

An Innovative Model for Heart Disease Prediction using Machine Learning

[¹] Gotlur Kalpana, [²] M.Pawan, [³] M.Arun, [⁴] Sai Kiran, [⁵] Tafazzul Shareef

[¹] Assistant Professor, Dept. of CSE, VJIT, Hyderabad, Telangana, India

[²] [³] [⁴] [⁵] Student, Vidya Jyothi Institute of Technology, Hyderabad, Telangana, India

Corresponding Author Email: [¹] gkalpana@vjit.ac.in, [²] pawanmartha02@gmail.com, [³] arunmaripelly0999@gmail.com,

[⁴] saikiranlingampally14an@gmail.com, [⁵] tafazzulshareef65@gmail.com

Abstract— One of the leading causes of death in the modern world is heart disease. The terms "heart disease" and "cardiovascular disease" are frequently used interchangeably. Heart attacks, chest pain (angina), strokes, and other illnesses caused by restricted or obstructed blood vessels are together referred to as cardiovascular disease. Clinical data analysis faces a significant problem when predicting cardiovascular disease. The main challenge in today's healthcare is provision of best quality services and effective accurate diagnosis. A machine learning model can be very helpful in the early detection and providing treatment for people with cardiovascular disease or who are at high cardiovascular risk. In this paper we have developed a Machine Learning Model with Random Forest Algorithm to detect Heart Disease accurately. We have also compared Random Forest with Logistic Regression and K-Nearest Neighbors algorithm and our experimental results shown that the accuracy parameter is high in Random Forest based heart disease detection and low in K Nearest Neighbor approach.

Keywords: Machine Learning, SVM, Random Forest, KNN.

I. INTRODUCTION

The main organ of the human body is the heart. In essence, it controls the flow of blood throughout our body. Any heart irregularity can exacerbate pain in other body areas. Any type of interruption to the heart's normal function might be categorised as a heart ailment. In the modern world, heart disease is one of the main causes of most fatalities. Heart disease can be brought on by living a sedentary lifestyle, smoking, drinking alcohol, and eating a lot of fat, which can raise blood pressure. The World Health Organization estimates that more than 10 million people worldwide pass away each year as a result of heart disease. The only means of stopping heart-related ailments are a healthy lifestyle and early identification. With the use of machine learning (ML), it has been demonstrated that it is possible to make predictions from the vast amount of data generated by the healthcare sector. Several feature combinations and well-known categorization methods can be used to introduce the prediction model. With the data, we can determine whether a patient has cardiac disease or not. The accuracy of cardiovascular disease prediction is improved by the application of machine learning techniques. The primary goal of this research is to give clinicians a tool to identify cardiac disease at an early stage. In turn, this will support giving patients effective care and averting negative outcomes. To uncover hidden discrete patterns and analyse the provided data, machine learning (ML) plays a critical role. ML approaches aid in the early detection and prediction of cardiac disease after data processing. This study examines the effectiveness of different machine learning (ML) methods for early heart disease prediction, including Naive Bayes, Decision Trees, Logistic Regression, and Random Forest.

II. LITERATURE REVIEW

This section is a review of the literature based on various methods for predicting heart disease that are currently in use. Prediction and Diagnostics of Heart Disease was created by Mamatha Alex P and Shaicy P Shaji[1]. Patients are input into SVM, Random forest, KNN, and ANN classification algorithms for the prediction of cardiac illnesses utilising data mining techniques in these attributes. where ANN produced the best outcome with the greatest degree of accuracy. 92.21%. Using big data, Abderrahmane Ed-daoudy and Khalil Maalmi[2] created "Real-time machine learning for early identification of cardiac disease." This uses a single node cluster for all of the heart disease observations and is based on the machine learning library MLlib. Scala was used to create the computer-aided categorization system.

Simulated applications was generated more than 500000 data streams (heart disease attributes) per second, Four classification algorithms, including Random Forest, Decision Tree, Logistic Regression, and Naive Bayes, are employed in the system described by Apurb Rajdhan et al. [3] to forecast the patient's health. Data are divided into training data and testing data, respectively. It was decided to build a confusion matrix that would show true and false negatives as well as true and false positives. The highest accuracy achieved with Random Forest classification was 90.16%. Francesco Mercaldo, Luca Brunese, Fabio Martinelli, Antonella Santone, and [4] This feature vector is used as the input for a deep neural network that uses "Deep learning for heart disease identification through cardiac sounds" to determine whether a cardiac sound belongs to a healthy person or a person who has a cardiac condition. The experiment we carried out established the viability of the suggested method

in a real-world setting. Galgotias University, Greater Noida, Uttar Pradesh, I [5] The four different categories of chest discomfort are taken into consideration while predicting heart disorders in the study "Heart Disease prediction using Exploratory data analysis." One of the most straightforward and well-liked unsupervised machine learning algorithms is K-means clustering. Here, the datasets are clustered, and the occurrence of chest discomfort is predicted using the clusters. , Gowsalya M., Logesh Kannan N., and Chithambaram T. [6] The major goal of the paper "Heart Disease Identification Using Machine Learning" is to improve the accuracy of heart-disease detection utilising algorithms in which the target output counts whether a person has heart disease or not.

III. HEART DISEASES

Globally, heart disease is regarded as the disease that claims the most lives from a human perspective. Particularly in this type of condition, the heart is unable to provide the necessary amount of blood to the other human body organs in order to carry out the normal activities[7]. Physical bodily weakness, poor breathing, swollen feet, etc. are some of the signs of heart disease. Techniques are necessary to spot difficult heart conditions that carry a high chance of negatively affecting human life [8]. The shortage of doctors and diagnostic tools that affect the treatment of cardiac patients makes diagnosis and treatment processes extremely difficult at the moment [9]. Early detection of heart disease is essential for minimizing heart-related problems and protecting against from serious risks. [10]. Based on a patient's medical history, a symptom analysis report from a professional, and a physical laboratory report, invasive techniques are used to identify cardiac disorders. Due to human intervention, it also delays and results in inaccurate diagnosis. It takes a lot of time, requires a lot of processing, and is expensive [11].

Several symptoms, including age, gender, pulse rate, and others, can be used to indicate the presence of heart disease. Data analysis in healthcare aids in disease prediction, better diagnosis, symptom analysis, provision of suitable medications, enhancement of care quality, reduction of costs, extension of life duration, and decrease in the death rate of cardiac patients. By placing an ECG (Electro Cardio Gram) on a patient's chest and monitoring their heartbeat, ECG (Electro Cardio Gram) aids in the early detection of irregular heartbeat and stroke. With the help of thorough clinical data, experts can forecast the development of heart disease. The heart's blood vessels must operate properly for human life to exist. Insufficient blood flow can also result in imminent death due to heart inactivity, kidney failure, and brain imbalance. Obesity, smoking, diabetes, blood pressure, cholesterol, inactivity, and poor diet are a few of the risk factors that might result in heart disease.

Cardiovascular disease known as acute myocardial infarction (AMI) occurs when blood flow or circulation to the

heart muscle is interrupted, resulting in damage or death to the heart muscle [12]. This condition is primarily brought on by a blockage, which results in decreased or obstructed blood flow to the heart muscle.

IV. MACHINE LEARNING ALGORITHM

The algorithms employed in this project are very beneficial in predicting the precise results to identify cardiac illness in which components that cause a disease can be identified. The project includes the following algorithms.

i. K –Nearest Neighbor algorithm:

KNN is a supervised classifier that carry-outs a observations from within a test set to predict classification labels. KNN is one of the classification technique used whenever there is a classification. It has a few assumptions includes dataset has little noise, labeled and it should contains relevant features. By applying KNN in large datasets takes long time to process.

ii. Random Forest Classifier:

One of the machine learning most effective technique is the random forest classifier. We should be able to obtain greater accuracy with this more refined, and training time ought to be reduced. To begin, we must divide the variables into a training set and a test set. Train the dependant variables and forecast the outcome after splitting the data.

iii. Decision tree classifier:

Preprocessing in this approach is done by first dividing the data into training and test sets. As a result of normalising the values prior to prediction, feature scaling is possible. Import a decision tree classifier to the training sets of dependent and independent variables, where the accuracy or response for the test set is predicted using the Gini index criterion.

iv. Support Vector Machine (SVM) :

SVM is another traditional machine learning technique that can predict results with higher accuracy. It is superior to other algorithms at accurately predicting accuracy in the way that is expected[5].

v. Logistic Regression Classifier :

A technique for supervised classification is logistic regression. Many independent factors are used to forecast a categorical dependent variable using this method. Logistic regression is used to forecast the result of a categorical dependant variable. The result has to be a discrete or categorical attribute as a consequence. It's less complicated to use, interpret, and train with. If there are less observations than features in the input data set, it is not advisable to use logistic regression since overfitting could result[13].

V. BACKGROUND WORK

Dataset Information

Input for the proposed model is Heart disease dataset. we took Heart.csv dataset from Kaggle database and the dimension of the date set is 1025X14 with 1025 rows and 14 columns and the names of the 14battributes are as follows.

- age
- sex
- chest pain type (4 values)
- resting blood pressure
- serum cholestorai in mg/dl
- fasting blood sugar > 120 mg/dl
- resting electrocardiographic results (values 0,1,2)
- maximum heart rate achieved
- exercise induced angina
- oldpeak = ST depression induced by exercise relative to rest
- the slope of the peak exercise ST segment
- number of major vessels (0-3) colored by flourosopy
- thal: 0 = normal; 1 = fixed defect; 2 = reversable defect
- The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.
- Target variable

Architecture:

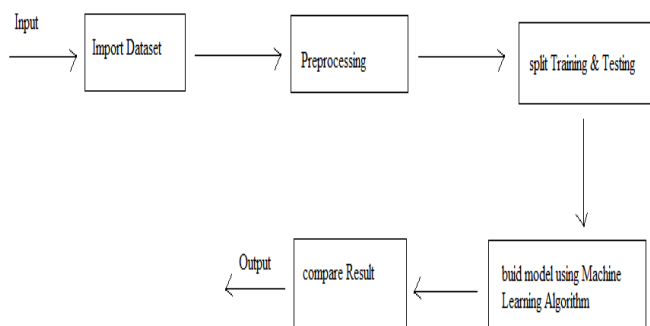


Figure1: Architecture of model

Proposed System

To build proposed system we have used python language in Jupyter Notebook. we have imported various libraries like Numpy,Pandas,Sklearn etc. this environment provides an easy-to-use visual representation of the dataset, working environment and building the predictive analytics. Then we load the dataset into environment . Here we have used the `_train_test_split_` to split the data in 80:20 ratio i.e. 80% of the data will be used for training the model while 20% will be used for testing the model that is built out of it.the following steps are applied to build ML model.

Implementation steps:

- i. in the first step we have imported dataset in to the Jupyter environment and then imported needed libraries in to the environment.

- ii. Dataset is forward to preprocessing phase .As part of preprocessing we removed noisy data, null values. and we replaced missing values by averaging existing data
- iii. After preprocessing splitting the dataset into training(80%) and testing(20%) sub datasets followed by feature selection. here we considered only needed columns which are highly correlated to predict the heart disease and we ignored irrelevant columns or fields
- iv. Fitting a ML Model using Machine Learning classification algorithms
- v. Finally Performance evaluation was done using metrics and measures and then compared 3 algorithms Logistic regression, KNN and Random forest technique. At the end we found that Accuracy improved with Random forest technique

Evaluation

Confusion matrix describes performance of a classification model; it contains information about actual and predicted classifications performed by a classifier.

Accuracy: Accuracy is defined as the ratio of correctly predicted predictions to all other types of predictions that were successfully made in the problem categorization. $Accuracy = [TP+TN/(TP+TN+FP+FN)]*100$

Precision: The proportion of accurately positive scores (Tp) to all the positive scores predicted by the classification algorithm (Tp + Fp) is known as precision..

$$Precision=[TP/(TP+FP)]*100$$

Recall: The Recall can be defined as the ratio of accurate TP to the total Tp "+" Fn.

$$Recall=[TP/(TP+FN)]$$

Specificity: The ratio of newly identified healthy people to all healthy people is known as specificity. It indicates that the person is well and that the forecast was wrong. The following is the formula for determining specificity:

$$Specificity=[TN/(TN+FP)] * 100$$

Sensitivity: The sensitivity is the proportion of newly diagnosed heart patients to all heart disease patients. It indicates that the model's prediction was accurate and that the person has heart disease. The following formula can be used to determine sensitivity:

$$Sensitivity= [TP/(TP+FN)] *100$$

F1-score: The weighted measure of sensitivity and recall precision is called F1. Its value lies between between 0 and 1. If it has a value of one, the classification method is performing well, and if it has a value of zero, the algorithm is performing poorly[14].

$$F1 = 2* (Precision*Recall) / (Precision+Recall)$$

VI. RESULTS AND DISCUSSION

Several machine learning techniques employ various datasets with independent specifications[15].These values are displayed in fig4,which illustrates how various ML

approaches' performance in detecting heart disease is evaluated with respect to various factors. According to the below-mentioned Figure 4, the accuracy parameter is high when heart disease is detected using a Random Forest model and low when using the K Nearest Neighbor method.

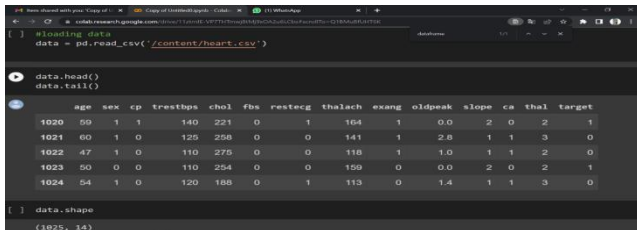


Figure 2: Loading Dataset Using Pandas

```
import numpy as np
from sklearn.ensemble import RandomForestClassifier
model=RandomForestClassifier()
inp_data=[49,1,80,0,30,1,427000,1,138,0,0,12]
inp_as_numpy=np.asarray(inp_data)
inp_data_reshape=inp_as_numpy.reshape(1,-1)
model.fit(X_train,Y_train)
prediction=model.predict(inp_data_reshape)
print(prediction)
```

[0.]

Figure 3: Building model with Random Forest classifier

```
scores = [score_lr,score_knn,score_rf]
algorithms = ["Logistic Regression","K-Nearest Neighbors","Random Forest"]

for i in range(len(algorithms)):
    print("The accuracy score achieved using "+algorithms[i]+" is: "+str(scores[i])+" %")
```

The accuracy score achieved using Logistic Regression is: 78.33 %
The accuracy score achieved using K-Nearest Neighbors is: 58.33 %
The accuracy score achieved using Random Forest is: 91.67 %

Figure 4: Accuracy Comparison of 3 ML algorithms

VII. CONCLUSION

In this research we have developed a Machine Learning Model to detect Heart Disease accurately. Initially we imported dataset and then performed preprocessing after that we have performed splitting the dataset into two subsets training and testing and then we have extracted relevant features for further processing finally we have used Machine Learning based Random forest algorithm to build model for prediction of Heart Disease. In this paper we have considered different attributes and implemented with three different ML algorithms ie. Logistic Regression, K Nearest Neighbor, Random Forest algorithms used to predict Heart Disease and to analyse the accuracy. Finally comparison was made among them. The experimental results shown that the proposed model with Random forest was more effective with better accuracy in achieving higher results and low in K Nearest Neighbor approach. This research can considerably improve the healthcare system and give medical practitioners an important tool for diagnosing and predicting heart failure survival. Further we can extend this work with comparing all other ML algorithms. we can also build model using advanced deep learning techniques to improve accuracy.

REFERENCES

[1]. Mamatha Alex P and Shaicy P Shaji, " Prediction and Diagnosis of Heart Disease Patients using Data Mining Technique", International Conference on Communication

and Signal Processing, April 4-6, 2019, India.

[2]. Abderrahmane Ed-daoudy, Khalil Maalmi," "Real-time machine learning for early detection of heart disease using big data approach" 978-1-5386-7850-3/19/\$31.00 ©2019 IEEE.

[3]. Apurb Rajdhan , 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) .

[4]. Luca Brunesea, Fabio Martinellib, Francesco Mercaldoa,b,*, Antonella Santonec "Deep learning for heart disease detection through cardiac sounds".

[5]. hool of Computing Science & Engineering, Galgotias University, Greater Noida, U.P., I"Heart Disease prediction using Exploratory data analysis"Procedia Computer Science 173 (2020) 130–139,ICITETM2020.

[6]. Chithambaram T , Logesh Kannan N,Gowsalya M ⁵ "Heart Disease Detection Using Machine Learning", Creative Commons Attribution 4.0 International License,Oct 27th 2020.

[7]. A. L. Bui, T. B. Horwich, and G. C. Fonarow, "Epidemiology and risk profile of heart failure," Nature Reviews Cardiology, vol. 8, no. 1, pp. 30–41, 2011.

[8]. J.Mourão-Miranda,A.L.W.Bokde, C. Born, H.Hampel, and M. Stetter, "Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data,"NeuroImage,vol.28,no.4,pp.980–995, 2005.

[9]. S.Ghwanmeh, A.Mohammad, and A.Al-Ibrahim, "Innovative artificial neural networks-based decision support system for heartdiseasesdiagnosis," Journal of Intelligent LearningSystems and Applications, vol. 5, no. 3, pp. 176–183, 2013.

[10]. Q. K. Al-Shayea, "Artificial neural networks in medical diagnosis," International Journal of Computer Science Issues, vol. 8, no. 2, pp. 150– 154, 2011.

[11]. K. Vanisree and J. Singaraju, "Decision support system for congenital heart disease diagnosis based on signs and symptoms using neural networks," International Journal of Computer Applications, vol. 19, no. 6, pp. 6–12, 2011.

[12]. Al Mamoon I, Sani AS, Islam AM, Yee OC, Kobayashi F, Komaki S (2013) "A proposal of body implementable early heart attack detection system", 1-4. Patterson K (2016) Matthias Nahrendorf. Circ Res 119: 790-793.

[13]. B Padmajaa, Chintala Srinidhib, Kotha Sindhuc, Kalali Vanajad, N M Deepikae, E Krishna Rao Patrof" Early and Accurate Prediction of Heart Disease Using Machine Learning Model" Turkish Journal of Computer and Mathematics Education Research Article , Vol.12 No.6 (2021), 4516-4528.

[14]. Yar Muhammad1, Muhammad Tahir1, Maqsood Hayat1* & Kil To Chong," Early and accurate detection and diagnosis of heart disease using intelligent computational model" Scientific Reports | (2020) 10:19747.

[15]. Umarani Nagavelli, Debabrata Samanta , and Partha Chakraborty " Machine Learning Technology-Based Heart Disease Detection Models" Hindawi Journal of Healthcare Engineering Volume 2022, Article ID 7351061, 9 pages <https://doi.org/10.1155/2022/7351061>.