STM JOURNALS

Journal of
Artificial Intelligence
Research & Advances

ISSN: 2395-6720
Volume 9, Issue 2, 2022
DOI Journal: 10.37591/JoAIRA

http://computers.stmjournals.com/index.php?journal=JoAIRA&page=index

Review JoAIRA

Experimental Study on Heart Disease Prediction Using Different Machine Learning Algorithms

Amrutha L.*, Rachita R.

Abstract

Heart disease which can also be referred to as the cardiovascular disease is one of the raising concerns in today's world. It is one of the major health problems causing death among humans irrespective of the age group and therefore has made it necessary to look into different medical factors that are required to predict the same in advance using the collected historical datasets of various patients. Thus, we have used various machine learning algorithms to predict the potential of person, to suffer from a heart disease with high precision and reliability so that we can admonish the patient in advance and take the required precautionary measures. In this study, an existing heart disease dataset accessible from the UCI Machine Learning Repository is used as the primary dataset. The experimental analysis and comparative study between different algorithms, helps in deciding the best suitable algorithm for the given problem statement by using results obtained which are very competitive and can be used for identification and treatment. The proposed work predicts the probability of heart condition and ranks the patient's risk level based on different supervised learning algorithms such as K-Neighbors, AdaBoost, Gradient Boosting, etc. The test results portray that K-neighbors has the highest accuracy score of 91% when compared with other algorithms.

Keywords: Heart disease prediction, algorithms, machine learning, CNN, KNN

INTRODUCTION

Heart disease encompasses a wide range of disorders that affect the heart and has been the primary cause of death for periods. The major reasons being the changes in the lifestyle, work-related stress, and poor eating habits, all contribute to increased rates of heart disease. Around 17.9 million individuals per year lose their lives to cardiovascular disease, most of which are caused by coronary heart disease and stroke. Since it takes more intelligence, time, and knowledge, it is not always possible to accurately monitor patients every day, and a doctor cannot consult with a patient for a whole 24 h. Numerous research have been conducted to pinpoint the most crucial heart disease risk factors and precisely calculate the overall risk. Heart disease is also referred to as a silent killer because it causes a person to pass away without any evident signs. In order to avoid complications in high-risk patients and make decisions about lifestyle changes, early detection of heart disease is crucial. By analyzing patient data that uses machine learning algorithms to categorize whether a patient has heart disease or not, this study seeks to predict future cases of heart disease.

***Author for Correspondence**
Amrutha L.
E-mail: 05amruthal@gmail.com

Student, Department of Computer Sciences, Global Academy of Technology, Bangalore, Karnataka, India

Received Date: August 10, 2022
Accepted Date: August 18, 2022
Published Date: August 22, 2022

Citation: Amrutha L, Rachita R. Experimental Study on Heart Disease Prediction Using Different Machine Learning Algorithms. Journal of Artificial Intelligence Research & Advances. 2022; 9(2): 18–24p.

Machine learning is one of the most quickly evolving subfields of artificial intelligence. Machine learning has recently emerged as the most advanced, dependable, and supportive technology in the medical field, offering the greatest assistance for disease prediction with the proper case of training and testing. These algorithms can evaluate

large quantities of data in variety of fields, one of which is medicine. It is a computer-based alternative to ordinary prediction modeling for comprehending complicated and nonlinear interactions among many components by reducing mistakes in the projected and actual results [1–5].

Healthcare practitioners examine this data so that they can make good diagnostic choices. To forecast heart illness in patients, it tests categorization systems. The purpose of this study is to determine whether a patient's medical history suggests that they are likely to develop cardiovascular heart disease. The existence of cardiac disease is indicated by high cholesterol levels, high blood pressure, and rapid heartbeats. Machine Learning (ML), a subset of data extraction, aids in managing massive datasets with numerous properties. It has the ability to improve accuracy by utilizing complex relationships between risk factors and can be used to detect and diagnose a wide range of illnesses in the medical sector. These big datasets can be quickly investigated for comprehension using machine learning techniques. As a result, recent research has shown that all of these algorithm technologies are highly helpful in accurately detecting the presence or absence of heart-related illnesses. The major purpose of this research is to give doctors a tool for the early detection of heart disease. As a result, it is easier to provide proper treatment to patients while preventing severe causes. Machine learning is crucial in discovering hidden discrete patterns and interpreting data. As a result, it is simpler to treat patients appropriately and stop serious causes. In order to understand the data and find hidden discrete patterns, machine learning is essential [6–9].

RELATED WORK

Bharti *et al*. employed several machine learning and deep learning algorithms to predict heart disease in their study "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning" [5]. The outcome derived is that machine learning techniques outperform this analysis. The comparison methods are confusion matrix, precision, specificity, sensitivity, and F1 score. When preprocessing data was used, the classifier K-Neighbors outperformed the ML technique for the 13 characteristics in the dataset. The computation time has also been reduced, which is very useful when implementing a model. The challenge here is that the dataset's sample size is small. If there is a large data set, the results can also increase significantly in deep learning and ML. The algorithm they applied in the ANN architecture increased the accuracy of which was compared with different researchers. This can increase the size of the dataset, then deep learning can be used with many other 1087 optimizations and more promising results can be obtained. Machine learning and a variety of other optimization approaches can be utilized to enhance the evaluation results yet again.

Senthilkumar implemented conjugate machine learning algorithms to predict heart disease [10]. The dataset used is the Cleveland dataset. The first step is the data preprocessing step. In this case, the tuples removed from the dataset have missing values. The authors also did not use the dataset's age and gender features since they believe this is personal information that has no effect on the forecast. The remaining 11 properties are considered important because they contain important clinical records. Linear Hybrid Random Forest Method (HRFLM) is a hybrid of the Random Forest (RF) and the Linear Method (LM), The first algorithm is concerned with splitting the input data set. It operates on a decision tree for each sample in the data set. The data set is partitioned into leaf nodes after the feature space has been defined. The first algorithm produces a partition of the data set. The rules are then applied to the data set in the second process, and the outcome is a data classifier using those rules. The features are extracted using the least error classifier in the third algorithm. This algorithm is concerned with determining the classifier's minimum and maximum error rates, and the result is entities with classified attributes. They employ Classifier in the fourth algorithm, which is an association approach based on the error rate on the extracted features. Finally, they compare the results achieved by using HRFLM to those obtained by using other classification techniques such as decision trees and support vector machines. As a result, since RF and LM gave better results than other algorithms, the two algorithms were combined and a new HRFLM algorithm was generated. The authors propose to further improve accuracy using a combination of different machine learning algorithms.

In a paper by Rajdhan *et al.*, they have done a comparative study by pointing accuracy of Decision Trees, Logistic Regression, Random Forest, and Naive Bayes algorithms to predict heart disease using the UCI machine learning benchmark dataset [1]. The results of this study indicated that the Random Forest algorithm was the most effective one.

In a paper by Jindal *et al.* named "Heart disease prediction using machine learning algorithms", a cardiovascular disease detection model was developed using three modeling techniques of the ML classifier [3]. This project predicted people with cardiovascular illness by extracting the history of patients who had fatal heart disease from a dataset that included the patient's medical history, such as chest discomfort, blood pressure, and sugar levels. Logistic regression, random forest classifier, and KNN are the algorithms utilized to build the presented models. Using more training data ensures a higher chance that the model will correctly predict whether the given person will have heart disease. To conclude, this article aids in the prediction of patients with heart disease by cleaning the dataset and employing logistic regression and KNN to reach an average accuracy of 87.5%.

METHODOLOGY

Dataset Collection

The data set considered here includes the histories of different patients from various age groups. The dataset offers a wide range of details, such as medical traits including age, height, weight, resting blood pressure, chest discomfort, cholesterol, and maximum heart rate of the patient, which help in establishing whether a patient has been diagnosed with a heart disease. We can identify persons at risk by neatly analyzing these factors based on the correlation of each parameter in determining the heart disease and classify the patients (Table 1). The next step after data collection is preprocessing of the dataset which is essential set in any machine learning which helps in cleaning the dataset by removing the noise from it.

Table 1. Attributes and details of dataset of heart disease.

S. N.	Attribute	Representative	Details
1	Age	Age	Patient's age in years
2	Sex	Sex	0=female, 1=male
3	Chest pain	Cp	4 types of chest pain (1–typical angina; 2–a typical angina; 3–non-anginal pain; 4–asymptomatic)
4	Res blood pressure	Trestbps	Resting systolic blood (in mg Hg on admission to the hospital)
5	Scrum cholesterol	Chol	Serum cholesterol in mg/dl
6	Fasting blood pressure	Fbs	Fasting blood sugar>120 mg/dl (0–false; 1–true)
7	Rest electrocardiograph	Resteeg	0–normal; 1–having ST-T wave abnormality; 2–left ventricular hypertrophy
8	Max Heart rate	Thalch	Maximum heart rate achieved
9	Exercise induced angina	Exang	Exercise induced angina (0–no; 1–yes)
10	ST depression	Oldpeak	ST Depression induced by exercise relative to rest
11	Slope	Slope	Slope of the peak exercise ST segment (1–upsloping; 2–flat; 3–down sloping)
12	No. of vessels	Ca	No. of major vessels (0-3) colored by fluoroscopy
13	Thalassemia	Thal	Defect types; 3–normal; 6–fixed defect; 7–reversible defect
14	Num (class attribute)	Class	Diagnosis of heart disease status (0–nil risk; 1–low risk; 2–potential risk; 3–high risk; 4–very high risk)

Dealing with Outliers

Outliers may suggest test error, measurement variation, or anomalies. Outliers in the data were detected using the Z-score and IQR method (interquartile range). Outliers are found using the Z-score

approach. This approach is typically employed when a variable's distribution resembles a Gaussian distribution. The Z-score calculates how far a variable's value deviates from its mean by standard deviation.

$$(X\text{-mean})/(\text{Standard deviation})=Z\text{-Score}$$

By dividing the data set into quartiles, IQR is used to measure variability. The information is sorted in ascending order and divided into equal parts. As values that separate equal parts, Q1, Q2, and Q3 are referred to as the first, second, and third quartiles. In this case, an IQR with a cut-off value of 1.5 and unchanged outliers is used, according to our experience. The steps used are to first sort the dataset in ascending order, then calculate the first and third quartiles (Q1, Q3) and IQR=Q3Q1 before calculating the lower bound =(Q1–1.5*IQR) and upper bound =(Q3 + 1.5 * IQR). The algorithm then iterates through the values in the dataset, flagging those that are below the lower bound and above the upper bound as outliers.

By investigating the three classification methods listed below and doing performance analysis, the intended work anticipates cardiac disease. Effectively predicting whether a patient will have heart disease was the aim of this investigation. The Figure 1 depicts the full procedure involved in feeding the various characteristic data into a machine that is used to estimate the possibility of heart disease where we can observe different process like the pre-processing of the data and then dividing it to test and train datasets on which the final calculations are done.

Algorithms

The input dataset initially is divided into two parts: 80% for training and 20% for testing. A training dataset is a data set of variety of attributes that is utilized to teach a model of machine. The performance of the trained model is validated using the test dataset. The performance of each algorithm is calculated and analyzed using various attributes metrics like procedure, precision, prediction, accuracy and measure scores, as explained in more detail. The attributes mentioned in are fed into various ML algorithms such as K-Neighbors, Adaboost Classifier and Gradient Boosting. An outline about different algorithms used is defined below in the following section:

K-Nearest-Neighbors Classifier

The KNN algorithm is a supervised classification technique. It categorizes entities based on their nearest neighbors. It is a case-based learning. KNN is mainly based on feature similarity, that is, it assumes that similar things exist in close proximity. The Euclidean distance is used to calculate the distance between a classifier attribute and its neighbors. It takes a set of assigned points and based on which other points are marked. The information is clustered; clustering helps with grouping similar attributes and helps with fill up in the absent values of information using KNN. After the missing values are filled in, the information set is subjected to a variety of prediction techniques. It is possible to enhance accuracy value by combining all those algorithms in different ways. The KNN algorithm is easy to implement without the need of reference model and assumptions. This algorithm is flexible and can be used for many alternative methods like regression, classification, clustering and searching.

Adaboost Classifier

The AdaBoost algorithm, also known as Adaptive Boosting, is an ensemble method used in machine learning. Here, ensemble models means those models take the onus of combining different models and later produce a more advanced meta model which has higher accuracy which can be achieved in two different steps. The initial stage in which numerous weak learners are permitted to learn from practice data. The second phase involves combining these models to create a meta-model with the goal of correcting errors made by the weak learners individually. The influences are redistributed to each occurrence, with higher weights assigned to instances that were incorrectly classified, thus the term "adaptive boosting". Boosting is used in supervised learning to decrease bias and variance. It works on the basis that students advance in stages. With the exception of the first, each student after the first is developed from a previous learner. Simply put, weak students become strong students. The AdaBoost method is similar in concept to boosting, but it differs slightly.

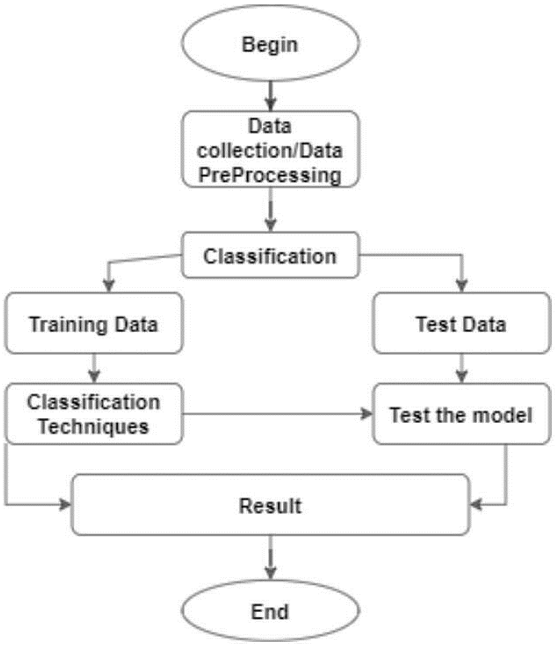

Figure 1. Structural flow.

Gradient Boosting

It is a machine learning technique that is employed in a variety of applications involving classification and regression. It offers a decision model in the form of a weak prediction model for the prediction of results. Each predictor in this case corrects the error of the preceding predictor, and each predictor is trained using the residual error of the preceding predictors as labels. Gradient Boosted Trees, whose base learner is CART (Classification and regression trees) which is the resulting method of methodology, typically outperforms the random forest algorithm when a decision tree is the weak model representation. As with prior boosting techniques, a Gradient Boosted Trees model is constructed in stages, but it simplifies those procedures by allowing the optimization of any differentiable loss function.

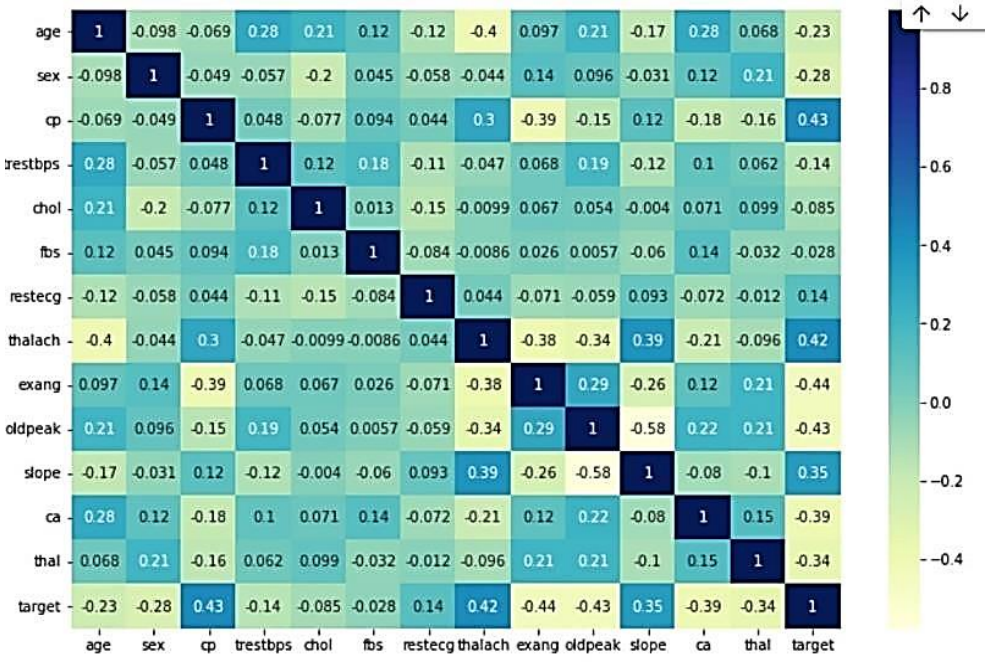

Figure 2. Correlation graph.

RESULTS

The results of using KNN, Adaboost and Gradient Boosting are shown in this section (Figure 2). An algorithm performance analysis is carried out using the accuracy score, precision (P), recall (R), and F-measure. Precision metrics provide an accurate measure of a positive analysis (Tables 2–4). The number of correct true positives is defined by recall. F-measure is a precision metric.

During the experimental analysis of the selected dataset, a correlation graph is plotted. A correlation attribute is much needed because each factor depends on another factor up to certain degree and it becomes necessary to analyze these relations to consider the important parameters because not all the values are necessary for prediction and not all the values are dependent on the factors and hence by plotting this graph, we can understand the dependencies of each attribute. The Figure 2 represents the correlation graph for our dataset with 14 different attributes, each plotted along x-y axis.

Table 2. Analysis of machine learning algorithm.

Algorithms	Accuracy
K-Neighbors	91%
AdaBoost	86%
Gradient Boosting	89%

Table 3. For patient without heart disease (0).

Algorithm	Precision	Recall	F1-score
K-Neighbors	1.00	0.94	0.97
AdaBoost	0.78	0.88	0.82
Gradient Boosting	0.79	0.94	0.86

Table 4. For patient with heart disease (1).

Algorithm	Precision	Recall	F1-score
K-Neighbors	0.97	1.00	0.98
AdaBoost	0.93	0.87	0.90
Gradient Boosting	0.96	0.87	0.91

Table 5. Confusion matrix.

Algorithms	True Negative	False Positive	False Negative	True Positive
K-Neighbors	15	1	0	30
AdaBoost	14	2	4	26
Gradient Boosting	15	1	4	26

Confusion matrix is plotted to understand the performance of the classification model (Table 5). For our problem statement for heart disease prediction, the extremely important aspect of the confusion matrix is the false negative. False negative is when we predict no for the patient who actually might suffer from the disease, which would cost heavy on the patient's health. So, from the above table we can observe that the false prediction count for K-Neighbor is zero when compared to other models and hence from this also we can conclude that K-Neighbors is better when compared to other algorithms

CONCLUSION

Given the rise in heart disease fatalities, it is critical to build a system that can forecast heart diseases precisely and effectively. This study is the driving force to discover the extremely effective Machine Learning algorithm used for detecting heart disease. Overall, the study compares all the results with the accuracy scores of the algorithms K-Neighbors, AdaBoost, and Gradient Boosting. Result of heart

disease prediction is predicted using data attributes from the UCI machine learning depository. The K-neighbors procedure is the best effective algorithm for expecting heart disease, according to the study's findings, with a score of 91% accuracy.

In the future, the study could be improved with developing a web-based application on the K-Nearest-Neighbors algorithm and using a higher attribute of dataset than the dataset used in this analysis of prediction. This would aid medical professionals in accurately and efficiently forecasting cardiac disease, resulting in better results.

Acknowledgement

We, the authors would like to thank Dr. B.S. Ajaykumar, Professor and Editorial member for your journal, who motivated and encouraged us to publish our article.

REFERENCES

1. Apurb Rajdhan, Milan Sai, Avi Agarwal, Dundigalla Ravi. Heart Disease Prediction using Machine Learning. Int J Eng Res Technol (IJERT). 2020; 9(4): 659–662.
2. Rindhe BU, Ahire N, Patil R, Gagare S, Darade M. Heart Disease Prediction Using Machine Learning. Heart Disease. 2021 May; 5(1): 267–273.
3. Jindal H, Agrawal S, Khera R, Jain R, Nagrath P. Heart disease prediction using machine learning algorithms. In IOP Conf Ser: Mater Sci Eng. 2021; 1022(1): 012072. IOP Publishing.
4. Shah D, Patel S, Bharti SK. Heart disease prediction using machine learning techniques. SN Comput Sci. 2020 Nov; 1(6): 1–6.
5. Bharti R, Khamparia A, Shabaz M, Dhiman G, Pande S, Singh P. Prediction of heart disease using a combination of machine learning and deep learning. Comput Intell Neurosci. 2021 Jul 1; 2021: 8387680.
6. Golande A, Pavan Kumar T. Heart disease prediction using effective machine learning techniques. Int J Recent Technol Eng. 2019 Jun; 8(1): 944–50.
7. Beyene C, Kamat P. Survey on prediction and analysis the occurrence of heart disease using data mining techniques. Int J Pure Appl Math. 2018 Jan; 118(8): 165–74.
8. Soni J, Ansari U, Sharma D, Soni S. Predictive data mining for medical diagnosis: An overview of heart disease prediction. Int J Comput Appl. 2011 Mar 8; 17(8): 43–8.
9. Seckeler MD, Hoke TR. The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease. Clin Epidemiol. 2011; 3: 67–84.
10. Mohan Senthilkumar, Chandrasegar Thirumalai, Gautam Srivastava. Effective heart disease prediction using hybrid machine learning techniques. IEEE Access. 2019; 7: 81542–81554.

