Machine learning-based system to predicting the diagnosis of coronary artery disease

A dissertation submitted for the Degree of Master of Computer Science

D C Wagachchi
University of Colombo School of Computing
2021

DECLARATION

I hereby declare that the thesis is my original work, and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis. This thesis has also not been submitted for any degree in any university previously.

Student Name: D. C. Wagachchi

Registration Number: 2018MCS094

Index Number: 18440946

30-11-2021

Signature of the Student & Date

This is to certify that this thesis is based on the work of Ms. D.C Wagachchi under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.
Certified by,

Supervisor Name: Prof G.K.A. Dias

Signature of the Supervisor & Date 30/11/2021

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor Prof. G. K. A. Dias for his guidance, motivation, and immense knowledge towards the successful completion of my research work. Moreover, his encouragement, insightful comments and guidance helped me in all the time of research and writing of this thesis.

Besides my supervisor, I my sincere gratefulness goes to our project coordinator Dr. Randil Pushpananda for his guidance given throughout the research work. Moreover, my thanks also delivered to the university staff and all the lecturers for the support given to successfully complete this research.

And finally, I would like to thank all the people who helped, supported, and encouraged me throughout this research work.

Thank you for all your encouragement!

ABSTRACT

Heart disease is now a regular occurrence and one of the leading causes of death all over the world. Among these diseases, coronary artery disease (CAD) is one of the common diseases around the world. This necessitates a prompt and precise identification of cardiac disease. Heart disease can be managed effectively with a combination of lifestyle changes, medicine, in some cases surgery. Heart disease symptoms can be decreased, and the heart's function can be enhanced with the correct treatment. But in recent times, heart disease prediction is one of the most complicated tasks in medical field. Because predicting cardiac illness is a difficult undertaking, it is necessary to automate the process in order to avoid the risks connected with it and to inform the patient well in advance.

The proposed work predicts the chances of coronary artery disease and classifies patient's risk level by implementing different machine learning techniques such as Random Forest Tree Classification, Decision Tree Algorithm and K -Nearest Neighbor Algorithm (KNN). And also discusses the viable machine learning algorithm-based web-based system and mobile application for the prediction of coronary artery disease (CAD) diagnosis accurately predict the diagnosis of coronary artery heart disease using only a few tests and features. And also, these project outcomes can be used to avoid surgical treatment and other costs.

As a result, this study provides a comparative analysis of the performance of several machine learning algorithms. The experiment results verify that the Random Forest Tree algorithm has the highest accuracy of 86 percent when compared to other machine learning algorithms.

Keywords: Coronary Artery Disease, Machine Learning, Random Forest Tree Classification, Decision Tree Algorithm and K -Nearest Neighbor Algorithm

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

CAD	Coronary Artery Disease
DT	Decision Tree
ECG	Electrocardiogram
FN	False Negative
FP	False Positive
ICT	Information and Communications Technology
ISH	International Society of Hypertension
KNN	K-Nearest Neighbor
ML	Machine Learning
NB	Naïve Bayes
RBF	Radial Basis Function
SCRL	Single Conjunctive Rule Learner
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
WHO	World Health Organization

1.0 INTRODUCTION

1.1 Overview

Heart attacks are the most common cause of death among all deadly disorders. Medical professionals undertake many surveys on heart disorders in order to collect data about heart patients, their symptoms, and the progression of the condition.

People in today's fast-paced society aspire to live a luxurious lifestyle, so they work like machines to make a lot of money and live comfortably. As a result, people neglect to look for themselves, and their food habits and overall lifestyle shift as a result. As a result, people are more tense, have high blood pressure and sugar levels at a young age, and take their own medication. As a result of all of these tiny mistakes, heart disease becomes a serious concern.

Heart Diseases remain the biggest cause of death for the last two decades. Among these diseases, coronary artery disease (CAD) is one of the common diseases around the world. Coronary artery disease (CAD) is one such disease with an annual mortality rate of about 7 million. Thus, early diagnosis of Coronary artery disease (CAD) is of vital importance. (Abdar, et al., 2019).

1.2 Problem Statement

Heart disease can be managed effectively with a combination of lifestyle changes, medicine, in some cases surgery. Heart disease symptoms can be decreased and the heart's function can be enhanced with the correct treatment. The projected outcomes can be used to avoid surgical treatment and other costs.

The ultimate goal of my research will be to accurately predict the diagnosis of coronary artery heart disease using only a few tests and features. Attributes are thought to provide the primary foundation for testing and, more or less, provide accurate findings.

Rather than the knowledge-rich data hidden in the data set and databases, decisions are frequently designed entirely on doctors' intuition and expertise. This practice results in unintended biases, errors, and exorbitant medical costs, all of which have an impact on the quality of care offered to patients.

Angiography is the preferred method for detecting coronary artery disease at the moment (CAD). However, because of the complexities and expense, academics have turned to machine learning algorithms as an alternative. (Kirmani, 2017) As the amount of data grows, machine learning is becoming more popular. Machine learning assists in gaining insight from a vast volume of data that is difficult for humans to process and often impossible.

Recently, computer technology and machine learning approaches have been used to produce software that can help doctors diagnose cardiac problems early on. Clinical and pathological data are used to diagnose cardiac disease. Based on the clinical data of patients, a heart disease prediction system can assist medical professionals in forecasting heart disease status. (Ngure, 2019) (Pradeep Gupta, 2020 July).

1.3 Aim and objectives.

1.3.1 Aim of the project

The project's purpose is to find a viable machine learning algorithm-based web-based system for the prediction of coronary artery disease diagnosis using effective clinical data aspects. The suggested technology would be used to differentiate between persons who have cardiac disease and those who are healthy. Developing a mobile application to collect data as well.

1.3.2 Objectives

The Research's objectives are listed below.

- Detection of features effective when it comes to detecting coronary artery disease (CAD) using the Non-acute chest pain and other pain features have diagnostic relevance for coronary artery disease in patients referring to cardiology clinics that treat outpatients.
- Creation of a database on coronary artery disease (CAD), including effective features and their weights.
- Finding an effective analytical algorithm or method for the evaluation of the collected dataset.
- The anticipation of the incidence of coronary artery disease (CAD) via data mining methods.
- Evaluation of the method for the diagnosis of coronary artery disease (CAD) through training and test sets.
- Develop a web-based system (decision support system) that integrates the coronary artery disease (CAD) detection machine learning algorithm (Saadatfar, et al., 2020).
- Create an Android application that allows people to monitor their heart health at any time by identifying different types of cardiac arrhythmias.

1.3.3 Scope

The scope of the project to find a suitable machine learning algorithm-based web-based system for the prediction of the diagnosis of coronary artery disease using effective features of the clinical data set. The proposed system will be developed to classify people with heart disease and healthy people. Also develop a mobile application for collecting data.

The following are the features of the web-based system.

- The system allows users to create an account and login
- The system allows users to update their profile and password.
- The system allows users to add training data.
- The system allows users to remove inappropriate training data.
- The system allows users to view training data.
- The system allows users to predict/analysis disease.
- The system allows users to generate reports.

The following are the features of the android application.

- The app allows users to create an account and login
- The app allows users to add data using questionaries.
- The app allows users to check their heart health status.

1.4 Research Questions

The research questions were as follows:

- Which machine learning algorithms are best for predicting the diagnosis of coronary artery disease?
- How can ICT be used to check whether a person has coronary artery disease in Sri Lanka?

1.5 Methodology

Acquiring specific data by providing a questionnaire to clinical patients and using documents and records in previous research and the processed dataset is used to train the predictive models using variety of machine learning approaches (Algorithms). Using the web application and the mobile application, based on the classification models generated by these ML techniques, The user will be able to determine if a patient who exhibits the underlying characteristics of CAD is really suffering from CAD or not.

There have been three main components in the proposed solution.

1. Database
2. Mobile application and web-based system
3. Machine learning model (Refer Figure 1.1 from the main text)

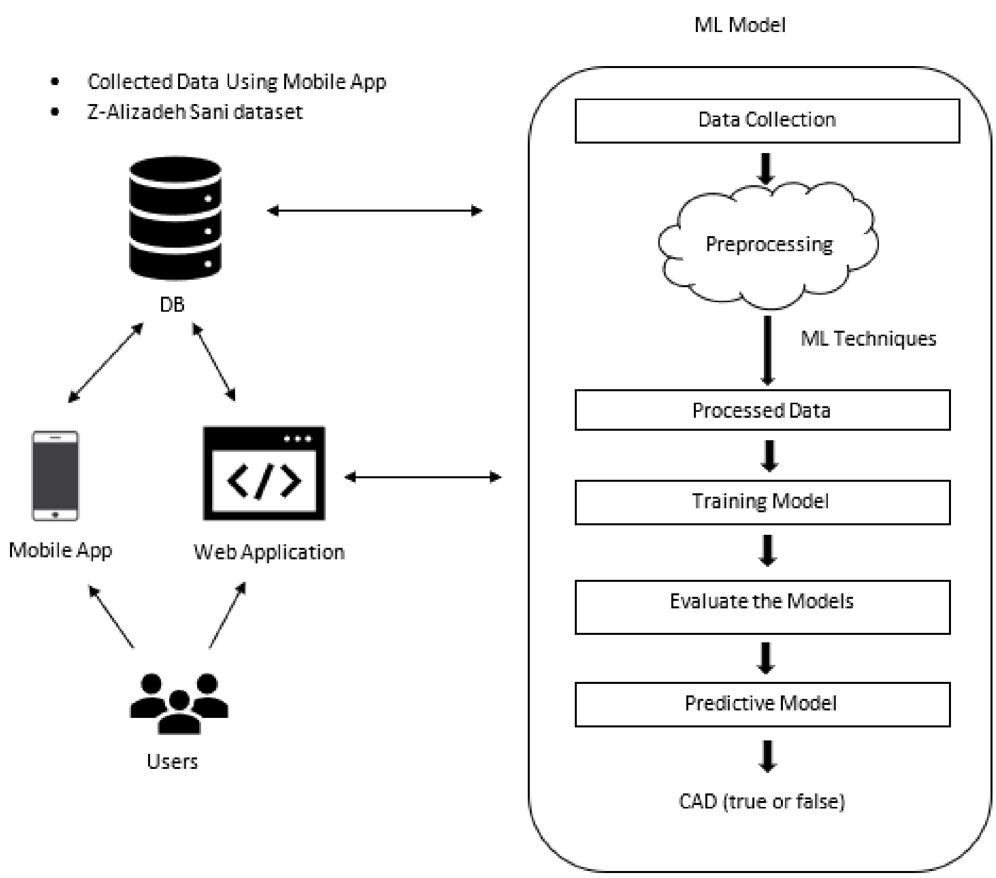

Figure 1:1 Methodology of Proposed Solution

1.6 Challenges

Medical diagnosis is an inherently difficult undertaking that necessitates extreme precision while considering a variety of circumstances.

Collecting clinical data in the COVID 19 setting is also a challenge. Furthermore, given the level of experience and knowledge required for accurate results, predicting the diagnosis of coronary artery disease is a far more difficult task.

1.7 Outline of thesis

The following is a breakdown of the thesis's structure. The second chapter examines existing machine learning-based methods to coronary artery disease (CAD) and other heart disease prediction systems. The proposed research design and technique are described in the third chapter. In this chapter, potential solutions to the research problem are discussed. Chapter four demonstrates the implementation details of the proposed methodologies. The assessment model and results of the proposed methodologies are presented in Chapter five. The final chapter, Chapter six, explains the thesis' conclusion and outlines further work.

2.0 LITERATURE REVIEW

There has previously been many good researches in the field of heart disease prediction using machine learning algorithms, but most of them have focused on machine learning in medical labs. Several studies on the diagnosis of CAD using data mining methods have been undertaken in previous years on various datasets. In the realm of cardiac disease, the Z-Alizadeh Sani dataset is the most recent dataset that researchers have used. Z-Alizadeh Sani has proposed using data mining approaches to diagnose CAD using ECG symptoms and features. (Senthilkumar Mohan, 2019) (Kirmani, 2017) (Kiruthika Devi, et al., 2016). To find the diagnosis of CAD disease, the majority of the researchers used sequential minimum optimization and nave Bayes methods.

For the noninvasive diagnosis of CAD, Z-Alizadeh Sani developed a feature engineering approach that used the nave Bayes and SVM classifiers. They grew their dataset from around 500 sample records to over a thousand. They achieved 86% accuracy for the naïve Bayes algorithms algorithm and 96% accuracy for the SVM algorithm (Saadatfar, et al., 2020) (Soni, et al., 2011).

Mursal Furqan, Hiba Rajput (Mursal Furqan, 2020 December) discusses a statistical model of heart disease that, based on basic parameters of the patients' health history, will help medical examiners and cardiac practitioners forecast heart disease. They applied three Machine Learning Classifier Models to create this prediction model: Logistic Regression Classifier, K-Nearest Neighbors Classifier, and Random Forest Classifier. Different important clinical features of a patient, critical for deciding a patient's heart disease, are taken by them in the first section and, secondly, they calculated the accuracy using different ML Classifiers are defined on the given dataset. When the proportion of test data is 0.2 percent, the maximum accuracy achieved is 87 percent by logistic regression.

Joloudari, et al. (Saadatfar, et al., 2020) for CAD diagnosis, use both the above methods and neural networks. They had the best results, with a 94% accuracy rate.

Abdar et al. (Abdar, et al., 2019) established a new optimization technique called N2Genetic optimizer. The nuSVM was then utilized to determine whether or not the patients had CAD. On the Z-Alizadeh Sani dataset, the suggested detection approach outperformed existing methods with an accuracy of 93.08 %.

Mohan et al. (S. Mohan, et al., 2019) developed a hybrid strategy for predicting heart disease based on a random forest and a linear model (HRFLM). On the Cleveland dataset, the proposed technique improved performance with an accuracy of 88.7%.

K. Polara j et al (Polaraju & D, 2017) Prediction of related heart disease using several algorithms Multiple Linear Regression is good for predicting the risk of heart disease, according to the model. The research is based on a raw data collection of 1000 cases with 10 different features that were previously established. Since it is clear by looking at the result, the conclusion of the registration algorithms is maximal in comparison to various algorithms, there are two phases of data division where 70% of the data is used to train the machine and 30% of the data is utilized for testing purposes.

Makwana & Patel (Makwana & Patel, 2015) had done the research using a sequential minimal optimization algorithm and achieved 86% accuracy, and using naïve Bayes algorithms algorithm they achieved 87% accuracy. After combining both algorithms as a hybrid algorithm, they achieved 88% accuracy.

Marjia et al. (Marjia & Afrin Haider, 2017) employing WEKA software for KStar, j48, SMO, Bayes Net, and Multilayer perception to construct a projection framework for heart disorders. Multilayer perception and J48 approaches achieve superior performance than KStar based on results from particular factor SMO and Bayes Net using kfold cross-validation. The precision of the algorithms is still unacceptably low. As a result, the efficiency of the accuracy is improved even more, allowing for the correct diagnosis to be made.

Existing CAD detection systems suffer from some of the following flaws, as evidenced by the above-mentioned pieces of evidence. To begin with, most researchers have validated their suggested approach on a single dataset, with only a few investigations including at least two CAD datasets. As a result, the prediction results are unreliable. To demonstrate the generalizability of the suggested strategy, numerous CAD datasets should be used.

Comparison between similar projects and proposed model in feature-wise.

- Feature 1 – Dataset

Several studies on the diagnosis of CAD using data mining algorithms have been undertaken in recent years on various datasets. In the realm of cardiac disease, the Z-Alizadeh Sani dataset has lately been utilized by researchers (Senthilkumar Mohan, 2019), but they have not used clinical data. By contacting specialist doctors and medical professionals with clinical data sets, the proposed system would include some 40 criteria significant to a heart attack, including their weight, age, and priority levels.

Comparison between related research and proposed model according to the dataset they had used in table 2.1 below.

Reference	Related Projects Dataset Details	Proposed Model Dataset Details
(S. Mohan, et al., 2019)	There are 303 records in the collection, each having 13 properties.	Z-Alizadeh Sani Data set (303 records with 55 attributes) and local clinical data set with 40 attributes
(Abdar, et al., 2019)	There are 303 records in the collection, each having 54 properties.	
(Saadatfar, et al., 2020)	There are 500 records in the collection, each having 54 properties.	
(Makwana & Patel, 2015)	There are 303 records in the collection, each having 76 properties.	
(Polaraju & D, 2017)	There are 3000 records in the collection, each having 13 properties.	

Table 2:1 Comparison between related research and proposed model according to the dataset they had used.

- Feature 2 – Algorithm

Several research on the diagnosis of CAD utilizing various algorithms have been undertaken in recent years. We are using Random Forest Tree Classification, Decision Tree Algorithm, and K -Nearest Neighbor Algorithm (KNN) techniques to develop an effective heart attack prediction system in this system. Because of using the above algorithms, delineate mining techniques can be understood with special adaptability. The algorithms are applied after the input is taken. The operation is carried out after accessing the data set, and an accurate prediction of CAD level is provided.

Comparison between related research and proposed model according to the algorithm they used in table 2.2 below.

Reference	Related Projects Algorithm Details	Proposed Model Algorithm Details
(Mursal Furqan, 2020 December)	K -Nearest Neighbor Algorithm (KNN), Logistic Regression, Random forest	Random Forest Tree Classification, Decision Tree Algorithm, K -Nearest Neighbor Algorithm (KNN)
(S. Mohan, et al., 2019)	K -Nearest Neighbor Algorithm (KNN), Decision Trees (DT)	
(Abdar, et al., 2019)	Naïve Bayes (NB), K-NN, Radial Basis Function (RBF), Single Conjunctive Rule Learner (SCRL),	
(Saadatfar, et al., 2020)	Naïve Bayes-SMO, Random Forest Algorithm	
(Makwana & Patel, 2015)	K-means algorithms, MAFIA algorithms and decision tree classification	
(Polaraju & D, 2017)	J48 algorithm, logistic model tree algorithm, Random Forest decision tree algorithm	

Table 2:2 Comparison between related research and proposed model according to the algorithms they had used

3.0 METHODOLOGY

Acquiring specific data by providing a questionnaire to clinical patients and using documents and records in previous research and using a variety of machine learning approaches (Algorithms), the processed dataset is utilized to train predictive models. Using the web application and the mobile application, the user will be able to predict whether a patient who exhibits the underlying characteristics of CAD suffers from CAD or not based on the classification models built by these ML approaches. In the proposed solution, there were three essential components.

1.0 Database

2.0 Machine learning model

3.0 Mobile application and web-based system

3.1 Database

This study uses a data source that contains the medical histories of approximately 500 people of various ages. This dataset provides us with much-needed data. The patient's medical characteristics, such as age, resting blood pressure, fasting sugar level, and so on, assist us in determining whether or not the patient has been diagnosed with heart disease.

This dataset contains 40 medical variables from around 500 patients that help us determine if a patient is at risk of developing heart disease or not, as well as categories individuals who are at risk and those who are not.

The disease dataset comes from the Z-Alizadeh Sani collection, as well as clinical data from the local.

The pattern that leads to the detection of people at risk for CAD disease is retrieved from this dataset. These records are split into two parts. Training and Testing.

Figure 3.1 shows a description of the sample features utilized in the Z-Alizadeh-Sani dataset, along with their valid range.

Feature Type	Feature Name	Range	Measurement			
			Mean	Std. Error of Mean	Std. Deviation	Variance
Demographic	Age	(30–80)	58.90	0.6	10.39	108
Demographic	Weight	(48–120)	73.83	0.69	11.99	143.7
Demographic	Length	(140–188)	164.72	0.54	9.33	87.01
Demographic	Sex	Male, Female	—	—	—	—
Demographic	BMI (body mass index Kb/m^2)	(18–41)	27.25	0.24	4.1	16.8
Demographic	DM (diabetes mellitus)	(0, 1)	0.3	0.03	0.46	0.21
Demographic	HTN (hypertension)	(0, 1)	0.6	0.03	0.49	0.24
Demographic	Current smoker	(0, 1)	0.21	0.02	0.41	0.17
Demographic	Ex-smoker	(0, 1)	0.03	0.01	0.18	0.03
Demographic	FH (family history)	(0, 1)	0.16	0.02	0.37	0.13
Demographic	Obesity	Yes if MBI > 25, No otherwise	—	—	—	—
Demographic	CRF (chronic renal failure)	Yes, No	—	—	—	—

Figure 3:1 Features used in the Z-Alizadeh-Sani dataset with their valid range

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Age	Weight	Length	Sex	BMI	DM (Diabetes Mellitu	HTN (Hypertension)	Current Smoker	EX-Smoker	FH (Family History)	Obesity	CRF	CVA	Airway disease	Thyroid Disease	CHF	DLP
2	53	90	175	Male	29.3877551	0	1	1	0	0	Y	N	N	N	N	N	Y
3	67	70	157	Fmale	28.398718	0	1	0	0	0	Y	N	N	N	N	N	N
4	54	54	164	Male	20.07733492	0	0	1	0	0	N	N	N	N	N	N	N
5	66	67	158	Fmale	26.83864765	0	1	0	0	0	Y	N	N	N	N	N	N
6	50	87	153	Fmale	37.16519287	0	1	0	0	0	Y	N	N	N	N	N	N
7	50	75	175	Male	24.48979592	0	0	1	0	0	N	N	N	N	N	N	N
8	55	80	165	Male	29.38475666	0	0	0	1	0	Y	N	N	N	N	N	N
9	72	80	175	Male	26.12244898	1	0	1	0	0	Y	N	N	N	N	N	Y
10	58	84	163	Fmale	31.61579284	0	0	0	0	0	Y	N	N	N	N	N	N
11	60	71	170	Male	24.56747405	1	0	0	0	0	N	N	N	N	N	N	N
12	58	75	168	Male	26.57312925	0	1	0	1	0	Y	N	N	N	N	N	N
13	80	67	153	Fmale	28.62147037	0	1	0	0	0	Y	N	N	N	N	N	Y
14	70	70	151	Fmale	30.70040788	1	1	0	0	0	Y	N	N	N	N	N	Y
15	67	63	154	Fmale	26.56434475	1	1	0	0	0	Y	N	N	N	N	N	Y
16	66	63	155	Fmale	26.2226847	1	1	0	0	0	Y	N	N	N	N	N	Y
17	59	81	167	Male	29.04370899	1	0	0	0	0	Y	N	N	N	N	N	Y
18	41	68	169	Male	23.80869017	0	0	1	0	0	N	N	N	N	N	N	N
19	68	59	161	Fmale	22.76146754	0	0	0	0	1	N	N	N	N	N	N	Y
20	60	89	163	Fmale	33.49768527	1	1	0	0	0	Y	N	N	N	N	N	N
21	65	72	150	Fmale	32	1	1	0	0	0	Y	Y	N	N	N	N	Y
22	47	84	170	Fmale	29.06574394	0	0	0	0	1	Y	N	N	N	N	N	N
23	66	89	151	Fmale	39.03337573	0	1	0	0	0	Y	N	N	N	N	N	N
24	66	75	170	Male	25.95155709	1	1	0	0	0	Y	N	N	N	N	N	N
25	72	66	161	Fmale	25.46198063	1	1	0	0	0	Y	N	N	N	N	N	N
26	50	66	164	Fmale	24.5389649	1	0	0	0	1	N	N	N	N	N	N	N
27	65	74	164	Male	27.51338489	0	0	0	0	0	Y	N	N	N	N	N	N
28	56	73	173	Male	24.39105884	0	0	0	0	0	N	N	N	N	N	N	Y
29	50	81	165	Male	29.75206612	0	1	0	0	0	Y	N	N	N	N	N	Y
30	80	51	148	Fmale	23.28341855	0	1	0	0	0	N	N	N	N	N	N	Y

Figure 3:2 Sample Dataset

3.2 Machine Learning Model

ML techniques enable the application of intelligent procedures to a variety of datasets in order to uncover important insights. Because of ML's programmability in exploring, analyzing, and interpreting datasets, it's a good fit for decision-makers in fields like medical diagnostics.

The proposed methodology (in figure 3.2.1 below) includes steps, where the first stage is referred to as the collection of the data than in. The second stage extracts significant values, and the third stage is preprocessing, which involves data exploration. The initial dataset will be preprocessed to remove any potential noise that could affect the predictive analysis results.

Depending on the techniques utilized, data preprocessing deals with missing values, data cleansing, and standardization. The preprocessing step is performed after the data is classified. In general, a set of processes leads to the generation of a set of cleaned data that may be utilized on the dataset, a procedure known as data preparation.

The classifier is used to classify the pre-processed data once it has been pre-processed. K closest neighbors (KNN), Decision tree classification, and Random Forest tree classification are the classifiers employed in the suggested model. Finally, we put the proposed model to the test, evaluating it for accuracy and performance using multiple performance measures.

During the training phase, a number of independent variables, such as age, gender, medical history, and symptoms, will be utilized in conjunction with a dependent variable to build a classification model.

The user will therefore be able to predict whether a patient who exhibits the underlying characteristics of CAD suffers from CAD or not based on the classification models built by these ML approaches.

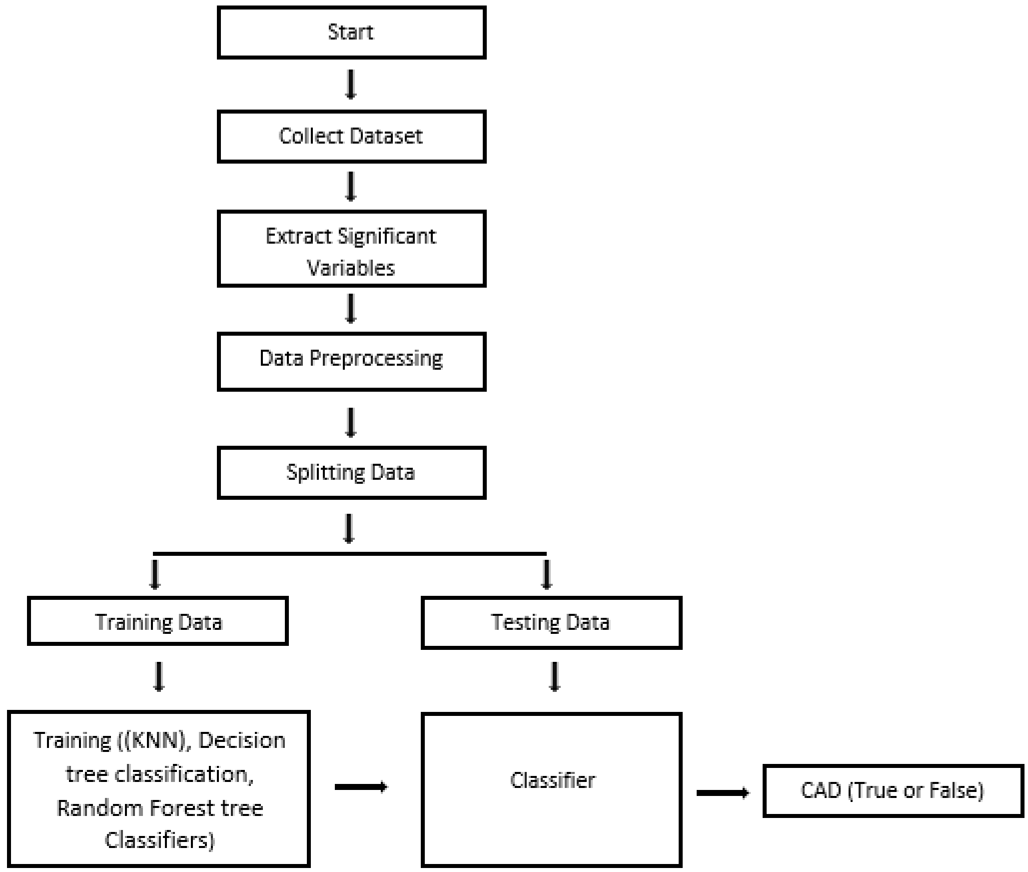

Figure 3:3 Proposed Machine Learning Model

3.3 Mobile Application and Web-based System

A web application that uses an intelligent algorithm to provide users with real-time advice on their risk of heart disease.

Various details are provided into the application which are related to the heart disease. The application allows users to share their heart-related difficulties with others. It then checks for various risks linked with the CAD using user-specific information. Using some smart data mining algorithms to predict the most accurate condition that might be linked to the patient's information.

This web-based system accesses the machine learning model and the local database. Data retrieval, risk prediction algorithm, and generating a report is implemented by the system. The patients who supposed to enter the data using the mobile application can retrieve their own risk profile. (Saadatfar, et al., 2020). And also used to WHO/ ISH risk prediction chart parameters for the region of southeast Asia to calculate the CAD risk range based on these charts for the web application.

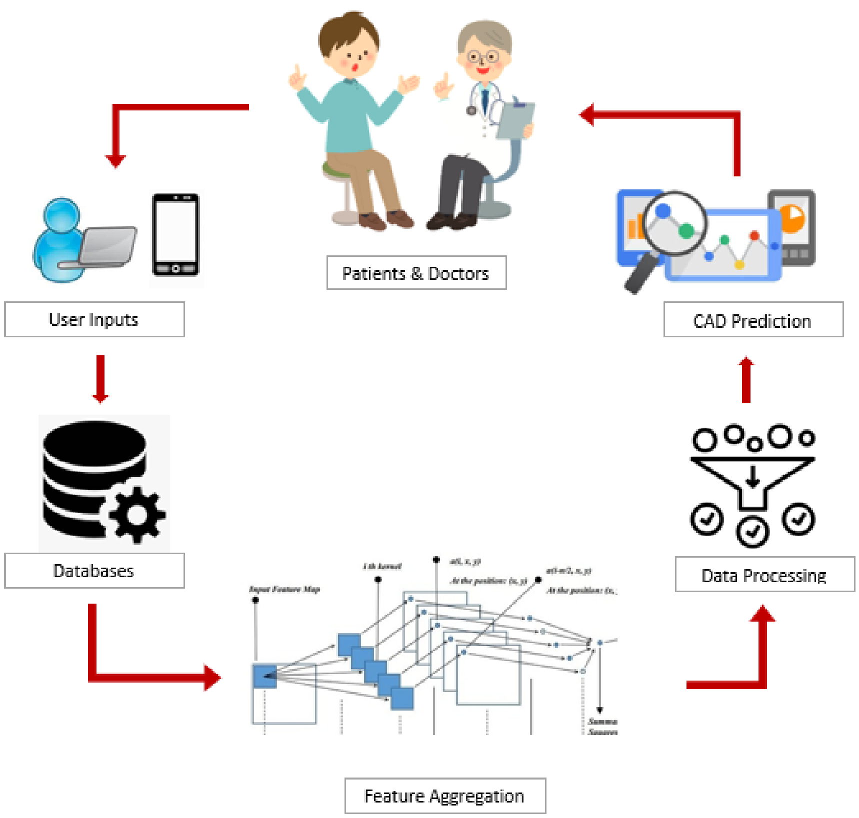

Figure 3:4 Methodology of Proposed Solution

3.3.1 WHO/ISH risk prediction chart for South-East Asia

This is the WHO/ ISH risk prediction chart for the region of southeast Asia. The charts provide approximate estimates of cardiovascular disease risk. Gender, Smoker or non-smoker, Age, blood pressure, total blood cholesterol (HDL/LDL/Triglyceride) are the parameters WHO used. I used these parameters to calculate the CAD risk range based on these charts for the web application.

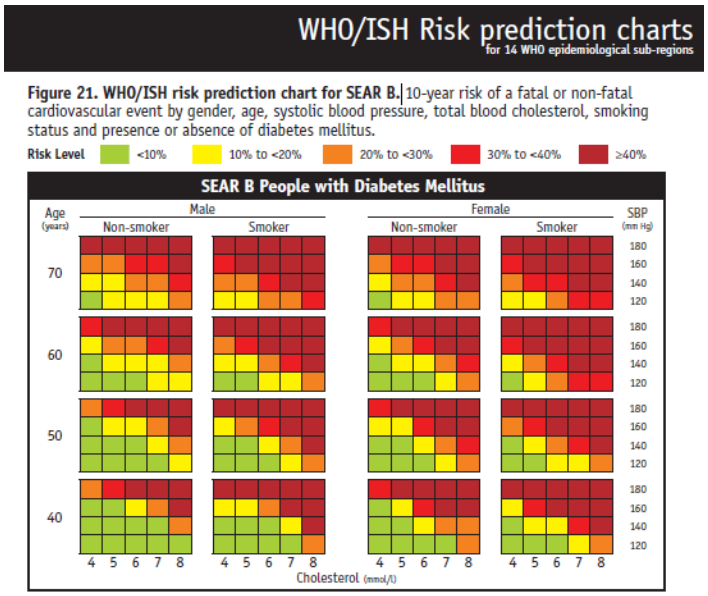

Figure 3:5 WHO/ISH risk prediction chart for SEAR B people with diabetes mellitus

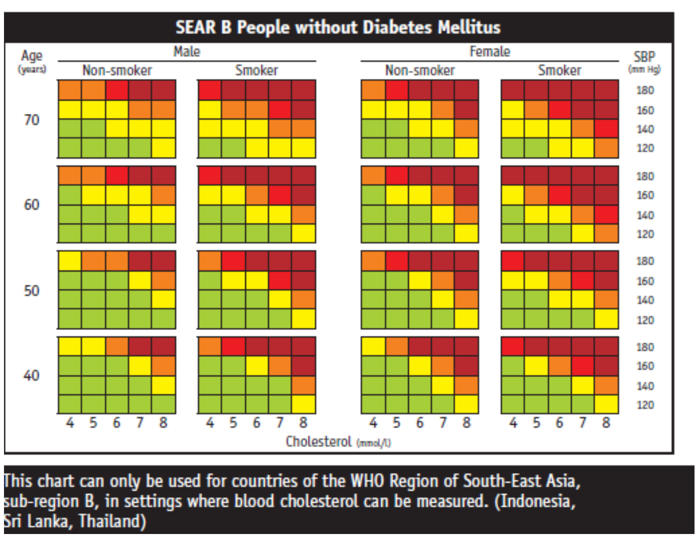

Figure 3:6 WHO/ISH risk prediction chart for SEAR B people without diabetes mellitus

4.0 IMPLEMENTATION

4.3 Machine Learning Model

- Collect data set

```
import itertools
import numpy as np
import matplotlib.pyplot as plt
from matplotlib.ticker import NullFormatter
import pandas as pd
import numpy as np
import matplotlib.ticker as ticker
from sklearn import preprocessing
%matplotlib inline

dataset = pd.read_csv("/content/drive/MyDrive/Colab Msc/Z-Alizadeh-sani-dataset-Colabs.csv")
dataset.head()
```

	Age	Weight	Le0gth	Sex	BMI	DM	HTN	Current Smoker	EX-Smoker	FH	Obesity	CRF	CVA	Airway disease	Thyroid Disease	CHF	DLP	BP	PR	Edema	Weak Peripheral Pulse	Lung rales
0	53	90	175	0	29.387755	0	1	1	0	0	1	0	0	0	0	0	1	110	80	0	0	0
1	67	70	157	1	28.398718	0	1	0	0	0	1	0	0	0	0	0	0	140	80	1	0	0
2	54	54	164	0	20.077335	0	0	1	0	0	0	0	0	0	0	0	0	100	100	0	0	0
3	66	67	158	1	26.838648	0	1	0	0	0	1	0	0	0	0	0	0	100	80	0	0	0
4	50	87	153	1	37.165193	0	1	0	0	0	1	0	0	0	0	0	0	110	80	0	0	0

Figure 4:1 ML Model-Collect data set

- We need to check if there are any Null values in the dataset. If there are any Null, then we need to impute the values.

```
dataset.isnull().values.any()
```

```
False
```

- Check the datatypes to see if we need to perform encoding categorical data.

```
dataset.dtypes
```

```
Age                    int64
Weight                 int64
Le0gth                 int64
Sex                    int64
BMI                  float64
DM                     int64
HTN                    int64
Current Smoker         int64
EX-Smoker              int64
FH                     int64
Obesity                int64
CRF                    int64
CVA                    int64
Airway disease         int64
Thyroid Disease        int64
```

Figure 4:2 Dataset -datatypes

- One of the most important stages in machine learning is feature selection. Irrelevant Parameters will lower the performance of the model.

```
Corr = dataset.corr()
Corr
```

	Age	Weight	Le0gth	Sex	BMI	DM	HTN	Current Smoker	EX-Smoker	FH	Ol
Age	1.000000	-0.264585	-0.163753	0.045769	-0.161414	0.072543	0.246690	-0.143879	0.076608	-0.183900	-0.1
Weight	-0.264585	1.000000	0.460631	-0.234529	0.725005	-0.003531	-0.028532	0.157385	0.068977	0.021963	0.5
Le0gth	-0.163753	0.460631	1.000000	-0.700279	-0.269356	-0.052318	-0.153668	0.335248	0.079034	0.004488	-0.1
Sex	0.045769	-0.234529	-0.700279	1.000000	0.284088	0.194348	0.149278	-0.336330	-0.156932	0.071098	0.2
BMI	-0.161414	0.725005	-0.269356	0.284088	1.000000	0.045360	0.091652	-0.089398	0.005016	0.014045	0.7
DM	0.072543	-0.003531	-0.052318	0.194348	0.045360	1.000000	0.217864	-0.208458	-0.120087	-0.064434	0.0
HTN	0.246690	-0.028532	-0.153668	0.149278	0.091652	0.217864	1.000000	-0.169000	0.041045	-0.098467	0.1
Current Smoker	-0.143879	0.157385	0.335248	-0.336330	-0.089398	-0.208458	-0.169000	1.000000	-0.094652	0.089532	-0.0

Figure 4:3 Dataset correction

- Plotting heatmap to analyze the correlation of all the parameters

```python
import seaborn as sb
sb.heatmap(Corr,vmin=0, vmax=1, center=0,
           square=True, linewidths=1, cbar_kws={"shrink": .5})
```

<matplotlib.axes._subplots.AxesSubplot at 0x7f2210b0b750>

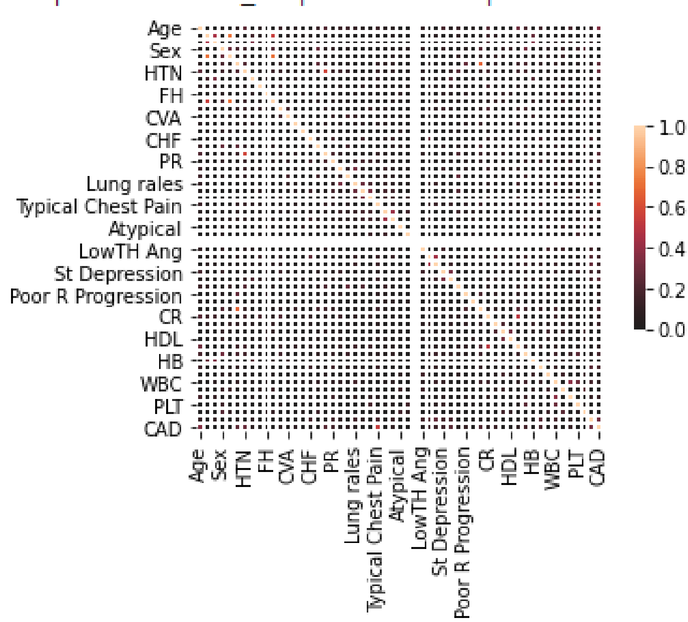

Figure 4:4 Heatmap

- Feature selection - Univariate Selection

```python
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2

X = dataset.drop(['CAD'], axis = 1)
y = dataset['CAD']

#apply SelectKBest class to extract best features
parameters = SelectKBest(score_func=chi2, k=54)
fit = parameters.fit(X,y)
dfscores = pd.DataFrame(fit.scores_)
dfcolumns = pd.DataFrame(X.columns)

#concat two dataframes for better visualization
featureScores = pd.concat([dfcolumns,dfscores],axis=1)

#naming the dataframe columns
featureScores.columns = ['Specs','Score']

#print features
print(featureScores.nlargest(54,'Score'))
```

Figure 4:5 Feature Selection

```
                 Specs        Score
48                 WBC  1167.316183
40                  TG   381.017191
38                 FBS   290.382804
44                 ESR   125.493829
0                  Age    70.675365
53         Region RWMA    62.344117
17                  BP    47.262174
51                 PLT    45.376214
24  Typical Chest Pain    40.978986
27            Atypical    36.328101
52              EF-TTE    27.904458
28         nonanginal     21.575700
49               Lymph    14.938850
5                   DM    13.622818
34          Tinversion    11.957221
6                  HTN    10.267988
18                  PR     9.048199
50                neut     8.015657
43                 BUN     6.600525
31              Q Wave     6.444444
```

```
23      Diastolic Murmur     6.333812
32          St Elevation     5.638889
33         St Depression     4.839249
26        Function Class     4.570598
36     Poor R Progression    3.625000
25               Dyspnea     2.649543
1                 Weight     2.625386
11                   CRF     2.416667
37                   BBB     2.273104
13        Airway disease     2.069140
20  Weak Peripheral Pulse    2.013889
41                   LDL     2.002821
42                   HDL     1.517846
7         Current Smoker     1.296669
4                    BMI     1.138373
19                 Edema     0.850734
30            LowTH Ang      0.805556
3                    Sex     0.791037
35                   LVH     0.741762
```

Figure 4:6 Feature Selection Result

- Feature selection - Feature Importance

```
from sklearn.ensemble import ExtraTreesClassifier
model = ExtraTreesClassifier()
model.fit(X,y)

#plot graph of feature importances for better visualization
feat_importances = pd.Series(model.feature_importances_, index=X.columns)
feat_importances.nlargest(54).plot(kind='barh')
plt.show()
```

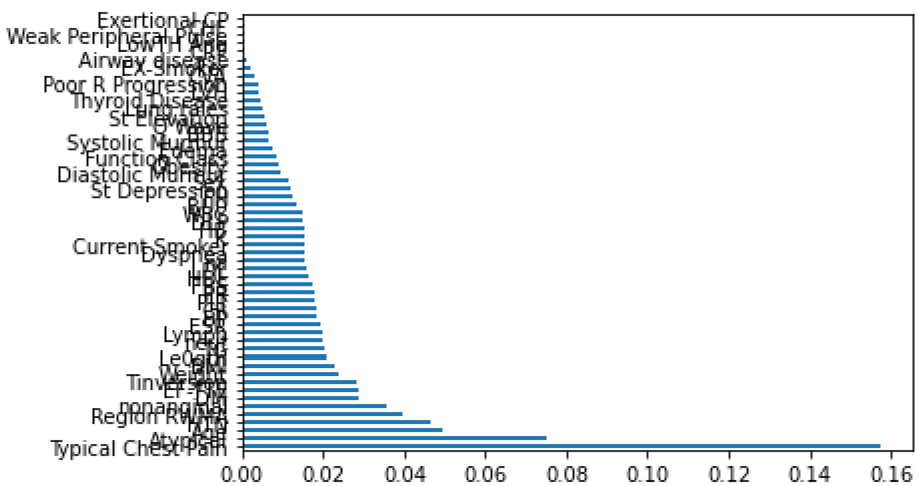

Figure 4:7 Graph of feature importance

- Choose test and training set

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split( X, y, test_size=0.2, random_state=4)

print ('Train set:', X_train.shape,  y_train.shape)
print ('Test set:', X_test.shape,  y_test.shape)
```

```
Train set: (242, 54) (242,)
Test set: (61, 54) (61,)
```

Figure 4:8 Choose test and training set

4.4 Mobile Application

Figure 4:9 Mobile App Icon

The users can choose this 'CAD Prediction' android application and can check their CAD risk using this app.

4.4.1 Mobile app login and sign up

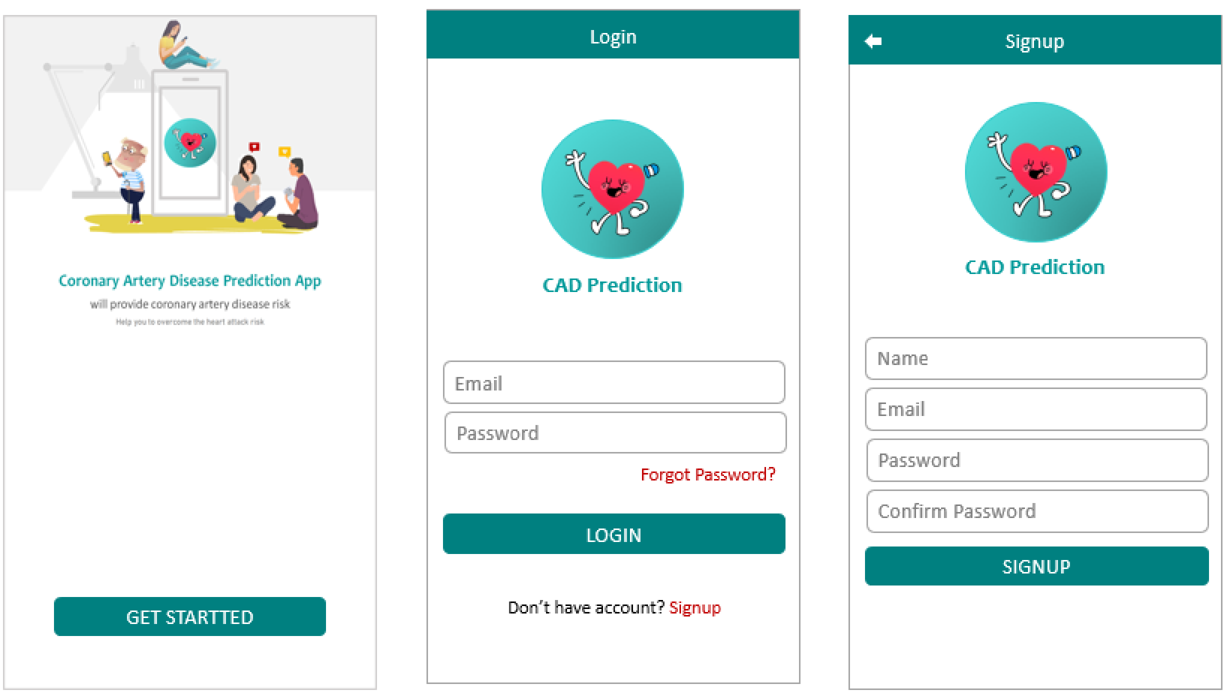

Figure 4:10 Mobile app login and sign-up screens

To use the app, user need to login using the valid email address and password of your account. If you do not have an account yet, click 'Signup' and register with information like name, email and password.

4.4.2 Data Entry Page

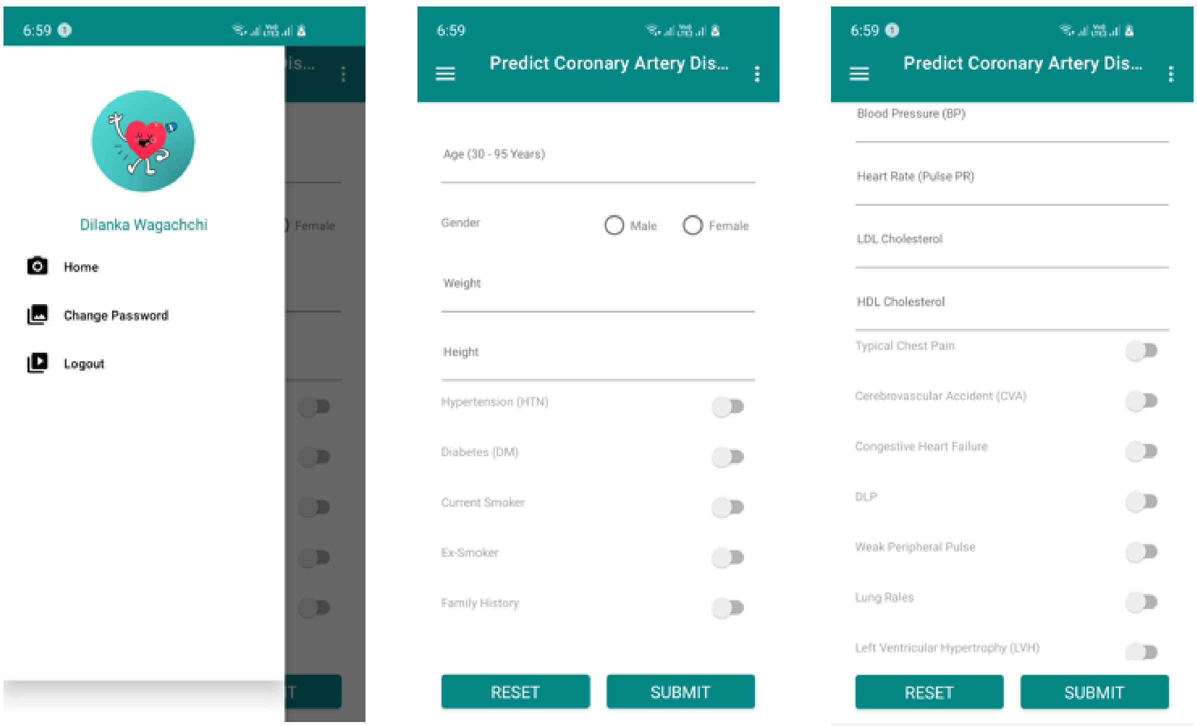

Figure 4:11 Data entry page

After login to the app user can fill the above form with fields related to their details of cardio. After submitting the form, the application will calculate the percentage of CAD risk.

4.4.3 Result Page

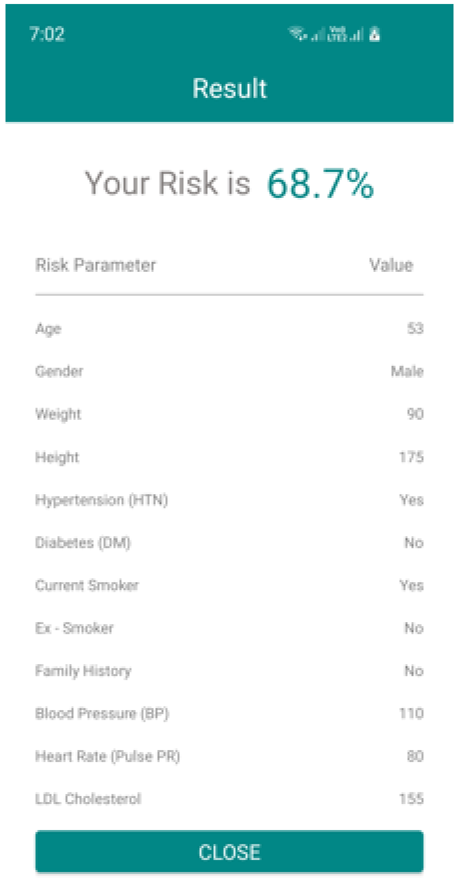

Figure 4:12 Result Screen

4.5 Web-based System

4.5.1 Login Page

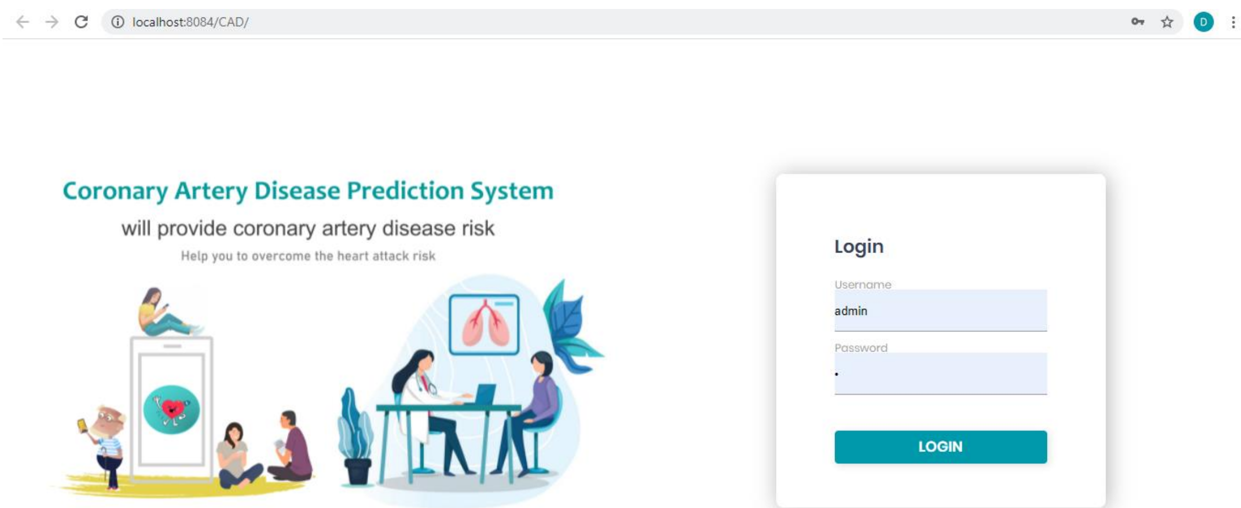

Figure 4:13 Login Page

4.5.2 Menu

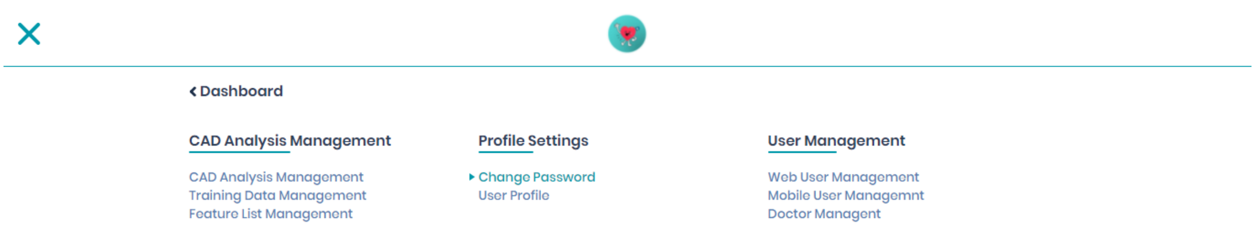

Figure 4:14 Menu

4.5.3 CAD Analysis Management

4.5.3.1 Data Grid and Search

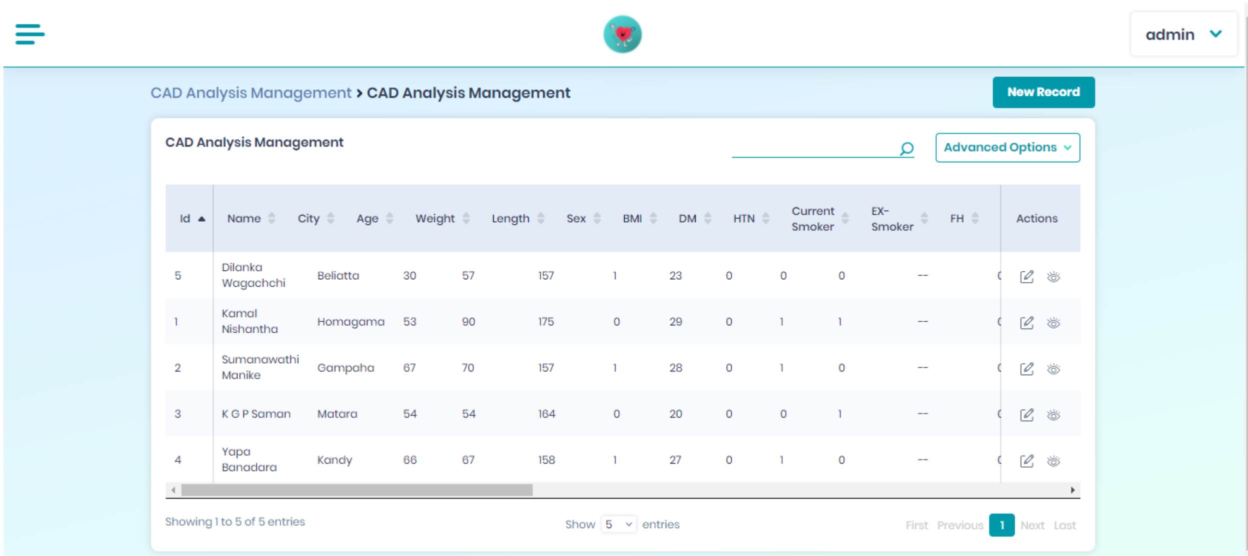

Figure 4:15 Data grid and search

4.5.3.2 Add new record for analysis CAD Risk

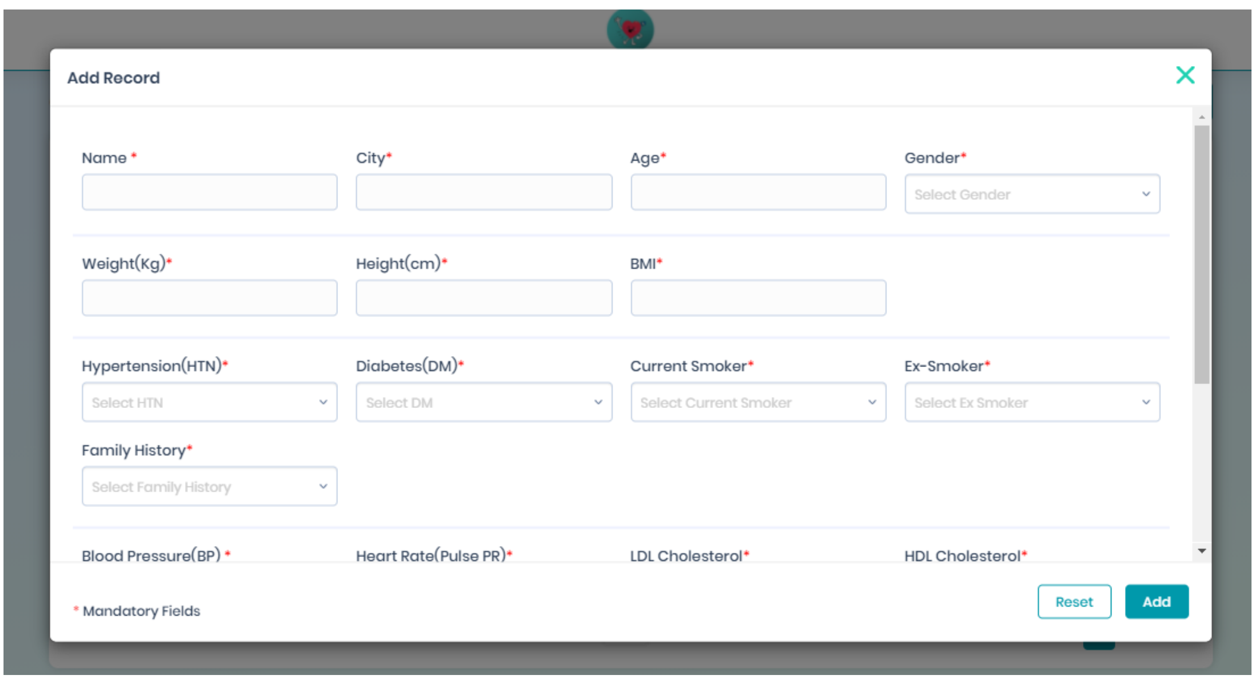

Figure 4:16 CAD Analysis new record

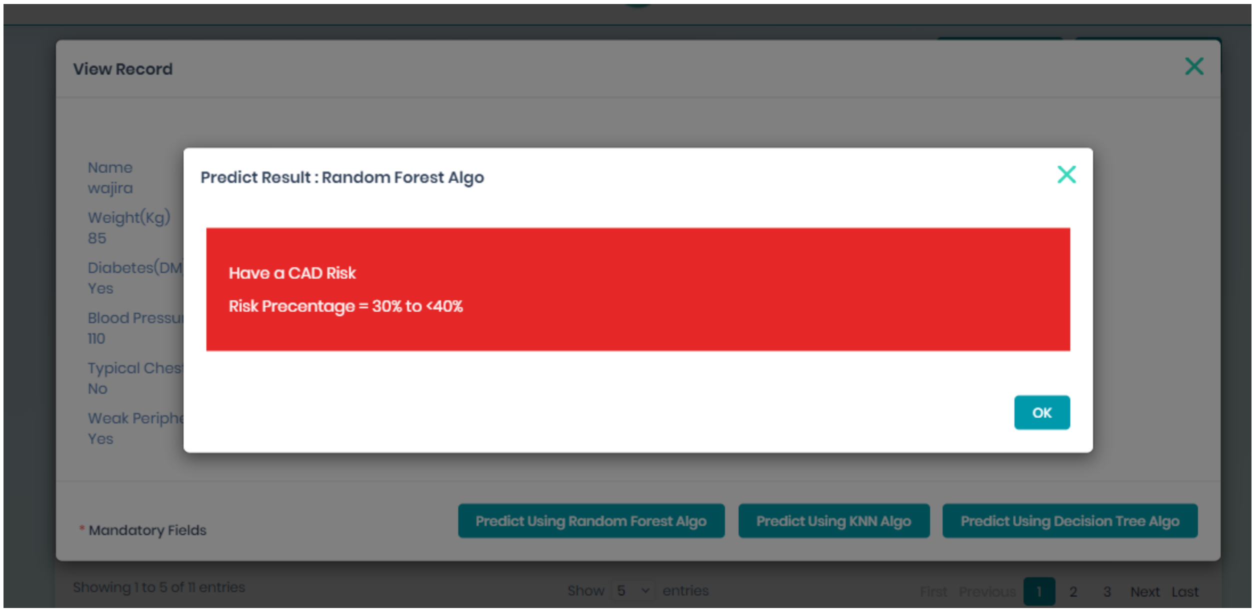

Figure 4:17 The risk color and the percentage based on the WHO chart

4.5.4 Training Data Management

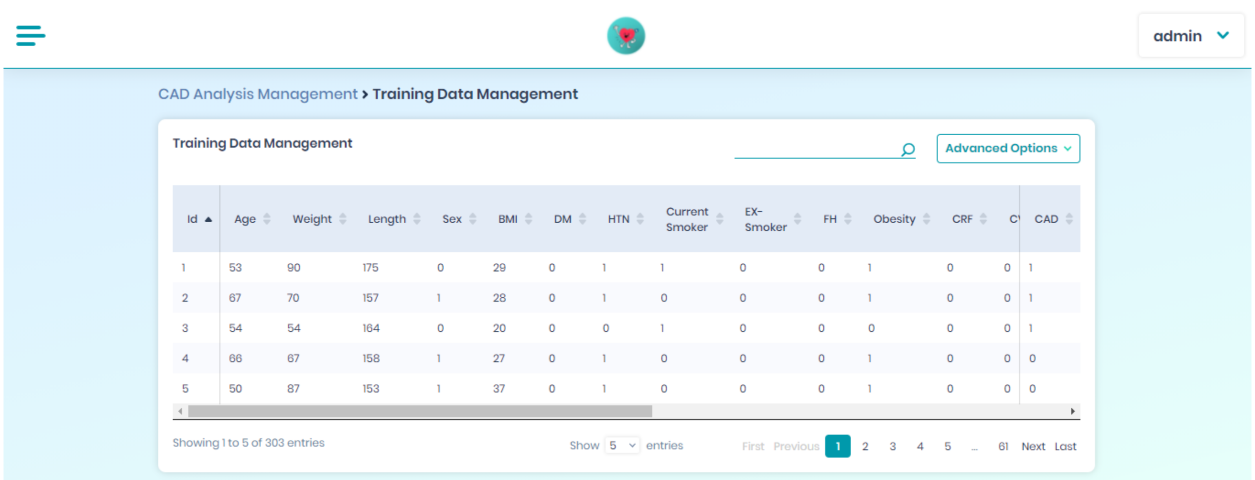

Figure 4:18Training data management

4.5.5 Feature List Management

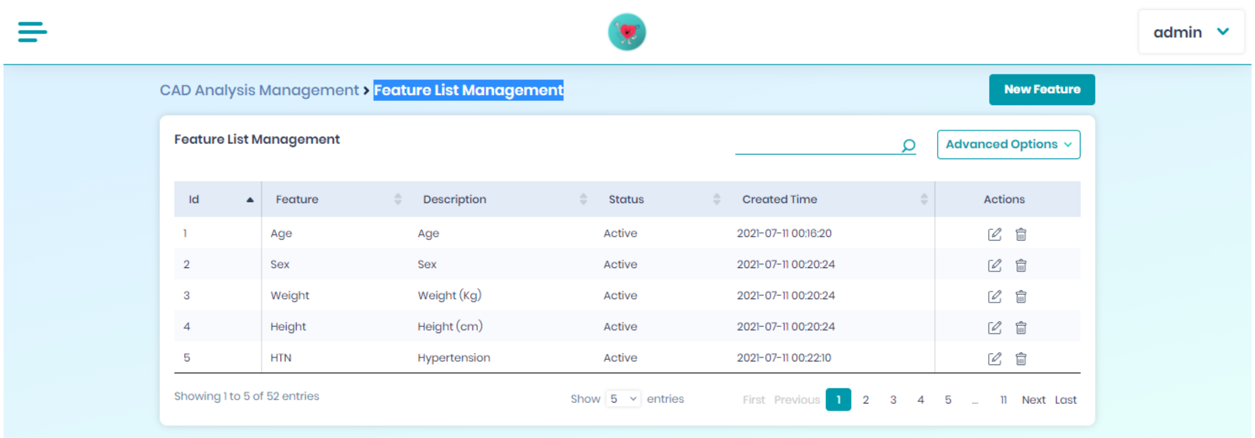

Figure 4:19 Feature list management

4.5.6 Doctor Management

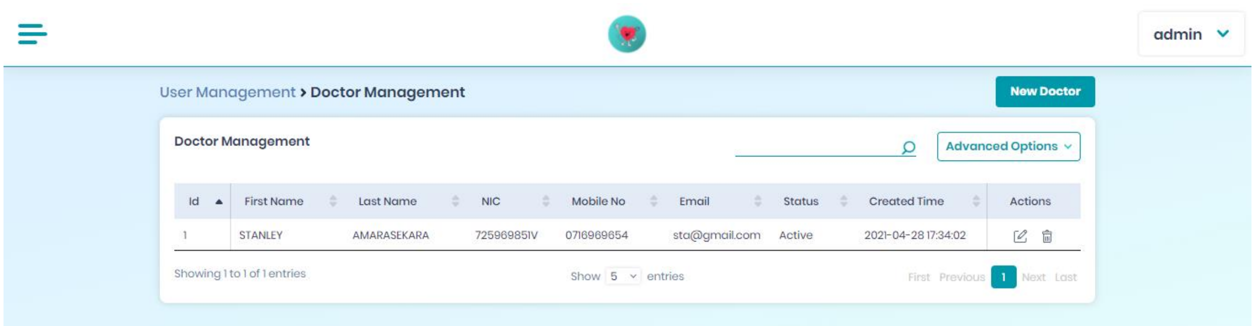

Figure 4:20 Doctor management

4.5.7 Change Password

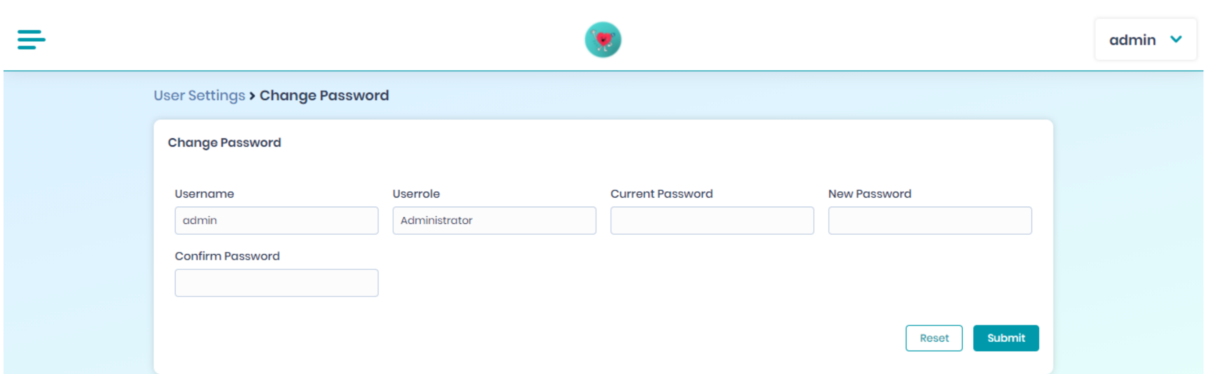

Figure 4:21 Change password

5.0 EVALUTION

Heart disease can be effectively managed with a combination of lifestyle changes, medications, and, in rare circumstances, surgery. Heart disease symptoms can be decreased, and the heart's function can be enhanced with the correct treatment. The predicted outcomes can be used to avoid and thereby lower the cost of surgical therapy as well as other costs.

The overarching goal of my research will be to accurately predict the diagnosis of coronary artery heart disease using only a few tests and features. Characteristics that are thought to be the basic foundation for testing and, more or less, provide correct findings. Rather than the knowledge-rich data hidden in the data set and databases, decisions are frequently made based on doctors' intuition and expertise. This practice leads to unintended biases, errors, and exorbitant medical costs, all of which have an impact on the quality of care offered to patients.

The evaluation can be done by testing the accuracy of the classification using a different set of data. Case studies that have been carried out before as well as the data collected from portals will be used as the baseline to train the model and evaluation. The datasets will be split into training and test datasets. Analysis metrics such as accuracy, precision, the recall will be used for the evaluation of the model. The results obtained by the proposed model data and those of currently available data will be compared against each other.

The confusion matrix is used to assess the performance of machine learning-derived classification models. The confusion matrix is a contingency table that shows how many instances are assigned to each class, allowing us to calculate classification accuracy, sensitivity, specificity, true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs), among other things. (Makwana, 2015) Although there can be two or more classes involved, the dataset only contains two, resulting in a 2x2 confusion matrix for each classification model (Figure 1). In the case of the experiment in question,

- Has CAD - YES
- No CAD – NO

Actual Class		Predicted Class	
		Has CAD	No CAD
	Has CAD	True Positive (TP)	False Negative (FN)
	No CAD	False Positive (FP)	True Negative (TN)

True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN) are defined as given in the table 1. At the same time, we observed the confusion matrix which was built according to the above table.

In the confusion matrix, the terms TP, FP, FN, and TN are used.

- True Positive (TP) - Number of patients that are predicate to have CAD and actually have CAD.
- False Positive (FP) - Number of patients that are predicate to have CAD and do not actually have CAD.
- False Negative (FN) - Number of patients that are predicate to not have CAD and actually have CAD.
- True Negative (TN) - Number of patients that are predicate to not have CAD and do not actually have CAD.

Accuracy is one of the most used performance comparison measures in classification analysis. The number of true predictions made by the model out of the total number of predictions generated by the model.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

Error Rate is another performance metric that goes hand in hand with accuracy. After training the classifier with a given dataset, it is the total number of inaccurate predictions made by the model as a percentage of the total number of predictions.

$$Error\ Rate = 1 - Accuracy$$

Sensitivity is the next performance indicator. It counts how many cases the classifier accurately predicted as positive out of the total number of instances that are actually positive. Recall or True Positive Rate are other terms for sensitivity.

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity is the last performance metric that is used. It counts how many of the classifier's negative predictions were correct out of the total number of cases that were truly negative. True Negative Rate is another name for it.

$$Specificity = \frac{TN}{TN + FP}$$

To predict the presence of CAD, I started with three fundamental machine learning models: Random Forest Tree Classification, Decision Tree Algorithm, and K-Nearest Neighbor Algorithm (KNN). The intuition was that the outputs of the basic models would be simple to interpret and explain to a non-machine learning audience.

5.3 Apply K -Nearest Neighbor Algorithm (KNN) Algorithm

Apply K -Nearest Neighbor Algorithm (KNN) algorithm for the model and calculate the confusion matrix to see how many correct and incorrect predictions are through the confusion matrix.

```
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, Pred_y_1)
plot_confusion_matrix(cm, normalize = False,target_names = ['Has CAD', 'NO CAD'], title = "Confusion Matrix")
```

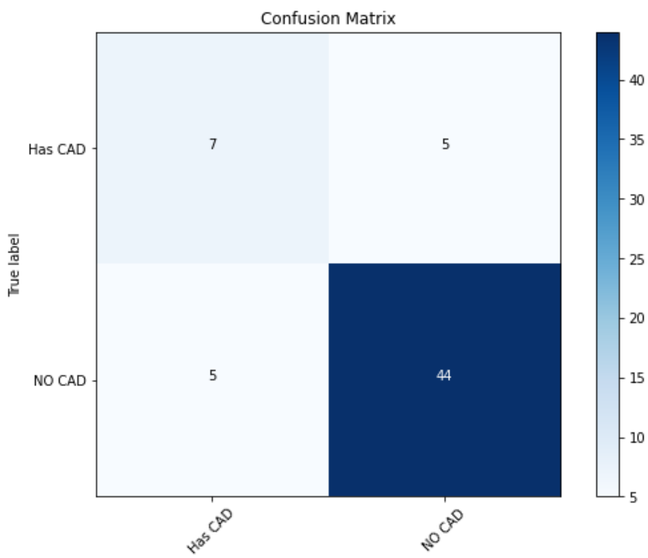

Figure 5:1 Confusion matrix -KNN Algorithm

```
Accuracy = (cm[0,0]+cm[1,1])/(cm[0,0]+cm[0,1]+cm[1,0]+cm[1,1])
print('Accuracy : ', Accuracy )

Error_Rate = 1-Accuracy
print('Error_Rate : ', Error_Rate )

Sensitivity = cm[0,0]/(cm[0,0]+cm[0,1])
print('Sensitivity : ', Sensitivity)

Specifity = cm[1,1]/(cm[1,1]+cm[0,1])
print('Specifity : ', Specifity)
```

```
Accuracy :  0.8360655737704918
Error_Rate :  0.16393442622950816
Sensitivity :  0.5833333333333334
Specifity :  0.8979591836734694
```

Figure 5:2 KNN Algorithm - Accuracy, Error rate, Sensitivity, Specificity

5.4 Apply Random Forest Tree Classification Algorithm

Apply random forest tree classification algorithm for the model and calculate the confusion matrix to see how many correct and incorrect predictions are through the confusion matrix.

```
plot_confusion_matrix(cm,normalize= False,target_names = ['Has CAD', 'NO CAD'],title = "Confusion Matrix")
```

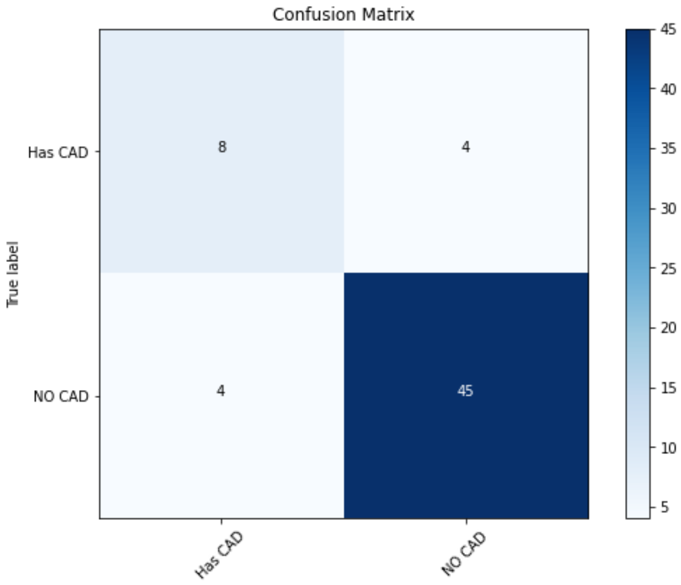

Figure 5:3 Confusion matrix -Random Forest Algorithm

```
Accuracy = (cm[0,0]+cm[1,1])/(cm[0,0]+cm[0,1]+cm[1,0]+cm[1,1])
print('Accuracy : ', Accuracy )

Error_Rate = 1-Accuracy
print('Error_Rate : ', Error_Rate )

Sensitivity = cm[0,0]/(cm[0,0]+cm[0,1])
print('Sensitivity : ', Sensitivity)

Specifity = cm[1,1]/(cm[1,1]+cm[0,1])
print('Specifity : ', Specifity)
```

```
Accuracy :  0.8688524590163934
Error_Rate :  0.1311475409836066
Sensitivity :  0.6666666666666666
Specifity :  0.9183673469387755
```

Figure 5:4 Random Forest Algorithm - Accuracy, Error rate, Sensitivity, Specificity

5.5 Apply Decision Tree Classification algorithm

Apply decision tree classification algorithm for the model and calculate the confusion matrix to see how many correct and incorrect predictions are through the confusion matrix.

```
plot_confusion_matrix(cm, normalize = False,target_names = ['Has CAD', 'NO CAD'],title = "Confusion Matrix")
```

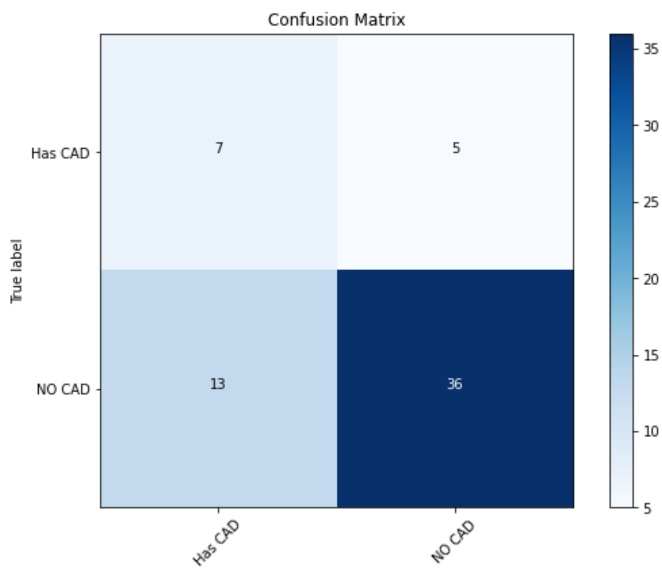

Figure 5:5 Confusion matrix - Decision Tree Algorithm

```
Accuracy = (cm[0,0]+cm[1,1])/(cm[0,0]+cm[0,1]+cm[1,0]+cm[1,1])
print('Accuracy : ', Accuracy )

Error_Rate = 1-Accuracy
print('Error_Rate : ', Error_Rate )

Sensitivity = cm[0,0]/(cm[0,0]+cm[0,1])
print('Sensitivity : ', Sensitivity)

Specifity = cm[1,1]/(cm[1,1]+cm[0,1])
print('Specifity : ', Specifity)
```

```
Accuracy :  0.7049180327868853
Error_Rate :  0.29508196721311475
Sensitivity :  0.5833333333333334
Specifity :  0.8780487804878049
```

Figure 5:6 Decision Tree Algorithm - Accuracy, Error rate, Sensitivity, Specificity

However, after analyzing the information, I discovered that the Random Forest Tree Classification model is more susceptible to over-fitting than the other two algorithms.

Figure 1 summarizes the classification accuracy findings for the three classification techniques Random Forest Tree Classification, Decision Tree Algorithm, and K -Nearest Neighbor Algorithm (KNN).

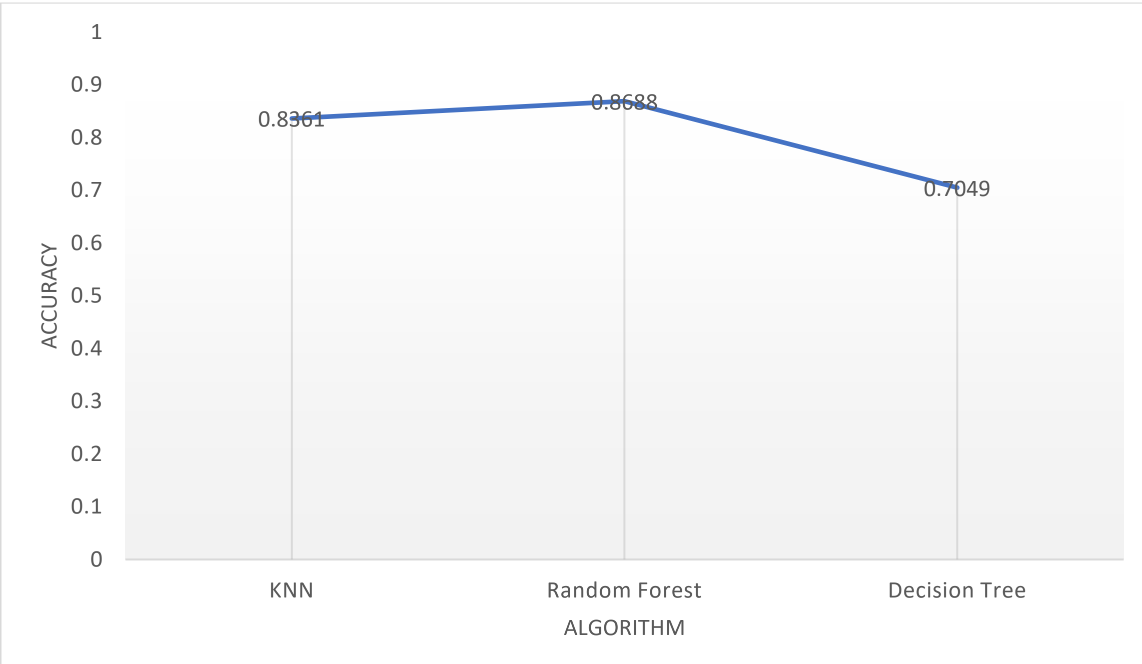

Figure 5:7 Accuracy Summarization

Random Forest Tree Classification surpasses the Decision Tree Algorithm and the K-Nearest Neighbor Algorithm (KNN) in terms of accuracy, according to the results. While all three models have accuracy rates of over 80%, accuracy cannot be used as the sole performance criterion for the underlying research.

Because the labeled class's bi-variable response is uneven, this is the case. Only 216 of the original 303 patients are believed to have CAD, whereas the remaining 87 are said to be free of the disease. This mismatch may have a significant impact on the accuracy rate since the model can anticipate all values in the majority class, resulting in a high overall accuracy while blinding out mispredictions in the minority class. We recorded other parameters such as sensitivity and specificity to avoid this imbalance impacting our performance measurement.

6.0 CONCLUSION AND FUTURE WORK

6.3 Conclusion

With the increasing number of deaths due to coronary artery disease (CAD), 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 detection of coronary artery disease (CAD) and also discusses the viable machine learning algorithm-based web-based system and mobile application for the prediction of coronary artery disease (CAD) diagnosis accurately. This study compares the accuracy score of Random Forest Tree Classification, Decision Tree Algorithm and K -Nearest Neighbor Algorithm (KNN) algorithms for predicting heart disease using UCI machine learning repository dataset. The result of this study indicates that the Random Forest Tree algorithm is the most efficient algorithm with accuracy score of 86% for prediction of coronary artery disease (CAD).

6.4 Future Work

In the future, the work can be improved by enhance this web application for various types of heart disease prediction based on different algorithms and using a larger dataset than the one used in this analysis, which will help to provide better results and assist health professionals in effectively and efficiently predicting heart disease.

REFERENCES

Abdar, M., Książek, W., U. R. Acharya & R. S. Tan, 2019. A new machine learning technique for an accurate diagnosis of coronary artery disease. *Computer Methods and Programs in Biomedicine,* Volume 179.

Alizadehsani, R. et al., 2012. Diagnosis of Coronary Artery Disease Using Data Mining Based on Lab Data and Echo Features. *Journal of Medical and Bioengineering,* 1(1), pp. 26-29.

Alizadehsani, R. et al., 2012. Diagnosis of coronary arteries stenosis using data mining. *Journal of Medical Signals & Sensors,* 2(3).

Haq, A. U. et al., 2018. A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms. *Mobile Information Systems,* 2 12.pp. 1-21.

Kirmani, M., 2017. Cardiovascular Disease Prediction using Data Mining Techniques. *Oriental journal of computer science and technology,* 26 5, 10(2), pp. 520-528.

Kiruthika Devi, S., Krishnapriya, S. & Kalita, D., 2016. Prediction of Heart Disease using Data Mining Techniques. *Indian Journal of Science and Technology,* 24 10.9(39).

Makwana, A. & Patel, J., 2015. Decision Support System for Heart Disease Prediction using Data Mining Techniques. *International Journal of Computer Applications,* 20 5, 117(22), pp. 1-5.

Marjia, S. & Afrin Haider, 2017. *Heart disease prediction using WEKA tool and 10-Fold cross-validation.* s.l., The Institute of Electrical and Electronics Engineers.

Mursal Furqan, H. R. N., 2020 December. Heart Disease Prediction using Machine Learning Algorithms.

Ngure, K., 2019. Heart Disease Prediction System.

Polaraju, K. & D, D. P., 2017. *Prediction of heart disease using multiple linear,* s.l.: s.n.

Pradeep Gupta, S. V., 2020 July. Heart Disease Prediction System Using Classification Algorithms.

S. Mohan, C. Thirumalai & G. Srivastava, 2019. Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access,* Volume 19 June 2019, pp. 81542 - 81554.

Saadatfar, H. et al., 2020. Coronary Artery Disease Diagnosis; Ranking the Significant Features Using a Random Trees Model. *International Journal of Environmental Research and Public Health,* 23 1, 17(3), p. 731.

Senthilkumar Mohan, C. T., 2019. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *Computing and Cybersecurity for Information-Centric Internet of Things.*

Soni, J., Ansari, U., Sharma, D. & Soni, S., 2011. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. *International Journal of Computer Applications,* 31 3, 17(8), pp. 43-48.Soni, J., Ansari, U., Sharma, D. & Soni, S., 2011. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. *International Journal of Computer Applications,* 31 3, 17(8), pp. 43-48.