DEEP LEARNING ALGORITHM FOR ANALYSIS AND PREDICTION OF CARDIOVASCULAR DISEASE: A CASE STUDY OF CARDIOVASCULAR DISEASE AFTER COVID-19

Project Reference No.: *45S_BE_3063*

College **:** *Bapuji Institute of Engineering and Technology, Davanagere*
Branch **:** *Department of Computer Science Engineering*
Guide(s) **:** *Mr. Arjun H*
Dr. Pradeep N
Student(S) **:** *Ms. Swetha M*
Mr. Srujan H J
Ms. Tejaswini T P
Ms. Akshatha S P

Keywords:

Deep Learning, Machine learning, Cardiovascular disease, Covid-19, Prediction.

Introduction:

The most important part of human body is the heart which pumps blood to every other part. Heart diseases have developed as one of the most indispensable reasons for death all around the globe. If cardiovascular disease is predicted earlier, it becomes easier to find or apply a cure before it gets dangerous. This type of prediction problem, related to medical diagnosis comes under the branch of computer science i.e., bioinformatics.

Post-acute sequelae of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)—the virus that causes coronavirus disease 2019 (COVID-19) can involve the pulmonary and several extrapulmonary organs, including the cardiovascular system. Although COVID-19 is primarily a respiratory or lung disease, the heart can also suffer.

Temporary or lasting damage to heart tissue can be due to several factors like lack of oxygen, inflammation of the heart, Stress cardiomyopathy etc. In some people, heart rates can vary from fast to slow, unrelated to exertion, for no apparent reason. But, Post says, shortness of breath, chest pain or palpitations after having COVID-19 is a common complaint. “Any of these problems could be related to the heart, but they could also be due to other factors, including the aftermath of being very ill, prolonged inactivity and spending weeks convalescing in bed.”

In the project cardiovascular disease is identified using different machine learning algorithms namely decision tree classifier, logistic regression, random forest classifier, gradient boost, support vector machine and deep learning algorithm TabNet. The performance of various algorithms used in the project was compared based on accuracy.

Objectives:

1. To collect a publicly available dataset of healthy and Cardiovascular diseased patients.

2. To clean the dataset by handling missing values and outliers.
3. To understand and analyse the correlation among the dataset features and feature selection for training and testing models.
4. To create a training model using various machine learning algorithm and deep learning approaches.
5. To predict heart diseases of the samples from the best trained models.
6. To derive the inferences from the performance of the algorithms considered.

Methodology:

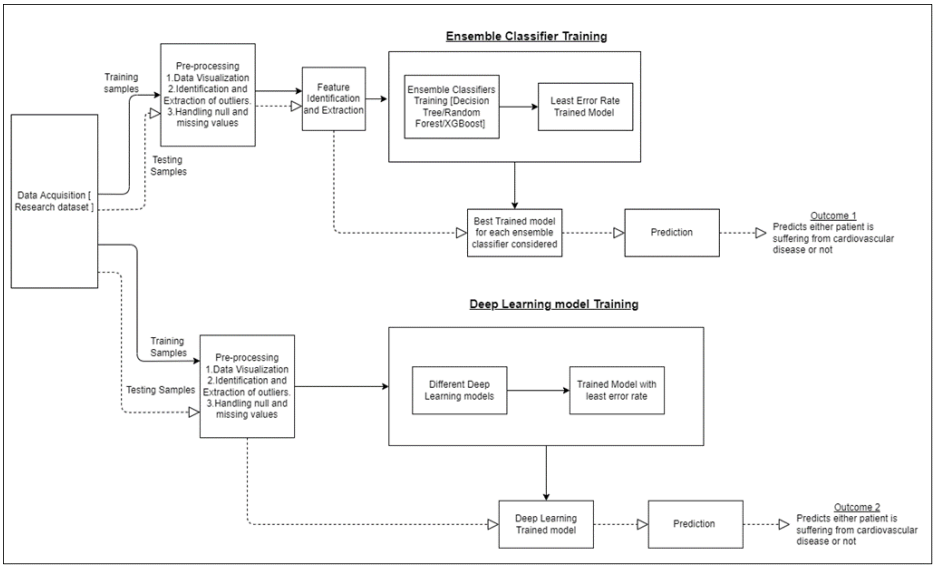

Figure 1: Proposed Methodology for Analysis and Prediction of Cardiovascular Diseases.

1. Dataset Acquisition:

Dataset Acquisition is the process of collecting required data for the proposed project. Here we are collecting cardiovascular disease dataset of patients. There are many publicly available dataset from Kaggle, UCI, Dataworld. Here we considered Kaggle dataset.

2. Pre-processing:

Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

- **Handling null and missing values:** By calculating the mean of a particular column or row which contains any missing value and we will put it on the place of missing value. This strategy is useful for the features which have numeric data such as age. Missing values can also be handled by deleting the rows or columns having null values. If columns have more than half of the rows as null then the entire column can be dropped.
- **Removing of outliers:** Outliers as samples that are exceptionally far from the mainstream of the data. The simplest way to detect an outlier is by graphing the features or the data points. Data visualization is one of the best and easiest ways to have an

inference about the overall data and the outliers. Scatter plots and box plots are the most preferred visualization tools to detect outliers.

- **Feature selection:** Feature selection is primarily focused on removing non-informative or redundant data from the model.

3. Training of the model:

A training model is a dataset that is used to train an ML algorithm. The entire dataset is divides into training data and testing data. It consists of the sample output data and the corresponding sets of input data that have an influence on the output. The training model is used to run the input data through the algorithm to correlate the processed output against the sample output. The result from this correlation is used to modify the model. Training a model simply means learning (determining) good values for all the weights and the bias from labelled examples. We have trained the model using pre-processed training data. Here, we have used Scikit-learn library which contains various libraries for building machine learning models.

4. Testing of the model:

In machine learning, model testing is referred to as the process where the performance of a fully trained model is evaluated on a testing set. The testing set consisting of a set of testing samples should be separated from the both training and validation sets, but it should follow the same probability distribution as the training set. We have tested the model using pre-processed test data and evaluate the model.

5. Prediction:

Return the result to user's display. Prediction of heart diseases of the samples from the best trained models.

Results and Conclusion:

Multiple feature selection algorithms were developed as part of the project to ensure proper selection of high-risk factor features. The significant Bio-markers identified which are of high-risk factor for cardiac diseases are- Age, Systolic BP, BMI, Platelet count, Creatinine, Glucose, Triglycerides, Cholesterol. Using several machine learning techniques, the project developed many classification models for the classification of cardiac disease.

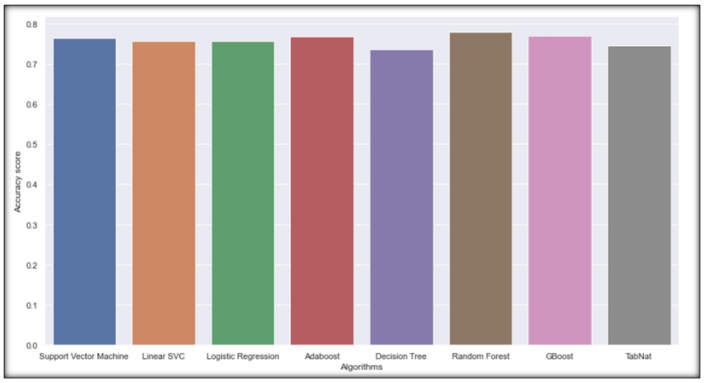

Figure 2: Comparison of Accuracy for Analysis and Prediction of Cardiovascular Diseases.

Table 1: Accuracy Comparison for Analysis and Prediction of Cardiovascular Diseases

MODELS	ACCURACY
DECISION TREE CLASSIFIER	73.6
LOGISTIC REGRESSION	75.5
RANDOM FOREST CLASSIFIER	77.8
ADABOOST	76.7
GRADIENT BOOST	76.8
SUPPORT VECTOR MACHINE	75.5
TABNET	74.4

From the analysis of different performance parameters, the Random Forest Classifier has the highest performance among the other models, and is good for early detection of cardiac diseases.

Scope For Future Work

This analysis and classification can be further extended to predict cardiovascular diseases and various other chronic or fatal diseases like Cancer, Diabetes, etc. The dataset size can be increased and then deep learning with various other optimizations can be used and more promising results can be achieved. Machine learning and various other optimization techniques can also be used so that the evaluation results can again be increased. More different ways of normalizing the data can be used and the results can be compared. And more ways could be found where we could integrate heart-disease-trained ML and DL models with certain multimedia for the ease of patients and doctors.

New algorithms with specific hyper-tuning can be proposed to achieve even more accuracy and reliability. Further this could be improved with the collection of more reliable and recent data to further improve the model accuracy.