Bigdata Analytics In Healthcare For Heart Disease Prediction Using Machine Learning

Abinaya R[1], Maruthupandian M[2], Akash A[3], Nazih J[4]
[1]Assistant Professor, Dept of CSE
[2, 3, 4]Dept of CSE
[1, 2, 3, 4] Sri Ramakrishna Institute of Technology, Coimbatore

Abstract- *In the present era, heart-related illnesses are quite prevalent. This occurs because of changes in eating and living patterns. The heart-related disease can be life-threatening, As the population grows to make a diagnosis is challenging. It must be carried out accurately and effectively then only will it assist in providing patients with care at the appropriate time. Clinicians can personalize their treatment and diagnosis for each patient if a disease is caught in its initial stages or if it can be diagnosed well in advance. It reduces the mortality rate and the risk factor. With the help of a learning algorithm, cardiovascular disease detection.*

The system includes electronic health data to make a diagnosis of the condition. Researchers are focusing on creating intelligent algorithms to precisely diagnose patients using electronic health data. Using the random forest approach, the model is trained. The test accuracy of 84.71% and train accuracy of 85.25% is obtained. Before constructing the simulations, steps in data, which was before, and feature extraction were taken based on various accuracy measures the models were assessed. This model has the best accuracy, 87%.

Keywords- Heart disease, Random Forest, Electronical data, Simulations, Prediction model, Test data, Train data

I. INTRODUCTION

The principal contributor to the worldwide death rate is heart-related diseases. 17.9 million individuals every year died from heart-related conditions. Mortality is more prevalent in low- and moderate-income nations.

Major Modifications in dietary and lifestyle habits are often linked to health disorders like hypertension, hyperlipidaemias, and diabetes.These kinds of health problems eventually cause complications with the heart.Cardio illnesses are caused by a wide range of lifestyle factors, such as cigarettes, frequent caffeine, and alcohol abuse, stress, and inadequate physical activity. The disease is usually only identified when it is already progressed.So, it’s very hard to treat the patients it may cause a serious problem and leads to death.

If indeed the disease is identified at a preliminary phase, both the general mortality rate and the participant's risk factors are decreased. To take action to save death, an early, accurate, and effective medical diagnosis. The health information dataset and computational modelling algorithms are used to determine the condition. The data set has a variety of features, and with the help of a prediction algorithm, the system is readily trained.

Machine learning is one of the primary technologies used to build and evaluate the system. Computers are imitating abilities, a subset of the reaction to current learning algorithms, within the larger field of research called ai technology. The phrase "intelligent machines" implies the integration of evolving systems.Computational technologies, on the other hand, are taught to identify how to analyze and make use of data. It acquires knowledge from the natural world as well as some biological factors like blood pressure, sex, age, etc., and compares. This scientific study provides a concise illustration of how learning algorithms could be used to predict cardiac dysfunction.

II. MATERIALS AND METHODS

2.1 MACHINE LEARNING.

Automation is one of the fundamental methods used to teach and analyze the system. The system was quickly educated from the data via machine learning techniques. There seem to be three key computational model algorithms.

1. Supervised
2. Unsupervised
3. Reinforced.

1.Supervised Learning- The labeled data is involved in this technique.
2.Unsupervised Learning- Non-labelled data is involved in this technique.
3.Reinforcement Learning-There is some kind of feedback loop in place, and a parameter needs to be optimized

2.2 METHODOLOGY

The first step in the process of this system is data collection. The information is obtained from a reliable and authorized area.

Using the collected data, the system is evaluated and trained. After gathering the dataset, suitable variables are selected for our system, such as age, gender, and other pertinent elements [1]. The logistics model is calibrated using the data. By exploring multiple techniques of machine learning on the information in a bid to achieve our aim. The information is split into two groups after processing. the model's training and test data Seventy percent of the data are utilized for training. After becoming trained with the learning set of data, the algorithm can anticipate the ailment.

2.2.1 ARCHITECTURE OF THE HEART DISEASE PREDICTION MODEL

With the aid of the data gathered from the various datasets, the model's primary goal is to forecast cardiac illness in its early stages. The data is utilized to train and test the prediction system. The prediction system is a multiuser online application that allows users to log in using their login credentials. After filling out the form's required information, the predictive model then begins analyzing the data. Following analysis, the model will forecast the result and display the result as a target value.The prediction model is composed of three basic layers: authentication, analysis, and target value creation.

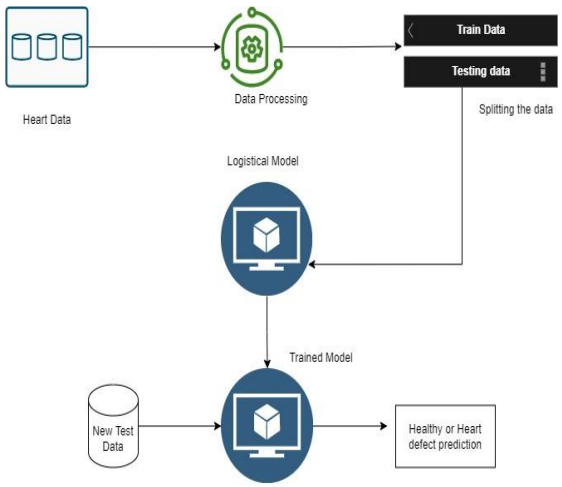

Figure :1 Workflow Diagram of Prediction System

2.2.2 DATA SOURCE

In this system, the UCI Repository is used to select our dataset the set contains 14 various attributes more than 200 examples of data like the age and gender of a person along with their health condition like chest pain and ECG results.

2.2.3 DATA PRE-PROCESSING

Hard data can include extremely large figures, inconsistent data, and invalid information. To avoid these problems and offer accurate predictions, these data have been already analyzed. Data cleaning usually includes removing noise and missing values. The voids in the data for hypertension, lipid profile, albumin, etc. must be fulfilled to obtain a precise and useful result. To make data more comprehensible, it is transformed, or altered, from one structure to another. Activities such as grading, filtering, and aggregating are addressed. Resampling is applied to enhance the accuracy.

The flowchart for data. The dataset is put through a pre-processing step that includes managing missing values, feature selection and deletion, normalization, standardization, and resampling. The data is utilized for training and testing after pre-processing. The trained model is then used to forecast cardiac problems.

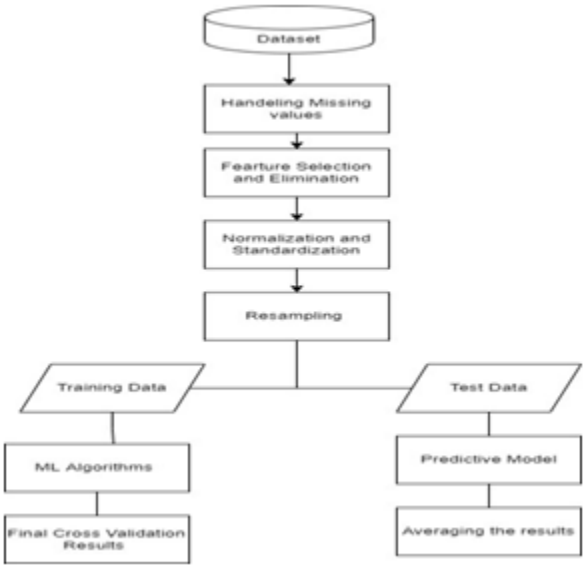

Figure:2 Data Flow Diagram

S.NO	Attribute	Representative Icon	Details
1	Age	Age	Age in years
2	Sex	Sex	0-female,1-male
3	Chest pain	CP	Types of chest pain
4	Rest blood pressure	TRES	Resting systolic blood pressure serum
5	Serum cholesterol	CHOL	Serum cholesterol
6	Fasting blood sugar	FBS	Fasting blood sugar>120 mg/dl
7	Rest electrocardiograph	RESCG	Electronic data of heart wave
8	Max cardiac rate	THAL	Maximum heart rate achieved.
9	Exercise-induced angina	EXG	Exercise-induced angina (0-no; 1—yes)
10	ST depression	OLD PEAK	ST depression induced by exercise
11	Slope	SLP	Slope of the peak exercise ST segment
12	No. of vessels	CA	No. of major vessels (0–3)
13	Thalassemia	THALI	Three types of defect can occur
14	Target	TAR	Target=0(Yes)and Target=1(NO)

Figure 3. Cardiovascular diseases statistic traits and characteristics.

Incorporation Since the data may come from more than one source and not just one, it must be combined before processing. To acquire successful results, the collected data reduction is to be arranged. After that, the data are labeled and divided into training data sets.

2.2.4 COLLECTION OF DATASETS

Gathering the data subsequently is the base for our prediction of the cardiovascular disease system. That after the dataset was acquired, Data for training and evaluation are separated from the entire dataset. The forecasting model learns from the developing dataset, while the testing dataset serves as the foundation for its evaluation. 70 percent of the total data used in this project is used for training, whereas only 30 percent is used for testing. The dataset for the study was cardiovascular disease. 14 of the dataset's 76 variables are utilized by the system throughout the operation.

2.3 METHODS AND ALGORITHMS USED

2.3.1 PRECISION AND RECALL

Accuracy is the main performance indicator in machine learning for pattern identification and categorization. To create the ideal machine learning model, it is essential to understand these ideas. it results in more precise and accurate results. In learning algorithms, some models require higher recall while others require more accuracy.

2.3.2 CONFUSION MATRIX

When real values are known, a classification technique tabulated representation of the anticipated outcomes of any binary classifier is used to illustrate how well a classifier worked on a set of test values. The words used in this matrix may be puzzling to novices, even though it is simple to use. A typical confusion matrix for a binary classifier is shown in the image below (However, it can be extended to use for classifiers with more than two classes). Because accuracy can be deceiving when applied to skewed datasets, other metrics based on the contingency table are likewise pertinent for assessing performance.

2.3.3 RANDOM FOREST

The most significant computational method the labeled data mining approach includes Random Forest. It may be used to solve classification and linear interpolation ML tasks. It is based on the notion of ensemble learning, a strategy for combining several classifiers to handle challenging problems and improve model performance. The Random Forest classifier, as its name suggests, averages the results from numerous decision trees applied to various subsets of an input dataset to improve the projected accuracy of the dataset. As opposed to relying just on one decision tree, the random forest receives forecasts from each tree and predicts based on the votes of numerous forecasts.

III. RESULT AND DISCUSSION

Whether the patient seems to have a cardiac problem will be shown as a Sure or Even no answer in the system's output. If the person is predisposed to have heart disease, the reaction will be Yes, and vice versa. The Random Forest method of machine learning is used to build the model while taking into account the evaluations and results of big data analytics in the healthcare field. If the prognosis is accurate, the patient should see a cardiologist for a more thorough check-up. statistics of the results obtained after testing the data.

Table 1:Percentage accuracy results of classification techniques.

S.NO	Algorithm	Test Accuracy	Train Accuracy
1	Nave Bayes	85.25%	83.47%
2	Random Forest	86.96%	87.64%
3	Decision Tree	82.00%	84.3%
4	K-Nearest Neighbour	78.10%	63.93%

IV. CONCLUSION

To understand the dataset, look for missing data, and identify the most important aspects, The data is explored first. Then visualizations were made throughout this process using Plotly, Seaborn, and Matplotlib. During the data preparation

step, then the attributes are converted into numeric ones, categorize values and build a few more features. After that, create machine-learning models and cross-validate them. Improvements might be made by undertaking a more complete feature engineering approach that entails contrasting and graphing the characteristics against one another as well as locating and eliminating the noisy features. Then the model is trained. After that the input is given as input then the model is predicting the output.

REFERENCES

[1] Archana Sing, Rakesh Kumar, "Heart Disease Prediction system using machine learning Algorithms", International Journal of Engineering Research & Technology (2020).

[2] Chaimaa Boukhatem Heba Yahia"Heart disease prediction using machine learning" International Conference of Electronics and Engineering (2018)

[3] VijetaSharma, Shrinkable Yadav, Manjari Gupta "Heart disease prediction using Machine Learning Techniques", IEEE 41st Annual Computer Software and ApplicationsConference (2019).

[4] Reldean Williams, Thokozani Shongwe, Ali N Hasan, Vikash Rameshwar "Heart Disease Prediction Using Machine Learning Techniques", IEEE 8th R10 Humanitarian Technology Conference (2020).

[5] Senthil Kumar Mohan, Chandrasekar Tirumala, Gautam Srivastava " Effective "Heart Disease Prediction Using Hybrid Machine Learning Techniques" InternationalConference on Computing Methodologies and Communication (2019).

[6] Sunitha Guruprasad,ValeshLevis Mathias,Winslet Dcunha "Heart Disease Prediction Using Machine Learning Techniques" IEEE International conference on Electrical, Electrons Communication (2021).

[7] Seckeler MD, Hoke TR. "The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease". Clin Epidemiol. 2011; 3:67.

[8] Gaziano TA, Bitton A, Anand S, Abrahams-Gessel S, Murphy A. "Growing epidemic of coronary heart disease in low-and middle-income countries."CurrProblCardiol. 2010;35(2):72–115.

[9] Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. "Can machine learning improve cardiovascular risk prediction using routine clinical data?" PLOS ONE. 2017;12(4): e0174944.

[10] Ramalingam VV, Dandapath A, Raja MK. "heart disease prediction using machine learning techniques: a survey". Int J Eng Tech nol. 2018;7(2.8):684–7.

[11] Patel J, TejalUpadhyay D, Patel S. "heart disease prediction using machine learning and data mining technique". International Journal for Research in Applied Science and Engineering Technology2015;7(1):129–37

[12] Pahwa K, Kumar R. "Prediction of heart disease using the hybrid technique for selecting features." In: 2017 4th IEEE Uttar Pradesh section international conference on health and science.

[13] Pouriyeh S, Vahid S, Sannino G, De Pietro G, Arabnia H, Gutier rez J. "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease." In: 2017 IEEE symposium on computers and communications (ISCC). IEEE.p. 204–207.

[14] Chauhan R, Bajaj P, Choudhary K, Gigras Y. "Framework to predict health diseases using attribute selection mechanism". In: 2015 2nd international conference on computing for sustainable global development (INDIACom). IEEE. p. 1880–84.

[15] Devansh Shah, Samir Patel & Santhosh Kumar Bharti "Heart Disease Prediction system using machine learning technique", International Conference on Engineering and Technology (2020)

[16] IsrealUfumaka, "Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction", International Journal of Scientific and Research Publications" DOI:10.29322.11.01.2021.P10936.(2020).