IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 5, May 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005292 |

A Review on Cardiac Infractions Observed Using Machine Learning

RAMYA S L , DR.HARISH B G

Master of Computer Application, UBDT College of Engineering, Davangere, India

HOD, Master of Computer Application, UBDT College of Engineering, Davangere, India

ABSTRACT: Heart disease is one of the deadliest problems in the world, which cannot be seen with the naked eye and occurs instantly when its limits are reached. Therefore, he needs an accurate diagnosis at a specific time. The healthcare industry produces an enormous amount of patience and disease-related data every day. However, these data are not used effectively by researchers and practitioners. Today's healthcare industry is data-rich but knowledge-poor. There are various data mining and machine learning techniques and tools available to extract effective knowledge from databases and use this knowledge for more accurate diagnosis and decision making. With the increase in research on heart disease prediction systems, it becomes important to summarize the completely incomplete research on this topic. The main objective of this research paper is to summarize recent research with comparative results that have been done on the prediction of heart disease and also to draw analytical conclusions. From the study, Naive Bayes is observed with the genetic algorithm; Decision tree and artificial neural network techniques improve the accuracy of the heart disease prediction system in different scenarios. In this article, commonly used data mining and machine learning techniques and their complexities are summarized.

KEYWORDS: Data Mining, Machine learning, Heart disease, Classification, Naïve Bayes, Artificial Neural Networks, Decision Trees, Associative Rule.

I. INTRODUCTION

Delicate heart disease a range of conditions that affect your heart. Previously, people were unaware of the food they ate which was supposed to be oily, sweet, salty and sour, which in the future may affect our body system by consuming more. Previously, people were unaware of this, but now that they are educated and the healthcare system is also improvised, they tell more about this in earlier phases. The main challenge in healthcare today is the provision of the highest quality services and accurate and efficient diagnostics.

People have routine and busy schedules that lead to stress and anxiety. In addition to this, the percentage of obese, stressed and addicted to cigarettes increases drastically [13]. This is a main contributory factor that leads to heart disease. Cardiovascular disease (CVD) is the leading cause of death worldwide, claiming an estimated 17.9 million lives each year, accounting for 31% of all deaths worldwide [14]. Four out of five CVD deaths are due to heart attacks and strokes, and one-third of these deaths prematurely in people under the age of 70 [14]. Heart failure is a common event enhanced by cardiovascular disease and this article can be used to predict possible heart disease by using various algorithms to find their accuracy and select the one that provides the best results. The number of people with heart disease increases regardless of age in both men and women [13]. Even though heart diseases have proven to be the number one cause of death worldwide in recent years, they are also the ones that can be controlled and controlled effectively. All the precision in the management of the disease lies in the right moment of detection of this disease. The proposed work tries to detect these heart diseases at an early stage to avoid disastrous consequences.

Recorded data of a large number of medical data created by medical experts is available to analyze and extract valuable knowledge. Data mining techniques are the way to extract valuable and hidden information from a large amount of available data. Most of the time, the medical database consists of discrete information. Therefore, decision making using discrete data becomes a complex and challenging task. Machine learning (ML), which is the subfield of data mining, efficiently handles well-formatted large-scale datasets.

Finding coronary heart disease is long-lasting due to many contributing risk factors including diabetes, high blood pressure, excess cholesterol, atypical pulse, and many other factors.

Various statistical and neural network mining strategies have been engaged to discover the severity of coronary heart disease in humans. The severity of the disease is classified mainly according to various techniques such as K-Nearest Neighbor Algorithm (KNN), Decision Trees (DT), Genetic set of rules (GA) and Naive Bayes (NB). The nature of coronary artery disease is complex and therefore the disease must be treated with care. Failure to do so can also affect the heart or cause premature death.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005292 |

II. REVIEW OF LITERATURE

In a study undertaken by Gupta R et al., "Epidemiology and Causality of Coronary Heart Disease and Stroke in India" in 2008, they predicted that India would host more than half of the heart disease cases in the world in over the next 15 years. Estimates of deaths from cardiovascular diseases have been predicted and their prevalence in rural areas ranges from 1.6% to 7.4% and from 1% to 13.2% in urban areas. A study by Rajiv et al., "Prevalence of Coronary Heart Disease and Risk Factors in an Urban Indian Population" in 2002 concluded that risks leading to heart disease have become widespread. Therese Prince. R, et al, conducted a survey including different classification algorithms used to predict heart disease. The classification techniques used were Naive Bayes, KNN (KNearestNeighbor), Decision Tree and Neural Network, and the accuracy of the classifiers was analyzed for the different number of attributes.
The secondary related work is one of the famous articles published in 2018, [2] written by Minnie. Coronary Heart Disease Prediction is a medical device knowledge acquisition project. Early detection of humans susceptible to the disease is key to stopping its progression. This article offers in-depth knowledge of methods to harvest advanced prediction of coronary artery disease. A larger community of Stacked Sparse Autoencoders (SSAE) is developed to harvest knowledge about the green function. The community includes more than a sparse autoencoder and a SoftMax classifier.
Frederic Commandeur et al., 2019 in their study were able to predict heart disease better than traditional clinical assessments. Another example of machine learning used in healthcare is McKinney et al., (2020). They developed an ML algorithm that detects cancerous tumors on mammograms. Likewise, this article focuses on using basic anatomical factors to create a machine learning model that predicts whether an individual is vulnerable to chronic heart disease. Several studies have incorporated the Naive Bayes algorithm and decision trees.

III. PROBLEM DEFINITION

India is overwhelmed by the demographic explosion. With the COVID 19 pandemic, it was evident that our healthcare field needed to be restructured with the best services and smart solutions to deal with the overwhelming influx of patients. Additionally, there is a need to mitigate excessively increased cardiac risks. A cure for heart disease is not a magic pill but a lifestyle improvement. A healthy lifestyle would reduce the risk of heart disease. And thus, this study becomes extremely crucial to achieve the said goal. Researchers are using the heart disease dataset to organize a machine learning model that accurately predicts whether an individual has a chance of being diagnosed with chronic heart disease.

IV. RELATED WORK

A lot of work has been done to predict heart disease using the UCI Machine Learning dataset. Different levels of precision have been achieved using various data mining techniques which are explained below.
Avinash Goland and. Al.; investigate various different ML algorithms that can be used for heart disease classification. The research was conducted to study the Decision Tree, KNN and K-Means algorithms that can be used for classification, and their accuracy was compared[1]. This research concludes that the accuracy obtained by the decision tree was the highest and it was inferred that it can be made efficient by a combination of different techniques and parameter tuning.

V. CLASSIFICATION

The attributes mentioned in Table 1 are provided as input to the different ML algorithms such as Random Forest, Decision Tree, Logistic Regression, and Naive Bayes classification techniques [12]. The input dataset is split into 80% of the training dataset and the remaining 20% into the test dataset. The training dataset is the dataset that is used to train a model. The testing dataset is used to check the performance of the trained model. For each of the algorithms, the performance is computed and analyzed based on different metrics used such as accuracy, precision, recall, and F-measure scores as described further. The different algorithms explored in this paper are listed below.

i. Random Forest

Random Forest algorithms are used for classification as well as regression. It creates a tree for the data and makes predictions based on that. The Random Forest algorithm can be used on large datasets and can produce the same result even when large sets of record values are missing. The generated samples from the decision tree can be saved so that they can be used on other data. In a random forest there are two stages, firstly create a random forest and then make a prediction using a random forest classifier created in the first stage.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005292 |

ii. Decision Tree

The Decision Tree algorithm is in the form of a flowchart where the inner node represents the dataset attributes and the outer branches are the outcome. Decision Tree is chosen because they are fast, reliable, and easy to interpret, and very little data preparation is required. In the Decision Tree, the prediction of class labels originates from the root of the tree. The value of the root attribute is compared to the record's attribute. On the result of the comparison, the corresponding branch is followed to that value, and a jump is made to the next node.

iii. Logistic Regression

Logistic Regression is a classification algorithm mostly used for binary classification problems. In logistic regression instead of fitting a straight line or hyperplane, the logistic regression algorithm uses the logistic function to squeeze the output of a linear equation between 0 and 1. There are 13 independent variables which make logistic regression good for classification.

iv. Naive Bayes

Naïve Bayes algorithm is based on the Bayes rule[]. The independence between the attributes of the dataset is the main assumption and the most important in making a classification. It is easy and fast to predict and holds best when the assumption of independence holds. Bayes' theorem calculates the posterior probability of an event given some prior probability of event B represented by P(A/B)[10] as shown in equation 1 :

$$P(A|B) = (P(B|A)P(A)) / P(B) \quad (1)$$

The results obtained by applying Random Forest, Decision Tree, Naive Bayes, and Logistic Regression are shown in this section. The metrics used to carry out performance analysis of the algorithm are Accuracy score, Precision (P), Recall (R), and F-measure. Precision (mentioned in equation (2)) metric provides the measure of positive analysis that is correct. Recall [mentioned in equation (3)] defines the measure of actual positives that are correct. F-measure [mentioned in equation (4)] tests accuracy.

$$\text{Precision} = (TP) / (TP + FP) \quad (2)$$

$$\text{Recall} = (TP) / (TP + FN) \quad (3)$$

$$\text{F– Measure} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (4)$$

- TP True positive: the patient has the disease and the test is positive.
- FP False positive: the patient does not have the disease but the test is positive.
- TN True negative: the patient does not have the disease and the test is negative.
- FN False negative: the patient has the disease but the test is negative.

In the experiment, the pre-processed dataset is used to carry out the experiments and the above-mentioned algorithms are explored and applied. The above-mentioned performance metrics are obtained using the confusion matrix.

VI. ALGORITHMS

K-Nearest neighbor: It's a classification algorithm. The class of a particular data point is determined based on the class which is most common among its k nearest neighbors where k is a small positive integer.

Support vector machine: It's an algorithm that is used in machine learning for classification and regression techniques. It is regularly used as a classification technique due to its efficiency when compared with the other algorithms. This technique plots a hyperplane for every attribute as a coordinate that is present in the dataset.

Logistic regression: It's a predictive analysis technique that is used when the target variable is dichotomous (binary). The logistic Regression model explains the relationship between one dependent binary variable and one or more independent variables.

Decision Tree Classifier: It organizes the characteristics to inferences about the target value. The classification trees are the tree models in which the target parameter can acquire a finite set of values. In these, the class labels are signified by the leaves, and the branches describe the concurrences of features that guide those class labels. The regression trees are the decision trees in which the target parameter can take the continuous value.

a)**Dataset Description:**

The proposed research will work on the dataset repository, It has 303 anatomical records of patients across 14 variables. Following the variable description:-

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005292 |

1. **Age –** Age description of the patient.
2. **Sex –** Gender of the patient
3. **Chest pain -** 1: typical angina, 2: atypical angina, 3: non-anginal pain, and 4: asymptomatic.
4. **trestbps -** The blood pressure when resting
5. **chol –** Cholesterol level in mg/dl
6. **FBS -** Fasting blood sugar > 120 mg/dl; Yes or No.
7. **restecg -** Electrocardiographic results while resting (values 0,1,2)
8. **thalach -** maximum heart rate achieved
9. **exang –** If pain was induced due to exercise.
10. **old peak -** ST depression induced by exercise relative to rest
11. **slope -** the slope of the peak exercise ST segment
12. **ca -** number of major vessels (0-3) colored by fluoroscopy
13. **Thal-** 3= normal; 6 = fixed defect; 7 = reversible defect
14. **target –** Dependent variable

b)ADVANTAGES

By trying multiple algorithms we've increased accuracy and thus giving more effective heart disease prediction.
Early and online prediction can prove very useful in case of a medical emergency. A free web platform provides a cost-effective diagnosis for patients.

c)DISADVANTAGES

A computerized system alone does not ensure accuracy since the prediction system is not fully automated, we still need the user to enter a wide variety of data for diagnosis, and the warehouse data is not faultless and substantiate.
The model cannot handle immeasurable datasets of patient records and data processing for prediction.

VII. CONCLUSION

With the increasing number of deaths due to heart diseases, it has become mandatory to develop a system to predict heart diseases effectively and accurately. The motivation for the study was to find the most efficient ML algorithm for the detection of heart diseases. This study compares the accuracy score of Decision Tree, Logistic Regression, Random Forest, and Naive Bayes algorithms for predicting heart disease using the UCI machine learning repository dataset. The result of this study indicates that the Random Forest algorithm is the most efficient algorithm with an accuracy score of 90.16% for prediction of heart disease. In future the work can be enhanced by developing a web application based on the Random Forest algorithm as well as using a larger dataset as compared to the one used in this analysis which will help to provide better results and help health professionals in predicting the heart disease effectively and efficient.

REFERENCES

1. Cardiovascular diseases-(cvds) https://www.who.int/en/newsroom/factsheets/detail/cardiovascular-diseases-(cvds)
2. Murray CJ, Lopez AD. Alternative projections of mortality and disability by cause 19902020: Global Burden of Disease Study. Lancet. 1997 May 24;349(9064):1498-504. DOI: 0.1016/S0140-6736(96)07492-2. PMID: 9167458.
3. Gupta R, Joshi P, Mohan V, Reddy KS, UCI, —Heart Disease Data Set.[Online]. Available (Accessed on May 1 2020): https://www.kaggle.com/ronitf/heart-disease-uci. Yusuf S. Epidemiology and causation of coronary heart disease and stroke in India. Heart. 2008 Jan;94(1):16-26. DOI: 10.1136/hrt.2007.132951. PMID: 18083949.
4. T.Nagamani, S.Logeswari, B.Gomathy," Heart Disease Prediction using Data Mining with Mapreduce Algorithm", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-3, January 2019.
   5. H. Alkeshuosh, M. Z. Moghadam, I. Al Mansoori, and M. Abdar, ''Using PSO algorithm for producing best rules in diagnosis of heart disease,'' in International Conference Computer Application (ICCA), Sep. 2017.
5. Hongzu Li, Pierre Boulanger, "A Survey of Heart Anomaly Detection Using Ambulatory Electrocardiogram (ECG)", May 2020.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005292 |

6. Jafar Alzubi, Anand Nayyar, Akshi Kumar. "Machine Learning from Theory to Algorithms: An Overview", Journal of Physics: Conference Series, 2018.
7. C. B. Rjeily, G. Badr, E. Hassani, A. H.,and E. Andres, ―Medical Data Mining for Heart Diseases and the Future of Sequential Mining in Medical Field,‖ in Machine Learning Paradigms, 2019, pp. 71–99.
8. Rajeev Gupta, V P Gupta, MukeshSarna, Smita Bhatnagar, JyotiThanvi, Vibha Sharma, A K Singh, J B Gupta, Vijay Kaul., (2002) Prevalence of coronary heart disease and risk factors in an urban Indian population: Jaipur Heart Watch-2. Available at - https://pubmed.ncbi.nlm.nih.gov/11999090/

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details