International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 issue: 05 | MAY - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Study on Heart disease predication using Artificial Intelligence

PALLAVI, ALAMMA B H

Department of MCA, Dayananda Sagar College of Engineering
Department of MCA, Dayananda Sagar College of Engineering

----------**----------

Abstract

Several disorders may cause cardiovascular disease, or heart disease. Heart disease has several risk factors and requires quick diagnosis for proper treatment. A symptom Data mining is a common way to analyse healthcare data. Researchers employ data mining and machine learning to analyse complex medical data for healthcare practitioners. Cardiologists predict heart disease. This study's model uses heart disease-related parameters. Supervised learning uses Naive Bayes, decision tree, K-nearest neighbour, and random forest. It uses UCI's Cleveland database of heart disease patients. The 303 instances of data include characteristics. Only 14 of the 76 traits are verified, but they're crucial to confirming the claim. Algorithm performance. This research estimates heart disease risk in a population. Patients. K-nearest neighbour has the highest accuracy, as shown in result and discussion.

1.INTRODUCTION

In the human body, the heart plays a crucial role. Everything in our bodies is oxygenated and nourished thanks to it. In a matter of minutes, the brain and other vital organs will cease to function, and the individual will die.

Heart-related ailments are on the rise as a result of lifestyle changes, workplace stress, and poor dietary habits. Cardiovascular disease has risen to prominence as a leading cause of mortality among the world's population. More than a third of all deaths throughout the globe are caused by heart-related conditions, according to the World Health Organization. Not just in the United States, but also in India, cardiovascular disease is becoming the leading cause of mortality. Heart disease claimed the lives of 1.7 million Indians in 2016, according to the World Health Organization's 2016 Global Burden of Disease Report, released on September 15th.. An individual's productivity suffers as well as their health care costs when they suffer from heart-related disorders.

Over the previous ten years, cardiovascular disease has been the leading cause of death. Cardiovascular disease kills 17.9 million people worldwide each year, according to the World Health Organization. Cardiovascular disease and cerebral stroke account for 80% of these deaths. Thousands of individuals die each year in low- and middle-income countries [2]. Heart disease may be influenced by a person's lifestyle, work habits, and underlying genetics. Smoking, drinking too much alcohol or coffee, being stressed, and not exercising all contribute to the development of heart disease. Other physiological variables including high blood pressure, obesity, and pre-existing cardiac diseases are also risk factors. Preventive steps may be taken to avoid mortality if heart disease is diagnosed quickly and accurately.

Doctors are increasingly relying on information technology to help them make decisions in the healthcare business. Disease management, treatment, and pattern and link identification among diagnostic data are all made easier with the assistance of this tool. There are many persons who may benefit from preventative medication who are not identified by current methods of predicting cardiovascular risk. The intricate connections between risk variables may be exploited by machine learning to increase accuracy. If machine learning may enhance cardiovascular risk prediction, we investigated this possibility.

2. LITERATURE SURVEY

A study by ChalaBeyene et al [3] suggests using data mining techniques to predict and analyse the development of heart disease. The major goal is to forecast the emergence of cardiovascular illness in order to quickly and automatically diagnose the condition. A healthcare organisation with specialists who lack current knowledge and expertise can benefit greatly from the suggested technique. A variety of medical features, including blood sugar and heart rate, age, and sex, are used to determine whether or not a person has heart disease. WEKA software is used to compute analyses of the dataset.

To predict cardiac disease, Senthilkumar Mohan et al [4]. devised a mixed machine learning approach. There is a Cleveland data set utilised in this research project. Data pre-processing is the initial stage. The tuples with missing values are eliminated from the data set. The authors also do not utilise the data set's attributes of age and sex since they believe that these are personal information that has no bearing on the prediction process. All of the following 11 qualities are crucial since they include critical clinical information. RF and Linear methods have been combined in their own Hybrid Random Forest Linear Method (HRFLM) (LM). The inventors of the HRFLM algorithm employed four different algorithms. The input dataset is partitioned in the first algorithm. Each sample of the dataset is run through a decision tree to arrive at the final result. After determining the feature space, the dataset is broken down into its leaf nodes. The first algorithm's output is a partition of the data set. Next, the data is analysed using a second algorithm, which

generates a classification based on those principles. The Less Error Classifier is used to extract algorithm characteristics in the third step. This approach focuses on determining the classifier's minimal and maximum error rates. Classified characteristics are the output of this algorithm. The Classifier, a hybrid technique based on the error rate of the Extracted Features, is used in the fourth algorithm. Their next step was to assess whether or not HRFLM outperformed other classification algorithms, such as decision trees and support vector machines, on the data collected after using it. As a consequence of the superior performance of RF and LM, the two algorithms are combined to produce the new unique algorithm HRFLM. Combining several machine learning methods may lead to significant improvements in accuracy, according to the authors.

Two linear SVM-based models are proposed by Ali, Liaqat, and colleagues[5] (SVM). First, L1 regularised, then L2 regularised. In the first model, the coefficients of superfluous characteristics are zeroed out. The second model is utilised for forecasting purposes. This section is where illness diagnosis is performed. Both models were optimised using a hybrid grid search technique. This approach uses parameters like as accuracy, sensitivity, septicity, the Matthews correlation coefficient, ROC chart, and the area under the curve to improve two models. " They drew on the Cleveland data set for their study. Holdout validation was utilised to divide the data into 70 percent training and 30 percent testing. To test the hypotheses, two experiments are run, each with a different set of values for the hyperparameters (C1, C2, and k) and subsets of features (k) for each model (L1, L2, and K). L1-linear SVM model layered with L2-linear SVM model has a maximum testing accuracy of 91.11 percent and training accuracy of 84.05 percent. Using an RBF kernel, we combined an L1-linear SVM model with an L2-linear SVM model in our second experiment. A maximum testing accuracy of 92.22 percent and a training accuracy of 85.02 percent may be achieved by using this method In comparison to standard SVM models, they have improved accuracy by 3.3 percentage points.

Using Data Mining classification algorithms, this research attempts to estimate the probability that a person would develop heart disease. Decision Tree and KNN algorithms are employed in this system to forecast cardiac illness in an existing system [6].

As a result of this research, 81.8 percent of heart disease estimates and all potential measurements of heart activity were predicted by the kernel and were equal to the kernel's measurements [7].

Heart illness may be predicted using a variety of parametric parameters, including both continuous and non-constant variations in Heart Rate. Neural networks may be used to predict heart disease based on Blood Stress and Sugar levels. Thirteen different pieces of information, such as blood pressure, menstruation, and age, are used to test the neural system [8].

3. PROPOSED SYSTEM

The flow of the proposed system is as shown in the below figure, which includes data collection, pre-processing – where the data cleaning and missing data handling is performed, train the machine learning algorithm and perform the predication once the model is fully trained.

Figure 1: Proposed system.

3.1 Employed dataset

We utilised data from the UCI Machine Learning repository for this investigation. 300 examples of actual data with 14 different features (13 predictors; one class) including blood pressure, kind of chest discomfort, ECG result, etc. are included. The information of the dataset is as shown in the below figure,

Attribute	Type	Description
Age	Continuous	Age of the patient in days
Gender	Discrete	1: women, 2: men
Height (cm)	Continuous	Height of the patient in cm
Weight (kg)	Continuous	Weight of the patient in kg
Ap_hi	Continuous	Systolic blood pressure
Ap_lo	Continuous	Diastolic blood pressure
Cholesterol	Discrete	1: normal, 2: above normal, 3: well above normal
Gluc	Discrete	1: normal, 2: above normal, 3: well above normal
Smoke	Discrete	whether patient smokes or not
Alco	Discrete	Alcohol intake-Binary feature
Active	Discrete	Physical activity-Binary feature
Cardio	Discrete	Presence or absence of cardiovascular disease

Figure 2: Details of the dataset. [9]

3.2 Data cleaning and handling missed data

Real-world data comprises enormous numbers of errors and omissions, as well as a great deal of erratic data. Such difficulties may be solved by pre-processing this data and making forecasts robustly.

3.3 Cleaning of the data

In most cases, the data that has been obtained contains noise and missing numbers. These data must be cleansed of noise and missing values filled in in order to get an accurate and effective outcome.

3.4 Transformation – For easy interpretation of data by the user there is need of the technique to convert from one format to readable format. Smoothing, standardisation, and aggregation are some of the activities are involved.

3.5 Splitting of dataset

For the training and testing purpose the dataset need to be divided to make the machine to understand and it allows the

user to evaluate the model. For training and testing the dataset is divided into 70:30 ratio.

3.6 Employed classifiers

3.6.1 Classification by Naive Bayes

A supervised method, the naive Bayes classifier is used to classify data. The Bayes theorem is used to classify the data. It is predicated on a strong (Naive) assumption of attribute independence. To determine the likelihood, one might use the Bayes theorem, which is a mathematical notion. The predictors are unrelated to one another and do not show any kind of link. In order to optimise it, each and every characteristic must be taken into account separately. The naïve bayes is represented mathematically as shown below,

$P(X/Y) = P(Y/X) \times P(X)\ P(Y)$ [10]

- Where, P(y/x)-> probability of the classifier on the predicated output
- P(x) -> Prior knowledge on the dataset
- P(y)-> probability of the prior classifier's output.

Decision tree

It is possible to use a decision tree method to classify both category and numerical data using this technique. Create tree-like structures by using decision trees. To deal with medical datasets, decision trees are a common and straightforward tool. Tree-shaped graphs are simple to construct and analyse. Three nodes make up the decision tree model, which is used to analyse data.

- Root node: This node is the mother node based on this node rest of the model's node is dependent.
- Interior node: The features of the dataset is taken care by this node.
- Leaf node: The result of the every node is represented by the leaf node.

3.6.2 K-Nearest Neighbourhood

It is one of the algorithm which works on the supervised machine learning algorithm. The classification is performed by the model by considering the number of nearest neighbours to the test value. The distance between the features of the dataset is calculated by using the Euclidian distance [3]. The know features are employed to label the unknown feature values. Based on the similarity between the neighbouring nodes the clusters are performed. The choose of the value K should be very much optimistic which decides the accuracy percentage.

3.6.3 Random forest

Using a random forest method, a supervised classification approach, is possible. Multiple trees form a forest in this method. There are many trees in a random forest, and each one produces a prediction for a certain class. The accuracy of the model is directly proportional to the number of nodes in the trees.

4.CONCLUSION

This research reveals a great deal about the use of machine learning approaches to classify cardiac disorders. To accurately forecast the best course of therapy for each patient, classifiers play a critical role in the healthcare business. It is our objective to provide accurate and efficient predictions using fewer features and tests. Only 14 of the most important characteristics were taken into account. K-nearest neighbour, Naive Bayes, decision tree, and random forest were all used in the classification of my data. Before being fed into the model, the data has to be pre-processed. According to this model, the best algorithms are K-nearest neighbour, Nave Bayes, and random forest After constructing four methods, later discovered that K-nearest neighbours (k=7) had the greatest accuracy.

REFERENCE

1. Seckeler MD, Hoke TR. The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease. Clin Epidemiol. 2011;3:67.
2. Gaziano TA, Bitton A, Anand S, Abrahams-Gessel S, Murphy A. Growing epidemic of coronary heart disease in low-and middleincome countries. Curr Probl Cardiol. 2010;35(2):72–115
3. Mr. ChalaBeyene, Prof. Pooja Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Technique", International Journal of Pure and Applied Mathematics, 2018.
4. Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava, "Effective heart disease prediction using hybrid machine learning techniques" IEEE Access 7 (2019): 81542-81554.
5. Ali, Liaqat, et al, "An optimized stacked support vector machines based expert system for the effective prediction of heart failure" IEEE Access 7 (2019): 54007-54014.
6. N. Prabakaran, R. Kannadasan, Prediction of Cardiac Disease Based on Patients Symptoms, in: 2018 International Conference on Inventive Communication and Computational Techniques (ICICCT 2018). IEEE, 2018.
7. Shahab Tayed, Matin Pirouz, Johann Sun, Kaylee Hall, Andrew Chang, Jessica Li, Connor Song, Apoorva Chauhan, Medical Ferra, Therasa Sugar, Justin Zhan, Shahram Latifi, Towards Predicting Medical Condition Using K-Nearest Neighbor, in: 2017 International Conference on big Data. IEEE, 2017.
8. Ahmad Ashari, A. Iman Paryudi, Min Tjoa, Performance Comparison between Naive Bayes, Decision Tree, K-Nearest Neighbor in searching alternative Design in an Energy Simulation Tool,

2013 International Journal of Advanced Computer Science and application. IEEE, 2013.

9. Padmanabhan, Meghana, et al. "Physician-friendly machine learning: A case study with cardiovascular disease risk prediction." Journal of clinical medicine 8.7 (2019): 1050.
10. Kaviani, Pouria, and Sunita Dhotre. "Short survey on naive bayes algorithm." *International Journal of Advance Engineering and Research Development* 4.11 (2017): 607-611.