Data Science And Its Application In Heart Disease Prediction

Mohammed Jawwad Ali Junaid
Research Scholar
Bhagwant University
Ajmer, India
jawwad.junaid@gmail.com

Dr. Rajeev Kumar
Associate Professor
Bhagwant University
Ajmer, India
drkiit1985@gmail.com

Abstract*:-This paper is the attempt to make the heart disease prediction at very early stage. As it is well known fact that most of the death causing disease all over the world is heart disease then comes cancer which is also very chronic and dangerous disease which has haunted the human being all over the globe. This disease and problem do not occur all of a sudden. Scientist and Doctors reveals that it is a continues process and is the result of being on a particular lifestyle for long time and also results after giving some basic and common symptoms being occurring all of a sudden. Eventually what does happen in the heart attacks is, the heart is not able to pump the required amount of blood to the parts of the body and more over it itself also does not get enough blood supply due to blocked arteries in the heart chambers there, for it results in heart failure and deaths. This paper brings the concepts of data science and its algorithms to make a hybrid model which can predict the disease of heart in the patient in comfortable ample of time advance. Moreover, the system must suggest useful and precautionary steps to the patient which are of globally accepted standards well in advance. The hybrid model is to predict and suggest the heart patient with world class heart solutions made with the help of data science algorithms namely Naïve Bayes, ANN, SVM and Hybrid Naïve Bayes, SVM, and ANN. The Accuracy, specificity, and sensitivity of Naïve Bayes, ANN, SVM and Hybrid Naïve Bayes, SVM, and ANN has been measured. The results nearly 2% increase in the accuracy in predicting the heart disease in hybrid model. Hybrid model also shows higher specificity and sensitivity by having 82.11% and 91.47 % respectively. This paper also considers different attributes and has shown positivity in the connections with the heart disease like genetics, physical activity, total fat consumption, stress and working conditions. This research shows new direction to the research area where these concepts can be applied to the smart devices, data science to revolutionize the heart disease diagnosis and cures. This project can be effective to the make awareness of the complexities of the heart disease.

Keywords— *Coronary artery disease (CAD), heart disease, Naïve Bayes, Artificial neural network (ANN), Support vector machine (SVM), Data science algorithms, accuracy, specificity and sensitivity*

I. INTRODUCTION

The heart attacks are often being common and heart disease became the deadliest human diseases around the globe. This disease and problem do not occur all of a sudden. Scientist and Doctors reveals that it is a continues process and is the result of being on a particular lifestyle for long time and also results after giving some basic and common symptoms being occurring all of a sudden. Eventually what does happen in the heart attacks is, the heart is unable to pump the required amount of blood to the parts of the body and more over it itself also does not get enough blood supply due to blocked arteries in the heart chambers there for it results in heart failure and deaths. The rate of heart disease is very high in United States when compare to other countries unfortunately India is also having comparably having very high rate of deaths due to the heart diseases [S. Shylaja, & Muralidharan S. (2019)]. Many symptoms of heart diseases are included in this paper that could become the features that could be used to find the accurate diagnosis of the patient. The accurate and timely diagnosis of the heart diseases is very much necessary to improve the security of the heart and life. The Europeans Society of Cardiology has been reporting that every year 26 million adults worldwide are diagnosed with heart disease. It also reported that 50% of the heart disease patients die with in the 1-2 years of diagnosis.

PROBLEM STATEMENT

The invasive based techniques are normally performed when patient come with certain symptoms which normally are the basic symptoms where normal person also having little knowledge can understand that patient is suffering from heart disease or stroke right at that time. Moreover, the techniques mostly are very expensive and computationally complex and takes time in assessments. Meanwhile in the research we found that we do not have system which must analyze certain features and symptoms related to the patients, living style and parental history that could become precautionary to the patients. In advance we would like to make awareness to the patients to be care full and take necessary preventive steps to avoid such complex disease to enter the body and flourish. [S. Shylaja, & Muralidharan S. (2019).

CONTRIBUTION

There is a proposal and design of the hybrid algorithms using the different classifiers of machine learning. Some different features other than normally used to find the heart disease are to be examined to their existence and role heart disease formation. In order to prevent normal man to go with complex invasive based diagnosis of the heart disease, a non-invasive medical decision support system based on

978-1-7281-4097-1/20/$31.00 ©2020 IEEE

machine learning predictive models (data science classifiers) is to be designed. Different classifiers namely SVM, KNN, ANN, DT, LR, AB, NB, and FL are to be understood and a hybrid model is proposed so that it could reach peak accuracy and satisfactory [TY Jourti 2018].
Normal attributes and features used predictive diagnosis of heart diseases are Age, Sex, type of chest pain, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, number of major vessels colored by fluoroscopy, Thallium scan. We add some other features like parental history, lifestyle, sleeping hours, mental stress test results, history of depression, physical activity, food habits etc.

Key words: SVM, KNN, ANN, DT, LR, AB, NB, and FL.. [2]

II. EXISTING RESEARCH

Diagnosis is a critical part of the patient care cycle as it determines the nature of the treatment to be provided. Diagnosis is the first step to examine any disease correctly and further can provide nature of the treatment to be provided. Even in this era of peak technology, era of 21st century diagnosis is far from perfect.

The errors caused in clinical alone in USA caused 40,000 - 80,000 deaths in the year 2018. Data science, analysts with the deep learning techniques can be more precise diagnosis. With certain features data analysis can allow detect early signs and can advise preventive care well in advance. Data Made Simple. (2019) and [Robby Gupta 2019].
Machine learning can play vital role to predict heart diseases and some disorders in heart diseases. Here this project with mentioned algorithms K Neighbors classifiers, SVM, DTC, RFC (Random forest classifier) has worked on predicting potential heart diseases in the people. Finally, this project of heart diseases prediction with different classifiers and models were examined. The percentage of accuracy are written after the name of the models K Neighbors Classifier: 87%. Support Vector Classifier: 83%. Decision Tree Classifier: 79%. Random Forest Classifier: 84%

Declining cardiac functions is the measure of the heart disease. Doctors measure the function of the heart by different parameters and ejection fraction of LV. Accurate measure of heart function is done by MRI and reading MRI images is slow process and requires skilled Doctors. Data science can make measurement process more efficient and can enhance doctors to diagnose heart conditions predicatively. Tencia Lee and Qi Liu, along with team of data scientist have worked unanimously and tested different algorithms and finally developed an algorithm to read the heart scans (datasciencebowl. 2020). Data science and Big data tools has been shown positively in prediction, prevention and can come up with best treatment plans. Mobile devices, smart devices, sensors along with data science can become heart attack predictors. Data science and cloud computing has reduced the distance between the hospitals and global experts of the heart diseases. These technologies can be used to develop data driven health care system where the forefront medical field angels (nurses) can assist to give first aid till the expert doctors are available. Data science can revolutionize the heart treatment and can make awareness of being on wrong living style that could lead to heart diseases. (Alexander, Cheryl & Wang, Lidong. (2017).)

PRELIMINARIES

There have been worked done related to the heart disease prediction using data science classifiers. SVM Support vector Machine and ANN Artificial Neural Network on the heart disease data set collected from the secondary source. The intentions of the research work done to get accurate prediction. There exists research on the prediction of the heart diseases with hybrid data science classifiers to get better results. The prediction seems to be accurate and has shown nearly 87% accuracy in predicting heart diseases.

CONSIDERATIONS

Several researchers have developed prediction model using their own individual techniques and few of them also attempted by combining multiple techniques by making hybrid models in order to reach the accuracy. This research paper is to consider the work done by combining two or more techniques to make hybrid model in order to predict the heart disease. The problem which is found is the attributes which researchers has used been using to make a relevant data in prediction is incomplete. Therefore, this paper is to propose two things, to predict the data as well as come up with different suggestion which are related to living style.
The increased attributes shall be 1. Genetic 2. History of heart disease in ancestors. 3. Body mass Index. 4 physical activity or footsteps counted by step counter smart devices in daily life for longer period 7 days. (N. K. S. Banu and S. Swamy 2016).

III. PROPOSED ARCHITECTURE OR FRAMEWORK

The present system which is to be developed with the aim to find the heart disease in advance of its actual chronic occurrence where the people are endangered with their life. This research is to aim to predict the heart disease with the help of the data science classifier Naive Bayes NB, Support Vector Machine SVM and other useful algorithms are to be used to develop predictive mode for heart disease. The model is to predict the heart disease and follow up by useful suggestions regarding the healthy heart.

CONSIDERATIONS

The problem of heart diseases is that it needs first aid at the time of heart attack of stroke. The number of cases encounter throughout the world is “the death due to heart attack occurs because to lack of knowledge and first aid given to them”. As the living style of the human being all round the globe has been changed which is the basic foundation for the cause of different heart complexities, There is evident research done to predict the heart disease well in advance and has shown nearly 90% and above

accuracy in prediction. The problem here we incurred that the prediction alone cannot over all rule out the disease from the body. It needs to be cured by three important basic things. 1. Medicine 2. Precautions 3. Changing living style by suggesting physical activity by considering patients different attributes. Therefore, our model is to predict the level of heart disease and suggest them with medicinal and non- medicinal ways to get rid of the heart disease.

IV. EVALUATION

Quantitative research needs numerical data that can come out either from the numerical data itself or otherwise graphs. Statistical methods are applied on it to get usefulness from the data. The results of statistics will serve as the results for future scope. Qualitative research is in words and in thoughts. There must be expert opinion that can bring useful information through the thoughts and feeling of the examinee. Qualitative research is to understand concepts, thoughts, experiences and feelings of the patients. This research paper uses both quantitative and qualitative data. The different attributes and clinical research which are directly in numerical values are the quantitative data and ultrasound images, e.g. curves and tread mill test curves are qualitative data. The attributes for the heart disease prediction are mentioned below:

Patient identification number, age: age in years, Gender, chest pain location, chest pain type, resting blood pressure, serum cholesterol in mg/dl, smoke parameter:

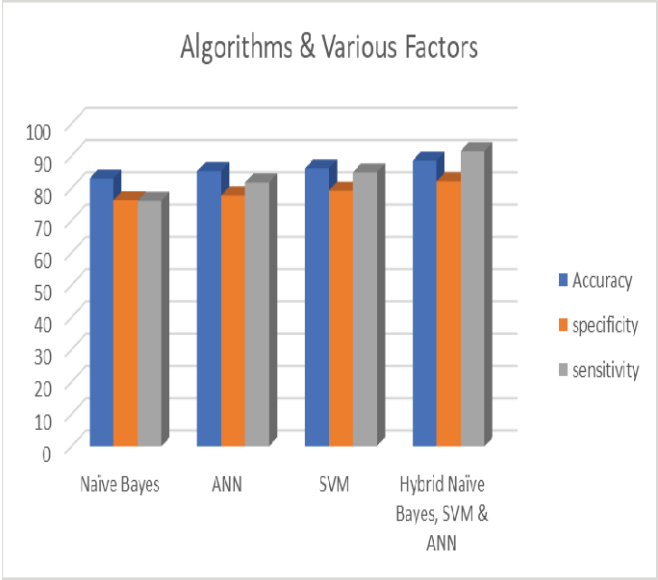
Where 1= Smoker and 0 = Non-Smoker, cigarettes/per day), number of years as a smoker, (fasting blood sugar parameters > 120 mg/dl) (1 = true; 0 = false), If Diabetic (History of diabetes = 1;No History = 0), Coronary artery disease family history (yes=1; No= 0), and also the results of resting electrocardiographic

ST-T wave abnormality Value = 1 (‘T wave inversions’ and/or ‘ST elevation’ or ‘depression’ of > 0.05 mV) ST-T wave abnormality Value = 2: showing probable or definite left ventricular hypertrophy by Estes' criteria exercise ECG, exercise protocol, exercise test duration in minutes, time of ST measure depression, achieving of maximum and resting heart rate, Blood Pressure at peak exercise (first of 2 parts), ST depression induced by exercise relative to rest the slope of the peak exercise ST segment, up sloping, flat, down sloping height at rest, height at peak exercise number of major vessels, colored by fluoroscopy rest radionuclide, ejection fraction rest wall motion abnormality, ejection fraction. (Dua, D. and Graff, C. 2019) In the different researches it is found that study of significant population of both men and women has shown BMI and Waist circumference are strong associated with the risk of coronary heart disease. (Flint, Alan J et al.) Threshold limit of WC is 84 cm and 71 cm for men and women must be used along with the increased attributes to predict the heart disease in early stages.

SPECIFIC RESULTS WITH GRAPH, TABLE etc.

Algorithms	Accuracy	specificity	sensitivity
Naïve Bayes	82.97	76.27	76.10
ANN	85.30	77.73	81.75
SVM	86.12	79.21	84.87
Hybrid Naïve Bayes, SVM & ANN	88.54	82.11	91.47

Table shows the accuracy, specificity, and sensitivity of Naïve Bayes, ANN, SVM and Hybrid Naïve Bayes, SVM, and ANN.



The above table shows the accuracy of the algorithms in predicting the heart disease.

The results in the table shows that there is nearly 2% increase in the accuracy in predicting the heart disease in hybrid model. Hybrid model shows higher specificity and sensitivity by having 82.11% and 91.47 % respectively. Accuracy, specificity and sensitivity formulas are given in appendix I.

In the different researches it is found that study of significant population of both men and women has shown BMI and Waist circumference are strongly associated with the risk of coronary heart disease. (Flint, Alan J et al.) Threshold limit of WC is 84 cm and 71 cm for men and women must be used along with the increased attributes to predict the heart disease in early stages.

Among different attributes specifically critical attributes are classified by using our hybrid classifiers. Each attribute has its low and high threshold limit. Higher the number of attributes greater than the threshold limit of the risk factor then higher is the risk of having heart disease in near time. According the different attributes and risk factors mentioned above the system provides with the risk factors. Below are the 5 stages of the risked factor heart disease prediction model.

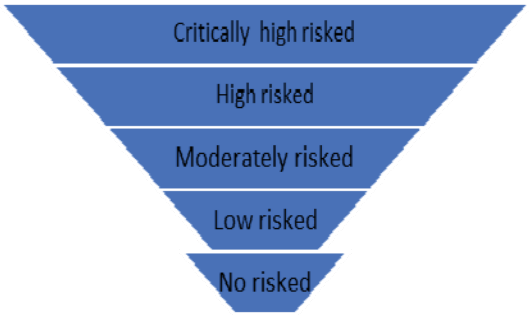


Figure: Shows the flow and order of risk chances of having coronary heart disease by observing the risked attributes.

This research deals with the prediction of the heart disease and calculation of risk factors with classified attributes. The extension of the research is to make such data science application which can be used in highly sophisticated computer aided smart devices which could allow the patient with certain suggestion which are prescribed and are internationally accepted. Here the model consists of 3 phases where the data is entered manually or dumped from data bases.

These data usually comes from personal experience, small clinical results, and Ultrasound images (echo cardiograph). Second stage is to select classify different attributes and calculate and count risk factors of heart diseases.

lastly the system provides with the precautionary suggestions that the patient can be follow up to reduce the risk of having the deadliest disease near him. Below are the precautionary suggestions provided by our model.

V. FUTURE SCOPE

Today we are familiar with the smart devices, portable and wearable devices to track health statistics. How much physical activity is done and number of steps that a person has walked can be counted. Therefore, near future all smart devices with the data science algorithms and suggestions model can revolutionize the health care industry. This model can be helpful to reduce the deaths occurred due to heart diseases complexities. Even the smart devices with the help of internet connectivity can get expert opinion for the patients who are very critically risked.

VI. CONCLUSIONS AND LIMITATIONS

In this research study, a hybrid algorithm-based prediction of heart disease and following with suitable suggestion to the patient is proposed. Among different individual algorithms hybrid models are effective and has given best accuracy, specificity, and sensitivity. This model even has increased attributes namely BMI, waist circumference, family history, and physical activities in routine life. This Hybrid model also shows higher specificity and sensitivity by having 82.11% and 91.47 % respectively. This research shows new direction to the research area where these concepts can be applied to the smart devices, data science to revolutionize the heart disease diagnosis and cures. This project can be effective to make awareness of the complexities of heart disease.

Appendix I

Accuracy Classification: shows the overall performance of the classification system as specified:

$$\text{Accuracy \%} = \frac{TP+TN}{TP+TN+FP+FN} \text{ X } 100$$

Classification Sensitivity: Ratio of the recently classified heart patients to the total number of heart patients. The sensitivity of classifier for detecting positive instances is called “true positive rate.” In other words, we can say that true positive sensitivity fraction confirms that if a diagnostic test is positive and the subject has the disease. It can be written as:

$$\text{Sensitivity \%} = \frac{TP}{TP+FN} \text{ X } 100$$

Specificity: If a diagnostic test is not positive, and the person is healthy and is mathematically written as follows:

$$\text{Specificity \%} = \frac{TN}{TN+FP} \text{ X } 100$$

REFERENCES

[1] S. Shylaja, & Muralidharan S. (2019). 2R. International Journal of Engineering Development and Research (www.ijedr.org), Volume 7, (Issue 3 | ISSN: 2321-9939), p.365.

[2] Ty - Jourti - A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms VL - 2018PY - 2018DA - 2018/12/02DO - 10.1155/2018/3860146 UR - https://doi.org/10.1155/2018/3860146AB

[3] Data Made Simple. (2019). Top 6 use cases of data science in healthcare. [online] Available at: https://bigdata-madesimple.com/top-6-use-cases-of-data-science-in-healthcare/ [Accessed 29 Jan. 2020].

[4] Towardsdatascience.com. 2020. Predicting presence of Heart Diseases using Machine Learning. [ONLINE] Available at:

https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning-36f00f3edb2c. [Accessed 28 January 2020]

[5] Datasciencebowl. 2020. Transforming how me measure heart disease. [ONLINE] Available at: https://datasciencebowl.com/transforming-how-we-diagnose-heart-disease/. [Accessed 2 February 2020]

[6] Rashmee U. Shah, John S. Rumsfeld, Big Data in Cardiology, European Heart Journal, Volume 38, Issue 24, 21 June 2017, Pages 1865–1867, https://doi.org/10.1093/eurheartj/ehx284

[7] Alexander, Cheryl & Wang, Lidong. (2017). Big Data Analytics in Heart Attack Prediction. Journal of Nursing & Care. 06. 10.4172/2167-1168.1000393

[8] N. K. S. Banu and S. Swamy,(2016) "Prediction of heart disease at early stage using data mining and big data analytics: A survey," 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), Mysuru, 2016, pp. 256-261.doi: 10.1109/ICEECCOT.2016.7955226.

[9] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[10] Flint, Alan J et al. “Body mass index, waist circumference, and risk of coronary heart disease: a prospective study among men and women.” Obesity research & clinical practice vol. 4,3 (2010): e171-e181. doi:10.1016/j.orcp.2010.01.001