated. Finally, the best plane will be selected. Navebayes computes the probability to make the decision about a hypothesis. By using random-forest, the attributes are selected in the random manner to generate decision trees and then the split is determined.

Results and Discussion

MBE-Mean-Absolute-error: The difference between the predicted value(yp) and actual value(ya) is called absolute error(AE)(yp-ya). The average of AE is said to be MBE, $e_i = y_i - x_i$. The decision tree shows the low error rate and MLP shows high rate of MBE. Table 2 and Figure 2 describe the MBE investigation of the proposed model. RMSE: The standard deviation of the prediction errors is said to be Root-Mean-Square Error. Table 2 and

Table 2 MAE

Classification Algorithm	Mean Absolute Error
Decision Tree	0.383
MLP	0.402
SVM	0.398
Random Forest	0.392
Bayesnet	0.389

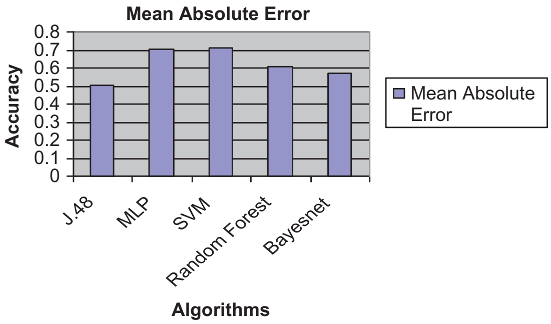

Figure 2 Comparison chart of MAE

Table 3 RMSE

Classification Algorithm	Root Mean Square Error
Decision Tree	0.398
MLP	0.417
SVM	0.409
Random Forest	0.412
Bayesnet	0.403

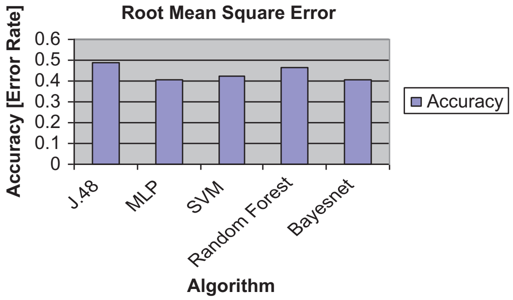

Figure 3 Comparison chart of RMSE

Table 4 RSE

Classification Algorithm	Relative Squared Error
J.48	65.32
MLP	65.87
SVM	65.58
Random Forest	65.73
Bayesnet	65.33

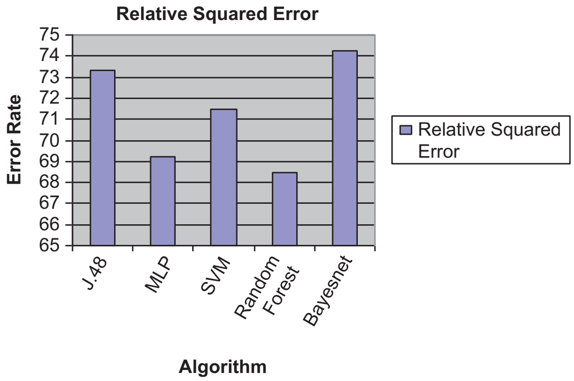

Figure 4 Comparison Chart of RSE

Table 5 Accuracy

Classification Algorithm	Proposed Accuracy
J48	96.04
MLP	77.54
SVM	73.44
Random Forest	80.22
Bayesnet	90.33

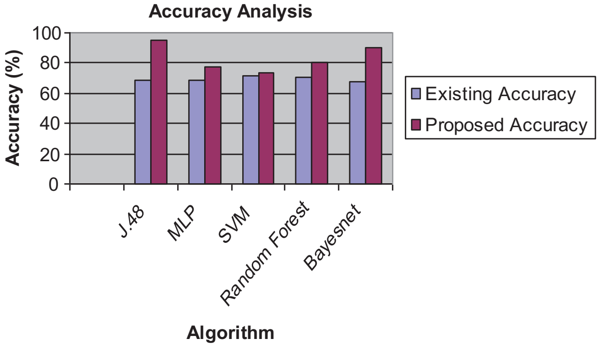

Figure 5 Comparison Chart of Accuracy

Figure 3 express Root-Mean-Square Error (RMSE) analysis of the proposed system. Relative Squared Error(RSE): The total squared error is divided by the predictor's total-squared error. Table 4 and Figure 4 describes a relative squared error analysis. The Table 5 and Figure 5 describes accuracy of all classifiers. Classification techniques such as Multilayer Perceptron, Random Forest, Support Vector Machines and J48 reads the data from the data frame for testing and training. Testing and training results must be differing because of split function used in the data set. J48, Random Forest and Bayes theorem gives the higher accuracy. SVM provides the low accuracy than other classifiers.

Conclusion

To envisage the hear-disease, we inspect the concert of five well-liked ML methods. Among these five methods, Navebayes and J48 confers superior outcomes than remaining three methods. This valuable calibration methods can prop up the correctness of heart disease. The machine erudition methods based on Navebayes and J48 in this study can enhance the efficacy of potential health studies associated with heart disease.

Related Works

N. Satish Chandra Reddy et al.[2] proposed a prediction model for heart disease prediction. They used random forest algorithm for both feature selection and classification. They provides 90–95% accuracy with the datasets Cleveland and statlog. Five datasets (V.A. Medical, Switzerland, Cleveland, Hungarian, and Statlog) are used in this study. All these five databases were merged into one dataset to acquire good performance model. The final dataset contains 1190 observations with 14 features.

Sundas Naqeeb Khan et al.[10] performed an analysis of different classifiers used in various papers with heart disease prediction and they showed that performance of classifiers like Decision tree c4.5, Nave bayes, SVM, RIPPER, KNN, GA and ANN. Among these algorithms, Decision tree, Navebayes and SVM are concluded as best classifiers than others.

Hager Ahmed et al.,[9] developed a system based on big data framework for the prediction of heart disease using streaming data collected from social network activities and Cleveland dataset to achieve high accuracy. Feature selection algorithms are used to select the features from dataset. They implemented SVM, RF, LR and decision tree Classifiers on all the features and a subset of the selected features. To enhance the accuracy, they used hyper parameter tuning and cross-validation with machine learning techniques. The random forest classifier performs better compared to other models by achieving the highest accuracy of 94.9%.

C. Beulah Christalin Latha, S. Carolin Jeeva[11] proposed a model to reduce the weakness of weak classification algorithms by merging them with other classification algorithms. To improve the accuracy of weak algorithm they used a method called ensemble classification techniques like bagging and boosting. Data mining techniques are powerful in predicting heart disease such as decision trees, neural networks, Naive Bayes and associative classification. They used Cleveland heart database for the experiments. Support vector machine produce very accurate result for heartbeat classification. Particle Swarm optimization is used to improve the performance of the classifier and also used to optimize the parameters. The weak classifier accuracy is increased upto 7% when it is used with ensemble classification.

The accuracy is one of the most important metric related with system performance, So we compute and evaluate the correctness of five classifiers (MLP, SVM, Nave Bayes, j.48 & RF) with the UCI dataset.

References

1. https://www.who.int/en/news-room/fact-sheets/detail/ cardiovascular-diseases-(cvds) (referred on 19.11.2019).
2. Satish Chandra Reddy N., Song Shue Nee, Lim Zhi Min & Chew Xin Ying (2019). Classification and Feature Selection Approaches by Machine Learning Techniques: Heart Disease Prediction. International Journal of Innovative Computing, 9(1):3–46.
3. Desai S.D., Giraddi S, Narayankar P., Pudakalakatti N.R, Sulegaon S (2019). Back propagation neural network versus logistic regression in heart disease classification. Advanced Computing and Communication Technologies, Springer, 133–144.
4. Burse K., Kirar V.P.S., Burse A, Burse R.(2019) .Various preprocessing methods for neural network based heart disease prediction. Smart Innovations in Communication and Computational Sciences, Springer, 55–65.
5. Enriko I.K.A., Suryanegara M., Gunawan D. (2016). Heart disease prediction system using k-nearest neighbor algorithm with simplified patient's health parameters. Journal of Telecommunication, Electronic and Computer Engineering, 8(12): 59–65.
6. Chau A., Li X., Yu Liu W., (2014). Support vector machine classification for large datasets using decision tree and fisher linear discriminant, Future Generation Computer Systems, 36, 57–65.
7. Nguyen T., Khosravi A., Creighton D., Nahavandi S.(2015). Classification of healthcare data using genetic fuzzy logic system and wavelets. Expert System Applications, 42(4): 2184–97.
8. Maji S., Arora S, (2019). Decision tree algorithms for prediction of heart disease, in: Information and Communication Technology for Competitive Strategies. Springer, 447–454.
9. Hager Ahmed, Eman M.G. Younis, Abdeltawab Hendawi, Abdelmgeid A. Ali d(2019). Heart disease identification from patients' social posts, machine learning solution on Spark. Future Generation Computer Systems, 111: 714–722.
10. Sundas Naqeeb Khan et al (2017). Comparative Analysis for Heart Disease Prediction. International Journal On Informatics Visualization, 1:227–231.
11. Beulah Christalin Latha C., Carolin Jeeva S, (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine, 16.