IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 3, March 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 3, March 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1103021 |

Survey on Prediction of Heart Attack using Machine Learning Algorithms

Prof.Jyotsna Nanajkar[1],Priti Kadav[2], Umakant Dudhwad[3], Durgesh Mamdapure[4], Harshal Pohakar[5]

Professor, Department of Information Technology, Zeal College of Engineering and Research, Pune, Maharashtra, India [1]

Students, Department of Information Technology, Zeal College of Engineering and Research, Pune, Maharashtra, India [2,3,4,5]

ABSTRACT: Heart attacks are a major cause of death and morbidity globally, and early prediction can play a critical role in reducingthe risk of complications and death. This project report presents a machine learning-based approach for predicting heartattacks using a Random Forest Classifier. The model was trained on a data-set of patient medical records and demographic information to predict heart attacks. The results of the study demonstrate the high level of accuracy achieved by the model in prediction,with metricssuch as accuracy, precision, recall, and F1score being used to evaluate the model. Theproposed system has the potential to significantly improve heart attack predictionand inform clinicaldecision- makingbyproviding healthcare providers with an accurate and efficient tool for predicting heart attacks. The model was then deployed onthe Amazon WebServices (AWS) platform tomake it accessible to healthcare providers and patients.The deployment on AWS allowed for the easy and efficient use of the model by healthcare providers and patients, and facilitated the integration of the model into clinicalpractice.

KEYWORDS: Machine Learning, Random Forest Classifier, Cardiovascular, AWS

I. INTRODUCTION

Heart attacks, also known as myocardial infarctions, area major cause of death and morbidity globally. They occur when the blood supply to the heart muscle is blocked, leading to damage or death of the heart muscle. Early prediction of heart attacks is crucial for reducing the risk of complications and death, as prompt treatment can prevent or minimize damage to the heartmuscle.

Conventional methods for predicting heart attacks, such as statistical analysis and risk factor analysis, have limitations as they rely on manually selected variables and may miss important predictors. This has led to a need for a more accurate and efficient method of predicting heart attacks. Machine learning provides a powerful tool for predicting heart attacks by leveraging large amounts of patient data and identifying complex patterns and relationships. In this project, we used a dataset of patient medical records and demographic information to train and test a machine learning model for heart attack prediction.

The model was implemented using a Random Forest Classifier, a popular machine learning algorithm known for its ability to handle complex and non-linear relationships in data. The results of the study demonstrate the potential of machine learning and predictive models to improve heart attack prediction and inform clinical decision-making.

II. LITERATURE REVIEW

Apurv Garg et al. utilized KNN and Random Forest and found that Chest Pain and Maximum heart rate achieved positively correlated with the target attribute, achieving an accuracy of 86.885% with KNN and 81.967% with Random Forest. Rishabh Magar et al[1]. developed a web applicationthatusedLogisticRegressiontoobtainanaccuracyof82.89%, followed by SVM at 81.57% and Naive Bayes and Decision Tree at 80.43% each. ApurbRajdhanet al[2]. proposed a system that used Random Forest, Decision Tree, Logistic Regression, and Naive Bayes, resulting in a maximum accuracy of 90.16% using Random Forest. Devansh Shah et al.[3] built a system using four individual classification techniquesincludingNB, KNN, RF, DT and obtained the highest accuracy through KNN. Harshit Jindal[4] et al. implemented a system with KNN, RF, and LR, achieving an accuracy of 87.5% with KNN providing the highest accuracy of 88.52%. AadarPandita[5]et al. proposed a web application that used five machine learning algorithms and achieved the highest accuracy using KNN at 89.06%. N. Saranya[6] et al. proposed a time and money efficient model using Random Forest and KNN, with an accuracy of 100% using Random Forest and 91.36% using KNN. An ensemble model with and without Logistic Regression was also used, resulting in an accuracy of 98.77% and 95.06% respectively. Aravind Akella[7] et al. applied six predictive models on the UCI dataset and achieved a maximum accuracy of 93.03% with Neural Networks. RavindharNV [8] et al. implemented five algorithms including Logistic Regression, Naive Bayes, Fuzzy

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 3, March 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1103021 |

KNN, K-Means Clustering, and back propagation Neural-Network,obtainingthehighestaccuracyof98.2%usingbackpropagation NeuralNetwork.

Year	Author	Paper Name	Algorithms Used	Accuracy Obtained
2020	Rishabh Magar etal. [1]	Heart disease prediction using machine learning	Logistic Regression SVM Naive Bayes Decision Tree	82.89% 81.57% 80.43% 80.43%
2020	Apurb Rajdhanet al. [2]	Heart disease prediction using machine learning	Logistic Regression Decision Tree Random Forest Naive Bayes	85.25% 81.97% 90.16% 85.25%
2020	Devansh Shah et al. [3]	Heart disease prediction using machine learning techniques	Naive Bayes KNN Random Forest Decision Tree	88.157% 90.789% 86.84% 80.263%
2021	Harshit Jindal et al.[4]	Heart Disease prediction using machine learning algorithms	KNN Logistic Regression KNN & LR based model	88.52% 88.5% 87.5%

III. METHODOLOGY

In our study, we evaluated the accuracy and performance of two machine learning algorithms, K nearest neighbors (KNN) and Logistic Regression, for predicting heart diseases. The process of this study was divided into three phases: data collection, value extraction, and data exploration. In the data preprocessing phase, we dealt with missing values, cleaned, and normalized the data. After preprocessing the data, we used a classifier to identify the heart disease based on the preprocessed data. Finally, we tested the proposed model using 20% of the full dataset, evaluating its accuracy and performance using various performance metrics.

IV. PROPOSED MODEL

The aim of this study is to accurately predict heart disease by using four different classification algorithms. Health professionals will input data from the patient's health report into the model, which will then generate the likelihood of the patient having heart disease.

A. *Data Collection and Preprocessing*

The Heart Disease Dataset, which is a combination of four databases, was used as the data source. However, only the UCI Cleveland dataset was utilized. This database has 76 attributes, but for our analysis, we used a subset of 14 features, as mentioned in [9]. The processed UCI Cleveland dataset found on Kaggle was used for our analysis. Table 1 below provides a comprehensive description of the 14 attributes used in this study.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 3, March 2023 ||

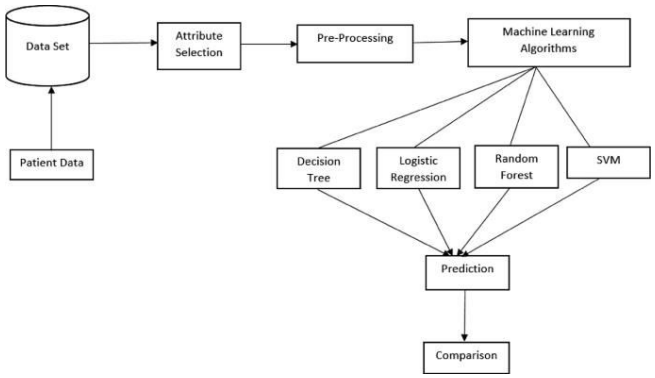
| DOI: 10.15680/IJIRCCE.2023.1103021 |

Sl. No.	Attribute Description	DistinctValues of Attribute
1.	*Age*- represent the age of a person	Multiple values between 29 & 71
2.	*Sex*- describe the gender of person (0- Female, 1-Male)	0,1
3.	*CP*- represents the severity of chest pain patient is suffering.	0,1,2,3
4.	*RestBP*-It represents the patient's BP.	Multiple values between 94& 200
5.	*Chol*-It shows the cholesterol level of the patient.	Multiple values between 126 & 564
6.	*FBS*-It represent the fasting blood sugar in the patient.	0,1
7.	*Resting ECG*-It shows the result of ECG	0,1,2
8.	*Heartbeat*- shows the max heartbeat of patient	Multiple values from71 to 202
9.	*Exang*- used to identify if there is an exercise induced angina. If yes=1 or else no=0	0,1
10.	*OldPeak*-describespatient's depression level.	Multiple values between 0 to 6.2.
11.	*Slope*- describes patient condition during peak exercise. It is divided into three segments (Unsloping, Flat, Down sloping)	1,2,3.
12.	*CA*- Result of fluoroscopy.	0,1,2,3
13.	*Thal*- test required for patient suffering from pain in chest or difficultyin breathing. There are 4 kinds of values which represent Thallium test.	0,1,2,3
14.	*Target*-It is the final column of the dataset. It is class or label Colum. It represents the number of classes in dataset. This dataset has binary classification i.e., two classes (0,1). In class "0" represent there is less possibility of heart disease whereas "1" represent high chances of heart disease. Thevalue "0" Or "1" depends on other 13 attribute.	0,1

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 3, March 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1103021 |



B. Classification

The attributes listed in Table 1 are inputted into various machine learning algorithms including Random Forest, Decision Tree, Logistic Regression, and Naive Bayes classification techniques [12]. The input dataset is divided into 80% for training and 20% for testing. The training dataset is used to train the model, while the testing dataset is used to evaluate the performance of the trained model. For each algorithm, the performance is evaluated using metrics such as accuracy, precision, recall, and F-measure scores. The algorithms explored in this study are as follows:

i. RandomForest

The Random Forest algorithm is utilized for both classification and regression purposes. It creates a decision tree based on the data and makes predictions accordingly. This algorithm is effective for large datasets and can produce consistent results even when a significant number of records are missing. The generated samples from the decision tree can be saved for use on other data. The Random Forest process consists of two stages: the creation of a Random Forest and the prediction using the Random Forest classifier created in the first stage.

ii. DecisionTree

The Decision Tree algorithm is depicted as a flowchart, with inner nodes representing the dataset attributes and outer branches as the results. Decision Trees are chosen due to their speed, reliability, ease of interpretation, and minimal data preparation requirements. In a Decision Tree, class label predictions originate from the root of the tree. The value of the root attribute is compared to the record's attribute, and based on the comparison results, the record is directed to the next appropriate node. The corresponding branch is then followed to its corresponding value, leading to the next node in the tree.

iii. LogisticRegression

Logistic Regression is a popular classification technique that is commonly used for solving binary classification problems. Unlike traditional regression, which uses a straight line or hyperplane to fit the data, Logistic Regression uses the logistic function to map the output of a linear equation between 0 and 1. With 13 independent variables in the data, Logistic Regression is well-suited for the classification task.

iv. Naive Bayes

The Naive Bayes algorithm is based on Bayes' rule and operates under the assumption of independence between the attributes in the dataset. This assumption is crucial for making accurate classifications. The algorithm is simple and efficient in predictions and performs well when the independence assumption holds true. Bayes' theorem calculates the probability of an event (A) given the prior probability of event B, represented as P(A/B), as shown in equation 1.

$P(A|B) = (P(B|A)\ P(A)) / P(B)$

V. RESULTS AND ANALYSIS

The performance of four machine learning algorithms, Random Forest, Decision Tree, Naive Bayes, and Logistic Regression, was evaluated using accuracyscore,precision(P),recall(R),andF-measure.Precisionmeasures the correct positive predictions, recall measures the correct positive results out of actual positive results, and F-measure combines precision and recall to indicate the accuracy of the model. These metrics were obtained using a confusion matrix, which summarizes the performance of the model. The results of the experiments on the pre-processed dataset

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 3, March 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1103021 |

are shown in Table III and Table IV, which display the confusion matrix and accuracy score for each algorithmrespectively.

F-measure [mentioned in equation (4)] tests accuracy. Precision = (TP) / (TP +FP)

Recall = (TP) / (TP+FN)

F– Measure = (2 * Precision * Recall) / (Precision +Recall)

TP True positive: the patient has the disease and the test is positive.

FP False positive: the patient does not have the disease but the test is positive.

TN True negative: the patient does not have the disease and the test is negative.

FN False negative: the patient has the disease but the test is negative.

TABLE III. VALUES OBTAINED FOR CONFUSION MATRIX USING DIFFERENT ALGORITHM

Algorithm	True Positive	False Positive	False Negative	True Negative
Logistic Regression	22	5	4	30
Naive Bayes	21	6	3	31
Random Forest	22	5	6	28
Decision Tree	25	2	4	30

TABLE IV. ANALYSIS OF MACHINE LEARNING ALGORITHM

Algorithm	Precision	Recall	F-measure	Accuracy
Decision Tree	0.845	0.823	0.835	81.97%
Logistic Regression	0.857	0.882	0.869	85.25%
Random Forest	0.937	0.882	0.909	90.16%
Naive Bayes	0.837	0.911	0.873	85.25%

VI. CONCLUSION

Due to the high number of fatalities caused by heart disease, it has becomeessentialtocreateasystemforaccuratelypredictingheartdiseases. This study aimed to determine the most effective machine learning algorithm for heart disease detection. Using the UCI machine learning repository dataset, the accuracy scores of Decision Tree, Logistic Regression, Random Forest, and Naive Bayes algorithms were compared. The results showed that the Random Forest algorithm was the most efficient, with an accuracy score of 90.16% for heart disease prediction. In the future, the work could be improved by developing a web application based on the Random Forest algorithm and using a larger dataset, which would lead to more accurate results and help healthcare professionals effectively predict heartdisease

REFERENCES

1. https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 3, March 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1103021 |

2. Garg Apurv& Sharma, Bhartendu& Khan, Rizwan. (2021). Heart disease prediction using machine learning techniques. IOP ConferenceSeries:MaterialsScienceandEngineering.1022.012046. 10.1088/1757-899X/1022/1/012046

3. HEART DISEASE PREDICTION USING MACHINE LEARNING", International Journal of Emerging Technologies and Innovative Research (www.jetir.org), ISSN:2349-5162, Vol.7, Issue 6 , page no .2081-2085 , June-2020,Available: :http://www.jetir.org/papers/JETIR2006301.pdf

4. ApurbRajdhan, Avi Agarwal, Milan Sai, Dundigalla Ravi, Dr. Poonam Ghuli, 2020, Heart Disease Prediction using Machine Learning, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH &TECHNOLOGY (IJERT) Volume 09, Issue 04(April 2020).

5. Shah D., Patel, S. & Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. SN COMPUT. SCI. 1, 345 (2020).https://doi.org/10.1007/s42979-020-00365

6. Harshit Jindal et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1022 012072

7. Aadar Pandita, Siddharth Vashisht, Aryan Tyagi, Prof. Sarita Yadav. "Prediction of Heart Disease using Machine Learning Algorithms", Volume 9, Issue V, International Journal for Research in Applied Science and Engineering Technology (IJRASET) Page No: 2422-2429, ISSN 2321-9653, www.ijraset.com.

8. N. Saranya, P. Kaviyarasu, A. Keerthana, C. Oveya. Heart Disease Prediction using Machine Learning International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-9 Issue-I, May 2020, Page No: 700-70

9. Akella, Aravind and Akella, Sudheer. Machine learning algorithms for predicting coronary artery disease: efforts toward an open source solution. Future Science OA Volume 7, Number 6, Pages FS0698, 2021, https://doi.org/10.2144/fsoa-2020-0206

10. Ravindhar NV, Anand, Hariharan Shanmugasundaram,Ragavendran, Godfrey Winster. Intelligent Diagnosis of CardiacDisease Prediction using MachineLearning. Volume-8 Issue-I l, September 2019, ISSN: 2278-3075(Online). Page No: 1417-1421.DOI: 10.35940/ijitee.J9765.0981119

11. https://archive.ics.uci.edu/mVdatasets/Heart+Disease.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details