International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Review on Heart Disease Prediction Using Machine Learning Algorithms

Ms. Madhuri Thorat[1]*, Shruti Munot*[2]

[1]*Assistant Professor, Department of Information Technology, AISSMS's Institute of Information Technology, Pune-411001, INDIA*
[2]*TE. (Information Technology), AISSMS's Institute of Information Technology, Pune-411001, INDIA*

ABSTRACT

In the modern era, heart disease is one of the leading causes of death around the globe. High blood pressure, obesity, high cholesterol, drinking, and smoking, are some of the risk factors for heart disease. It is more difficult to predict cardiac disease than it is to make an automatic diagnosis of illness. Because healthcare facilities retain a vast amount of data that is complicated to interpret Because of these factors' modern approaches like Machine Learning and Data Mining are used for predicting the disease. These methods have proved effective in making predictions from a huge quantity of data produced by the healthcare industry. Here, we have used the dataset available in the Cleveland database of the University of California Irvine (UCI) machine learning repository. Parameters such as age, sex, height, weight, smoking, alcohol consumption, cholesterol, ECG, etc. are considered for prediction. We are analyzing and comparing various Machine Learning approaches to predict heart disease. Different machine learning approaches based on such as Decision Tree, Logistic Regression, Linear Regression Random Forest, Support Vector Machine, etc. Using these parameters and algorithms, we analyze the best algorithm to predict whether the patient has heart disease or not.

Keywords: classification algorithms, heart disease, machine learning

1. Introduction

Heart disease also known as cardiovascular disease, could be various conditions that impact the heart and is the reason for death worldwide. Heart disease term includes several heart-related diseases such as coronary artery disease, arrhythmia, and heart defects such as congenital heart defects [7]. Cardiovascular disease is a condition in which the blood vessel is narrowed or blocked, which leads to heart attack, or stroke. According to WHO, an estimated 17.9 million people have died of cardiovascular disease in 2019 [1]. Of these, 85% of deaths were due to heart attack and heart stroke. Various risk factors such as smoking, alcohol consumption, overuse of caffeine, mental and physical stress, as well as physiological factors such as high blood pressure, high blood cholesterol, obesity, diabetes, family history of coronary illness and not being physically active can cause heart disease [7]. Identification of disease at the primary stage can reduce the death risk. ML techniques are used to make medical aid software for early diagnosis of heart-related illnesses. By using machine learning techniques, we can predict heart disease with help of the medical history of the patient. Machine learning incorporates various classifiers of supervised and unsupervised learning which are used for predicting and finding the accuracy of a given dataset [1]. ML algorithms use past data to make predictions. We require training data, to learn ML algorithms. After learning a model is produced as an output of the ML algorithm. This model is tested on real-time data set. The accuracy of a model is compared with the actual value and can find the overall accuracy of the predicted result

2. Literature Survey

This literature survey consists of five research papers. The papers are studied and the implementation of various machine learning algorithms is observed. The papers along with the algorithms used and the accuracy obtained are mentioned in the below table 1.

Table 1: Literature Survey

Sr. No.	Author/Year	Paper Name	Algorithms Used	Accuracy Obtained
1.	Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain, and Preeti Nagrath, 2021 [1]	Heart Disease Prediction using Machine learning algorithms.	KNN Logistic Regression LR, RF Classifier and KNN Model	88.52% 88.5% 87.5%
2.	Vijeta Sharma, Shrinkhala Yadav, Manjari Gupta, 2020 [2]	Heart Disease Prediction using Machine Learning Techniques	Random Forest SVM Naïve Bayes	99% 98% 90%
3.	Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava, 2019 [3]	Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques	SVM Logistic Regression Random Forest Decision Tree HRFLM Model	86.1% 82.9% 86.1% 85% 88.4%
4.	Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin, 2019 [4]	Design And Implementing Heart Disease Prediction Using Naive Bayesian	Naïve Bayesian	89.77%
5.	Pahulpreet Singh Kohli, Shriya Arora, 2018 [5]	Application of Machine Learning in Disease Prediction	Logistic Regression Decision Tree Random Forest Support Vector Machine	87.12% 70.97% 77.42% 83.87%
6.	Liaqat Ali, Awais Niamat; Javed Ali Khan; Noorbakhsh Amiri Golilarz; Xiong Xingzhong, 2019 [6]	An Optimized Stacked SVM Based Expert system for the Effective Prediction of Heart Failure	Two SVM models were used. The first model is to eliminate irrelevant features and the second is for the predictive model.	92.22%
7.	Santhana Krishnan J., Geetha S., 2019 [7]	Prediction of Heart Disease Using Machine Learning Algorithms	Decision Tree Naïve Bayes	91% 87%
8.	Sai Bhavan Gubbala, 2022 [8] DOI An Intelligent Learning System based on Random Search Algorithm and	**Optimized Random Forest Model for Improved Heart Disease Detection** Heart Disease Prediction Using Machine Learning Techniques**.**	Random Forest Logistic Regression Decision Tree Support Vector Machine Naïve Bayes SVM	85.22% 81.23% 74.34% 67.43% 84% 88%

Methodology

3.1 Data Collection and Pre-Processing

In this study, we collected a heart disease dataset known as the Cleveland heart disease database from an online machine learning and data mining repository of the University of California, Irvine (UCI). The dataset consists of 303 subjects. However, the data of 6 subjects have missing values. Thus, the data of 297 subjects are considered for experiments. The original dataset has 76 raw features per subject. But, most of the previous studies used only 13 of them. Hence, in this study, the commonly used 13 HF features are considered. The features are shown in the below table 2. Null value verification, loading python libraries, and splitting the dataset into training and testing data are all included in the preprocessing of the dataset [8].

Table 2. Features of Dataset

Sr. No.	Author/Year	Paper Name
Attribute	**Description**	**Range**
Age	Patient’s age in complete years	29 to 77
Sex	Patient’s gender (male =0, female=1)	0 to 1
Cp	Type of chest pain – 4 values: 1. Typical angina 2. atypical angina 3. non-anginal pain 4. asymptomatic	1 to 4

Trestbps	Level of blood pressure at resting (in mm/Hg at the time of admitting to hospital)	94 to 200
Chol	Serum cholesterol in mg/dl	126 to 564
FBS	Blood sugar level on fasting > 120 mg/dl (1 in case of true & 0 in case of false)	0 to 1
RestECG	Results of ECG while at rest	0 to 2
Thalach	The accomplishment of a maximum rate of heart	71 to 202
Exang	Angina induced by exercise (0 = no, 1 = yes)	0 to 1
Oldpeak	Exercise-induced ST depression in comparison with the state of rest	0 to 6.20
Slope	ST segment as slope during peak exercise, 3 values: 1. Unsloping 2. Flat 3. Downsloping	1 to 3
Ca	Fluoroscopy colored major vessels from 0 to 3	5 levels (Categorical)
Thal	Status of heart given numbered values. Normal =3, fixed defect = 6, reversible defect = 7	4 levels (Categorical)

3.2 Architecture Diagram

Following fig.1 [3], shows the experimental workflow with the UCI dataset. The UCI dataset is used for pre-processing of data. The pre-processed data is then used for feature selection. Various machine learning algorithms are used to prediction and model generation. The evaluation of these model is done and the accuracy, precision, F-measure, sensitivity and specificity is calculated.

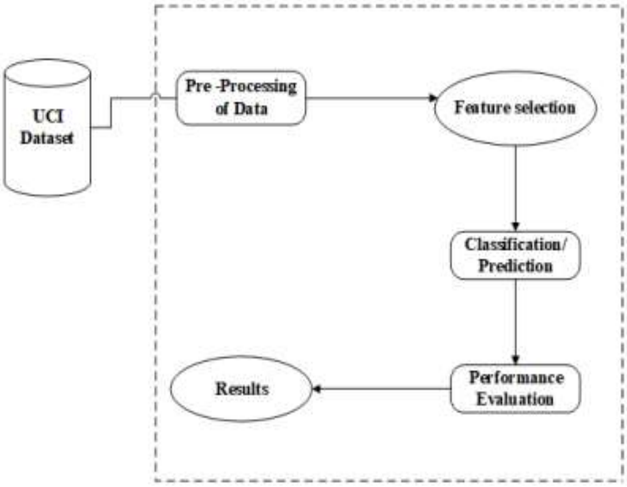

Fig.1. Experiment workflow with the UCI dataset

3.3 Algorithms

3.3.1 Decision Tree

A Decision Tree is a supervised learning algorithm, which is used for classification and regression. It can handle both categorical and numerical data. The output of this algorithm consists of vertical and horizontal line splits. It analyses the dataset in a tree-shaped format. The tree consists of nodes. Each node indicates an attribute in a group that is to be classified along with each branch symbolizes a value that the node can take.

3.3.2 Naïve Bayes

Naïve Bayes classifier is a supervised algorithm that classifies the dataset on the basis of the Bayes theorem. It is based on the probability theorem. It determines the class of the feature vector for a given class. Based on the conditional probability of each vector, the new vector is found. It is used for text classification.

3.3.3 Support Vector Machine (SVM)

SVM is a machine learning algorithm and is used to categorize the dataset. It is used for classification-related problems. SVM discovers the hyperplane which divides the two classes. It divides the dataset into these two categories and finds maximum marginal hyperplanes

3.3.4 K-Nearest Neighbor

KNN is a supervised learning algorithm. It is used for classification problems. It classifies objects depending on the nearest neighbor. KNN finds a predetermined number of training samples closest to a new point. The data is divided into training and testing datasets. It is the simplest algorithm but has noisy features which reduce the accuracy.

3.3.5 Random Forest

It is a supervised learning algorithm. Random forest is used for both classification and regression problems. In this, several trees create a forest. The tree created based on data is used for prediction. It is used for large datasets. The Random Forest can handle large datasets and can predict the same results even if the values in the dataset are missing. It uses the random forest classifier to make prediction.

3.3.6 Logistic Regression

Logistic regression is a machine learning algorithm, used for binary classification-related problems. It predicts the probability of the target variable. The given output of the linear equation is between 0 and 1. The 0 is for failure and the 1 represents success. Logistic regression uses most probability Estimation (MLE) to find the logistic regression coefficients. To squash the predicted value between 0 and 1, we use the sigmoid function.

Conclusion

From the study of various recent research papers written on heart disease prediction using various data mining and machine learning techniques and algorithms. We find that different techniques of data mining and machine learning are used to predict heart disease. Different datasets of heart disease patients are used in different experiments. In most experiments dataset used is taken from the online Cleveland database of UCI. In this survey, we learned how to apply different machine learning techniques to predict heart attacks. In some research studies, we learned that more accuracy can be obtained by the hybridization of two or more different algorithms.

References

[1] Harshit Jindal, “Heart Disease Prediction Using Machine Learning Algorithms”, IOP Conf. Ser.: Mater. Sci. Eng. 1022 012072, 2021

[2] Vijeta Sharm, Shrinkhala Yadav, Manjari Gupta, “Heart Disease Prediction using Machine Learning Techniques”, 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020.

[3] Senthilkumar Mohan, Chandrasegar Thirumalai and Gautam Srivastava, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques” IEEE Access, DOI 10.1109/ACCESS.2019.2923707, pp. 81542-81554, 2019

[4] Anjan Nikhil Repaka, Sai Deepak Ravikanti, and Ramya G Franklin,” Design and Implementing Heart Disease Prediction Using Naives Bayesian” IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8, pp. 292-297, 2019

[5] Pahulpreet Singh Kohli and Shriya Arora, “Application of Machine Learning in Disease Prediction” IEEE, 978-1-5386-6947-1/18, pp. 1-4, 2018.

[6] Liaqat Ali, Awais Niamat; Javed Ali Khan; Noorbakhsh Amiri Golilarz; Xiong Xingzhong, "An optimized stacked support vector machines based expert system for the effective prediction of heart failure." IEEE Access 7 (2019): 54007-54014.

[7] Santhana Krishnan J., Geetha S., Prediction of Heart Disease Using Machine Learning Algorithms, (ICIICT) IEEE, 2019. DOI:10.1109/ICIICT1.2019.8741465.

[8] Sai Bhavan Gubbala, “Heart Disease Prediction Using Machine Learning Techniques”, IRJET, Volume 09, Issue 10, 2395-0072, 2022