# Early Detection and Prediction of Coronary Artery Diseases

Praveen A[1], Chaithanya Vardhan C M[2], Chandan Y C[3], Darshan V[4]

[1]*Assistant Professor, Department of Electronics and Communication, K.S. Institute of Technology, Bengaluru, India*

[2,3,4]*UG Student, Department of Electronics and Communication, K.S. Institute of Technology, Bengaluru, India*

*Abstract—* **As the millions of Coronary Artery Disease people have been observed worldwide over the past years. Coronary Artery Disease (CAD) prediction is a very hard and challenging task in the medical field. The early prediction in the medical field especially the cardiovascular sector is one of the virtuosi. The prior studies about the construction of the early prediction model developed an understanding of the recent techniques to find the variation in medical imaging. The prevention of cardiovascular can be fulfilled through a diet chart prepared by the concerned physician after early prediction. Our research paper consists of the prediction of CAD by the proposed algorithm by constructing of pooled area curve (PUC) in the machine learning method. Data pre-processing helps to format the data into useful form by removing redundancy and noise, eliminating missing and non-numerical values, and also by normalization. Data analysis and visualization are carried out to improve the statistical analysis of given data. Logistic regression is carried out on the data since it contains lot of columns with categorical values. Accuracy of the model have been measured. Various conclusions can be drawn from this interdependent data set and can be stored as historical data for future analysis. We then try out various other ML algorithms like Random Forest classifier, SVM and KNN algorithm. We then compare the models with Logistic Regression method.**

*Index Terms:* **Machine Learning method, Coronary Artery Disease, Medical Imaging, Data pre-processing, Logistic regression, accuracy score, data analysis and visualization, Random Forest classifier, KNN algorithm**

## I.INTRODUCTION

The main cause of Coronary Artery Disease is due to blockage in arteries. It has many other names such as cardiovascular disease, arterial hypertension and Heart attack. Approximately, there are almost 26 million people around the world affecting with heart disease.

The worry point is, this ratio is expected to increase rapidly in coming years, if precautions are not taken efficiently. Apart from making life style healthy and diet control, the right time diagnosing and comprehensive analysis are other essential factors, which can ultimately save the lives. Therefore, this paper has taken a small step towards saving the lives of Coronary Artery Disease patients and describes a way to improve the performance of diagnosing the patients on the bases of their medical history.

Most of the time patients goes for several tests, which can overburden them with extra physical activities, time, and for sure additional financial charges. As previous studies suggested the common reasons behind heart disease can be unhealthy food, tobacco, excessive sugar, overweight or extra body fat, whereas the common symptoms can be pain in arms and chest. Noticeably, these reasons are independent from each other; proper analysis on this kind of dataset can improve the process of diagnosing and can assist the heart surgeons as well. Previously, different researches used number of techniques to improve the CAD diagnosis process such as Extreme Learning Machine, heart disease classification and machine learning classifiers.

Therefore, this research attempts to improve the performance of the classifiers by doing experiments using multiple machine-learning models to make better use of the dataset collected from different medical databases.

## II. LITERATURE SURVEY

Detailed Analysis of Coronary Artery Disease(CAD): Data Analytics and Machine Learning.
In this paper, they have used data analysis and Machine Learning algorithms to make predictions in seconds. Data analysis and visualization have been carried out for statistical and graphical analysis of the acquired data. Logistic regression has been carried out on the data set (categorical). Accuracy, precision, and f1 score of the model were measured. Drawback: Only logistic regression is given importance and other algorithms are neglected even though they are a better fit to obtain the data set and analyze them.

Comparative Analysis on Linear Regression and Support Vector Regression.
In this paper, they have analyzed linear and support vector regression models in order to use the correct model for prediction based on its requirements. In this paper, the Linear regression and support vector regression model is compared using the training data set. The analytical results concluded that Lease Med Sq is the best model for linear regression. Drawback: Linear regression is not a good fit for this project as a combination of many other algorithms that can give a better prediction. Support vector machine requires support from other classification algorithms to give a more efficient and effective result.

Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning.
This paper proposes a system that acquires data that classifies if patients have heart disease or not according to some parameters. The data which is collected is converted into knowledge by data analysis by using various combinations of algorithms. Nowadays, most of the data is computerized and everything is in the cloud which can be accessed although it is not utilized properly. So, the prediction of heart diseases can be enhanced by using advanced techniques and algorithms in less time complexity Daily Forecasting using Artificial Neural Network.

Heart Disease Prediction using Machine Learning.
The proposed work predicts the chances of Heart Disease and classifies patient's risk levels by implementing different machine learning techniques.

This paper presents a comparative study by analyzing the performance of different machine learning algorithms such as Naive Bayes, Decision Tree, Logistic Regression. Drawback: The work can be enhanced by developing a web application based on the Random Forest algorithm and by using a larger dataset which will give a better result could be considered a server as well. Servers are used to manage network resources. The services or information in the servers are provided through the Internet that are connected through LAN and made available for users via smart phones, web browser or other web browser devices to make the system more intelligent, adaptable and efficient.

Prediction of Cardiovascular Disease Using Machine Learning Algorithms
In the research, data pre-processing uses techniques like the removal of noisy data, removal of missing data, filling default values if applicable, and classification of attributes for prediction and decision making at different levels. This project proposes a prediction model to predict whether people have heart disease or not and to provide awareness or diagnosis on that. Drawback: The performance of this method is not up to the mark. Different ensemble methods of these algorithms can be used which can help to advance to better performance with more parameter settings for these algorithms.

## III. PROPOSED WORK

The objectives of the proposed work are:
- To minimize human effort in detection and prediction of coronary artery disease.
- To build a robust system for analysis.
- To design a machine learning model which includes KNN, Random forest, SVM for categorical classification as disease and no disease.
- To implement coronary artery disease analysis system using libraries such as pandas, numphy, matplot, seaborn, sklearn.

METHODOLOGY
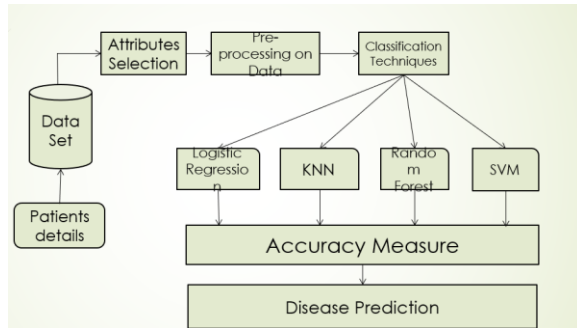
The dataset used in this research is collected from Kaggle platform, the dataset is also known as Heart Disease Dataset. The database file contains the record of 1190 patients. The complete description of each attribute and the number of values for each attribute is shown

Patients details:-
Here the patients details contains the information such as age, sex,height, weight, Blood pressure, cholesterol, sugar level.

Classification of Techniques :
Random Forest Algorithm

- Random Forest algorithm is less prone to overfitting than Decision Tree and other algorithms
- Random Forest algorithm computations may go far more complex compared to other algorithms.
- Bagging is used for making the model more stable and accurate by approaching averaging model technique.

Logistic regression Algorithm

- Logistic regression is easier to implement, interpret, and very efficient to train
- It can only be used to predict discrete functions. Hence, the dependent variable of Logistic Regression is bound to the discrete number set.
- It uses Logit Link function, in which the data values are fitted, for prediction. The mathematical interpretation defines Logit function as the natural log of the odds that Y equals one of the categories.

K-nearest neighbor Algorithm

- The KNN algorithm requires no training before making predictions, new data can be added

seamlessly which will not impact the accuracy of the algorithm.
- In large datasets, the cost of calculating the distance between the new point and each existing point is huge which degrades the performance of the algorithm.
- The algorithm is based on the principle of Euclidean distance that is the instances within a dataset generally exist in close proximity to other instances that have similar properties.

SVM

- Machine Support Vector Machines is also called Support Vector networks it is supervised learning algorithms used for both classification and regression analysis.
- It classifies the data points plotted in a multidimensional space into categories by parallel lines called the hyperplane.
- The classification of data points involves the maximization of margin between the hyperplane. There are different kernels available for mapping of linear or no linear data points in a multidimensional space for separation.

## IV.CONCLUSION

On the successful implementation of the project we have used data analysis and Machine Learning algorithms to make predictions in seconds Also the accuracy of these data can be verified. Data analysis and visualization was carried out for statistical and graphical analysis of the acquired data.
Logistic regression was carried out on the data set (categorical). Accuracy, precision, and f1 score of the model was measured. Other classification Machine Learning algorithms like SVM, Random Forest Classifier and KNN algorithm were applied on the prepared and split data.

## REFERENCES

[1] Interactions between kidney disease and diabetes- dangerous liaisons- Roberto Pecoits-Filho, Hugo Abensur, Carolina C.R. Betônico, Alisson Diego Machado, Erika B. Parente, Márcia Queiroz, João Eduardo Nunes Salles, Silvia Titan and Sergio Vencio- 2016- article 50.

[2] The Python Standard Library — Python 3.7.1rc2 documentation https://docs.python.org/3/library/

[3] Rishabh Magar, Rohan Memane, Suraj Raut , Prof. V. S. Rupnar""HEART DISEASE PREDICTION USING MACHINE LEARNING", International Journal of Emerging Technologies and Innovative Research (www.jetir.org | UGC and issn Approved), ISSN:2349-5162, Vol.7, Issue 6, page no. pp2081-2085, June-2020

[4] Data Warehousing Architecture and Pre-Processing- Vishesh S, Manu Srinath, Akshatha C Kumar, Nandan A.S.- IJARCCE, vol 6, issue 5, May 2017.

[5] Data Mining and Analytics: A Proactive Model - http://www.ijarcce.com/upload/2017/february-17/IJARCCE%20117.pdf

[6] A comparative analysis on linear regression and support vector regression- DOI: 10.1109/GET.2016.7916627- https://ieeexplore.ieee.org /abstract/document/7916627

[7] Ravindhar NV, Anand, Hariharan Shanmugasundaram, Ragavendran, Godfrey Winster"Intelligent Diagnosis of Cardiac Disease Prediction using Machine Learning", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-11, September 2019