International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT228686

A Review on Machine Learning-Based Algorithms for Heart Disease Diagnosis and Prediction

Aatiq Ali*, Dr. L C Manikandan
CSE, Valia Koonambaikulathamma College of Engineering and Technology, APJ Abdul Kalam Technological University, Thiruvananthapuram, Kerala, India

Article Info

Publication Issue :
Volume 8, Issue 6
November-December-2022
Page Number : 606-611
Article History
Accepted: 05 Dec 2022
Published: 22 Dec 2022

ABSTRACT

This work concentrates on different machine learning techniques that can be applied for the early diagnosis of heart diseases. Different Machine Learning Algorithms like KNN (K-Nearest Neighbour), Decision tree, Logistic Regression, Support Vector Machine, Random Forest and Naïve Bayes are used for prediction and detecting heart disease. The main purpose of this survey is to pay our attention towards machine learning techniques for diagnosing heart diseases in an efficient and effective manner.

Keywords: Machine Learning, KNN, Decision Tree, Logistic Regression, SVM, Random Forest, Native Bayers

I. INTRODUCTION

Heart is very important organ on the body and is responsible for normal blood flow throughout the body and so any problem to the organ cause serious issue in other parts of the body. Nowadays heart diseases are increasing like anything due to unhealthy lifestyle, smoking, taking alcohol etc. As per the records and according to World Health Organization the main cause for death is diseases related to heart. It is not even considering the case of India and it is true while considering the whole world.

Heart diseases include, heart infections, heart failure, cardiac arrest, hypertension, slow heartbeat, and stroke. Many factors for heart diseases are age, family history of coronary illness, blood pressure, and Cholesterol level. The heart related issues are responsible for millions and millions of death every year. Early detection and diagnosis of the disease is very important and helps in the treatment of the disease. Prediction plays an important role, it helps a lot for the successful treatment of heart disease. Artificial intelligence is a major field of computer science and engineering has been used for diagnosis of many diseases. Machine learning a part of AI is having a major role in the diagnosis of many diseases. Researches in the field of machine learning techniques are very useful for the early diagnosis, detection and prediction of these diseases and thereby a successful treatment can be given to the patient.

II. RELATED WORKS

Many researchers used machine-learning techniques that help in healthcare industry in previous years and specialists in the detection of heart-related illnesses. K-Nearest Neighbour (KNN), Logistic Regression (LR), Support Vector Machine, Decision Tree,

Random Forest, and Naive Bayes are some examples, Lot of work has been carried out in the existing system, by using Machine learning techniques and different datasets different accuracy have been attained.

Senthil Kumar et al, [2] improved prediction of cardiovascular disease using composite machine learning techniques that includes a methodology, which seeks to find significant implication through the applying machine learning, leading to enhanced accuracy within prediction of cardiovascular illness.

Abhay Kishore et al, [3] used Deep Learning to develop heart attack predictions Recurrent Neural Network which is a revolutionary characterisation approach that makes use of the Deep Learning technique in Artificial Neural Network, This paper goes into great detail on the framework's primary modules, and also the associated assumption. This suggested methodology uses deep learning with data mining to achieve the most accurate results with the lowest failures. This work provides a foundation and reference point for the building of a different variety of heart attack prediction platform.

Vembandasamy et al, [4] work was performed by using Naïve Bayes Algorithm which is a powerful independence assumption, the data was obtained from diabetic research institute and it consists of 500 patients record and Naïve Bayes Algorithm offers 86.919% of accuracy.

Mr. Santhana Krishnan. J and Dr.Geetha. S, [5] used classification techniques to predict cardiac disease in male patients. This work provides comprehensive data regarding Heart Diseases, covering Background, Prevalent Type, and Factors Associated. All three weak interfaces are used there; the key data mining methodologies are Naive Bayes, Artificial Neural Networks, and Decision Trees, and these techniques are used to predict heart disease.

The Table 1 shows the accuracy of various Machine Learning Algorithms used for diagnosis of heart diseases [1].

TABLE I. A COMPARATIVE STUDY OF VARIOUS ML ALGORITHMS

Author	Disease	Algorithm Used	Year	Accuracy %
Shan Xu	Heart disease	SVM	2017	98.9
Kamran Farooq	Heart disease	DT	2014	78.4604
Otoom	Heart disease	SVM	2015	88.3
Syed Muhammad Saqlain Shah	Heart disease	SVM	2017	91.30
Megha Shahi	Heart disease	SVM	2017	85
Vembanda-samy	Heart disease	Naïve bayes	2018	86.41
AbhuKidhorel	Heart attack	RNN	2019	92
Santhana Krishnan. J	Heart disease	NB, DT	2019	91
Senthilkumar Mohan	Heart disease	DT, SVM	2019	88.4

III. METHODOLOGY

Figure.1 shows the basic step which have been taken for all machine-learning models. Initially load the set of data. Pre-processing,it involve data cleaning to clean and format the data because raw data cannot be used directly. After pre-processing step the feature should be selected. Based on these features divide the data into different parts and testing is carried out. In Testing, a test was conducted to make sure we are getting the right answer. After cleaning and analyzing all dataset, apply machine-learning models such as SVM Logistic Regression, Naïve Bayas (NB), (DT) Decision Tree, and RD and then go for cross validation. Cross validation, it is the technique to evaluate predictive models by partitioning the original sample into a training set and test set. After these process predict the result. Then the result Analysis is done with suitable Normal and Abnormal metrics.

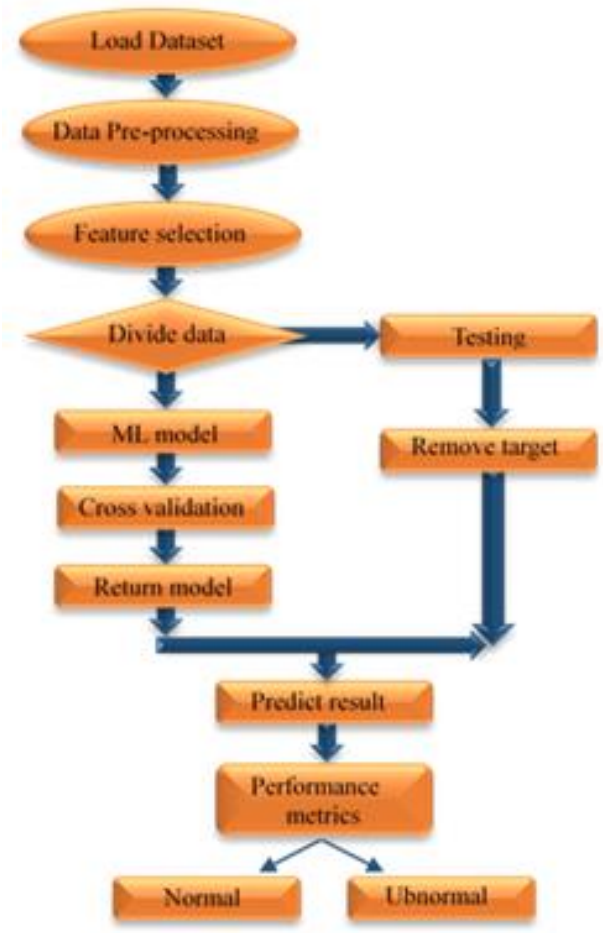

Figure 1. Machine Learning Model workflow

IV. ML ALGORITHMS AND TECHNIQUES

A. Support Vector Machine

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is usedfor Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes, so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Support Vector Machine is shown in Figure 2.

Figure 2. Support Vector Machine

B. Naïve Bayes

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

$$p\left(\frac{A}{c}\right) = \frac{p\left(\frac{A}{c}\right)p(c)}{p(A)} \quad \text{---------- Eqn 1}$$

C. Decision Tree Algorithm

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas

Leaf nodes are the output of those decisions and do not contain any further branches.

The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm. A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree in to sub-trees.

$$Entropys\ (S) = \sum_{i=1}^{c} -p_i\ log2\ p_i \qquad \text{---------- Eqn 2}$$

Information Gain:

$$Gain(S, A) = Entropys\ (S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \qquad \text{--------- Eqn 3}$$

D. Logistic Regression

The Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

E. Random Forest

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of over fitting. Figure 3 shows the random forest.

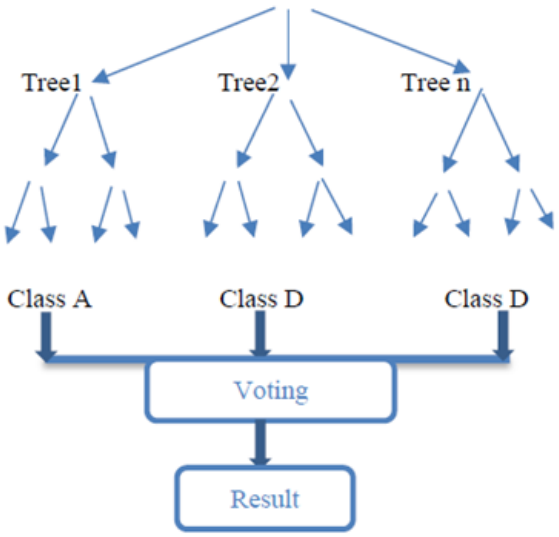

Figure 3. Random Forest

F. K-Nearest Neighbour

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and

put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data. KNN shown in Figure 4.

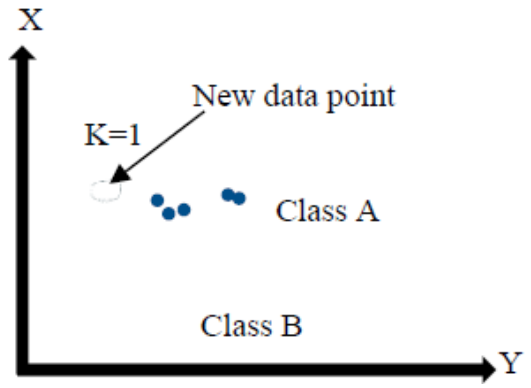

Figure 4. KNN

V. CONCLUSION

In this study, we briefly explain about the machine learning approach for early detection of heart malfunctioning. And this paper provides a comparative study on different machine learning approach like SVM, Decision tree algorithm, Logistic regression algorithm, KNN, Random Forest and Naive Bayes. This paper concludes that the SVM and Naves Bayes performed extremely well as compare with other techniques and the Decision tree algorithm is so poor because of large number of datasets.

VI. REFERENCES

[1] A. A. Ahdal, D. Prashar, M. Rakhra and A. Wadhawan, "Machine Learning-Based Heart Patient Scanning, Visualization & Monitoring, "International Conference on Computing Sciences (ICCS), pp. 212-215, 2021. doi: 10.1109/ICCS54944.2021.00049.

[2] Senthilkumar Mohan, ChandrasegarThirumalai, Gautam Srivastava, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques, Digital Object Identifier”, IEEE Access, Vol.7, 2019 Proc. Fourth Ann. Allerton Conf. Circuits and Systems Theory, pp. 8-16.

[3] Abhay Kishore1, Ajay Kumar2, Karan Singh3,Maninder Punia4, Yogita Hambir5, ”Heart Attack Prediction Using Deep Learning”, International Research Journal of Engineering and Technology(IRJET), Vol.05, Issue: 04, April 2018.

[4] Vembandasamy, K., Sasipriya, R. and Deepa, E, “Heart Diseases Detection Using Naive Bayes Algorithm”, International Journal of Innovative Science, Engineering & Technology, Vol.2, PP.441-444, 2015.

[5] SanthanaKrishnan.J, Dr.Geetha.S, “Prediction of Heart Disease Using Machine Learning Algorithms”, International Conference on Innovations in Information and Communication Technology(ICIICT), 2019. doi:10.1109/ICIICT1.2019.8741465.

[6] Victor Chang, Vallabhanent Rupa Bhavani, Ariel Qianwen Xu, MA Hossain, “An artificial intelligence model for heart disease detection using machine learning algorithms”, Healthcare Analytics, Vo.2, 2022.

[7] Mahesh, Batta, “Machine Learning Algorithms-A Review”, 2019.

AUTHOR PROFILE

Aatiq Ali is studying B.Tech., degree in Computer Science and Engineering at Valia Koonambaikulathamma College of Engineering and Technology under APJ Abdul Kalam Technological University, Thiruvananthapuram, Kerala, India. He trained on blockchain foundation program conducted by kerala blockchain academy. He also interested in Adobe Photoshop, DBMS and Web developing tools. He received award from college for securing good CGPA in academics.

Dr.L.C.Manikandan is working as Professor at Valia Koonambaikulathamma College of Engineering and Technology, Thiruvananthapuram, Kerala INDIA. He has received his Ph.D. and M.Tech. Degree in Computer and Information Technology from Manonmaniam Sundaranar University, M.Sc., and B.Sc. degree in Computer Science from Bharathidasan and Manonmaniam Sundaranar University. He has 18 years of teaching experience in reputed institutions. He has published several research papers in various reputed international journals and published three textbooks. He carries out research in Digital Image Processing, Video Surveillance and Video coding.

Cite this article as :

Aatiq Ali, Dr. L C Manikandan, "A Review on Machine Learning-Based Algorithms for Heart Disease Diagnosis and Prediction ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 6, pp. 606-611, November-December 2022. Available at doi : https://doi.org/10.32628/CSEIT228686
Journal URL : https://ijsrcseit.com/CSEIT228686