SN Computer Science (2023) 4:128
https://doi.org/10.1007/s42979-022-01512-3

ORIGINAL RESEARCH

Comparison of Different Classification Algorithms for Prediction of Heart Disease by Machine Learning Techniques

B. Harshitha[1] · P. Maria Rufina[1] · B. L. Shilpa[1]

Received: 17 October 2022 / Accepted: 15 November 2022 / Published online: 27 December 2022

Abstract

Cardiovascular disease commonly referred as heart disease, encompasses diverse conditions that the heart undergoes which in turn leads to sudden death or prolonged sickness worldwide over the past decades. More recently, foreseeing heart disease is the stimulating responsibility in the health arena. In recent eras, every minute approximately one person expires due to heart ailment. Data Science processes big volumes of healthcare data and researchers apply a variety of datamining and machine learning techniques to analyze vast and complex medical data to help health care professionals predict heart disease. This tabloid collects heart disease dataset from UCI machine learning source analyzing which envisages the accuracy of heart disease by considering major risk factors based on different classifier algorithms. This research paper objective is to diagnose imminent heart disease via scrutinizing data of patients and analyzing if heart disease is pestilent with machine-learning algorithm.

Keywords Logistic regression · Random Forest · Heart disease prediction · Classification algorithm

Introduction

At recent times, diseases related to heart is predominantly one of the main reasons for fatal and unexpected deaths that occur. A healthy lifestyle and primary finding are eminent and safest idea to prevent heart disease. The exertion projected herein focuses primarily on noticing the heart disease using machine learning approach with the help of medical test parameters. Human heart functions throughout the life span of a human being and considered to be the utmost vital part and any abnormality in its working develop a life-threatening health problem. Aberrations in heart is the basis for illness in other parts of the body which in turn is classified as can be classified as heart disease. Heart complications triggered due to insalubrious routine, smoking, liquor consumption and high cholesterol may cause high blood pressure [1]. The World Health Organization has a record where in around 10 million succumb due to heart disease per annum. A hale and hearty lifestyle and initial finding is the solitary means to avoid heart disease. The biggest challenge in health care today is delivering the best possible quality service and an accurate diagnosis. Even if still heart disease is considered as the leading cause of death, they are also the ones that can be managed and controlled effectively. The aim of this research paper is to test a patient for cardio related health issues by examining the on medical attributes such as gender, age, chest ache, abstaining blood sugar etc. A record of collected data is selected from the UCI source accompanied by the patient's medical past and characteristics. Taking this dataset into account, we foresee if the patient, in near future faces heart related complications or not. With 76 attributes in hand only 14 attributes are considered for testing. These medicinal features are skilled under two algorithms: logistic regression and Random Forest Classifier. Random Forest which gives us the accuracy of 90.16% can be considered as the most efficient algorithm than the former.

This article is part of the topical collection "Advances in Computational Intelligence for Artificial Intelligence, Machine Learning, Internet of Things and Data Analytics" guest edited by S. Meenakshi Sundaram, Young Lee and Gururaj K S.

✉ P. Maria Rufina
pmariarufina@gsss.edu.in

[1] GSSSIETW, Mysuru, India

Literature Survey

A quite substantial amount of work related to diagnosis of cardio vascular disease using machine learning algorithm has motivated this work.

Golande and Pavan Kumar studied on three different machine learning algorithms such as Decision tree, K-means clustering and KNN for classification of heart disease and found that accuracy produced by Decision Tree was best [2].

Alotaibi, carried out research to predict heart disease by implementing five different machine learning algorithms [1]. Their study used the Rapid miner tool which gave the highest accuracy than any other tool. The only drawback in their research was the size of the data set. From their analysis it was concluded that Decision tree produced the highest accuracy.

Lutimath et al. in their research worked on "Prediction Of Heart Disease using Naïve Bayes classification and SVM". In their research SVM with radial kernel produced better accuracy than Naïve Bayes classification. Mean Absolute Error, Sum of Squared Error and Root Mean Squared Error were the major parameters used in the performance analyses [3].

Rajdhan and Sai and many others worked on the same using four machine learning algorithms [4]. Their research contrasts the exactness of Decision Tress, Naïve Bayes, Logistic regression and Random Forest and of the four Random Forest algorithm was the most effective. The research included metrics such as Accuracy, Precision, Fmeasure and Recall to carry out the performance analysis which was applied on pre-processed dataset.

Machine learning algorithms partake effectual in producing results with a high level of accuracy. It has aided health scholars and medics all over the world in recognising patterns in the patients ensuing in initial detections of heart diseases [5].

Results verified with hybrid blend of PSO with SVM (PSO_SVM) completes superior predictive performance over additional replicas. The study will consequently allow doctors to detect cardiovascular diseases and pledge apt action without the intrusion of a qualified cardiologist [6].

Proposed Model

Our projected work forecasts heart disease by discovering two classification algorithms (mentioned above) and ensures performance. The objective of this study is to detect heart disease using machine learning approach with the help of medical test parameters.

A healthcare employee inputs the values from patient's health report which is fed into the model. The model using the data that is uploaded predicts if the patient is prone to heart disease or not using machine-learning algorithm. Figure 1 shows the complete process involved.

Dataset Assembly and Preprocessing

The data set used is a cardiology data set available in UCI Cleveland. This dataset is a collection of 76 attributes, out of which only 14 are well-thought-out for testing and even vital to substantiate the performance of above considered algorithms [7]. A detailed report of the 14 attributes specified in the proposed work can be found in the Table 1.

Classification

The attributes listed in Table 1 are provided as input to various ML algorithms such as random forest, Logistic Regression Classification technique. The dataset is divided as training (80%) and test (20%) data sets. The training dataset is the data set used to train the model. A test dataset is used to validate the performance of trained model. The performance of each algorithm is calculated and analysed using various indicators such as accuracy [9]. Various algorithms considered in this paper is as follows.

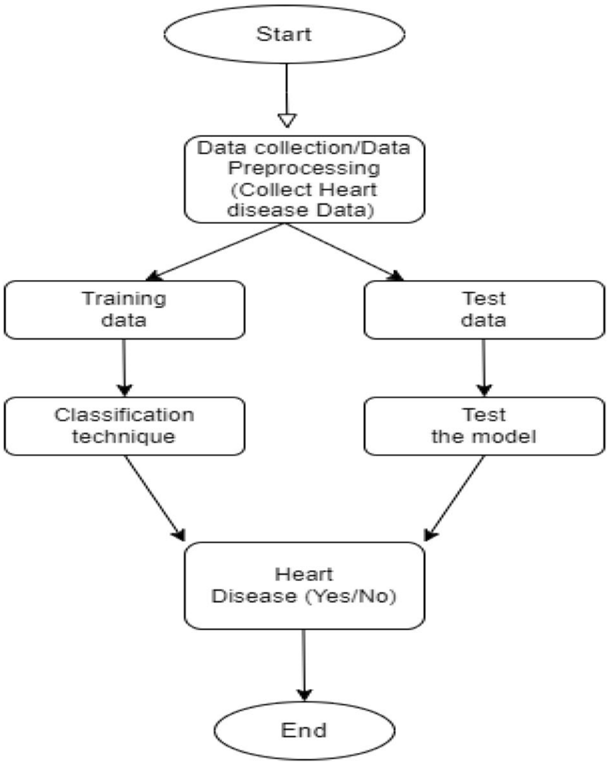

Fig. 1 Proposed model predicting heart disease

Table 1 Attribute description of UCI-ML Heart Disease Dataset Cleveland heart disease dataset is extracted from UCI-ML repository (Machine Learning Repository of University of California, Irvine; 2017) to comprehend the effect of hybrid feature selection and classification algorithms [8]

SL. No.	Description of the attribute
1	Age—A Person's current age (29–77)
2	Gender—(0 for Female, 1 for Male)
3	CP—chest pain severeness (0,1,2,3)
4	Restbp: Resting Blood pressure (Values b/w 94 and 400)
5	Chol: Cholesterol Level (Values between 164 and 564)
6	FBS: Fasting blood sugar (0,1)
7	Resting ECG: (0,1,2)
8	Thal (Thalium Test): test for patient who is treated for chest pain or breathing difficulty (0,1,2,3)
9	Exang—identifies exercise induced angina (yes = 1 or else no = 0)
10	OldPeak—depression level of the patient (between 0 and 6.2.)
11	Slope—peak exercise time what is the patient condition (1,2,3)
12	CA—fluoroscopy Result (0,1,2,3)
13	Thalch: Maximum heart rate achieved(values between 71 and 202)
14	Target—It is the last data in the data set. binary classification of dataset with two classes (0,1). "0" less possibility and "1" says high probability of heart disease

Logistic Regression

Despite its name, logistic regression is more of a classification model than a regression model. Logistic regression is a simple and efficient method for binomial and linear classification problems [10]. Data mining algorithms in hybrid mode outperforms algorithms when used individually in diagnosing cardiovascular diseases. This is a classification model that is very easy to implement and achieves very good performance on linearly separable classes. This is a classification algorithm widely used in the industry. The output of logistic regression can take values such as yes or no, 0 or 1, true or false.

Random Forest

Random Forest, a commonly chosen supervised machine learning algorithm is used for both classification and regression. A classifier such as random forest algorithm considers a dataset of decision trees over diverse subclasses of a given dataset and norms them to improve the prediction perfection of these dataset [11]. Decision trees are built by random forest diverse subsets takes mainstream vote for classification and average in case of regression.

Summary

- Classifiers like Support Vector Machines (SVM), Artificial Neural Network (ANN), Naïve Bayes, K-Nearest Neighbour (KNN), Random Forest and C5.0 Decision Tree are assumed to achieve improved results mutually with Particle Swarm Optimization (PSO) feature selection method.
- UCI-ML heart disease dataset consists of 76 parameters over-all. Yet, mainstream of the available researcher and research papers mentions to using a subset of 14 attributes.
- Accurateness, sensitiveness and specific metrics are the most common performance metrics for assessing models in spotting cardiovascular illnesses.

Analysis

The outcomes received by smearing Random Forest and Logistic Regression algorithms are depicted here. The Accuracy score is used to measure algorithm performance according to the values obtained. We use the recorded dataset and carry the experiments by implementing the above cited algorithms [12].

Our project projects us that Random Forest accomplishes good results than Logistic Regression Classifier by predicting whether the patient is diagnosed with a heart disease or not. This evidences that Random Forest is better in diagnosis of a heart disease [13]. The following Figs. 2, 3, 4, and 5 depicts a plot of the number of patients that are been separated and foreseen by the classifier reliant on the Chest Pain [14].

Conclusion and Future Enhancements

Cardiac related disease is presumed as the main cause of demise amongst patients with illness. From our study, we can conclude that there is a necessity to innovate an

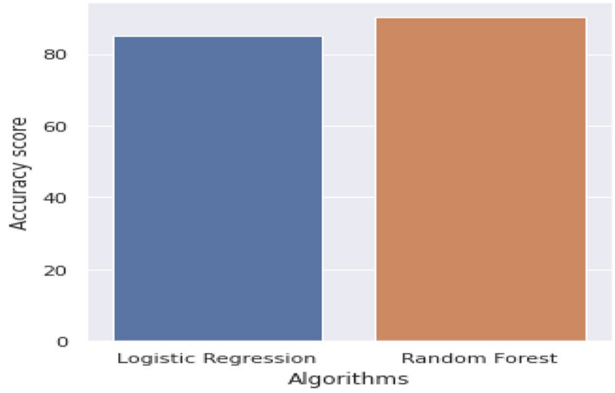

Fig. 2 The figure shows how we obtain the accuracy score of heart disease using the algorithm [14]

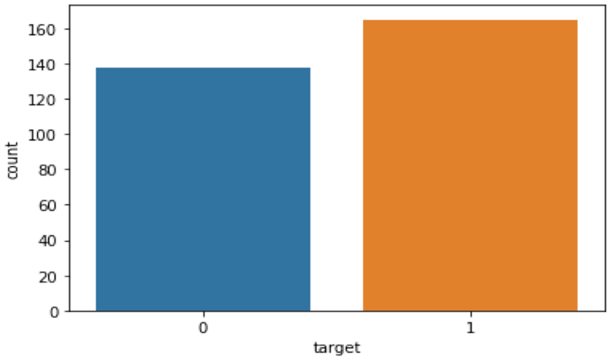

Fig. 3 Target count of logistic Regression(0) and Random forest(1) [14]

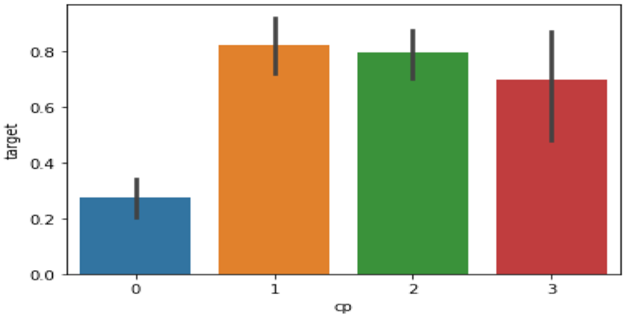

Fig. 4 Severity of chest pain denoted as cp in the above diagram [14]

automated system which predicts heart disease precisely [15]. The key challenge which motivated the learning is to choose the effective algorithm under ML for prediction of cardiovascular disease. This exertion correlates the correct notch of Logistic Regression and Random Forest algorithms in detecting heart disease as early as possible. From this study it is clear to consider the Random Forest algorithm as the utmost effective algorithm showing 90.2% accuracy rate in finding the heart disease [16]. Future enhancements, we can develop an online application that better displays the details of the patient along with the indication of possible heart attack. We can also reduce the number of parameters considered to predict the heart disease.

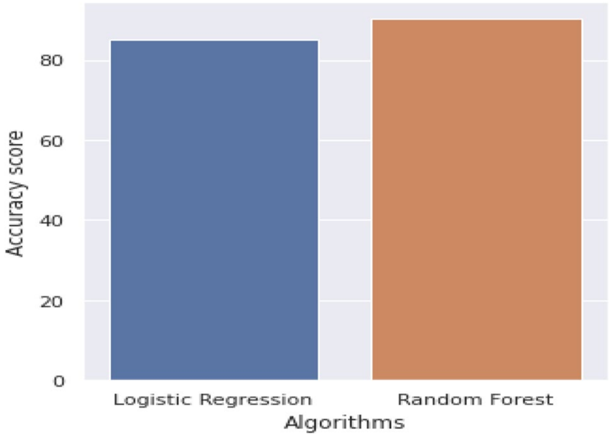

Fig. 5 Accuracy score achieved using Logistic Regression is: 85.25%. The accuracy score achieved using Random Forest is: 90.16% [14]

Funding No funding received for this paper.

Data availability Data regarding the research results are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. FS Alotaibi. Implementation of machine learning model to predict heart failure disease. (IJACSA) Int J Adv Comput Sci Appl. 2019;10(6).
2. Golande A, Pavan Kumar T. Heart disease prediction using effective machine learning techniques. International Journal of Recent Technology and Engineering. 2019;8:944–50.
3. Lutimath NM, Chethan C, Pol BS. Prediction of heart disease using machine learning. International journal Of Recent Technology and Engineering. 2019;8(2S10):474–7.
4. Rajdhan A, Sai M. Heart disease prediction using machine learning. Int J Eng Res Technol (IJERT). 2020;9(04). ISSN: 2278-0181 IJERTV9IS040614.
5. Rajdhan A, Sai M and others carried out the prediction of heart disease using four machine learning algorithms. Their research compares the accuracy of Decision Tress, Naïve Bayes, Logistic regression and Random Forest and the most efficient algorithm was Random Forest algorithm.
6. Mondal S. Diagnosis of cardiovascular diseases using hybrid feature selection and classification algorithms.
7. Xu S, Zhu T, Zang Z, Wang D, Hu J, Duan X, et al. Cardiovascular risk prediction method based on CFS subset evaluation

and random forest classification framework. In: 2017 IEEE 2nd international conference on big data analysis.
8. Rjeily CB, Badr G, Hassani EAH, Andres E. Medical data mining for heart diseases and the future of sequential mining in medical field. In: Machine learning paradigms. 2019. p. 71–99.
9. Ambekar S, Phalnikar R. Disease risk prediction by using convolutional neural network. In: 2018 fourth international conference on computing communication control and automation.
10. UCI, ——Heart Disease Data Set.[Online]. https://www.kaggle.com/ronitf/heart-disease-uci. Accessed 1 May 2020.
11. Pugazhenthi D, Quaid-E-Millath, Meenakshi, et al. Detection of ischemic heart diseases from medical images. In: 2016 international conference on micro-electronics and telecommunication engineering
12. Hodges J, et al. Discriminatory analysis, nonparametric discrimination: consistency properties. 1981.
13. Rajathi S, Radhamani G, et al. Prediction and analysis of rheumatic heart disease using kNN classification with ACO. 2016.
14. https://colab.research.google.com/drive/1HQnT4KhOKp51ZBBrFrkRRiI7H5VK8z6X#scrollTo=-Dk3GnhycCGg
15. Bansal P, Saini R, et al. Classification of heart diseases from ECG signals using wavelet transform and kNN classifier. In: International conference on computing, communication and automation (ICCCA2015).
16. Simge EKIZ, Erdogmus P, et al. Comparitive Study of heart Disease Classification. 978-1-5386-0440-3/17/$31.00 ©2017 IEEE.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.