Comparative Study on Heart Disease Prediction Using Machine Learning

Rikendra
Computer science and engineering
GL bajaj institute of technology and management
Uttar pradesh , India
rikendrapoonia@gmail.com

Dr.Savita kumari
Computer science and engineering
Galgotias University
Uttar pradesh , India
savitasangwan92@gmail.com

Krishan kumar
Computer science and engineering
Chandigarh university
Mohali , Panjab , India
kpuniya1976@gmail.com

2023 International Conference on Disruptive Technologies (ICDT) | 979-8-3503-2388-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICDT57929.2023.10150734

***Abstract*—Given the dramatic increase in the rate of heart attacks in young people, we need to develop a system that can detect and prevent heart attack symptoms early. A method that is practical and trustworthy in determining the likelihood of heart disease must be in place because it is unrealistic for the average person to routinely undertake expensive tests like the ECG. As a result, we suggest creating a program that, provided basic symptoms like Age, Gender, and heart rate, can forecast the vulnerability of a cardiac condition. Since neural networks are the most accurate and trustworthy machine learning algorithm is used.**

Keywords—Heart disease signs, Machine Learning, Disease Prediction, Heart Disease Dataset, Health care services

I. INTRODUCTION

The absence shortage of documentation about the symptoms experienced by patients having heart attacks. However, their full potential to aid in the prediction of associated possibilities in otherwise healthy persons is not being fully utilized. For illustration: According to the Indian Heart Association, 25% of all heart strokes in Indians happen before the age of 40, and 50% of heart attacks happen before the age of 50.

The risk of heart attacks is three times more likely in urban areas than in rural ones. In light of this, we advocate gathering relevant data regarding every facet of our research area and training the data using the indicated machine learning method, and then forecast the likelihood that a patient would develop a heart condition, for the benefit of the sufferers.

In section II, we go through all of the common sensors on the market and the symptoms that each one measures. Making the system user-friendly is our primary goal while building it, as it enables regular patient monitoring. Therefore, the input's required variables must be both highly precise and accessible.

The workflow and structure of our system are explained in detail in section III. We provide a more detailed description of the algorithm we utilized. The outcomes of our experiment utilizing a sample dataset and the suggested algorithm are then shown. Additionally, many statistical methodologies are explained. The paper is briefly described in the fina l paragraph, followed by the conclusion.

II. LITERATURE REVIEW

According to Tom Mitchell, a computer programme is said to learn from experience and from some tasks, and performance on particular activities is believed to get better with practise. The majority of machine learning algorithms in useare focused on identifying or utilizing relationship between datasets. Machine Learning is a combination of correlation sand relationships. If machine learningalgorithms can identify specific correlations, the model can either generalize the data to find intriguing patterns or use these links to forecast future observations. Regression, linear regression, logistic regression, Naive Bayes classifier, Bayes theorem, decision tree, entropy, ID3, SVM (Support Vector Machines), K-means algorithm, random forest, and other algorithms are only a few examples of the many different types of algorithms used in machine learning.With the use of numerous approaches, algorithms, and other tools, we have developed a system that can forecast a patient's disease based on their symptoms. This system is called the recommended system of heart disease prediction usingmachine learning. By using those symptoms, we may compare them to the dataset currently present in the system, and after that, we are showing a comparative study of the effectiveness of various algorithms.

The Heart Disease Prediction using Machine Learning aims to avoid diseases at an early stage because, as is well known, people are no longer concerned with their health due to the competitive environment of development.. According to study, 60% of people choose are unconcerned with their health. Python is used exclusively in the project "Comparative study on Heart Disease Prediction using Machine Learning". Here, the user must first choose the symptoms from a checkbox menu before entering all the available symptoms, at which point the algorithm will return anaccurate result The main tools used to make this prediction are machine learning algorithms like decision tree, random forest, so on. The user must press the submit button after entering all the symptoms in order to receive the expected outcomes.

III. METHODOLOGY

To begin the job, we have begun gathering data in all relevant areas to achieve the system's objectives. First, the research focused on the primary causes or variables which have a significant impact on heart health. Age, sex, and family history are unchangeable factors, but blood pressure, heart rate, and other parameters can be controlled according to certain guidelines. To maintain heart health, many experts advise a healthy diet and moderate exercise. The following are the elements that were taken into account when designing the system and have a high risk of developing CAD.

1. Chest pain type
2. Sex
3. Resting blood pressure
4. Age
5. FBS over 120
6. ECG results

7. Heart rate
8. Serum cholesterol in mg/dl
9. BP
10. Exercise angina.
11. ST depression.
12. Slope of ST.
13. Number of vessels fluro
14. Thallium.
15. Target
16. Body Mass Index(obesity)

The Dataset gathering came next. Because of this, we from the UCI library, using the Cleveland dataset. The dataset contains up to 76 parameters that can be utilized to define the overall condition of heart health.. These metrics are gathered through pricey clinical examinations like an ECG or CT scan, among others. The conventional approach for predicting heart disease employs 13 of these key variables.

Determining these parameters—such as the type of chest discomfort, ECG, ST depression, etc.—requires costly lab procedures. We chose the aforementioned parameters as measured help of market technology and simplifying the systems The most recent sensors on the market today are briefly described in the study paper that follows. These sensors are used to measure various parameters

A. AliveKor

It is available as a bracelet or a touch pad that connects to your cellphone over a wireless network. Through Bluetooth, the touch padsimulates the patient's ECG on his mobile device. As a result, al the relevant parameters, including blood pressure and heart rate, are readily available. On the wristband, however, the pulse function is shown on the dial through finger contact. Additionally,it may signal atrial fibrillation.

B. MyHeart

This system uses a variety of on body sensors to wirelessly transmit physio-logical data to a PDA. The data is processed, andthe analysis is used to provide the user with health suggestions. [1]

C. healthcare

An application called Health Gear tracks the most popular metrics, including physical and lab measurements. Blood pressure, hemoglobin, WBC, RBC, and platelets are among the [4] fields. [Physical] indices like height, weight, and BMI are also included.

- [Lipids]: Triglycerides, HDL, LDL, and VLDL
- [Sugar] Fasting Glucose, HbA1C, and After Meals.

D. Fitbit

Useful to monitor one's health and has functions for detecting heart rate, blood pressure, and calories burned.Post conducting this analysis, conclusion can be made that Fitbit would be the most convenient and cost-effective way to gather data, while Health-gear would be used for all other metrics.

IV. Proposed Mechanism

According to recent assessments, the machine learning field's most well-liked and rapidly developing sub-field is neural networks. The data-set was trained and tested in the proposed system using the multi-layer perceptron (MLP) neural network algorithm.

The supervised neural network approach known as the "multi-layer perceptron algorithm" has layers for the input, output, and one or more hidden layers between these two layers are layered. The hidden layers connect each input layer node to the output layer nodes. Giving the connection between two node weights, the input is calculated using the following formula.

$$Y_{in} = \sum(i = 0\ to\ i = n) w_i x_i \quad (1)$$

where w_i is the matching weight and xi is the input.

To balance the perceptron, a second identity input with weight b is introduced to the node and is referred to as Bias. Depending on the situation, the nodes connection can either be feed-forward or feedback. The flow-diagram for the input/output signal as shown in Figure 3.

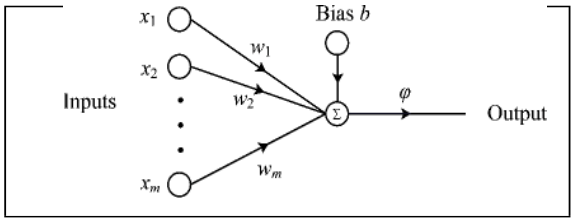

Fig. 3. Signalflow-graphof perceptron

The weighted input is used for the activation function to produce the output. Connecting the input and output layers and doing internal data processing are the duties of the hidden layer. The MLP logical framework.

The PyCharm IDE is used to create the system's Python code. This system is implemented utilizing the sci-kit learn Python module. These are the sample parameter settings used for the MLP function.

V. Machine Learning Algorithms

A. Naïve Bayes (NB):

Naïve Bayes Classification models relies on Bayes theorem that works on the principle of three kinds of probabilities called prior probability, likelihood probability, and posterior probability.

It is a supervised binary class or multi class classification algorithm. There is a family of Naïve Bayes classifiers based on a common principle. These algorithms classify for datasets whose features are independent and each feature is assumed to be given equal weight-age. It works for a large data-set and is very fast. It s one of the most effective and simple classification algorithms. Some important applications of these algorithms are text classification, recommendation system and face recognition.

$$P(E/F) = \frac{P(E/F)P(E)}{P(F)}$$

P(E/F) is the Posterior probability, and P(F/E) is the Likelihood probability.

B. Logistic Regression (LR):

Linear Regression predicts the numerical response but is not suitable for predicting the categorical variables. When

categorical variables are involved, it is known as classification problem. Logistic regression is suitable for binary classification problem. Here, the output is often a categorical variable.

For example, following below scenarios are instances of predicting categorical variables.

The student being pass or fail is based on marks secured.Is the mail spam or not spam? The answer is Yes or No. Thus, categorical dependent variable is a binary response of Yes or No. Thus, logistic regression is used as a binary classifier and works by predicting the probability of the categorical variable. In General, it takes one or more features x and predicts the response y.

The definition of the logit function for p is as follows if p is the probability:

$$Logit(q) = \ln\left(\frac{q}{1-q}\right)$$

C. *Decision Tree (DT):*

Decision tree learning model, one of the most popular supervised predictive learning models, classifies data instances with high accuracy and consistency.

A decision tree is a concept tree that presents a tree-like summary of the data from the training datasets.

This model can be used to classify both categorical target variables and continuous-valued target variables. Inputs to the model are data instances or objects with a set of features or attributes which can be discrete or continuous and the output of the model is a decision tree which predicts or classifies the target class for the test data object.

A decision tree has a structure that consists of a root node, internal nodes/decision nodes, branches, and terminal nodes/leaf nodes. Topmost node called as root node. Internal nodes are the test nodes and are also called as decision nodes.

Advantages of Decision Tree are they are simple to understand and easy to model and interpret, and quick to train easily.

Disadvantage is it is quite hard to know how deeply a decision tree can be grown or when to stop growing it, if training data has errors or missing attribute values, then the decision tree constructed may become unstable or biased, and a complex decision tree may also be over-fitting with the training data

D. *Random Forest (RF):*

A classifier called Random Forest "contains the number of decision trees on various subsets of the provided dataset and, takes average to improve the predictive accuracy of that specific dataset," as the name implies". Random Forest algorithm doesn't focus on a single decision tree, but uses forecasts from each tree and predicts the results using the votes of majority of predictions.

Higher accuracy and over fitting are prevented by large number of trees in the forest.

The Random Forest method should be used for the reasons listed below:

In comparison to other algorithms, it requires less training time.

Even with the enormous dataset, it operates effectively and predicts the outcome with a high degree of accuracy.

VI. RESULTS

The system's output will indicate, in the form of a Yes or No response, whether the subject has a heart condition. The approach provides a prediction of the heart condition that will lead to CAD. The outcome will be Yes if the subject is predisposed to having heart disease, and vice versa. If the results are favorable, he should see a cardiologist for a more thorough analysis.

The following table displays the statistics of the findings made while testing the dataset.

TABLE I. ACCORDING TO OUR PROPOSED MODEL, THE ACCURACY OF THE CLASSIFICATION OBTAINED BY DIFFERENT ALGORITHMS

Classification Algorithm	Test Size	Accuracy (%)
Naive Bayes (NB)	0.35	74.07%
Logistic Regression(LR)	0.36	83.33%
Decision Tree (DT)	0.35	85.19%
Random Forest (RF)	0.36	85.23%

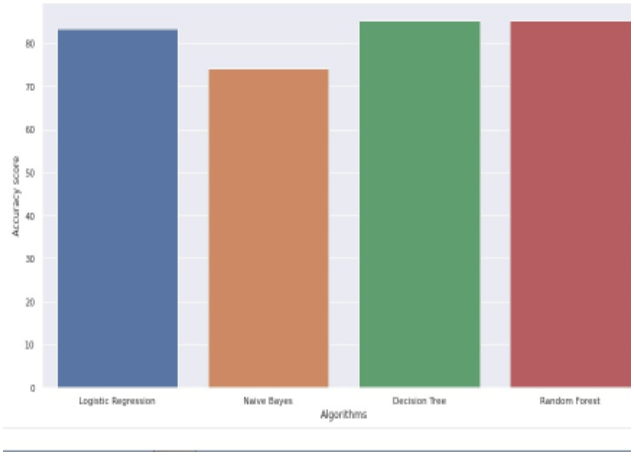

Fig. 5. Comparison of different algorithms for Heart Disease Data-set

VII. CONCLUSION

The machine learning algorithms have greatly improved as a result, we adopt Multi-Layered Perceptron (MLP) in the suggested system due to its effectiveness and precision. Additionally, based on the data inputted by the consumers, the algorithm outputs a trustworthy result that is close by. The accuracy achieved for Naïve Bayes (NB) is 74.07 %, for Logistic Regression (LR) is 83.33 %, for Decision Tree (DT) is 85.19%, and for Random Forest (RF) is 85.23 %.

Here, the Random Forest has good results as compared to other algorithms. We desire to apply ML to exhibit a connection between cardiovascular illness caused and air quality. The more people who use the system, the more people will be aware of their current heart health, which will eventually lead to a decline in the number of deaths from heart disease.

VIII. FUTURE WORK

Leveraging innovations in image processing, fuzzy logic, machine learning, and other fields, comparable prediction systems can be created for a variety of other deadly or chronic conditions like Cancer, Diabetes, Brain

Tumour prediction, etc. New algorithms may also be suggested to increase accuracy and dependability. Massive amounts of data from all users globally can be stored using big data technologies like Hadoop, and Spark, and Cloud computing technology can be utilized to manage the user's data or reports..

REFERENCES

[1] L. Ali et al., "An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure," IEEE Access, vol. 7, pp. 54007–54014, 2019, doi: 10.1109/ACCESS.2019.2909969.

[2] A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor, and R. Nour, "An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection," IEEE Access, vol. 7, pp. 180235– 180243, 2019, doi: 10.1109/ACCESS.2019.2952107.

[3] S. Sperandei, "Lessons in biostatistics Understanding logistic regression analysis," Biochem. Medica, vol. 24, no. 1, pp. 12–18,(2014) .

[4] J. R. Quinlan, "Induction of Decision Trees," Mach. Learn., vol. 1, no. 1, pp. 81–106,(1986) .

[5] T. M. Mitchell, "Decision Tree Learning," Machine Learning. pp. 52–80, (1997).

[6] L. Breiman, "Random Forest," pp. 1–33, (2001).

[7] M. Denil, D. Matheson, and N. De Freitas, "Narrowing the Gap: Random Forests In TheDenil, M., Matheson, D., & De Freitas, N. (2014). Narrowing the Gap: Random Forests In Theory and In Practice. Proceedings of The 31st International Conference on Machine Learning, (1998), 665–673. Retrieved from ht," Proc. 31st Int. Conf. Mach. Learn., no. 1998, pp. 665–673, 2014.

[8] V. Jakkula, "Tutorial on Support Vector Machine (SVM),"Sch. EECS, Washingt. State Univ., pp. 1–13, (2006).

[9] Ahmed Fawzi Otoom , Emad E. Abdallah , Yousef Kilani , Ahmed Kefaye and Mohammad Ashour Effective Diagnosis and Monitoring of Heart Disease ISSN: 1738- 9984 IJSEIA(2015)

[10] G. H. Tang, A. B. M. Rabie, and U. Hägg, "Indian hedgehog: A mechanotransduction mediator in condylar cartilage," J. Dent. Res., vol. 83, no. 5, pp. 434–438, 2004, doi: 10.1177/154405910408300516. [11] Y. Karaca and C. Cattani, "7. Naive Bayesian classifier," Comput. Methods Data Anal., pp. 229–250, 2018, doi: 10.1515/9783110496369-007.

[11] N. Suma, S.R. Samson, S. Saranya, G. Shanmugapriya, R. Subhashri, IOT based smart agriculture monitoring system, Int. J. Recent Innov. Trends Comput. Commun. 5 (2) (2017) 177–181.

[12] Piller L B, Davis B R, Cutler J A, Cushman W C, Wright J T, Williamson J D & Haywood L J . Validation of heart failure events in the Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) participants assigned to doxazosin and chlorthalidone. Current controlled trials in cardiovascular medicine(2002)

[13] Raihan M, Mondal S, More A, Sagor M O F, Sikder G, Majumder M A & Ghosh K . Smartphone based ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design. In 2016 19th International Conference on Computer and Information Technology (ICCIT) (pp. 299-303)(2016, December).

[14] Patel S & Chauhan Y . Heart attack detection and medical attention using motion sensing device- kinect. International Journal of Scientific and Research Publications, 4(1), 1-4(2014).

[15] Wolgast G, Ehrenborg C, Israelsson A, Helander J, Johansson E & Manefjord H . Wireless body area network for heart attack detection [Education Corner]. IEEE antennas and propagation magazine, 58(5), 84-92(2016).

[16] Ankita Dewan and Meghna Sharma, "Prediction of heart disease using a hybrid technique in data mining classification", 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)

[17] Takci H . Improvement of heart attack prediction by the feature selection methods.Turkish Journal of Electrical Engineering & Computer Sciences, 26(1), 1-10(2018).

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