Machine Learning Algorithms for Heart Disease Prediction

Sikha Suhani Bhuyan
Odisha University of Technology and Research,
Bhubaneswar, Odisha, India
bhuyansikhasuhani@gmail.com

Ashis Kumar Mishra
Odisha University of Technology and Research,
Bhubaneswar, Odisha, India
akmishracse@cet.edu.in

***Abstract*: Cardiovascular disease, otherwise known as heart disease, encompasses many diseases that affect the heart. Heart disease prediction is among the most complicated tasks in medical field. In the modern age, about one person dies every minute as a result of heart disease. In addition to many factors that contribute to heart disease, it's necessary at this point in time to acquire accurate, reliable, and sensible approaches to make an early diagnosis so that the disease may be managed appropriately. Due to the complexity of finding out the heart condition, the prediction process must be automated to avoid risks related to it and to alert the patient at an early stage. In the healthcare domain, data mining is commonly used to analyze huge, complex medical data and predict heart disease. Researchers apply a variety of data mining and machine learning approaches to analyse huge complex medical data and predict heart disease. In this study, various heart disease attribute are presented, and model is developed on the basis of supervised learning algorithm as K-nearest neighbor, Decision Tree, Random Forest, Logistic Regression, SVM, Light GBM and Naïve bayes. This Paper makes use of heart condition dataset available in Kaggle repository. The purpose of this study is to anticipate heart disease risk in patients. The results show that K-nearest neighbor provides the most accurate result.**

Keywords* :K-nearest neighbor, Decision Tree ,Random Forest, Logistic Regression , SVM, Light GBM, Naïve Bayes

I. INTRODUCTION

Basically, the heart is the principle organ in our body that regulates blood flow throughout. The research proposed in this paper is primarily concerned with various data mining techniques that are applied to heart disease prediction. In our modern world, heart disease is one of the primary causes for death. Any type of irregularity in heart function can cause distress to other parts of the body. The World Health Organization reports that more than 10 million people die due to heart disease in this world every single year because of unhealthy lifestyle, smoking, alcohol and a high fat diet [2]. A healthy lifestyle and earliest detection are only ways to prevent the heart related diseases.

Today's healthcare challenges are ensuring best-in-class services are delivered and accurate diagnoses are achieved[1]. In spite of heart disease being the world's primary cause of death in recent years, it is also a disease that can be managed effectively and is controllable. In order to properly manage a disease, you need to diagnose it at the proper time. To avoid disastrous consequences from heart disease, the proposed study aims to discover these diseases at an early stage.

A large set of medical data created by medical experts is available for analysis and exploitation of valuable knowledge. Data mining is the process of identifying information from large amounts of data and extracting valuable and hidden patterns. There are usually discrete pieces of information in a medical database. Due to this, decision making using discrete data becomes complex and challenging. As a subfield within data mining, machine learning (ML) is an efficient means of handling large, well-formatted datasets. Various medical diseases can be diagnosed, detected, and predicted using machine learning methods in the medical field. A key objective of this paper is to provide doctors with an early detection tool for heart disease [6]. As a result, patients will receive effective treatment and severe consequences will be avoided. Machine learning plays an important role in detecting discrete patterns hidden in the data and analyzing it accordingly. By analyzing data, machine learning techniques can be used to predict and diagnose heart disease. This paper presents performance analysis of various ML techniques such as K-nearest neighbor, Decision Tree, Random Forest, Logistic Regression, SVM, Light GBM and Naïve Bayes for predicting heart disease in an early stage.

II. RELATED WORK

Many studies have been conducted using the Kaggle dataset to predict heart disease. Different stages of accuracy have been attained using diverse records mining strategies which are explained as follows.

T. Nagamani and colleagues presented a machine [2] that used data mining methodologies in conjunction with the MapReduce set of rules. For the 45 times of checking out set, the accuracy gained using this article was higher than the accuracy obtained using a standard fuzzy artificial neural network. Because of the usage of dynamic schema and linear scaling, the accuracy of the method employed improved.

Avinash Golande and colleagues investigate a number of different machine learning techniques that could be used to diagnose cardiac disease. The study of Decision Tree, KNN, and K-Means algorithms that can be used for classification was finished, and their accuracy was compared[1]. The accuracy acquired by utilising Decision Tree became the highest, and it was inferred that it might be made efficient with the aid of a blend of unique procedures and parameter adjustment.

Fahd Saleh Alotaibi has created a machine learning version that compares five different algorithms [3]. In comparison to Matlab and Weka, the Rapid Miner device was employed, which resulted in higher accuracy. The accuracy of the classification algorithms Decision Tree, Logistic Regression, Random Wooded Area, Naive Bayes,

and SVM were compared in this study. The most accurate algorithm was the decision tree algorithm.

Anjan Nikhil Repaka, et al., proposed a machine in [4] that uses NB (Nave Bayesian) strategies for dataset type and the AES (Advanced Encryption Standard) algorithm for data transfer ease for disorder prediction.

Theresa Princy, R., et al. conducted a survey that included a set of unique category rules for predicting heart disease. Naive Bayes, KNN (K-Nearest Neighbour), Decision tree, and Neural community were the category techniques utilised, and the accuracy of the classifiers was evaluated for a limited set of attributes [6]. The prediction of cardiac disease using the Naive Bayes type and SVM has been completed by Nagaraj M Lutimath et al (Support Vector Machine). Mean Absolute Error, Sum of Squared Error, and Root Mean Squared Error were the performance measures employed in the analysis. In terms of accuracy, SVM has been shown to outperform Naive Bayes [7].

After analysing the above publications, the main idea behind the suggested machine was to develop a coronary heart disease prediction system based on the inputs presented in Table 1. Based on their accuracy, we evaluated the classification algorithms Decision Tree, Random Forest, Logistic Regression, Light GBM, K-nearest neighbour, and SVM and identified the best class algorithm for heart disease prediction.

III. PROPOSED MODEL

The suggested research investigates the above-mentioned seven classification algorithms and performs an overall performance analysis. The purpose of these studies is to see if the patient has a coronary artery condition. The health professional enters the input values from the patient's health document. The information is fed into a model that predicts the risk of getting heart disease. Figure 1 depicts the full system.

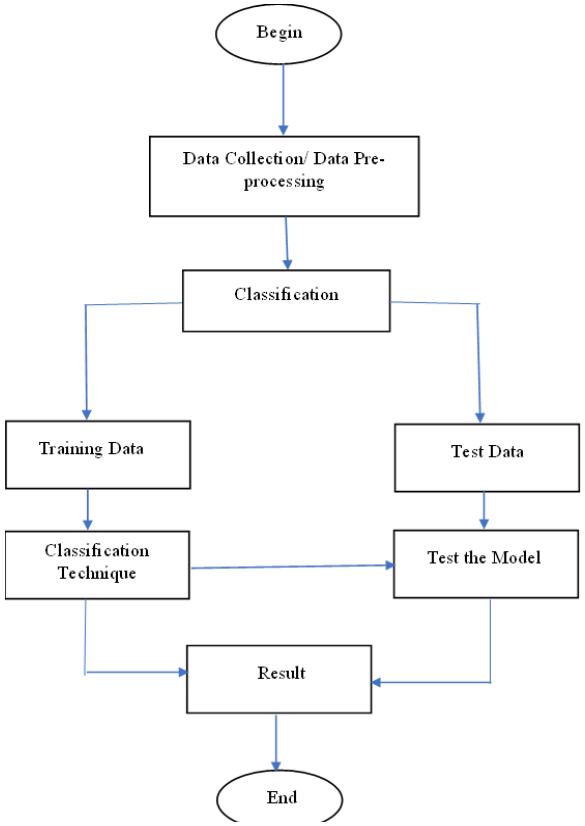

Fig. 1. Flow chart of Heart disease Prediction Model

A. Data Collection and Pre-Processing

The dataset utilized was the Heart disease Dataset which is a blend of 4 unique dataset, yet just the UCI Cleveland dataset was utilized. This data set comprises of an aggregate of 76 traits however totally distributed tests allude to utilizing a subset of just 14 highlights [8]. Subsequently, we have utilized the currently handled UCI Cleveland dataset accessible in the Kaggle site . The total depiction of the 14 ascribes utilized in the proposed work is referenced in Table 1 displayed underneath.

TABLE I. FEATURE SELECTED FROM DATASET

Sl.no	Attribute	Description
1	Age	Age in Years
2	Sex	1=male 0=female
3	CP	Chest Pain Type: 1=typical angina 2=atypical angina 3=non-angina pain 4=asymptomatic
4	Trestbps	Resting blood pressure (in mm Hg)
5	Chol	Serum cholesterol in mg/dl.
6	FBS	Fasting Blood Sugar > 120 mg/dl: 1=true 0=false
7	Rest ECG	Resting electrocardiographic results: 0 = normal 1 = having ST-T wave abnormality 2=showing probable or define left ventricular hypertrophy by Estes ‘criteria
8	Thalach	Maximum heart rate achieved
9	Exang	Exercise induced angina: 1 = yes 0 = no
10	OldPeak	Depression induced by exercise relative to rest
11	Slope	The slope of the peak exercise segment: 1 = up sloping 2 = flat 3= down sloping
12	CA	Number of major vessels colored by fluoroscopy that ranged between 0 and 3
13	Thal	3 = normal 6= fixed defect 7= reversible defect
14	Target.	Diagnosis classes: 0 = healthy 1= patient who is subject to possible heart disease

B. Classification

The attributes listed in Table 1 are fed into several machine learning algorithms such as Random Forest, Decision Tree, Logistic Regression, Light GBM, KNN, and SVM, as well as Nave Bayes classification approaches [1]. The input dataset is divided into two parts: 70% of the training dataset and 30% of the test dataset. A training dataset is a collection of data that is used to train a model.

Using the testing dataset, an evaluation of the trained model is performed. The performance of each method is computed and analysed using several metrics such as accuracy, precision, recall, and F-1 scores, as discussed below. The many algorithms investigated in this research are given below.

1) Decision Tree

The Decision Tree algorithm is shaped like a flowchart, with the internal node representing dataset properties and the outer branches representing the result. Decision Trees were chosen because they are quick, dependable, and simple to interpret, and they require little or no fact practise. In Decision Tree, the prediction of class label originates from root of the tree. The value of the root attribute is in comparison to report's attribute. On the result of comparison, the corresponding branch is followed to that value and jump is made to the next node.

2) K-Nearest Neighbor

KNN is one of the most straightforward and simple records mining techniques. Because the training examples want to be in memory at run-time, it's called Memory-Based Classification [9]. When working with continuous qualities, the difference between them is calculated using the Euclidean distance, which is Dist(p,q) =sqrt((p1-q1)2 + p2-q2)2 +....+ pn-qn)2. A predominant trouble while dealing with the Euclidean distance method is that the huge values frequency swamps the smaller ones. The KNN is generally used with continuous attributes, but it can also be used with discrete attributes. In dealing with discrete attributes, a difference between two instances is equal to one if the attribute values are different; otherwise, the difference is equal to zero.

3) LGBM Classifier

Gradient-Boosting Machines are used for a variety of tasks including ranking, classification, and machine learning. One example is Light GBM, which may be used for fast, distributed, and high-performance gradient boosting algorithms. This algorithm splits the tree leaf-wise with the simplest fit, while other algorithms split the tree level-wise or depth-wise. As a result, when growing on a leaf equivalent in Light GBM, the leaf-wise algorithm reduces more loss than the level-wise algorithm and leads to far better accuracy than all prevailing boosting algorithms.

4) Support Vector Machines

supervised machine learning algorithms called support vector machines (SVMs), which are widely used for classification and regression. Generally this algorithm used for classification problem. In multidimensional space, SVM models represent different classes via a hyperplane. SVM will generate the hyperplane iteratively to minimize the error.

5) Logistic Regression

It is a classification problem that used for binary classification problems. To fit a linear equation between 0 and 1, the logistic regression algorithm employs the logistic function instead of a straight line or hyperplane. As a result of the 13 independent variables, logistic regression works well for classification.

6) Random Forest

This is used for regression and classification. Predictions are made based on the tree generated by the algorithm. The Random Forest technique produces the same result even when there are missing entries in huge sets.The generated samples from the selection tree can be saved and used on additional data. There are two ranges in random forest: first, generate a random forest, and then make a prediction using the random forest classifier created in the first level.

7) Naïve Bayes

The Bayes rule is used. The fundamental and most important assumption in constructing a classification is that the dataset's properties are independent. It is simple and quick to anticipate, and it works best when the concept of independence is present. As seen in equation 1, Bayes' theorem determines the posterior probability of an event (A) given a few prior probabilities of event B indicated by P(A/B)[10].

$$P(A|B) = (P(B|A)P(A)) / P(B) \quad [1]$$

IV. OBSERVATION

From those consequences we can see that even though maximum of the researchers are the use of exclusive algorithms including SVC, Decision tree, KNN, Random Forest, Naïve Bayes, LGBM, Logistic Regression for the detection of patients recognized with Heart disease. KNN yield a better result to out rule them .

Accuracy score, Precision (P), Recall (R), and F-1 score measure are the metrics used to evaluate the algorithm's performance. Precision (referred to in equation (2) metric) is a correct measure of high quality evaluation. The measure of real positives that can be right is defined by recall [stated in equation (3)]. The F-1 score [given in equation (4)] is a test of accuracy.

$$\text{Precision} = (TP) / (TP + FP) \quad (2)$$

$$\text{Recall} = (TP) / (TP + FN) \quad (3)$$

$$\text{F–1 score Measure} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (4)$$

TABLE II. VALUES OBTAINED FOR CONFUSION MATRIX USING DIFFERENT ALGORITHM

Algorithm	TP	FP	FN	TN
DT	31	19	4	38
KNN	41	9	4	37
LGBM	38	12	5	36
SVC	38	12	5	36
LR	33	17	4	37
RF	37	13	5	37
NB	37	13	6	35

TABLE III. ANALYSIS OF MACHINE LEARNING ALGORITHM

Algorithm	Precision	Recall	F1- Score
DT	0.80	0.76	0.75
KNN	0.86	0.86	0.86
LGBM	0.82	0.81	0.81
SVC	0.83	0.81	0.81
LR	0.78	0.77	0.77
RF	0.82	0.80	0.80
NB	0.80	0.79	0.79

TABLE IV. ACCURACY SCORE OF DIFFERENT ML ALGORITHMS

Algorithm	Accuracy
DT	75.82%
KNN	85.71%
LGBM	81.31%
SCV	81.31%
LR	76.92%
RF	80.21%
NB	79.12%

Algorithms:

DT – Decision Tree,

KNN- KNearest Neighbor,

LGBM – Light GBM,

SVC – Support vector machine,

LR – Logistic Regression,

RF – Random Forest,

NB- Naive Bayes.

Confusion Matrix Attributes

TP- True Positive,

FP- False Positive,

FN- False Negative,

TN – True Negative.

V. RESULT AND ANALYSIS

The tests are carried out using a pre-processed dataset, and the above-mentioned techniques are examined and implemented. The confusion matrix is used to derive the above-mentioned overall performance indicators.

The model's performance is described by the Confusion Matrix. Table 2 shows the confusion matrix obtained using the proposed model for specific algorithms.

The Precision, recall and F1-Score value and the accuracy score obtained for DT, KNN, LR, LGBM, RF, NB and SVC is shown in Table 3 and Table 4 respectively.

Accuracy Score of KNN is more in comparison with other six machine learning algorithms that are used in this project. KNN gives 85.71% accuracy score which is more accurate and cost friendly than other machine learning algorithms used in this paper.

VI. CONCLUSION

With the rising number of deaths caused by heart disease, it has become necessary to create a technology that can accurately predict heart illness. The goal of the study was to discover the most effective ML set of criteria for detecting coronary heart disease. Using the UCI machine learning repository dataset, this research evaluates the accuracy score of Decision Tree, Logistic Regression, Random Forest, support vector machine, KNearest Neighbor, LGBM, and Naive Bayes algorithms for predicting coronary heart disease. The KNearest Neighbor algorithm is the most accurate green algorithm for predicting coronary heart disease, with an accuracy rating of 85.71 percent. In the future, the artwork may be more appropriate by means of developing an internet programme based KNN as well as the employment of a larger dataset than the one used in this analysis in order to help fitness experts anticipate heart ailment efficaciously and efficaciously.

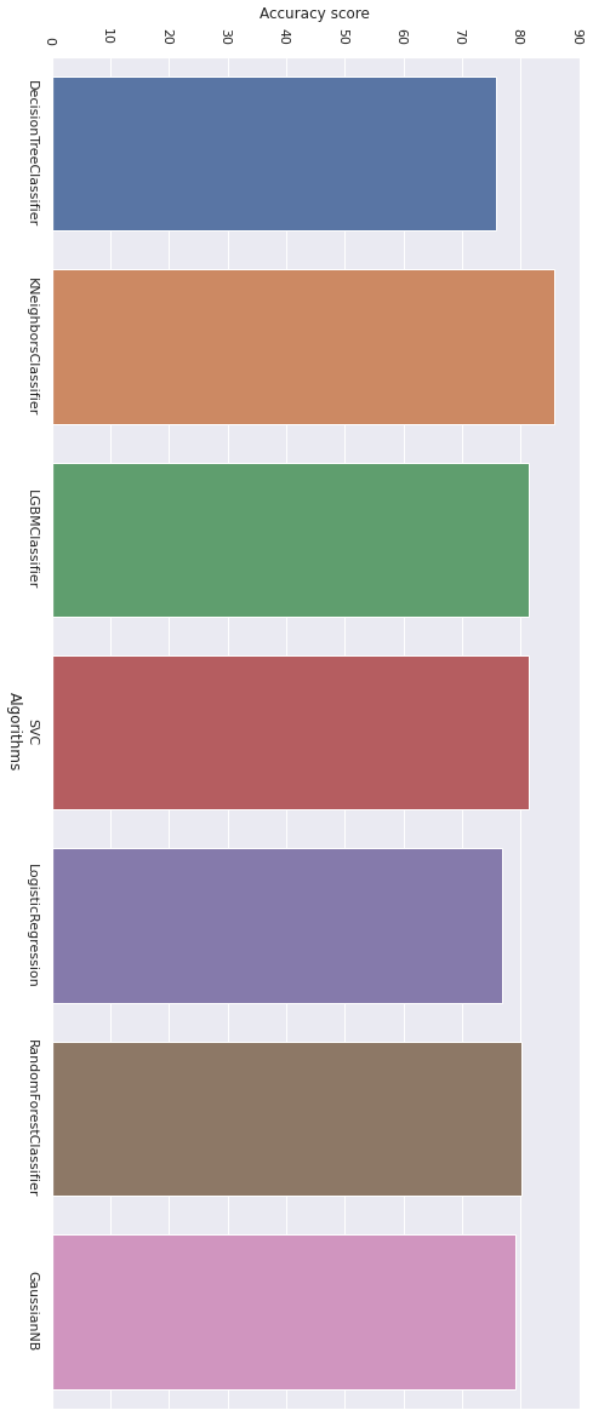

Fig. 2. Graphical Representations of accuracy score of different ML Algorithms

REFERENCES

[1] Avinash Golande, Pavan Kumar T, "Heart Disease Prediction Using Effective Machine Learning Techniques", International Journal of Recent Technology and Engineering, Vol 8, pp.944-950,2019.

[2] T.Nagamani, S.Logeswari, B.Gomathy," Heart Disease Prediction using Data Mining with Mapreduce Algorithm", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-3, January 2019.

[3] Fahd Saleh Alotaibi," Implementation of Machine Learning Model to predict Heart Failure Disease", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, 2019.

[4] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin, ”Design And Implementation Heart Disease Prediction Using Naives Bayesian”, International Conference on Trends in Electronics and Information(ICOEI 2019).

[5] UCI, ―Heart Disease Data Set.[Online]. Available (Accessed on May 1 2020): https://www.kaggle.com/ronitf/heart-disease-uci.

[6] Theresa Princy R,J. Thomas,’Human heart Disease Prediction System using Data Mining Techniques’, International Conference on Circuit Power and Computing Technologies,Bangalore,2016.

[7] Nagaraj M Lutimath,Chethan C,Basavaraj S Pol.,’Prediction Of Heart Disease using Machine Learning’, International journal Of Recent Technology and Engineering,8,(2S10), pp 474-477, 2019.

[8] C. B. Rjeily, G. Badr, E. Hassani, A. H., and E. Andres, ―Medical Data Mining for Heart Diseases and the Future of Sequential Mining in Medical Field,‖ in Machine Learning Paradigms, 2019, pp. 71–99.

[9] Fajr Ibrahem Alarsan., and Mamoon Younes ‘Analysis and classification of heart diseases using heartbeat features and machine learning algorithms’,Journal Of Big Data,2019;6:81.

[10] Internet source [Online].Available (Accessed on May 1 2021): http://acadpubl.eu/ap.

[11] Jafar Alzubi, Anand Nayyar, Akshi Kumar. "Machine Learning from Theory to Algorithms: An Overview", Journal of Physics: Conference Series, 2018.

[12] Sayali Ambekar, Rashmi Phalnikar,“Disease Risk Prediction by Using Convolutional Neural Network”,2018 Fourth International Conference on Computing Communication Control and Automation.