Webology (ISSN: 1735-188X)
Volume 19, Number 5, 2022

Diagnosis Of Heart Disease Using Machine Learning Algorithm

Pradeep Kumar[1], Dr. L.S. Maurya[2], Hiresh Kumar Gupta[3]

[1]Assistant Professor, [2]Professor, [3]Assistant Professor

[1,2,3]Department of Computer Science & Engineering

[1,2]Shri Ram Murti Smarak College of Engineering, Technology & Research, 13 Km, Bareilly-Nainital Road, Ram Murti Puram, Bareilly - 243202, (U.P.) – INDIA

[3]Shri Ram Murti Smarak College of Engineering & Technology, 13 Km, Bareilly-Nainital Road, Ram Murti Puram, Bareilly - 243202, (U.P.) – INDIA

Abstract— This Machine Learning techniques is involved in artificial intelligence. Today is the scenario being most problem of the heart disease in the world. So, Prediction of heart disease at the initial stage may reduce death ratio. Now a day's various healthcare organizations generate bulk amount of data but that data is unorganized. If the data are organized so there is various technique to used easily predict the heart disease. If this data is organized in a proper way using data mining technique it can be easily use for the prediction of heart diseases. Thus, the objective of this paper is to make a model for diagnosis the heart disease based on the various parameter. Using dataset of heart disease prediction for this work, which consist of 14 different types of parameters related to heart problem. Machine Learning algorithms such as supervised and unsupervised algorithm are such as Random Forest, Support Vector Machine, Naive Bayes and Decision tree, clustering but we used Logistic Regression algorithm have been used for the development of model. This model can be very helpful to the clinical diagnosis for early stage to detect the heart problem

Keywords Machine Learning, Heart disease, Artificial Intelligence, classification, Data Mining, SVM, Random Forest.

I. INTRODUCTION

Machine Learning is a branch of Artificial intelligence [1] and has become of hues amount of data in data science. The Machine Learning techniques are designed to perform a large number of tasks such as prediction, classification, decision making etc. To learn the ML algorithms, training data is required. After the training, a model is produced which is based on an output of the machine learning algorithm. This model is then tested and apply untrained data to diagnose the heart problem. Identification of any heart related illness at early stage can reduce the death rate. Various ML techniques are used in medical data to understand the pattern of data and making prediction from them. Healthcare data are generally big amount of and complex in structure. ML algorithms are capable to handle the big data

and diagnosis them to find the meaningful information [2]. Machine Learning algorithms learn from past data and do diagnosis on real time data. We use the Logistic regression algorithm to make a model to detect heart problem

II. REVIEW OF WORK

A number of studies evaluating the performance of machine learning algorithm like Decision Tree [3], Random Forest etc. using algorithm found to different accuracy. Devansh et al [4] has ML algorithm and diagnosis on using various parameters. In this paper used different algorithm for the experiment. This paper discussed Naïve Bayes, and random forest, K-nearest neighbour, are the algorithms showing the best results in this model. C. Beulah et al [5] evaluated using machine learning technique he was found feature selection accuracy up to approximate 85 percentages. Beulah was using parameter smoking, family history, Age, sex, cholesterol, poor diet, high blood pressure, obesity, physical inactivity, and alcohol intake are considered to be risk factors for heart disease, and heart disease can be caused by inherited risk factors such as high blood pressure and diabetes. Some risk factors are controllable. [6]]. There are different types of heart diseases, angina pectoris, congestive heart failure, cardiomyopathy, congenital heart disease [7].

III. CLASSIFICATION ALGORITHMS

Classification is a supervised learning technique that uses previously acquired data to anticipate the outcome. This study recommends that classification methods be used to diagnose heart illness, with an ensemble of classifiers being used to improve classification accuracy. The training dataset, which is divided into two sections: training and testing, is used to train individual classifiers. The performance of the classifiers is assessed using the test dataset

A. **Logistic Regression algorithm**:

Under the Supervised Learning approach, one of the most prominent Machine Learning algorithms is logistic regression. It's a method for predicting a categorical dependent variable from a set of independent variables. A categorical dependent variable's output is predicted using logistic regression. Logistic regression may be used to categorise observations based on many forms of data and can quickly identify the most useful factors for classification. The cost function used in Logistic Regression is more sophisticated, and it is known as the 'Sigmoid function' or the 'logistic function.' The Sigmoid function is given below

$$f(x) = \frac{1}{1+e^{-(x)}}$$

f(x) = output between 0 and 1 (probability estimate)
x = input to the function
e = base of natural log

A. Support Vector Machine

Support Vector Machine [11] is a machine learning classification technique for analysing data and detecting patterns in classification and regression analysis. When data is classified as a two-class problem, SVM is usually considered. Data is described in this technique by determining the optimum hyper plane that isolates all data points from one class from the other. The greater the separation or edge between the two classes, the better the model is thought to be. Support vectors are data points that are located near the margin's edge. The mathematical methods used to construct complex real-world situations form the foundation of SVM. Because our dataset - Cleveland Heart Disease Dataset CHDD - comprises multiple classes to forecast based on various factors, we picked SVM for this project. The mapping of training data in SVM is done via a function called a kernel (SVM kernels), which include linear kernels, quadratic kernels, polynomial kernels, Radial Basis Function kernels, Multilayer Perceptron kernels, and so on. In addition to the kernel's capabilities, SVM offers a few other methods such as quadratic programming, sequential minimal optimization, and least squares.

B. Decision Tree

The Classification models are created using the Decision Tree method [12] from Machine Learning. The tree-like structure underpins this classification methodology. This is classified as supervised learning because the desired outcome is already known. The Decision tree algorithm can be used with both category and numerical data. The root node, branches, and leaf nodes make up a decision tree. The traversal path from the root to a leaf node is used to evaluate the data. A total of 303 tuples were evaluated down the decision tree for our dataset - CHDD. They could have come to a favourable or negative conclusion about the risk of heart disease. These were compared to the actual parameters to see if there were any false positives or negatives.

C. Naive Baye

The Bayes' Theorem [13] underpins this supervised machine-learning approach, which assumes that features are statistically independent of one another. With high dimensionality of input data, the Nave Bayes Classifier [14] is utilised. In computer vision, the Naive Bayes approach is quite valuable. It has demonstrated itself to be a good classifier in particular.

D. Random Forest

Random Forest [15] is a group of classification-based trees that haven't been trimmed. It performs admirably in a variety of real-world issues since it is unaffected by noise in the dataset and the risk of overfitting is minimal. It is faster than many other tree-based algorithms and enhances accuracy for testing and validation data. The aggregation of the predictions of individual decision tree algorithms is known as random forests. When constructing a random tree, there are several options for tuning the random forest's performance.

E. K-Nearest Neighbours

Nearest neighbour (KNN) is a pattern recognition method that is relatively basic, widely used, and highly efficient and effective. KNN is a simple classifier that classifies samples based on the class of their nearest neighbour. The data is divided into training and test samples using KNN algorithms. The distance between the training point and the sample point is calculated, and the point with the shortest

distance is referred to as the nearest neighbour. When all of the attributes are continuous, nearest neighbour classification is utilised.

F. XG Boost

Extreme Gradient Boosting, or XGBoost, was proposed by academics at the University of Washington. XGBoost is a supervised learning algorithm that uses training data (with numerous features) xi to predict a target variable y_i . That XGBoost is a library for creating high-performance gradient boosting tree models in a short amount of time. That XGBoost outperforms the competition on a variety of tough machine learning tasks

G. Neural Network

Neural networks are a type of machine learning technique that uses numerous hidden layers and non-linear activation functions to describe complicated patterns in datasets. A neural network takes an input, runs it through numerous layers of hidden neurons, and then returns a prediction that represents all of the neurons' combined input. Iterative optimization techniques such as gradient descent are used to train neural networks. An error measure is calculated after each training cycle based on the difference between prediction and target. Using a technique known as backpropagation, the derivatives of this error metric are calculated and sent back through the network. Inside neural network layers, activation functions change the data before passing it on to the next layer. The power of neural networks comes from their activation functions, which allow them to simulate complicated non-linear interactions. Neural networks can simulate exceedingly complicated interactions between characteristics by changing inputs with non-linear functions. Relu and sigmoid are two popular activation functions.

IV. METHODOLOGY

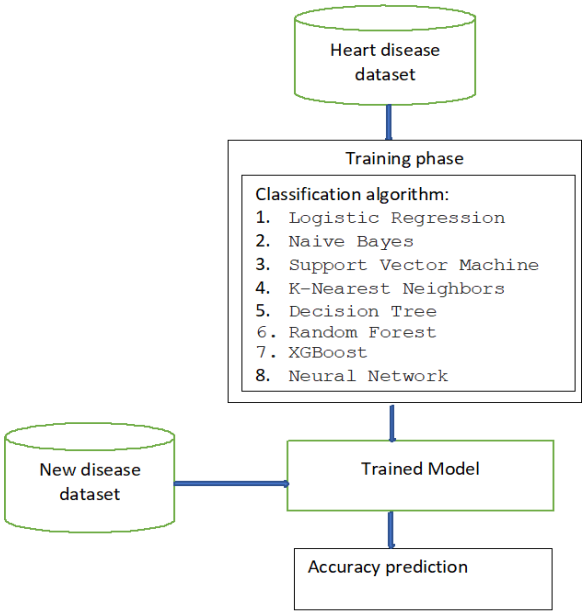

Fig:1 Methodology of the prediction model

The heart disease prediction can be performed by following the given fig1 which specifies the building a classification model required for the prediction of the heart disease in patients. This model first to making a training model using a known data to performance the accuracy. Using various classifiers such as logistic regression, Naïve Bayes, SVM, KNN, Decision tree, Random Forest, XGboost and Neural Network. And apply the heart disease dataset to obtain the trained model. And after we obtain the trained model, we apply new patient dataset and prediction is found.

In the percentage split, the training and testing data is split up in percentage of data such as 80% and 20% where the 80% is used for training and 20% is used for testing. In this work, the training phase includes training the eight classification algorithms namely logistic regression, Naïve Bayes, SVM, KNN, Decision tree, Random Forest, XGboost and Neural Network using the heart disease dataset and a classification model is built.

V. EXPERIMENTS RESULTS AND ANALYSIS

The database for this study was derived from a dataset in the UCI repository. It has 13 characteristics. There are 303 incidences of heart disease in the dataset used in this study, with no missing values. The dataset is commonly used for cardiac disorders such typical angina, atypical angina, non-anginal discomfort, and asymptomatic angina. The goal of this study is to predict cardiac disease regardless of the type of disease. The age of the patient is represented by a numeric data type that varies from 29 to 65 years. The Cp is a number that ranges from 1 to 4 and is used to determine the pain kind. The trestbpd is the resting blood pressure, which ranges from 92 to 100, and the fbs is the fasting blood sugar level, which is either a 1 or a 0, reflecting true or false Boolean values. The resting electrocardiographic result is displayed as three cases ranging from 0 to 2 in the restecg. The thalach is the maximal heart rate achieved, which can range from 82 to 185 beats per minute. The exang is a Boolean value that represents exercise-induced angina. The disease is the dataset's target class, with yes or no answers indicating the presence of heart disease. Similarly, all the attributes and their values are represented in Table.1

Table1: Attribute of the dataset use

S.N.	Attribute	Information
1	age	numeric
2	sex	1:male, 0:female
3	cp	chest pain type, 1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic
4	trestbps	resting blood pressure
5	chol	serum cholestoral in mg/dl

6	fbs	fasting blood sugar > 120 mg/dl
7	restecg	resting electrocardiographic results (values 0,1,2)
8	thalach	maximum heart rate achieved
9	exang	exercise induced angina
10	oldpeak	oldpeak = ST depression induced by exercise relative to rest
11	slope	the slope of the peak exercise ST segment
12	ca	number of major vessels (0-3) colored by flourosopy
13	thal	thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

We aaply above attribute on the trained model with classifier such that logistic regression, Naïve Bayes, SVM, KNN, Decision tree, Random Forest, XGboost and Neural Network we obtain different accuracy of the prediction of heart disease shown in the table 2 is given below

Table2: Accuracy of: heart disease prediction table

S.N0.	Classification algorithm	Accuracy (%)
1	Logistic Regression	85.25
2	Naive Bayes	85.25
3	SVM	81.97
4	KNN	67.21
5	Decision Tree	81.97
6	Random Forest	95.16
7	XGBoost	85.25
8	Neural Network	83.16

VI. CONCLUSIONS

In this research work analysis of the heart disease patient dataset with proper data processing. And using different models were trained and predictions are made with different algorithms KNN, Decision Tree, etc. Overall, it is found Random Forest is the best algorithm in the above classifier we obtain Random Forest is the best classifier to get the approximate 95 % achieved.

VII. REFERENCES

[1] K. M. Almustafa, "Prediction of heart disease and classifiers' sensitivity analysis," BMC Bioinformatics, vol. 21, no. 1, pp. 1–18, 2020, doi: 10.1186/s12859-020-03626-y.

[2] G. Choudhary and S. Narayan Singh, "Prediction of heart disease using machine learning algorithms," Proc. Int. Conf. Smart Technol. Comput. Electr. Electron. ICSTCEE 2020, vol. 7, pp. 197–202, 2020, doi: 10.1109/ICSTCEE49637.2020.9276802.

[3] P. Dutta, S. Paul, N. Shaw, S. Sen, and M. Majumder, "Heart Disease Prediction," Artif. Intell. Cybersecurity, pp. 1–18, 2021, doi: 10.1201/9781003097518-1.

[4] N. Foster, "The Future of Heart Attack Prediction," Mended Hear. Inc., 2021, [Online]. Available: https://mendedhearts.org/story/the-future-of-heart-attack-prediction/.

[5] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Knowledge discovery using associative classification for heart disease prediction," Adv. Intell. Syst. Comput., vol. 182 AISC, pp. 29–39, 2013, doi: 10.1007/978-3-642-32063-7_4.

[6] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," IOP Conf. Ser. Mater. Sci. Eng., vol. 1022, no. 1, 2021, doi: 10.1088/1757-899X/1022/1/012072.

[7] S. G. Kanakaraddi, K. C. Gull, J. Bali, A. K. Chikaraddi, and S. Giraddi, "Disease prediction using data mining and machine learning techniques," Lect. Notes Data Eng. Commun. Technol., vol. 64, pp. 71–92, 2021, doi: 10.1007/978-981-16-0538-3_4.

[8] V. Krishnaiah, G. Narsimha, and N. Subhash, "Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review," Int. J. Comput. Appl., vol. 136, no. 2, pp. 43–51, 2016, doi: 10.5120/ijca2016908409.

[9] C. Krittanawong et al., "Machine learning prediction in cardiovascular diseases: a meta-analysis," Sci. Rep., vol. 10, no. 1, pp. 1–11, 2020, doi: 10.1038/s41598-020-72685-1.

[10] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," Informatics Med. Unlocked, vol. 16, no. June, p. 100203, 2019, doi: 10.1016/j.imu.2019.100203.

[11] D. M and R. V, “Prediction Of Heart Disease Using Back Propagation MLP Algorithm,” Int. J. Sci. Technol. Res., vol. 4, no. 8, pp. 235–239, 2015.

[12] Z. Masetic and A. Subasi, “Congestive heart failure detection using random forest classifier,” Comput. Methods Programs Biomed., vol. 130, pp. 54–64, 2016, doi: 10.1016/j.cmpb.2016.03.020.

[13] Mehdi Khundmir Iliyas and I. S. Shaikh, “Prediction of Heart Disease Using Decision Tree,” Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 6, no. 3, pp. 530–532, 2016, [Online]. Available: https://www.researchgate.net/publication/339106269.

[14] S. Mohan, C. Thirumalai, and G. Srivastava, “Effective heart disease prediction using hybrid machine learning techniques,” IEEE Access, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[15] A. Rairikar, V. Kulkarni, V. Sabale, H. Kale, and A. Lamgunde, “Heart disease prediction using data mining techniques,” Proc. 2017 Int. Conf. Intell. Comput. Control. I2C2 2017, vol. 2018-Janua, no. October, pp. 1–8, 2018, doi: 10.1109/I2C2.2017.8321771.

[16] K. Srivastava and D. K. Choubey, “Heart Disease Prediction using Machine Learning and Data Mining,” Int. J. Recent Technol. Eng., vol. 9, no. 1, pp. 21–219, 2020, doi: 10.35940/ijrte.f9199.059120.

[17] R. Williams, T. Shongwe, A. N. Hasan, and V. Rameshar, “Heart Disease Prediction using Machine Learning Techniques,” 2021 Int. Conf. Data Anal. Bus. Ind. ICDABI 2021, pp. 118–123, 2021, doi: 10.1109/ICDABI53623.2021.9655783.

[18] C. S. Wu, M. Badshah, and V. Bhagwat, “Heart disease prediction using data mining techniques,” ACM Int. Conf. Proceeding Ser., vol. 10, no. 02, pp. 7–11, 2019, doi: 10.1145/3352411.3352413.

[19] H. Yang, K. Negishi, P. Otahal, and T. H. Marwick, “Clinical prediction of incident heart failure risk: A systematic review and meta-analysis,” Open Hear., vol. 2, no. 1, pp. 1–8, 2015, doi: 10.1136/openhrt-2014-000222.

[20] A. Newaz, N. Ahmed, and F. Shahriyar Haq, “Survival prediction of heart failure patients using machine learning techniques,” Informatics Med. Unlocked, vol. 26, no. August, p. 100772, 2021, doi: 10.1016/j.imu.2021.100772.

[21] R. Williams, T. Shongwe, A. N. Hasan, and V. Rameshar, “Heart Disease Prediction using Machine Learning Techniques,” 2021 Int. Conf. Data Anal. Bus. Ind. ICDABI 2021, no. 5, pp. 118–123, 2021, doi: 10.1109/ICDABI53623.2021.9655783.

[22] S. U. Ghumbre and A. A. Ghatol, “Heart disease diagnosis using machine learning Algorithm,” Adv. Intell. Soft Comput., vol. 132 AISC, pp. 217–225, 2012, doi: 10.1007/978-3-642-27443-5_25.

[23] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, “Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning,” Comput. Intell. Neurosci., vol. 2021, 2021, doi: 10.1155/2021/8387680.

[24] A. Awaysheh, J. Wilcke, F. Elvinger, L. Rees, W. Fan, and K. L. Zimmerman, “Review of Medical Decision Support and Machine-Learning Methods,” Vet. Pathol., vol. 56, no. 4, pp. 512–525, 2019, doi: 10.1177/0300985819829524.

[25] Mangesh Limbitote, “A Survey on Prediction Techniques of Heart Disease using Machine Learning,” Int. J. Eng. Res., vol. V9, no. 06, pp. 450–453, 2020, doi: 10.17577/ijertv9is060298.

[26] S. S. Yadav, S. M. Jadhav, R. G. Bonde, and S. T. Chaudhari, “Automated Cardiac Disease Diagnosis Using Support Vector Machine,” 2020 3rd Int. Conf. Commun. Syst. Comput. IT Appl. CSCITA 2020 - Proc., pp. 56–61, 2020, doi: 10.1109/CSCITA47329.2020.9137817.

[27] K. Srinivas, B. K. Rani, M. V. P. Rao, R. K. Patra, G. Madhukar, and A. Mahendar, “Prediction of heart disease using hybrid linear regression,” Eur. J. Mol. Clin. Med., vol. 7, no. 5, pp. 1172–1181, 2020.

[28] C. B. C. Latha and S. C. Jeeva, “Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques,” Informatics Med. Unlocked, vol. 16, 2019, doi: 10.1016/j.imu.2019.100203.

[29] A. Pandey, “A Heart Disease Prediction Model using Decision Tree,” IOSR J. Comput. Eng., vol. 12, no. 6, pp. 83–86, 2013, doi: 10.9790/0661-1268386.

[30] A. S. T. Nishadi, “Predicting Heart Diseases In Logistic Regression Of Machine Learning Algorithms By Python Jupyterlab,” Int. J. Adv. Res. Publ., vol. 3, no. 8, pp. 69–74, 2019.

[31] Y. a Sandhy, “Prediction of Heart Diseases using Support Vector Machine,” Int. J. Res. Appl. Sci. Eng. Technol., vol. 8, no. 2, pp. 126–135, 2020, doi: 10.22214/ijraset.2020.2021.

[32] A. Yazdani, K. D. Varathan, Y. K. Chiam, A. W. Malik, and W. A. Wan Ahmad, "A novel approach for heart disease prediction using strength scores with significant predictors," BMC Med. Inform. Decis. Mak., vol. 21, no. 1, pp. 1–16, 2021, doi: 10.1186/s12911-021-01527-5.

[33] Y. Muhammad, M. Tahir, M. Hayat, and K. T. Chong, "Early and accurate detection and diagnosis of heart disease using intelligent computational model," Sci. Rep., vol. 10, no. 1, pp. 1–17, 2020, doi: 10.1038/s41598-020-76635-9

[34] B. M. McLaren, R. Reilly, S. Zvacek, and J. Uhomoibhi, "Foreword," CSEDU 2018 - Proc. 10th Int. Conf. Comput. Support. Educ., vol. 1, pp. XIII–XIV, 2018.