# Prediction of Coronary Artery Disease Using Machine Learning

Ajay Gouda Yallappagoudar[1*], C. M. Arun[2], Shashank Gangadhar[3], Vishal Agarwal[4], M. Anitha[5]
[1,2,3,4,5]*Department of Computer Science Engg., Dayananda Sagar College of Engineering, Bangalore, India*

***Abstract*: For years, there has been a lot of onus on implementation of machine learning and its application techniques in the Medical field and is often referred to be a valuable rich information. Coronary Artery Disease (CAD) is one of the major causes of death all around the world and early detection can prevent it. The aim of this study is to predict CAD using historical medical data. The dataset is retrieved from the "Cleveland Dataset" which is openly available and provided by the University of California Irvine (UCI) Machine Learning Repository and it contains 303 instances with a total of 14 attributes. The algorithm is trained with the dataset and using many machine learning methods maximum accuracy will be found out. According to a survey by WHO, doctors can predict the disease with approximate of 67% and we here try to achieve accuracy greater than 90% of accuracy thus saving many lives. Using cutting edge ML technology of text to speech conversion, we improvise the quality of assistance. We also focus on a designing a webpage which receives attributes as input and give them an accurate result.**

***Keywords*: Coronary Artery Disease, Cleveland dataset, Machine Learning.**

## 1. Introduction

Coronary artery disease (CAD) is one of the most common type of heart disease affecting around the world. Present-days heart disease is considered as one of the major causes of human death in the world. 10% of the total death occurs in the world is due to heart disease only. Hence the disease has become one of the biggest concerns in various countries of the world. As per Japan death rate statistics, heart disease occupies the second position. Due to the improvement of technology and the availability of automation, people perform very less physical work and use the mental ability which makes them prone to get heart disease. Due to this people are getting addicted to smoking, alcohol which leads to have big bellies. As per University of Rochester's Medical Centre view the major source for Heart disease are overweight, lack of physical activity, fatness, consumption of malnutrition and tobacco. As heart disease is widely accepted as the major source of death hence medical analysis of heart disease becomes a regular need for every human being. Due to the number of ambiguity and risk factor the prediction of the disease became a very tough task for every physician. If the heart attack can be identified earlier then the life of the patient can be saved through proper medication and also harm to the heart can be saved up to a large extent. Heart diseases are of various types like Coronary artery disease, valvular heart disease, Cardiomyopathy. These diseases mainly affect the arteries of the heart, blood in and out valves of heart, heart muscle squeezing. Proper heart functioning is really a highly essential thing for a healthy life. In Coronary artery disease cholesterol, calcium and some other substance getting deposited in the veins through which blood circulation is done. As a result of which some blockage is being created against the smooth passage of blood. Due to this the heart muscles will not get an adequate amount of oxygen which creates discomfort in the patient's chest and results as chest pain with the patient. As per WHO report up to 2030 around 23.6 million individuals in the world will die due to heart disease. So there is a need to take some preventive steps to minimize the threats of heart disease. Practitioners mainly view the symptoms, expressions and medical test to identify the occurrence of the disease with the patient. For coronary artery disease identification doctors mainly uses SPECT and ECG methods. In SPECT method radioactive tracers are being injected in to blood for generating images of heart which are used by the doctors to know about the identification of coronary artery disease and also the prediction of the heart attack. ECG reports are used to know about abnormality in heart beating. The diagnosis made by the doctors about any disease is not always 100% correct. Hence various computerized tools are used in the healthcare domain. These tools are used to identify critical parameters for the diagnosis of the disease. Improvement in the health condition of any patient can be known by analyzing various critical parameters related to their disease. The main goal of the intelligent systems assisting medical diagnosis is to predict the presence of any disease accurately. A number of input symptoms are used to indicate the occurrence of Coronary artery disease. Out of all age, sex and family medical history cannot be changed. But others symptoms like smoking, blood pressure, cholesterol level, physical exercise can be changed to reduce the possibility of Coronary artery disease. As so many parameters are involved while diagnosing the disease so the practitioner uses the present medical report of the patient. After going through the report doctors adopt the same method which was used for any previous patient having a similar type of test report. Some modern genetic algorithms now days are trying to find out some important data which are mainly contributing towards the occurrence of the heart disease instead of analyzing a number of data. Through this the algorithms trying to identify the disease with an optimal number of parameters and consuming very less time. Recent development in the field of

*Corresponding author: ajayyallappagoudars@gmail.com

ML contributes a lot to medical science towards the development of intelligent systems. In last few decades various computational systems have developed which were helpful for the physicians in improving their diagnosis decisions. In the motivation towards the requirement of an expert system here an efficient heart disease prediction system is proposed.

The goal of the proposed system is –

- Understanding the data findings like finding out outliers, anomalies and data imputation in data set.
- Building an accurate predictive model by tuning hyper parameters in the cost function to avoid over fitting and under fitting.
- Finding the optimal value of features by minimizing the cost function through which the accuracy of the model can be improved.

## 2. Literature Review

Alberto Palacios Pawlovsky et al [1] has used an ensemble based KNN method which employs distance-based heart disease prediction. Here a two-phase method has been introduced. In the first phase different K values are chosen. For each K point we put all the 5 different distance formula like Euclid, Manhattan, Chebyshev, Canberra and Mahala Nobis and obtained distance value from the test instance and noted the class of each k neighbor. We also found the classification accuracy by taking different cross validation size from 10% to 90% of the total records. In second phase of the algorithm, we put majority based voting algorithm to assign the class to an unknown instance as per the majority class is chosen.

Yeshvendra K. Singh et al [2] used Random Forest method for detecting the heart disease. Introduced system uses all the 13 important input features for the prediction of heart disease provided UCI repository. The prediction system is implemented by removing the features between whom no correlation can be established. The enhancement in the accuracy is achieved by tuning various linearly dependent variables of the random forest like randomness, the number of trees, the minimum number of splits and the minimum number of leaf nodes. Initially the number of trees and the minimum number of splits are considered to find a better correlation with accuracy. Highest accuracy obtained when numbers of splits are 20 and the number of trees is 75.

TanmayKasbe et al [3] have employed an expert heart disease prediction system using fuzzy based logic. Here a fuzzy indicator functions like triangular and trapezoid are introduced for the implementation of a fuzzy expert system and fuzzy rule base. The fuzzy expert system first does categorization of independent features and dependent feature. In this stage, the features are experimented to observe its value range and its equivalent class. In the next stage Fuzzy rules over data is employed by the different combination of single or several features with AND, OR operator. Here the system development is done using total 86 fuzzy based rules with all possible arrangements. In the rule base, all the possible input variables with different combinations and its corresponding output value are used for the output level calculation. A relationship is established between all independent features value and their corresponding dependent feature values are set. This relation will be helpful to find out the class for an unknown instance.

Purushottam et al [4] put forward a well-organized heart disease prediction system which uses a large dataset from Cleveland which contains data about coronary diseases. First, the database pre-processing is done using all Possible-MV algorithms. It is used mainly to fill the missing data in the dataset. After that classification decision rules are generated to do the proper classification through pruned rules, original rules, classified rules and rules without duplicates.

Hongmei Yan and Jun Zheng [5], they have put forward a real genetic algorithm related system which selects the essential features for diagnosing heart disease. This system assists the diagnosing of hypertension, chronic pulmonale, coronary artery disease, congenital heart disease and rheumatic valvular heart disease. They considered the dataset which consisted of 352 entities and each entity recorded 40 attributes. Among the 352 entities, 24 major features were identified which helped majorly in diagnosing the heart disease and these features helped in securing high accuracy for diagnosing heart disease.

Ibrahim Turkoglu and Resul Das [6], put forward many methodologies and tools for developing medical decision support system. Many researchers tried helping the physicians by developing intelligent systems which would increase the chance of diagnosing a heart disease. They established a method which used Statically Analysis System software 9.1.3 to treat CAD. In these various neural networks were combined to practice on the matching task which used neural networks ensemble model and led to increase in performance. The results were amazing as the experimental results secured for diagnosing the heart disease.95.91% specificity values, 89% classification accuracy and 80.95% sensitivity.

Se-Hak Chun and Yoon-Joo Park, [7] have put forward a CSCBR - Cost-Sensitive Case-Based Reasoning, extraction technique which includes different misclassification cost into conventional case-based reasoning. A GA - genetic algorithm was introduced to examine the absence of disease. By using number of neighbors and boundary point they tried to reduce the misclassification errors costs into CBR. CSCBR helped to overpower the constant number of nearest neighbors. The best neighbors were selected by modifying maximum cut-off distance point and cut-off classification point. This technique was put forward in 5 various pharmaceutical datasets and then collated with CART and C5.0. It was deduced that the total misclassification cost of this technique was much lesser than the other cost-responsive methods.

Deo RC. Machine learning in medicine. [9] Has put forward 2 major disadvantages in the current diagnosing system that they take less number of factors into consideration in the dataset by each tool and impotency of these systems to deduce major information on the disease. They used a 2-phase strategy to reach this maxim. In the first phase, based on the model of Naïve-Bayes classifier they tried to introduce a common representation procedure and applied to the set of current features. In the second phase, based on various features of Bayes probabilistic judging and conditional probabilities they

introduced a combination programme which was optimized from the genetic algorithm.

K. Rajeswari and V. Vaithiyanathan [8] have addressed about the choosing the features when collecting data in order to reduce the number of inputs under assessment. They introduced a system which introduced the efficiency with respect to accuracy, time and cost. They used an artificial neural network to select important features from the dataset. They considered the IHD database which consisted 712 entities, initially they took 12 features and 17 attributes as input to neural networks. The predicted accuracy was around 82.25 for testing and 89.4% for training.

## 3. Future work

By reading many papers there were many issues that we came across that may be rectified doing research.

In spite of many tools and techniques used in the literature, there are still a need of good approach to solve the following research issues.

Data cleaning is an important problem for CAD database since some of the attribute values one of the major problems in CAD database is data cleaning since some of the attribute values cannot be obtained usually. Hence handling of unknown values for diagnosis problem is a tough assignment.

As a means to get a better accuracy it's important to select most suitable data for classification.

The measurements of the CAD is huge, hence identification of important attributes for better diagnosis of CAD is very tough task.

Selection of most suitable samples of data for classification instead of the whole data is another risk for. As a means to obtain best diagnosis report weighting of each attribute would be good although it’s a complex task.

Choosing a suitable classification technique without much computation complexity is another way and at a same time the effectiveness won’t be affected.

Handling of huge and complex dataset for diagnosis has been an issue because many of the classification algorithms are not fit it. Improvement on effectiveness is important research direction because the heart disease database is very

Improve on accuracy is important because the disease database is very sensitive, we can’t take risk on someone’s life. Considering many class labels for predicting output is also going to affect the performance significantly in the field of medical diagnosis.

## 4. Implementation Details

In Machine Learning Models, there’s no single algorithm is superior to the others. In Machine Learning, it's all about training and testing the models for accuracy if a datasets having known labels, similar to Cleveland dataset, then the learning is called “supervised”, similarly admitting into the “unsupervised” learning where instances are unlabeled. We will like to present on how some of the Machine Learning Algorithms work on the dataset and any conclusion that we will make in each case. In all the Algorithms dataset is first trained considering some of the datasets called as the “training set” and then tested on the “test set” that is considered as “unseen data” for finding the accuracy or whether algorithm is good fit to use.

### *A. Binary logistic regression*

Logistic Regression is extension of SLR Simple Linear Regression. It is just extended version of SLR Simple Linear Regression. Where dependent variable are binary hence we cannot use linear regression since it's used to predict relationship between independent variables and dependent variable. There should be two or more number of independent variables for logistic regression. All independent variables are tested to get their productiveness while considering the effect of prediction in the models. In this regression model. The dependent variable has two levels. If the output having more than two values then they are trained by multinomial logistic regression. This model itself simplifies model probability of output in term of input since it is not a classifier, although it can be used to make a classifier, taking into that consideration by choosing a cut-off range value and grouping inputs with probability greater than the cut-off range as one class, and below the cut-off range as the other this is a common way to make a binary classifier.

### *B. Decision tree*

Decision trees are built by splitting the training set into distinct nodes, where one node contains all of or most of one category of the data. Where each node represents a variable. The classification starts from the root node of the tree and then between their internal nodes, the process is terminated once it reaches the leaf nodes. The path is defined by classifying rules. There are reasons affecting the performance the data discretization method and type of tree used. It involves testing multiple classifiers and using different discretization methods and trees types to identify which combination will give better results.

### *C. Random forest*

It consists of multiple regression trees like a forest. As it is a combination of trees the dataset is would be splitted having slightly different sets. The final result is obtained by taking the average of all the individual tree prediction. It meets the aforementioned methods as it does not consider tuning parameters taking into consideration such that it does not enter overproduction phase. This method ensures the members of the ensemble dynamically taking into consideration combined performance.

## 5. Tools Used in Cad Prediction

### *A. Software Tools*

Anaconda Enterprise, V S Code, M S Excel, Flask, HTML and CSS.

### *B. Languages and Frameworks*

Python, JavaScript.

## 6. Conclusion

The designed system is under development and implementation phase. It is under training to achieve maximum accuracy. Once the proposed approach is fully implemented, it will surely help people to identify Coronary Artery Disease and take prior precautions to prevent it and live a smooth and healthy life.

## References

[1] Alberto Palacios Pawlovsky 'An Ensemble Based on Distances for a kNN Method for Heart Disease Diagnosis' 2018 International Conference on Electronics, Information, and Communication (ICEIC).

[2] Yeshvendra K. Singh, Nikhil Sinha., and Sanjay K. Singh., 'Heart Disease Prediction System Using Random Forest'; First International Conference, ICACDS 2016, pp. 613-623, November 2016.

[3] TanmayKasbe, Ravi Singh Pippal 'Design of Heart Disease Diagnosis System using Fuzzy Logic' 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017).

[4] Purushottam, Kanak Saxena,Richa Sharma, "Efficient Heart Disease Prediction System," Procedia, Computer Science 85 962 – 969, 2016.

[5] Yan, Hongmei & Zheng, Jun & Jiang, Yingtao & Peng, Chenglin & Xiao, Shouzhong. (2008). Selecting critical clinical features for heart diseases diagnosis with a real-coded genetic algorithm. Appl. Soft Comput.. 8. 1105-1111.

[6] Das, Resul & Turkoglu, Ibrahim & Sengur, Abdulkadir. (2008). Diagnosis of valvular heart disease through neural networks ensembles. Computer methods and programs in biomedicine. 93. 185-91.

[7] Park, yoon-joo & Chun, Se-Hak & Kim, Byung Chun. (2011). Cost-sensitive case-based reasoning using a genetic algorithm: Application to medical diagnosis. Artificial intelligence in medicine. 51. 133-45.

[8] Kannan, Rajeswari & Vaithiyanathan, V. & Neelakantan, T.R. (2012). Feature Selection in Ischemic Heart Disease Identification using Feed Forward Neural Networks. Procedia Engineering. 41. 1818-1823.

[9] Deo RC. Machine learning in medicine. Circulation. 2015;132(20):1920–30. pmid: 26572668.