International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 04 | Apr 2021 www.irjet.net p-ISSN: 2395-0072

Heart Disease Diagnosis System Using Machine Learning

Kalpana G[1], Snehitha G[2]

[1]*Assistant Professor (SS), Dept. of Computer Science and Engineering,Rajalakshmi Institute of Technology, Tamil Nadu, India.*
[2]*UG Student, Dept. of Computer Science and Engineering,Rajalakshmi Institute of Technology, Tamil Nadu, India.*

Abstract - *Cardiovascular Diseases (CVDs) could be the main cause of death globally. Nearly 17.9 million people died from CVDs in 2016, which accounted for 31% of all the deaths in that year. Out of the 17.9 million, 85%, i.e., 15.2 million died of heart attack and stroke[12]. The quantum of patients with heart ailments are on a rise because of the usage of tobacco, lifestyle, work environment and lack of physical activity across all spectrums of age. The high death rates are due to these conditions left undiagnosed till a critical stage. The traditional ways of establishing a heart ailment includes analyzing the results of multiple tests, inferring the collective effect of each of these reports and arriving at a conclusion. Since , the above process has multiple human interventions they are not error free. The paper critically examined applications using various Machine Learning algorithms to arrive at the best solution to detect heart ailments at an early stage with high precision. A solution with Random Forest, a supervised learning classification algorithm used to detect heart conditions using a result set of various other medical tests keyed in through a user-friendly interface has been found to be the most accessible and efficient system.*

Key Words: **Data Mining, Random Forest Algorithm, Supervised learning, Classification.**

1.INTRODUCTION

Heart diseases are critical conditions as a result of blockages that prevent blood from flowing to the heart. One of the key reasons for such blockades is the accumulation of fatty deposits on the inner walls of the blood vessels. Thus if the blockades are diagnosed in the right time and treatment is provided the fatality rate could be decreased proportionally. Usually a medical dataset consists of thousands of records and a number of attributes to be considered to diagnose or predict a particular disease. To achieve this data mining is used. Data Mining is a process of reading data, finding arithmetic relationships between the attributes and trying to predict values for an attribute with the knowledge developed from the data. In Machine Learning, there are many classification algorithms used to predict diseases. Also, the same, can be used to develop a treatment plan. The factors that are taken into account to diagnose the disease are age, gender, type of the chest pain, magnitude of the blood pressure, electrocardiographic result, max heart rate, exercise induced angina etc., The risk of stroke is higher in men at younger age than women. Usually the normal resting blood pressure is 120/80 mm Hg. The resting blood pressure higher or lower by 20/10 has higher risk of a stroke. Since such a huge number of features have to be considered to draw complex inferences, the margin of human error is often evident. Further, which large number of attributes to be taken into account to diagnose a disease and decide on the treatment to be given, the complexity and cost increases. Therefore, the paper critically analyses various systems which can be used to read data of the patient and diagnose whether the corresponding patient has disease using a data mining algorithm with minimal error. The central idea of the paper is to identify a system that would be easily accessible by underdeveloped or developing countries, with fewer multi-specialty hospitals and experienced doctors, to identify heart ailments at an early stage, low cost and high accuracy.

2.RELATED WORK

There have been multiple research papers carried out to diagnose heart diseases and strokes. Various Machine Learning algorithms have been put to use to detect possible cases with heart diseases. Naïve Bayes is one of the classification techniques, used in order to detect several diseases such as breast cancer, heart stroke, etc. In the context of heart disease prediction Naïve Bayes algorithm, provides an accuracy of 86%. In other studies ANN has also been used and the same yields a comparatively higher accuracy than the other algorithms used. Deep learning techniques are being keenly explored to be applied to these set of problems. In these studies, a wide range of features are taken into account, hence data cleansing and preprocessing becomes a critical part of the prediction activity. At these stages, the input feature data is cleaned, missing values treated/ removed and redundant features removed, standardizing the data.

3.SYSTEM METHODOLOGY

3.1 Principles and Feature Selection

Among many techniques used to diagnose Random Forest algorithm is selected due to its robustness among many classification algorithms that come under supervised

learning. The Random Forest is a technique where several decision trees are built, and input is given to all the decision trees and output is checked. The output which is predicted by most of the decision trees is taken as the overall output, thus the result being prominent and accurate.

3.2 Construction of Model

Now That the dataset used as input to the system has already been cleaned and ready to be worked on, the data must be split into training and testing datasets. A Random forest model is created to learn from the training dataset. While tuning the hyper parameter for the algorithm, the accuracy score for several values (0-2000) is compared and the value with highest accuracy is used for the diagnosis.

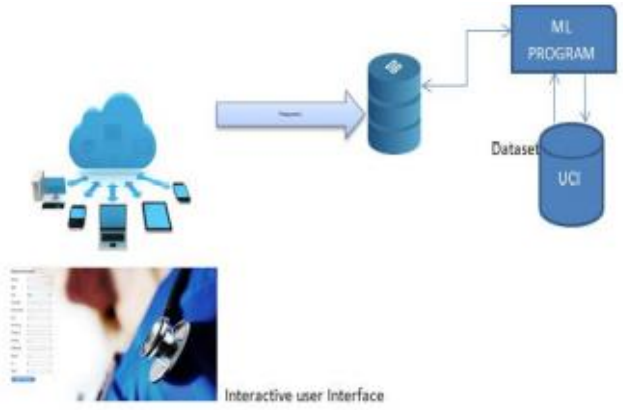

Fig -1: Architecture Diagram

Table -1: Dataset Attributes

S.No	Attribute Name	Description
1	Age	Age in years
2	Sex	Male=1,Female=0
3	CP	Chest pain Type
4	RBP	Resting Blood pressure upon Hospital admission
5	Cholesterol	Serum Cholesterol in mg/dl
6	Fasting Blood sugar	Fbs>120 mg/dl True=1 and False =0
7	Resting ECG	Resting Electrocardiographic Results
8	Thalach	Maximum Heart rate
9	Induced Angina	Whether the patient experience angina as a result of exercise(yes=1,no=0)
10	Old peak	ST depression induced by exercise relative to rest
11	Slop	Slope of the peak exercise ST segment
12	Thal	3=Normal,6=Fixed defect, 7=reversible defect
13	CA	Number of major vessels colored by fluoroscopy(0-3)

Table -2: Values of the features in Dataset

S.No.	Attributes	Heart disease patients	Non-heart disease patients
1	Chest Pain Type	4	1,2,3
2	RBP	134-153	142-154
3	EXANG	Yes	No
4	Oldpeaks	2.06-6.2	<2.06
5	Thalach	71-136	136-168
6	CA	1,2,3	0

Table -3: Comparison of accuracy of different Algorithms on the same dataset.

Algorithm	Training Set Score	Test Set Score	Difference
Random Forest	100.0	98.2	1.8
AdaBoost Classifier	75.21	78.69	3.48
Logistic Regression	84.71	85.25	0.54
Ridge Classifier	83.47	83.61	0.14
Linear SVC	84.30	83.61	0.69
Naïve Bayes	83.47	85.25	1.78

4.EXPERIMENTAL RESULTS

4.1.Accuracy Performance of the proposed model

The accuracy of the system is estimated by splitting the dataset into two parts namely training and testing datasets. Once the model is trained with the help of training dataset, it is verified using training dataset. The outcomes of the model on testing dataset and the actual outcomes in the testing datasets are compared and accuracy score is generated. This model when trained with the Random Forest algorithm provided 86% accuracy.

4.2 Robustness Performance of Proposed Model

The performance of the system is evaluated by checking the accuracy score against the testing dataset. The dataset has almost all the possible sets of data a patient could have. The robustness is achieved by dynamic hyperparameter tuning based on different datasets. Several values for the hyperparameter to be given for Random Forest Algorithm are checked by comparing the accuracy score and selecting the value with high accuracy. Such dynamism of the system makes it user friendly and understandable.

5. CONCLUSION AND FUTURE WORK

In the validated Heart disease diagnosis system, data mining techniques, machine learning algorithm and a user-friendly interface are seamlessly integrated. Random Forest Algorithm has been used because of the various factors to be taken into account to perform diagnosis with high accuracy. Other systems with different algorithms produce less accuracy and lack of interface makes it difficult to be used by local practitioners with less resources. Holistically the chosen model can be improved by adding time to time data as the data will always be scaled day by day. This improves the accuracy of the system from time to time with the change in lifestyle, age, BP, etc,.

REFERENCES

[1] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019.

[2] Deepikaverma, Nidhi Mishra, "Comparative analysis of breast cancer and hypothyroid dataset using data mining classification techniques", pp. 1624-1626, 2017.

[3] K.Aravinthan, "Heart attack prediction using data mining techniques", International Journal of Pure and Applied Mathematics, Volume 119 No. 12, 2018, 16119-16123.

[4] Purushottam, Kanak Saxena,Richa sharma,"Efficient heart disease prediction system",Procedia computer science 85, 2016, 962-969.

[5] Shamsher Bahadur Patel , Pramod Kumar Yadav , Dr. D. P.Shukla," Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques", IOSR journal of agriculture and veterinary science, VOLUME 4, Issue 2, PP 61-64,2013.

[6] Kiran Jyoti, Nidhi Bhatla,"An Analysis of Heart Disease Prediction using Different Data Mining Techniques," International journal of Engineering Research and Technology, vol. 1,Issue 8,ISSN. 2278-0181, 2012.

[7] A. H. Chen, S. Y. Huang, P. S. Hong, C. H. Cheng and E. J. Lin, "HDPS: Heart disease prediction system," 2011 Computing in Cardiology, Hangzhou, 2011, pp. 557-560.

[8] Asha Rajkumar, G. Sophia Reena, "Diagnosis of Heart Disease Using Datamining Algorithm" Global Journal of Computer Science and Technology, Page 38 Vol. 10 Issue 10 Ver. 1.0 September, 2010.

[9] Sunita Soni , Jyothi Pillai, O.P.Vyas," An Associative Classifier Using Weighted Association Rule" , IEEE proceedings of the World Congress on Nature and Biologically Inspired Computing (NaBIC'09), December 09- 11, 2009, 1492-1496.

[10] Carloz Ordonez," Association Rule Discovery with Train and Test approach for heart disease prediction", IEEE Transactions on Information Technology in Biomedicine, Volume 10, No. 2, April 2006.pp 334-343.

[11] Ankur Gupta, Rahul Kumar, Harkirat Singh Arora, Balasubramanian Raman, "MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis", Access IEEE, vol. 8, pp. 14659-14674, 2020

[12] CVD Details from WHO report. https://www.who.int/en/news-room/factsheets/detail/cardiovascular-diseases-(cvds).