Heart Disease Prediction Using Artificial Neural Network

Samiksha Jadhav
Student, Computer Science & Engineering
MIT World Peace University, Pune
1032190912@mitwpu.edu.in

Shambhavi Kumar
Student, Computer Science & Engineering
MIT World Peace University, Pune
1032191360@mitwpu.edu.in

Mustafa Mokashi
Student, Computer Science & Engineering
MIT World Peace University, Pune
1032191369@mitwpu.edu.in

Abstract— **Heart disease is the leading killer, according to WHO around 17.9 million deaths occur every year due to cardiovascular diseases (CVDs). The most important factors of heart disease and stroke are, raised blood pressure, obesity, unhealthy lifestyle etc. healthcare organizations facing issues due to poor clinical decisions which are leading to major consequences. Data mining can help to discover the rich knowledge hidden in hospital databases and can become an effective solution for poor clinical decisions and save excessive medical costs. The objective of this paper is to analyze the dataset and predict heart failure. We have used the largest heart disease dataset available on Kaggle and IEEEDataPort which is combined over 12 common features, this dataset is curated by combining Cleveland, Hungarian, Switzerland Long Beach VA, Statlog (Heart) Dataset. Data mining techniques such as Decision Tree, Random Forest, Hybrid (Decision Tree + Random Forest)[1], SVM, and KNN are used to predict heart disease. But Artificial Neural Networks is the most important data mining technique, by using this technique it provides accurate results with minimum execution time. In this paper we use ANN technique to predict heart disease.**

Keywords— Data Mining, Artificial Neural Network, Heart Disease, Attribute, Prediction.

I. Introduction

Heart disease is a critical health problem and the death rate due to heart failure is increasing year by year. The health of a person is dependent on eating habits and personal behavior [1]. Heart disease occurs with common signs or symptoms that indicate physical parameters of a person concluding whether the person is abnormal: cholesterol, blood sugar level, blood pressure, resting ECG, heart rate are some of physical parameters. Determining whether a person is prone to heart disease is a complicated task in the medical field, as it is a critical and most important task. Many researchers are trying their best to come up with feasible solutions which can help to reduce clinical errors and save a lot of money.

Early detection of heart disease is a very complex process [4] and lack of qualified doctors and diagnostic centers may prove to be disastrous when treating heart patients. With this concern, today the field of computer technology is widely adopted to help and provide feasible solutions with data mining and machine learning techniques. Data mining is a process dependent on huge databases to extract useful and meaningful information or patterns. However, to extract useful data patterns the data set needs to be organized and should be free from heterogeneous data, pre-processing is an important task which cannot be ignored in the data mining process.

Determining features or characteristics of data objects is an important task. Heart disease might be because of [3] weight, high blood pressure, absence of rest, hypertension, family ancestry, smoking etc. common symptoms include breathlessness, chest pain and angina. The characteristics such as bps, blood sugar level, resting ecg, age, sex and other common parameters can be considered for data mining. For analysis these characteristics are very much essential, by which data mining techniques can predict based on the data that is provided. Most of the researchers worked on techniques like support vector machine(SVM), decision tree, naïve bayes, random forest classification, KNN, ANN to predict heart disease most of them have used Cleveland dataset from UCI repository. But, In our research work we have used the largest heart disease dataset that is available on the internet, it is curated by combining five datasets namely Cleveland + Hungarian + Switzerland + Long Beach VA + Statlog (Heart) Data Set with 12 common attributes.

II. Problem Statement

Today, all hospitals use hospital management systems to store patients' data such as medical history, treatment requirements, and many other important factors. Some hospitals use this data to extract some useful statistics, like

- Analysis of heart disease based on high BP, diabetes, obesity, physical inactivity, etc.
- Average age of patients with heart disease.
- Survival rate of patients with heart disease.
- Recovery rates of patients with heart disease and a lot of such information.

But these systems cannot perform complex queries for predicting heart disease from a patient's medical history.

Data mining techniques have the potential to process data and perform analysis on raw data to generate meaningful and knowledgeable conclusions, which can help to prevent or improve clinical decisions for heart failure disease.

Purpose:

To detect the chances of having heart diseases at an early stage and prevent deaths resulting from it & To use various classification algorithms on heart records of the patients and finding out the most efficient classification technique.

Scope:

Integration of implemented techniques with the patient record system could reduce medical errors and improve clinical decisions & To generate meaningful information which can help to improve quality services at affordable costs.

Objectives:

To find the predominant symptoms leading to heart diseases in men and women, respectively & Reduce cost of medical tests.

III. LITERATURE SURVEY

TABLE I. LITERATURE SURVEY

Ref. No.	Techniques Used	Dataset Used	No. of features	Results Obtained	Advantages	Research Gap
[1]	Decision Tree, Random Forest, Hybrid (Decision Tree + Random Forest)	Cleveland Dataset	14	Decision Tree -79% Random Forest - 81% Hybrid - 88%	1.Tested with three different techniques.	Increasing sample size
[2]	SVM, Logistic Regression	Cleveland Dataset from UCI	14	79% Accuracy	1. Tested with five different ML algorithms. 2. Data visualization using various techniques(Heat Map, Box plot, confusion matrix)	Increasing sample size
[3]	Random Forest Classifier and PCA.	Cleveland and statlog Cleveland + Hungarian dataset from UCI machine learning repository.	14	92.85% Accuracy	1. Tested with 9 different techniques with 2 feature selection 2. Good performance with PCA feature selection	Old dataset is used
[4]	SVM, Random Forest, Naive Bayes and Decision Tree.	UCI Heart disease prediction - Cleveland	14	SVM - 98% Random Forest - 99% Naive Bayes - 90% Decision Tree - 85%	1. Tested with four different techniques. 2. Achieved highest accuracy of 99% with random forest	Old dataset is used
[5]	Naive Bayes, Decision trees and Neural Network	UCI repository	12	Naive Bayes-87.91%	1. Neural Network is providing a better result in such a non-linear application of health disease prediction.	it has drawback of being stuck in a local minima solution
[6]	Naive Bayes, Decision List and KNN.	UCI repository	11	Naive Bayes-85%	1.KNN provided better results.	Clustering of data was not performed.

IV. PROPOSED WORK

The proposed workflow has the following advantages

- Implemented five data mining algorithms
- Accuracy of all algorithm show model which is most suitable

The implementation is performed with the below given methodology:

a. Dataset is collected from Kaggle

b. Data Visualization is done

c. Splitting dataset into train data and test data

d. Train the model

e. Test the trained model and predict values

Statlog_cleveland_hungary dataset is considered. It is split into testing and training parts. We have assumed 80% of the dataset as training input to the machine learning algorithms and fit the model. the remaining 20% as testing data for heart disease prediction.

a. Logistic Regression

Logistic regression [2] is one among the foremost popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the specific variable employing a given set of independent variables.

Logistic regression predicts the output of a categorical variable. Therefore, the result must be a categorical or discrete value. It is often either 0 or 1, Yes or No, true or False, etc. but rather than giving the precise value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression may be a significant machine learning algorithm because it's the power to supply probabilities and classify new data using continuous and discrete datasets.

b. Random Forest

Random Forest a popular machine learning algorithm that belongs to the supervised learning technique. It is mostly used for both Regression and Classification problems in Machine Learning. it's supported the concept of ensemble learning, which may be a process of mixing multiple classifiers to unravel a posh problem and to enhance the performance of the model [4].

As the name suggests, "The Random Forest may be a subdivision containing a variety of decision trees in the various datasets provided and take the standard to improve the prediction accuracy of that database." rather than relying on a single decision tree, the random forest takes a prediction from each tree and supports multiple predictive votes, and it predicts the ultimate output.

The greater number of trees within the forest results in higher accuracy and prevents the matter of overfitting.

c. K-Nearest Neighbour (KNN)

K-Nearest Neighbour is one among the Machine Learning algorithms supported Supervised Learning technique [6]. The K-NN algorithm captures the similarities between new cases / data and available cases and places a new case in a category almost identical to the available categories.

The K-NN algorithm stores all available data and separates replacement data that supports similarity. this suggests when new data appears then it is often easily classified into a well suite category by using K- NN algorithm.

K-NN algorithm are often used for Regression also as for Classification but mostly it's used for the Classification problems. K-NN may be a non-parameter algorithm, suggesting that it does not make any assumptions about root data.

It is also called a lazy learner algorithm because it doesn't learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

The KNN algorithm in the training phase automatically backs up the database when it receives new data, and then classifies that data into a category that closely resembles new data.

d. Support Vector Machine

Support Vector Machine or SVM [4] is one among the foremost popular Supervised Learning algorithms, which is employed for Classification also as Regression problems. Mainly, however, it is used for Distribution Problems in Machine Learning.

The goal of the SVM algorithm is to create a simple line or decision line that will divide n-dimensional space into classes so that it can be easily set new datum within the correct category within the future. This best decision boundary is named a hyperplane.

SVM chooses the acute points/vectors that help in creating the hyperplane. These extreme cases are called supporting vectors, which is why the algorithm is called Vector Support Machine.

e. Neural Networks

Neural networks, also called artificial neural networks (ANNs) or mimic neural networks, are a subset of machine learning and are in the hands of deep learning algorithms. Their name and structure are inspired by the human brain, mimicking the way that biological neurons signal to at least one another.

Artificial neural networks (ANNs) [5] are comprised of a node layers, containing an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, connects to a different and has an associated weight and threshold. If the output of a person node is above the required threshold value, that node is activated, sending data to subsequent layer of the network. Otherwise, no data is passed along to subsequent layer of the network.

Neural networks depend on training data to find out and improve their accuracy over time. However, once these learning algorithms are fine-tuned for accuracy, they're powerful tools in computing and AI, allowing us to classify and cluster data at a high velocity. Tasks in speech recognition or image recognition can take minutes versus hours in comparison to the manual identification by human experts. one among the foremost well-known neural networks is Google’s search algorithm.

V. SYSTEM ARCHITECTURE

Fig 1 depicts how heart disease prediction system works with different data mining techniques.

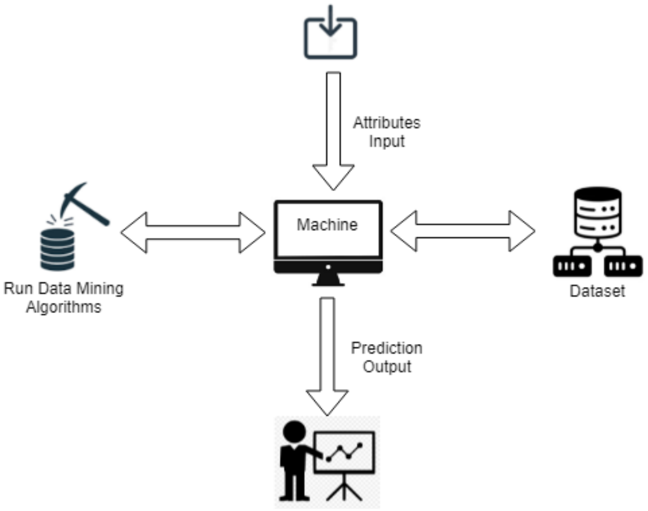

Fig 1: Heart Prediction System Architecture

Pre-processed data is fed into the machine, it is trained with different data mining algorithms. After you input the required attributes the machine learning algorithms analyse and predict the heart disease.

VI. DATASET DESCRIPTION

Dataset Description: The dataset used is taken from kaggle. It has 12 different attributes and 1190 instances. It does not have any missing values. The features are described in the table given below:

TABLE II. ATRIBUTES

S.No.	***Attribute***	***Code given***	***Data type***
1	age	Age	Numeric
2	sex	Sex	Binary
3	chest pain type	chest pain type	Nominal
4	resting blood pressure	resting bp s	Numeric
5	serum cholesterol	cholesterol	Numeric
6	fasting blood sugar	fasting blood sugar	Binary
7	resting electrocardiogram results	resting ecg	Nominal
8	maximum heart rate achieved	max heart rate	Numeric
9	exercise induced angina	exercise angina	Binary
10	oldpeak =ST	oldpeak	Numeric
11	the slope of the peak exercise ST segment	ST slope	Nominal
12	class	target	Binary

Fig 2 dataset is visualized to urge number of heart condition cases and number of normal cases from the dataset. It is shown as histogram plot.

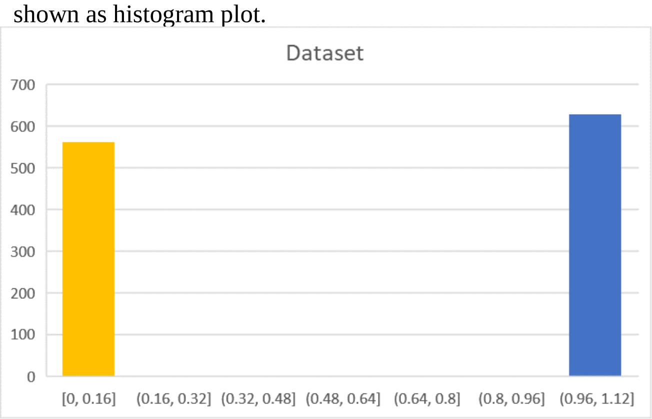

Fig 2: Data Visualization of Dataset

VII. EXPERIMENTAL WORK

Predicting diseases, especially heart disease, is a real issue because the monitoring and prediction of medical data is very important because it is related to people's lives and is an interesting activity and contains a very powerful challenge and other areas. Our current work consists of two main objectives: the first is data purification and the second is to identify patterns related to the diagnosis of heart disease in order to drive the process of incorporating cardiac predictors into a set of medical data. This work will help many medical professionals to diagnose and understand the underlying causes of the disease.

Tableau [7] Visualization:

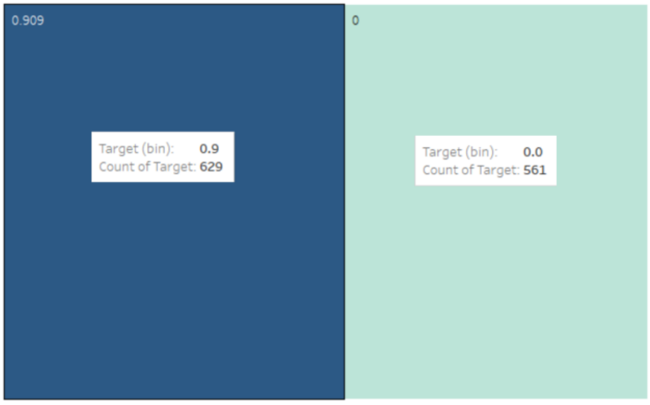

Fig 3: Analysis of Target Values

Fig 3 treemap shows target variables, left side area represents number of patients with heart disease, and right side area represents number of patients without heart disease.

Fig 4 bubbles shows male vs female heart disease patients, left bubble area represents number of female patients with heart disease, and right bubble area represents number of male patients with heart disease. Target value specifies count of patients with heart disease.

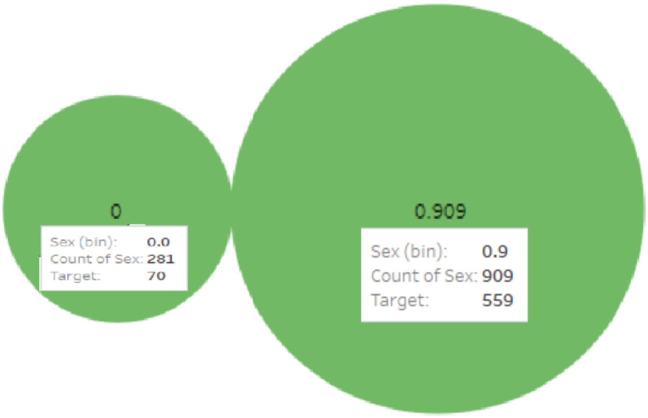

Fig 4: Male vs Female Comparison

The proposed algorithms are implemented in python, with libraries such as pandas, matplotlib, sklearn, and for neural networks tensorflow library is imported.

The dataset is downloaded from Kaggle. Machine learning techniques such as Linear Regression, Random Forest, K-Nearest Neighbour, Support Vector Machine and Neural Networks were used. These machine learning algorithms were used to predict the heart disease. The result shows that Heart disease detection is effective using the Logistic Regression, Random Forest And Neural Networks algorithm. Logistic Regression achieves around 89.13% accuracy, Random Forest achieves 89% accuracy, K-Nearest Neighbour achieves 85.32% accuracy, Support Vector Machine achieves 68.47% accuracy, Neural Networks model achieves 88.59% accuracy.

TABLE III. RESULTS OBTAINED

Algorithm	Accuracy
Logistic Regression	89.13%
Random Forest	89%
K-Nearest Neighbour	85.32%
Support Vector Machine	68.47%
Neural Networks	88.59%

VIII. COMPARATIVE ANALYSIS

The result shows that Logistic Regression achieves 89.13% accuracy, where as Random Forest yields almost same accuracy as Logistic Regression model that is 89%, and K-Nearest Neighbour achieves 3.81% less than Logistic Regression and Random Forest, Support Vector Machine achieved lowest accuracy rate among all other algorithms i.e., 68.47%, Neural Network achieved almost same accuracy as of Logistic Regression and Random Forest that is 88.59%.

Fig 5 shows the results obtained with various Machine Learning techniques, Logistic Regression, Random Forest and Neural Networks appear to be efficient as their results are high and almost same. Support Vector Machine seems to be inefficient as it has lowest accuracy rate and K-Nearest Neighbour achieved 85.32% accuracy.

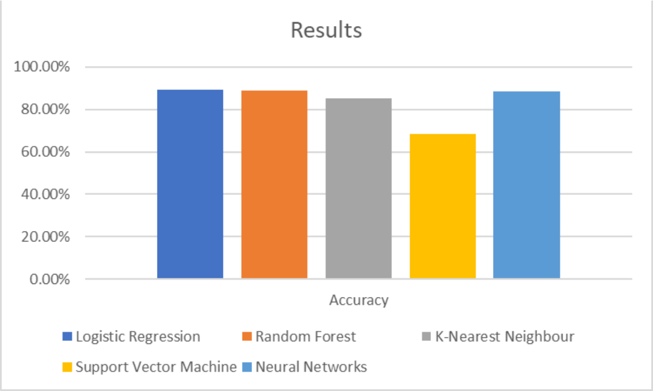

Fig 5: Analysis of different algorithms

IX. CONCLUSION

The main motivation for this project is to provide understanding about the diagnosis and treatment of heart disease using data mining process. Prediction the chances of a patient getting heart disease. These qualities are supplied to Logistic Regression, SVM, Random Forest, KNN, and Neural Networks. Arrangement of algorithms in which Neural Networks, Random Forest and Logistic Regression gave the best possible result with high accuracy. Valid performance is achieved uses the Neural Networks algorithm to diagnose heart disease and can further developed by increasing the number of attributes.

REFERENCES

[1] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 1329-1333, doi: 10.1109/ICICT50816.2021.9358597.

[2] V. Gupta, V. Aggarwal, S. Gupta, N. Sharma, K. Sharma and N. Sharma, "Visualization and Prediction of Heart Diseases Using Data Science Framework," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), 2021, pp. 1199-1202, doi: 10.1109/ICESC51422.2021.9532790.

[3] Farzana Tasnim and Sultana Umme Habiba, “A Comparative Study on Heart Disease Prediction Using Data Mining Techniques and Feature Selection”, 2021 2nd International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST), ISBN: 978-0-7381-3042-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICREST51555.2021.9331158

[4] Vijeta Sharma, Shrinkhala Yadav, Manjari Gupta, “Heart Disease Prediction using Machine Learning Techniques", 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), ISBN: 978-1-7281-8337-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICACCCN51052.2020.9362842

[5] Ankita Dewan,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)

[6] Sarath Babu, Vivek EM, Famina KP, Fida K, Aswathi P, Shanid M, Hena M, “Heart Disease Diagnosis Using Data Mining Technique”, 2017 ISBN: 978-1-5090-5686-6/17/$31.00 ©2017 IEEE

[7] Tableau Public: Free Data Visualization Software, Tableau Software, LLC, a Salesforce Company. | https://public.tableau.com/en-us