A Project Report

on

Heart Disease Prediction Web Application using Machine Learning

Submitted in partial fulfillment of the

requirement for the award ofthe degree of

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Under The Supervision of
Dr. Shrddha Sagar
Professor
Department of Computer Science and Engineering

Submitted By

Nikhil Singh
18021011655/18SCSE1010424

Rajat Kumar Soni
18021011803/18SCSE1010577

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT
OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA OCTOBER- 2021

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

I/We hereby certify that the work which is being presented in the thesis/project/dissertation, entitled "**Heart Disease Prediction Web Application using Machine Learning**" in partial fulfillment of the requirements for the award of the "**Bachelor of Technology**" submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of month, Year to Month and Year, under the supervision of **Dr. Shrddha Sagar(**Professor) Department of Computer Science and Engineering/Computer Application and Information and Science, of School of Computing Science and Engineering , Galgotias University, Greater Noida

The matter presented in the thesis/project/dissertation has not been submitted by me/us for the award of any other degree of this or any other places.

Nikhil Singh, 18SCSE1010424

Rajat Kumar Soni, 18SCSE1010577

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Dr. Shrddha Sagar

Professor

CERTIFICATE

The Final Thesis/Project/ Dissertation Viva-Voce examination of Nikhil Singh and Rajat Kumar Soni has been held on 20/12/2021 and his/her work is recommended for the award of Bachelor of Technology (CSE).

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

Date: December, 2021

Place: Greater Noida

Acknowledgement

We cannot express enough thanks to my committee for their continued support and encouragement: **Dr. Shrdha Sagar** (Project Mentor) and **Dr. Vishwadeepak Singh Baghela** (Project Reviewer). We offer my sincere appreciation for the learning opportunities provided by my committee.

Our completion of this project could not have been accomplished without the support of our project partners. This wouldn't have been possible if you people wouldn't have had my back at stressful times.

Finally, to our caring parents who took great care of us in these tough times. Your encouragement when the times got rough are much appreciated and duly noted. It was a great comfort and relief to know that you were willing to provide management of our household activities while We completed our work.

My heartfelt thanks!

Nikhil Singh 18SCSE1010424

Rajat Kumar Soni 18SCSE1010577

TABLE OF CONTENTS

S.No	Particulars	Page No
1	Abstract	
2	Literature Reviews/Comparative study	
3	Problem Formulation	
4	Required tools	
5	Complete work plan layout	
6	References	

ABSTRACT

The Heart Disease Prediction application is an end user support and online consultation project. Here, we propose an application that allows users to get instant guidance on their heart disease through an intelligent system web application online. The application is fed with various details and the heart disease associated with those details. As the death rate is increasing due to coronary diseases, the people of healthcare department depend largely on the patient's data to predict if the patient may have a risk of heart disease. Not every time can the doctors go through every minute detail of the data and predict accurately. It is time consuming and risky. The aim of this project is to find best predicting algorithm which can help the non-specialized doctors or medical technicians in predicting the risk of disease. The prediction system uses different machine learning algorithms like logistic regression, support vector machine, k-nearest neighbor, Gaussian naïve Bayes, decision tree classifier and random forest classifier. This whole system is introduced by the simple and easy integration of web integrated with Machine learning algorithm using Support vector machines using Python. In this work we will use Machine Learning Classification and ensemble techniques on a dataset to predict heart disease. Which is K-Nearest Neighbor (KNN), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM). The accuracy is different for every model when compared to other models. The Project work gives the accurate or higher accuracy model shows that the model can predict heart disease effectively Many researchers, in recent times, have been using several machine learning techniques to help the health care industry and the professionals in the diagnosis of heart-related diseases. Heart is the next major organ comparing to the brain which has more priority in the Human body. It pumps the blood and supplies it to all organs

of the whole body. Prediction of occurrences of heart diseases in the medical field is significant work. Data analytics is useful for prediction from more information and it helps the medical center to predict various diseases. A huge amount of patient-related data is maintained on monthly basis. The stored data can be useful for the source of predicting the occurrence of future diseases. Some of the data mining and machine learning techniques are used to predict heart diseases, such as Artificial Neural Network (ANN), Random Forest,and Support Vector Machine (SVM).Prediction and diagnosingof heart disease become a challenging factor faced by doctors and hospitals both in India and abroad. To reduce the large scale of deaths from heart diseases, a quick and efficient detection technique is to be discovered. Data mining techniques and machine learning algorithms play a very important role in this area. The outcomes are promising compared to existing machine learning approaches and other research models.

INTRODUCTION

According to the World Health Organization, every year 12 million deaths occur worldwide due to Heart Disease. The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researches have been conducted in attempt to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications. This project aims to predict future Heart Disease by analyzing data of patients which classifies whether they have heart disease or not using machine-learning algorithms.

1.1 Problem Definition

The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either they are expensive or are not efficient to calculate chance of heart disease in human. Early detection of cardiac diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time and expertise. Since we have a good amount of data in today's world, we can use various machine learning algorithms to analyze the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data.

1.2 Motivation

Machine learning techniques have been around us and has been compared and used for analysis for many kinds of data science applications. The major motivation behind this research-based project was to explore the feature selection methods, data preparation and processing behind the training models in the machine learning. With first hand models and libraries, the challenge we face today is data where beside their abundance, and our cooked models, the accuracy we see during training, testing and actual validation has a higher variance. Hence this project is carried out with the motivation to explore behind the models, and further implement Logistic Regression model to train the obtained data. Furthermore, as the whole machine learning is motivated to develop an appropriate computer-based system and decision support that can aid to early detection of heart disease, in this project we have developed a model which classifies if patient will have heart disease in ten years or not based on various features (i.e. potential risk factors that can cause heart disease) using logistic regression. Hence, the early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine.

1.3 Objectives

The main objective of developing this project are:

1. To develop machine learning model to predict future possibility of heart disease by implementing Logistic Regression.
2. To determine significant risk factors based on medical dataset which may lead to heart disease.
3. To analyze feature selection methods and understand their working principle.

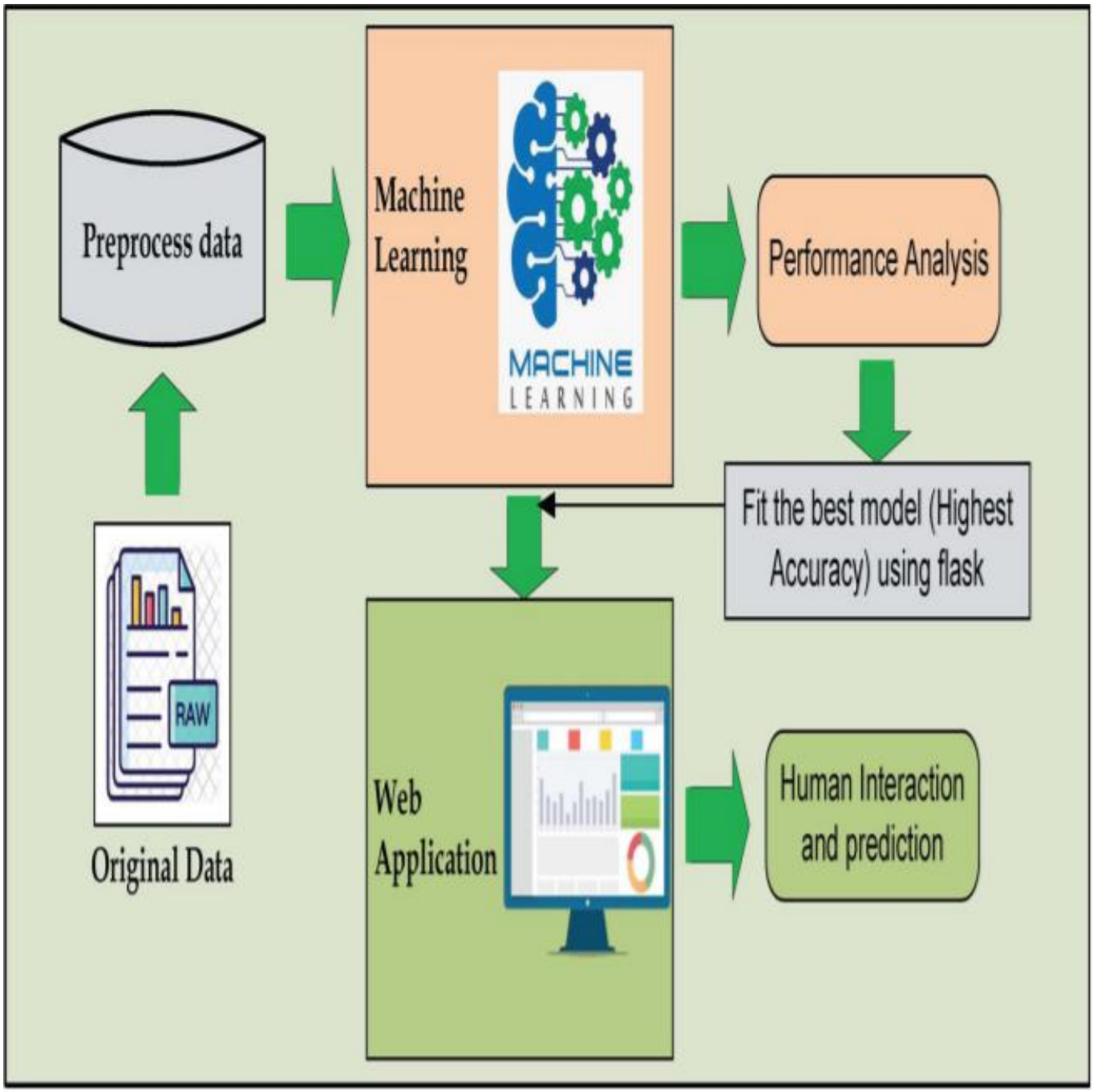


Fig. Overview of the Proposal

Medical organizations, all around the world, collect data on various health-related issues. These data can be exploited using various machine learning techniques to gain useful insights. But the data collected is very massive and, many times, this data can be very noisy. These datasets, which are too overwhelming for human minds to comprehend, can be easily explored using various machine learning techniques. Thus, these algorithms have become very useful, in recent times, to predict the presence or absence of heart-related diseases accurately The usage of information technology in the health care industry is increasing day by day to aid doctors in decision- making activities. It helps doctors and physicians in disease management, medications, and discovery of patterns and relationships among diagnosis data. Current approaches to predict cardiovascular risk fail to identify many people who would benefit from preventive treatment, while others receive unnecessary intervention. Machine-learning offers an opportunity to improve accuracy by exploiting complex interactions between risk factors. We assessed whether machine-learning can improve cardiovascular risk prediction

The work proposed in this paper focus mainly on various data mining practices that are employed in heart disease prediction. Human heart is the principal part of the human body. Basically, it regulates blood flow throughout our body. Any irregularity to heart can cause distress in other parts of body. Any sort of disturbance to normal functioning of the heart can be classified as a Heart disease. In todays contemporary world, heart disease is one of the primary reasons for occurrence of most deaths. Heart disease may occur due to unhealthy lifestyle, smoking, alcohol and high intake of fat which may cause hypertension [2]. According to the World Health Organization more than 10 million die due to Heart diseases every single

year around the world. A healthy lifestyle and earliest detection are only ways to prevent the heart related diseases.

The main challenge in today's healthcare is provision of best quality services and effective accurate diagnosis [1]. Even if heart diseases are found as the prime source of death in the world in recent years, they are also the ones that can be controlled and managed effectively. The whole accuracy in management of a disease lies on the proper time of detection

of that disease. The proposed work makes an attempt to detect these heart diseases at early stage to avoid disastrous consequences.

Records of large set of medical data created by medical experts are available for analysing and extracting valuable knowledge from it. Data mining techniques are the means of extracting valuable and hidden information from the large amount of data available. Mostly the medical database consists of discrete information. Hence, decision making using discrete data becomes complex and tough task. Machine Learning (ML) which is subfield of data mining handles large scale well-formatted dataset efficiently. In the medical field, machine learning can be used for diagnosis, detection and prediction of various diseases. The main goal of this paper is to provide a tool for doctors to detect heart disease as early stage [5]. This in turn will help to provide effective treatment to patients and avoid severe consequences. ML plays a very important role to detect the hidden discrete patterns and thereby analyse the given data. After analysis of data ML techniques help in heart disease prediction and early diagnosis. This paper presents performance analysis of various ML techniques such as Naive Bayes, Decision Tree, Logistic Regression and Random Forest for predicting heart disease at an early stage [3].

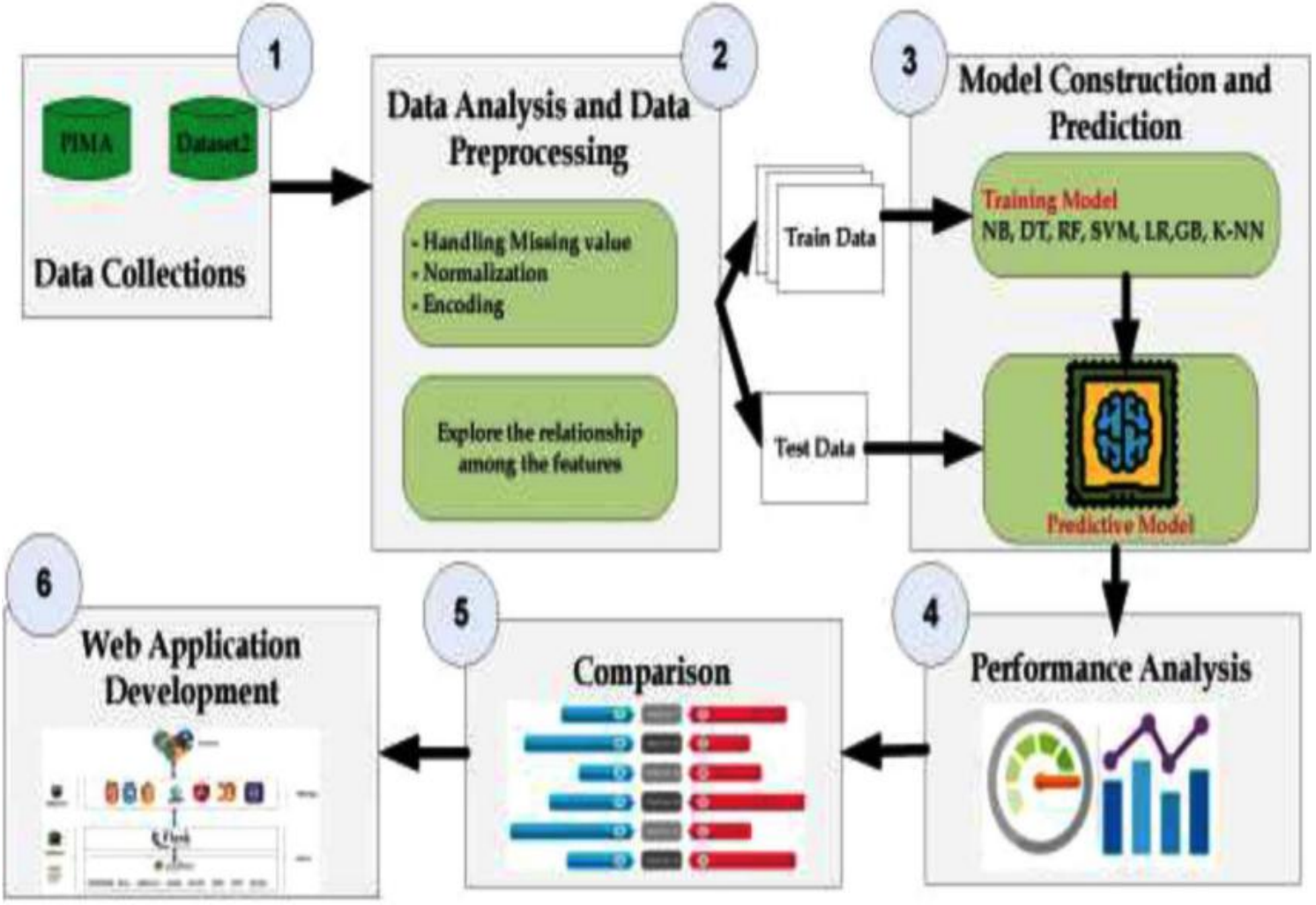


Fig. Workflow Diagram of the Proposal

Literature Reviews/Comparative Study

ChalaBeyene et al[1], recommended Prediction and Analysis of the occurrence of Heart Disease Using Data Mining Techniques. The main objective is to predict the occurrence of heart disease for early automatic diagnosis of the disease within result in a short time. The proposed methodology is also critical in a healthcare organization with experts that have no more knowledge and skill. It uses different medical attributes such as blood sugar and heart rate, age, sex are some of the attributes are included to identify if the person has heart disease or not. Analyses of the dataset are computed using WEKA software.

Senthilkumar Mohan et al[2],implemented hybrid machine learning for heart disease prediction. The data set used is Cleveland data set. The first step is data pre-processing step. In this the tuples are removed from the data set which has missed the values. Attributes age and sex from data set are also not used as the authors think that it's personal information and has no impact on predication. The remaining 11 attributes are considered important as they contain vital clinical records. They have proposed their own Hybrid Random Forest Linear Method (HRFLM) which is the combination of Random Forest (RF) and Linear method (LM). In the HRFLM algorithm, the authors have used four algorithms.

First algorithm deals with partitioning the input dataset. It is based on a decision tree which is executed for each sample of the dataset. After identifying the feature space, the dataset is split into the leaf nodes. Output of first algorithm is Partition of data set.

After that in second algorithm they apply rules to the data set and output

here is the classification of data with those rules. In third algorithm features are extracted using Less Error Classifier. This algorithm deals with finding the minimum and maximum error rate from the classifier. Output of this algorithm is the features with classified attributes.

In forth algorithm they apply Classifier which is hybrid method based on the error rate on the Extracted Features. Finally they have compared the results obtained after applying HRFLM with other classification algorithms such a decision tree and support vector machine. In result as RF and LM are giving better results than other, both the algorithms are put together and new unique algorithm HRFLM is created.

The authors suggest further improvement in accuracy by using combination of various machine learning algorithms. Ali, Liaqat, et al[3],propose a system containing two models based on linear Support Vector Machine (SVM). The first one is called L1 regularized and the second one is called L2 regularized. First model is used for removing unnecessary features by making coefficient of those features zero. The second model is used for prediction. Predication of disease is done in this part.

To optimize both models they proposed a hybrid grid search algorithm. This algorithm optimizes two models based on metrics: accuracy, sensitivity, septicity, the Matthews correlation coefficient, ROC chart and area under the curve. They used Cleveland data set. Data splits into 70% training and 30% testing used holdout validation. There are two experiments carried out and each experiment is carried out for various values of C1, C2 and k where C1 is hyperparameter of L1 regularized model, C2 is hyperparameter of L2 regularized model and k is the size of selected subset of features. First experiment is L1-linear SVM model stacked with

L2-linear SVM model which is giving maximum testing accuracy of 91.11% and training accuracy of 84.05%.

The second experiment is L1-linear SVM model cascaded with L2-linear SVM model with RBF kernel. This is giving maximum testing accuracy of 92.22% and training accuracy of 85.02.They have obtained an improvement in accuracy over conventional SVM models by 3.3%.

Many researches has been conducted by our researchers on this topic in order to investigate the disease and improve its accuracy. Some of them are as shown below.

One of them was “A Detail Set Cleveland Heart Disease. In this study the data was only for the testing purpose but not the training purpose. In this there were 303 cases and 75 marks, but the tests which were published contained only of the 14 subsets and not the whole of them.

But in comparison to this our work consisted of a number of preliminary processing of our data sets and the ones which were so termed as the missing values were removed for ignoring the inconsistency. Along width this we also provided the matrix data comparing our label outputs 0 and 1 describing the presence of heart disease and absence of it by 1 and 0 respectively.

The different Authors in the table provided their researches and their projects on the heart disease prediction and they used different algorithms like decision tree classifier, ANN, different data mining techniques and including the SV in parallel fashion. But were not very able to achieve the greater accuracy results.

Related Works

With growing development in the field of medical science alongside machine learning various experiments and researches has been carried out in these recent years releasing the relevant significant papers. The paper [1] propose heart disease prediction using KStar, J48, SMO, and Bayes Net and Multilayer perceptron using WEKA software. Based on performance from different factor SMO (89% of accuracy) and Bayes Net (87% of accuracy) achieve optimum performance than KStar, Multilayer perceptron and J48 techniques using k-fold cross validation. The accuracy performance achieved by those algorithms are still not satisfactory. So that if the performance of accuracy is improved more to give batter decision to diagnosis disease.

[2] In a research conducted using Cleveland dataset for heart diseases which contains 303 instances and used 10-fold Cross Validation, considering 13 attributes, implementing 4 different algorithms, they concluded Gaussian Naïve Bayes and Random Forest gave the maximum accuracy of 91.2 percent.

[3] Using the similar dataset of Framingham, Massachusetts, the experiments were carried out using 4 models and were trained and tested with maximum accuracy K Neighbors Classifier: 87%, Support Vector Classifier: 83%, Decision Tree Classifier: 79% and Random Forest Classifier: 84%.

Feature Selection Techniques

Feature selection is the process of selecting a subset of the most relevant features in the dataset to describe the target variable. It improves computation time, generalization performance, and interpretational issues in ML problems [26,27]. Feature selection techniques are categorized as filter based, wrapper based, and embedded type. Filterbased techniques screen out

features based on some specified criteria. Wrapper-based methods use a modelling algorithm that is taken as a black box to evaluate and rank features. The embedded methods have built-in feature selection approaches such as least absolute shrinkage and selection operator (Lasso) and random forest (RF) feature selection methods [28]. There are several types of feature selection techniques, including exhaustive search, Pearson correlation technique, chi-squared technique, recursive feature elimination, Lasso, and treebased feature selection techniques. In this study, we used a data-driven feature extraction technique, which combines the ANOVA test, chisquared test, and a tree-based recursive feature elimination technique.

Analysis of Variance

Analysis of variance (ANOVA) is a well-known statistical method to determine whether here is a difference in means between two groups [29]. In this study, the ANOVA test was utilized to select the significant numerical features in predicting the occurrence of T2D. The ANOVA test uses the F statistic for feature ranking. The larger the value of the F statistic, the better the discriminative capacity of the feature [30]. The degrees of freedom for mean square between and within is defined by d fb and d fw, respectively [31]. For all numerical features in the dataset, the F value was calculated using Equation (1) and the features with the larger value were selected.

Chi-Squared Test

The chi-squared test is a nonparametric statistical analyzing method. The technique calculates the chi-squared value using Equation (2) and selects the top n features [32]. In this work, the chi-squared test was employed to rank categorical features according to their significance in identifying the target class.

Recursive Feature Elimination

RFECV is greedy optimization algorithm which aims to find the best performing feature subset. Recursive Feature Elimination (RFE) fits a model repeatedly and removes the weakest feature until specified number of features is reached. The optimal number of features is used with RFE to score different feature subsets and select the best scoring collection of features which is RFECV. The main issue of this algorithm is that it can be expensive to run. So, it is better to reduce the number of features beforehand. Since correlated features provide the same information, such features can be eliminated prior to RFECV. To address this, correlation matrix is plotted and the correlated features are removed.

The arguments for instance of RFECV are:

a. estimator - model instance (RandomForestClassifier)
b. step - number of features removed on each iteration (1)
c. cv – Cross-Validation (StratifiedKFold)
d. scoring – scoring metric (accuracy)

Once RFECV is run and execution is finished, the features that are least important can be extracted and dropped from the dataset. Top 10 features ranked by the RFECV technique in our model listed below from least importance to highest importance.

1. prevalentStroke
2. diabetes
3. BPMeds
4. currentSmoker
5. prevalentHyp
6. male
7. cigsPerDay
8. heartrate
9. Glucose
10. diaBP

DATASETS

The dataset is publicly available on the Kaggle Website at [4] which is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. It provides patient information which includes over 4000 records and 14 attributes. The attributes include: age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting, sugar blood, resting electrocardiographic results, maximum heart rate, exercise induced angina, ST depression induced by exercise, slope of the peak exercise, number of major vessels, and target ranging from 0 to 2, where 0 is absence of heart disease. The data set is in csv (Comma Separated Value) format which is further prepared to data frame as supported by pandas library in python.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Figure 1: Original Dataset Snapshot

The education data is irrelevant to the heart disease of an individual, so it is dropped. Further with this dataset pre-processing and experiments are then carried out.

DataSet Features:

1. age - age in years
2. sex - (1 = male; 0 = female)

3. cp - chest pain type
    - 0: Typical angina: chest pain related decrease blood supply to the heart
    - 1: Atypical angina: chest pain not related to heart
    - 2: Non-anginal pain: typically esophageal spasms (non heart related)
    - 3: Asymptomatic: chest pain not showing signs of disease
4. trestbps - resting blood pressure (in mm Hg on admission to the hospital) anything above 130-140 is typically cause for concern
5. chol - serum cholestoral in mg/dl
    - serum = LDL + HDL + .2 * triglycerides
    - above 200 is cause for concern
6. fbs - (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
    - '>126' mg/dL signals diabetes
7. restecg - resting electrocardiographic results
    - 0: Nothing to note
    - 1: ST-T Wave abnormality
        - can range from mild symptoms to severe problems
        - signals non-normal heart beat
    - 2: Possible or definite left ventricular hypertrophy
        - Enlarged heart's main pumping chamber
8. thalach - maximum heart rate achieved
9. exang - exercise induced angina (1 = yes; 0 = no)
10. oldpeak - ST depression induced by exercise relative to rest looks at stress of heart during excercise unhealthy heart will stress more
11. slope - the slope of the peak exercise ST segment
    - 0: Upsloping: better heart rate with excercise (uncommon)
    - 1: Flatsloping: minimal change (typical healthy heart)
    - 2: Downslopins: signs of unhealthy heart
12. ca - number of major vessels (0-3) colored by flourosopy
    - colored vessel means the doctor can see the blood passing through
    - the more blood movement the better (no clots)
13. thal - thalium stress result
    - 1,3: normal
    - 6: fixed defect: used to be defect but ok now
    - 7: reversable defect: no proper blood movement when excercising
14. target - have disease or not (1=yes, 0=no) (= the predicted attribute)

Problem Fomulation

It might have happened so many times that you or someone yours need doctors help immediately, but they are not available due to some reason. The Heart Disease Prediction application is an end user support and online consultation project. Here, we propose a web application that allows users to get instant guidance on their Heart Disease through an intelligent system online. The application is fed with various details and the Heart Disease associated with those details. application allows user to share their Heart Disease related issues. It then processes user specific details to check for various illness that could be associated with it. Here we use some intelligent data mining techniques to guess the most accurate illness that could be associated with patient's details. Based on result, they can contact doctor accordingly for further treatment. The system allows user to view doctor's details too. The system can be used for free Heart Disease consulting online. In the pre-processing stage, correlation between attributes of the datasets is analyzed for finding useful features in detecting Heart Disease. After that, the data is divided into two sets: training and testing. The training set is utilized to develop predictive ML models using a variety of machine learning algorithms. Next, we assess the proposal's performance with respect to different metrics. Finally, the best ML model is deployed in a web application using flask. Following this, we describe the workflow of each part briefly:

1. **Data Collection:** We collected two alternative datasets, each with a different number of factors or features, to ensure the model's robust- ness. The datasets were compiled from a wide variety of sources, including Heart Disease statistics and health characteristics obtained from people around the world and from various health institutes.

2. **Data Analysis and Data Preprocessing:** Several pre-processing techniques are applied on the datasets before feeding these datasets into the machine learning model so that the performance of the model is improved. The pre-processing tasks include removing outliers and dealing with missing values, data standardization, encoding, and so on. a. Outliers Removal - Attributes' values that are beyond acceptable boundaries and have high variation from the rest of the respective attribute's value might be present in the dataset. Such at-tributes' value might degrade the machine learning algorithm's performance. To eliminate such outliers, we applied the IQR (Inter-quartile Range) approach. b. Missing value Handling - To improve model performance, the mean value of each attribute was employed for handling the missing values. c. Label Encoding - Label encoding is the process of converting the labels of text/categorical values into a numerical format that ML algorithms can interpret. For example, the categorical values of Junk food consumption status yes to '1 ′ and No to '0 ′ have been converted.

3. **Model Construction and Prediction:** To construct the predictive model, 80% of the pre-processed data has been used for training while the remaining 20% data is used for the testing purpose.

4. **Performance Analysis:** We have analyzed the results of the proposed model in terms of several performance metrics. The algorithm that provides highest prediction accuracy is selected as the best algorithm for the web application development.

5. **Performance Comparison:** In this step, the accuracy of the proposal has been compared with some recent works related to Heart Disease prediction. The performance results indicate that the proposal can improve the performance compared to the recent related research. 6. Web Application development: To develop a smart web application, we have used the Flask micro-framework and integrated the best model. To predict Heart Disease, a user is required to submit a form with necessary numbers of Heart Disease related parameters. The application uploaded in a server predicts the results using the adopted machine learning model. We describe the adopted machine learning algorithms in the following sections.

METHODS AND ALGORITHMS USED

The main purpose of designing this system is to predict the ten-year risk of future heart disease. We have used Logistic regression as a machine-learning algorithm to train our system and various feature selection algorithms like Backward elimination and Recursive feature elimination. These algorithms are discussed below in detail.

Now we've got our data split into training and test sets, it's time to build a machine learning model. We'll train it (find the patterns) on the training set. And we'll test it (use the patterns) on the test set. We're going to try 3 different machine learning models:

1. Logistic Regression
2. Nearest Neighbours Classifier
3. Support Vector machine
4. Decision Tree Classifier
5. Random Forest Classifier
6. XGBoost Classifier

Logistic Regression

Logistic Regression is a supervised classification algorithm. It is a predictive analysis algorithm based on the concept of probability. It measures the relationship between the dependent variable (TenyearCHD) and the one or more independent variables (risk factors) by estimating probabilities using underlying logistic function (sigmoid function). Sigmoid function is used as a cost function to limit the hypothesis of logistic regression between 0 and 1 (squashing) i.e. $0 \leq h_\theta(x) \leq 1$.

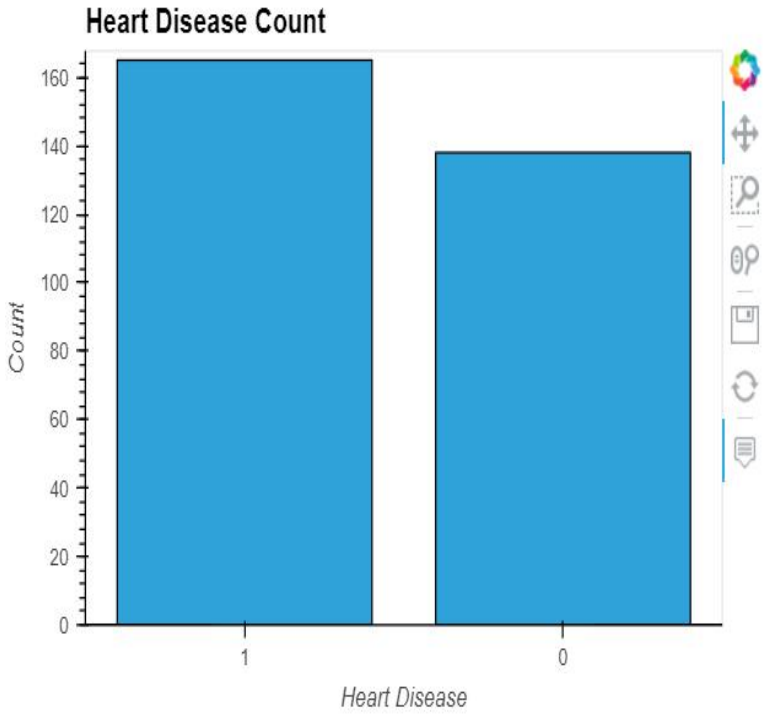
We in our project used SVM (Support Vector Machine) along with different algorithms and achieved greater result of "*93% training accuracy in SVM and of 100% in training accuracy of Decision Tree Classifier".*

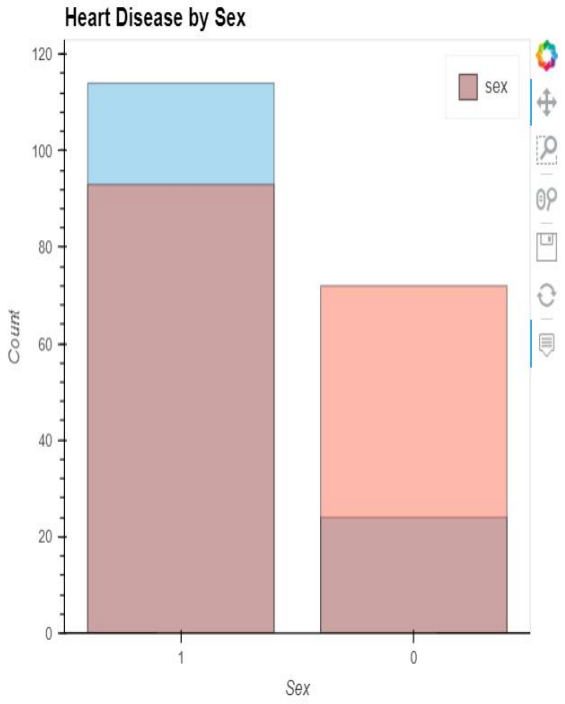
We have proposed this project to analyze the characteristics of diseases aiming to reduce physician's variabilities, ease of access and short amount of time taken.

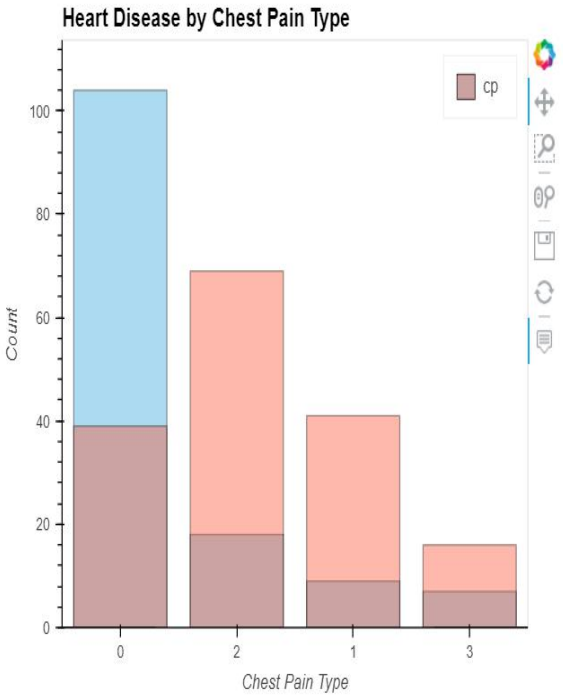
Data Preparation

Since the dataset consists of 4240 observations with 388 missing data and 644 observations to be risked for heart disease, two different experiments were performed for data preparation. First, we checked by dropping the missing data, leaving with only 3751 data and only 572 observations risked for heart disease.

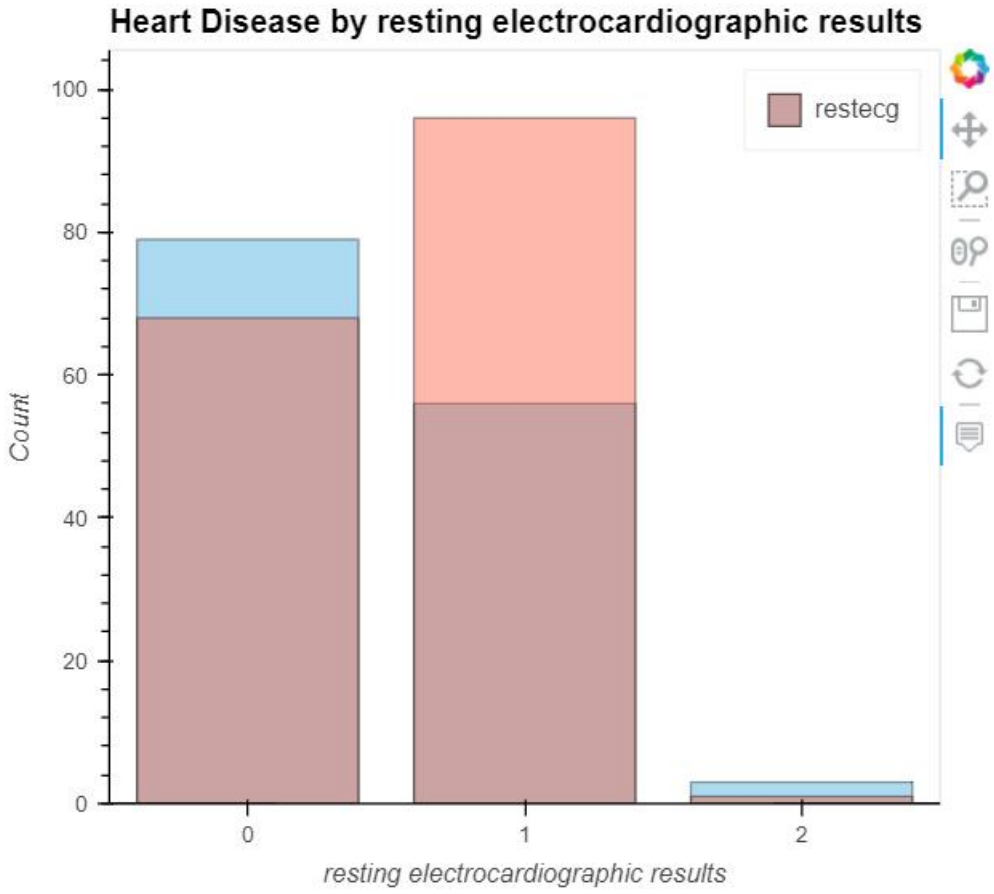
Data Exploration And Feature Extraction

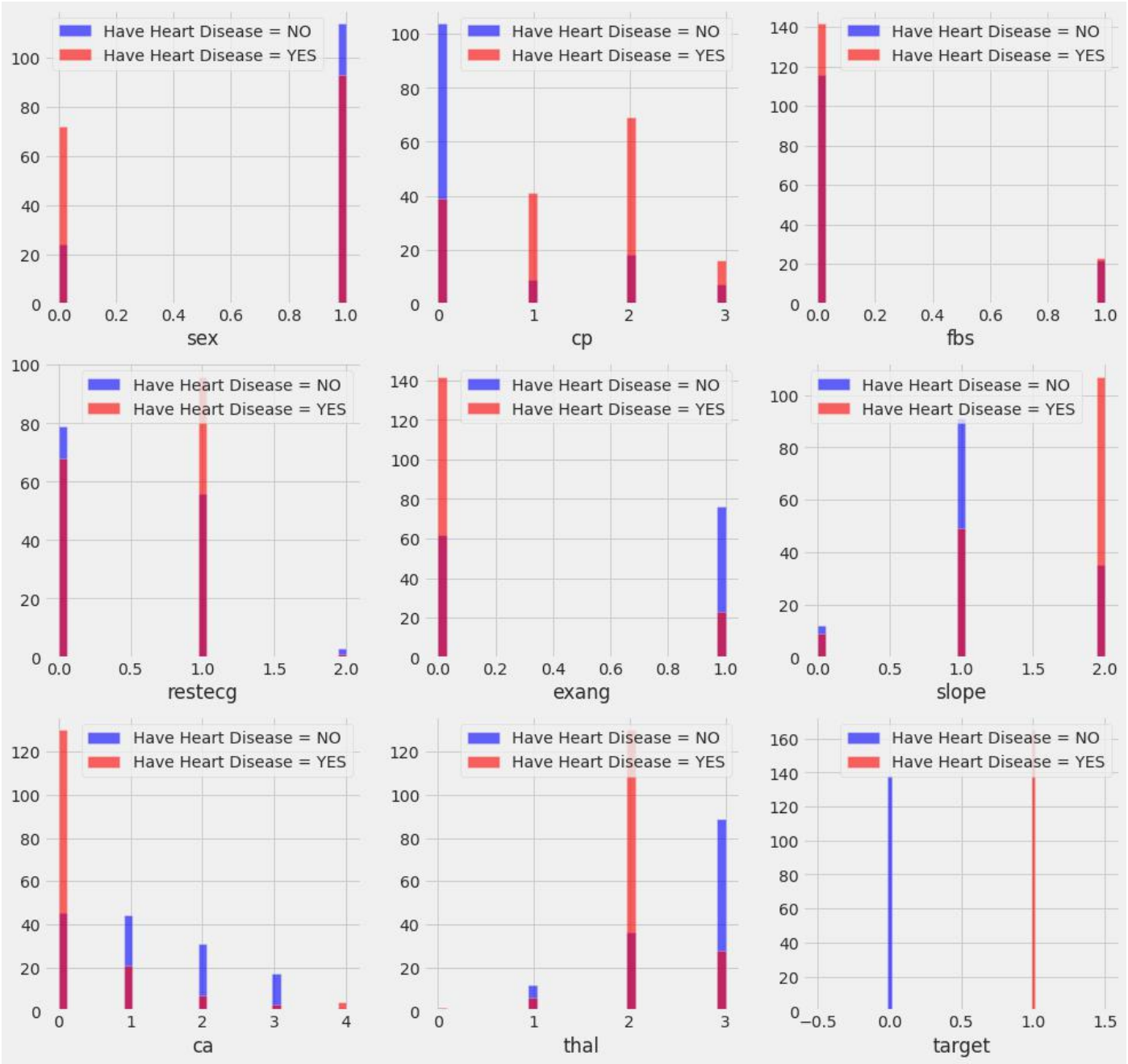

Heart Disease Count
160
140
120
100
80
60
40
20
0
Count
1
0
Heart Disease

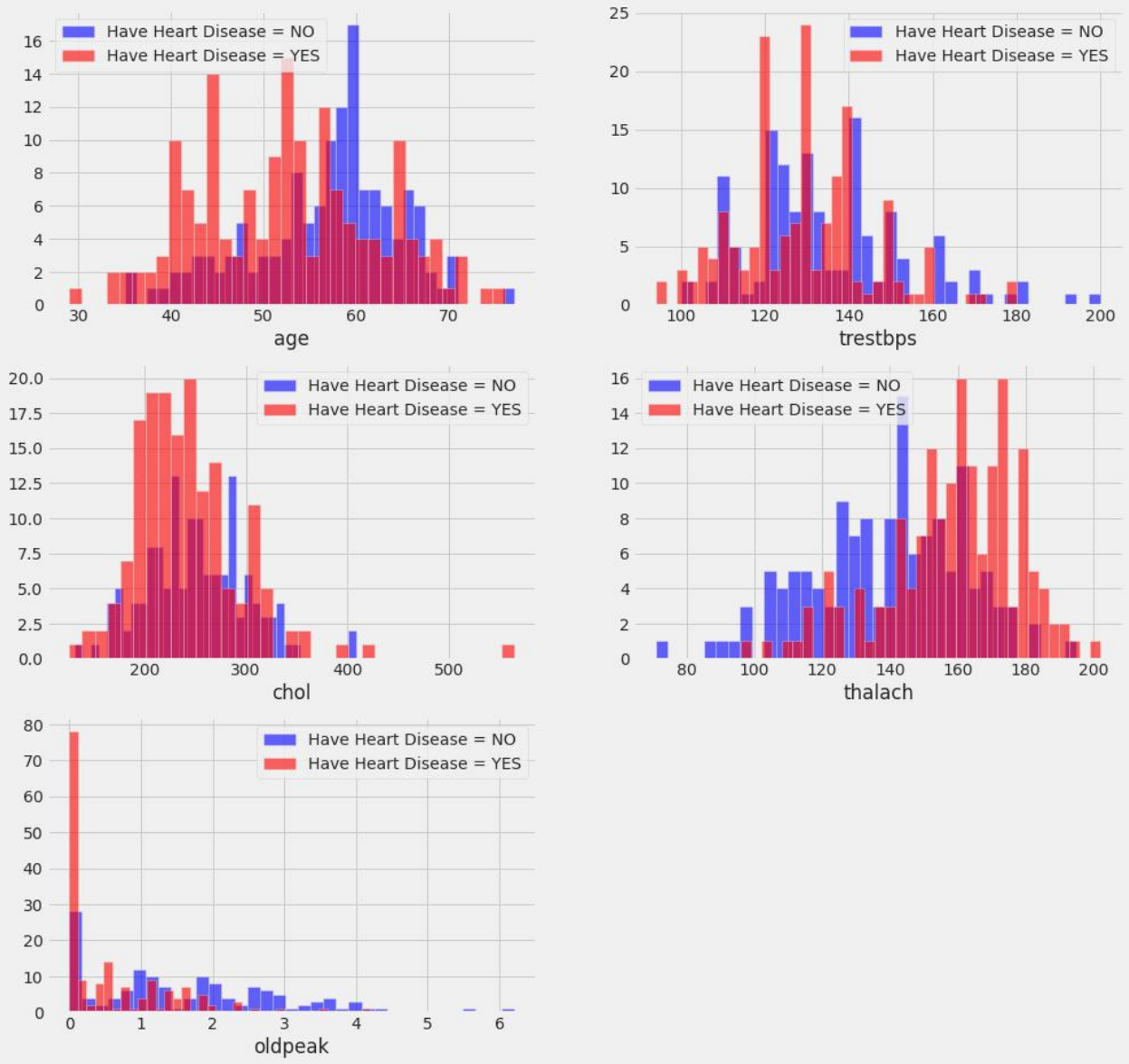

Heart Disease by Sex
sex
120
100
80
60
40
20
0
Count
1
0
Sex

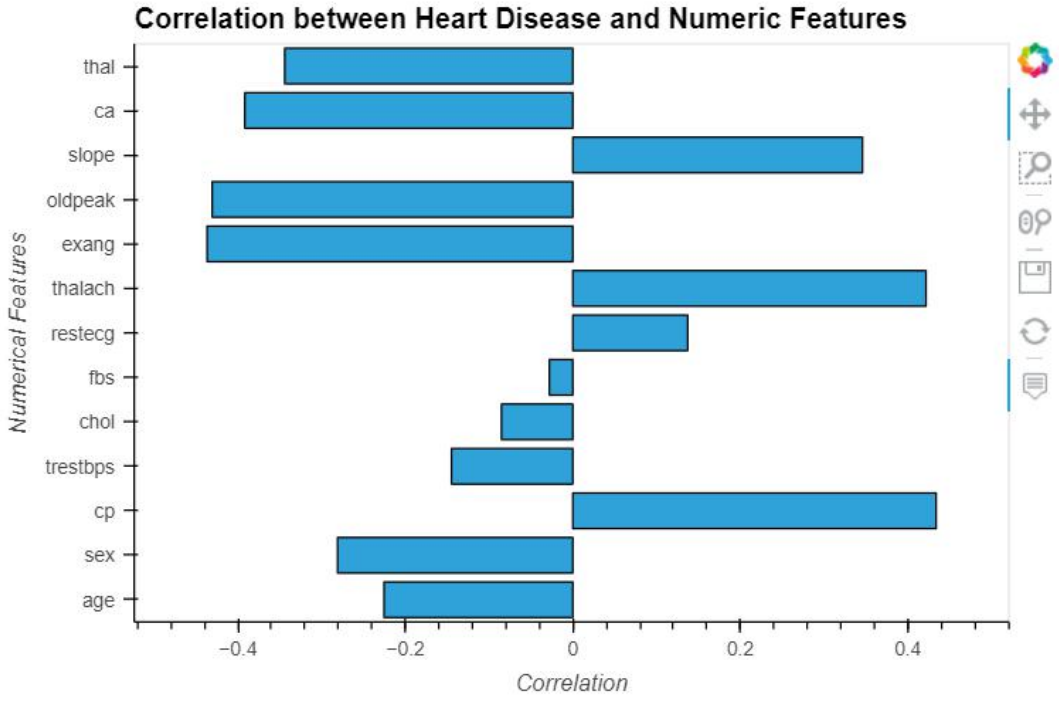

Heart Disease by Chest Pain Type
cp
100
80
60
40
20
0
Count
0
2
1
3
Chest Pain Type


Heart Disease by fasting blood sugar
fbs
Count
140
120
100
80
60
40
20
0
0
1
fasting blood sugar > 120 mg/dl (1 = true; 0 = false)


Heart Disease by resting electrocardiographic results
restecg
Count
100
80
60
40
20
0
0
1
2
resting electrocardiographic results


Have Heart Disease = NO
Have Heart Disease = YES
sex
cp
fbs
restecg
exang
slope
ca
thal
target


Have Heart Disease = NO
Have Heart Disease = YES
age
trestbps
chol
thalach
oldpeak


Correlation between Heart Disease and Numeric Features
thal
ca
slope
oldpeak
exang
thalach
restecg
fbs
chol
trestbps
cp
sex
age
Numerical Features
−0.4
−0.2
0
0.2
0.4
Correlation

Experimental results analysis

Our proposed model is tested and evaluated in this section using a variety of machine learning algorithms, including NB, DT, RF, SVM, LR, GB, and K-NN. To find the effectiveness we have used 2 different datasets and each of them contains different types and number of attributes.

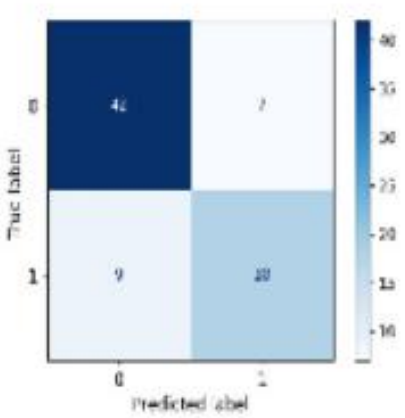
Experimental setup

The proposed model is built in Python and executes on a computer having an Intel Core i7 processor with a 4 GB graphics card, 16GB RAM and a 64-bit Windows operating system running at 1.80 GHz. To test the efficiency of our model, we have used a 10-fold cross validation process. The dataset is shuffled and divided into 10 segments at random, with one segment serving as the test set and the others serving as the training set in turn. The average of the results from multiple experiments is considered as the final output of the experiment.

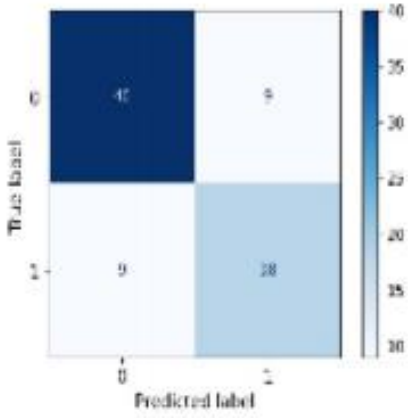
Performance metrics

The performance of the proposed approach has measured using con- fusion matrix. The confusion matrix has four different outcomes: true positive (TP), true negative (TN), false positive (FP), and false negative (FN), as follows: Next, we consider the following metrics to analysis the suggested model

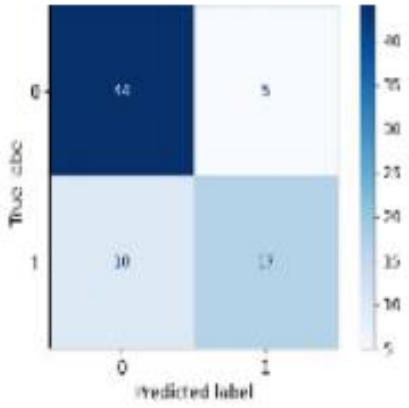
	Model	Training Accuracy %	Testing Accuracy %
0	Logistic Regression	86.79	86.81
1	K-nearest neighbors	86.79	86.81
2	Support Vector Machine	93.40	87.91
3	Decision Tree Classifier	100.00	78.02
4	Random Forest Classifier	100.00	82.42
5	XGBoost Classifier	100.00	82.42



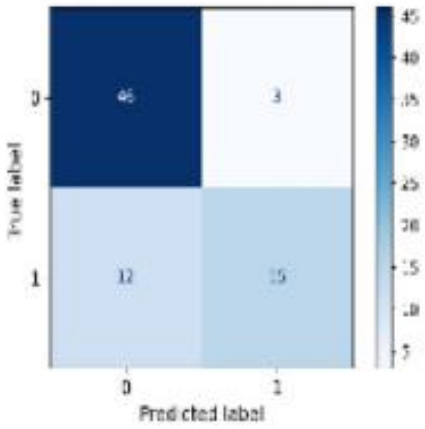
(a) Naive Bayes



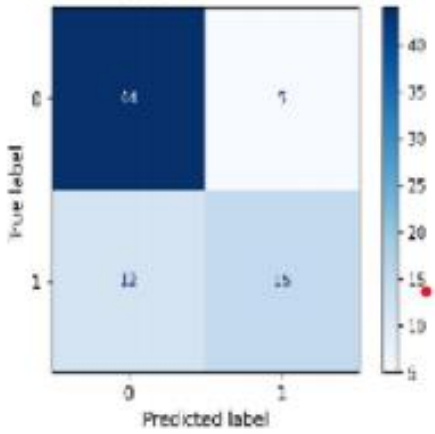
(b) Decision Tree



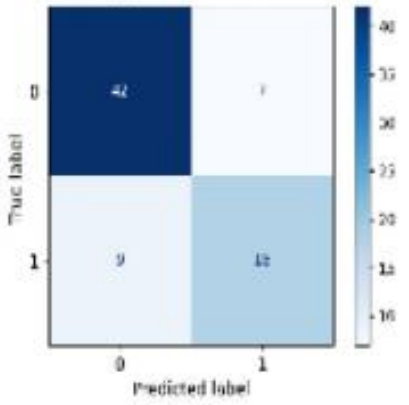
(c) Random Forest



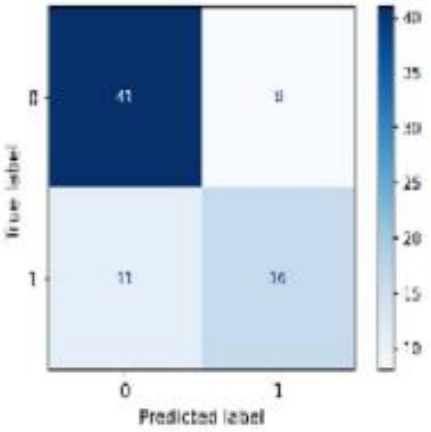
(d) Support Vector Machine



(e) Logistic Regression



(f)Gradient Boosting



(g) K-Nearest Neighbour

Web application development using flask

Flask is a Python-based microweb platform that allows users to add application functionality as if they were built into the framework itself. Fig. 11 shows the basic file structures of the developed application, and this development process comprises of four different program modules as follows: •model.pkl- This contains the machine learning model to predict diabetes. As SVM provided the highest accuracy of 78.125% with all the features, we will integrate this as predictive model in the model.pkl file.

•app.py- This package includes Flask APIs that receive Diabetes information through

GUI or API calls, compute the predicted value using our model, and return it.

•Template- The HTML form (index.html) in this folder allows the user to enter diabetes

information and shows the expected outcome.

•Static- This folder contains the css file which has the styling required for our HTML

form. The application workflow of the proposal is described in Fig. 12 has the following

steps:

•The user sends the necessary information required by the application in a Webpage

(Step-1).

•The information is sent to the back-end (Step-2:).

•The flask server adopted with the machine learning algorithm predict the results (Step-

3 and Step-4).

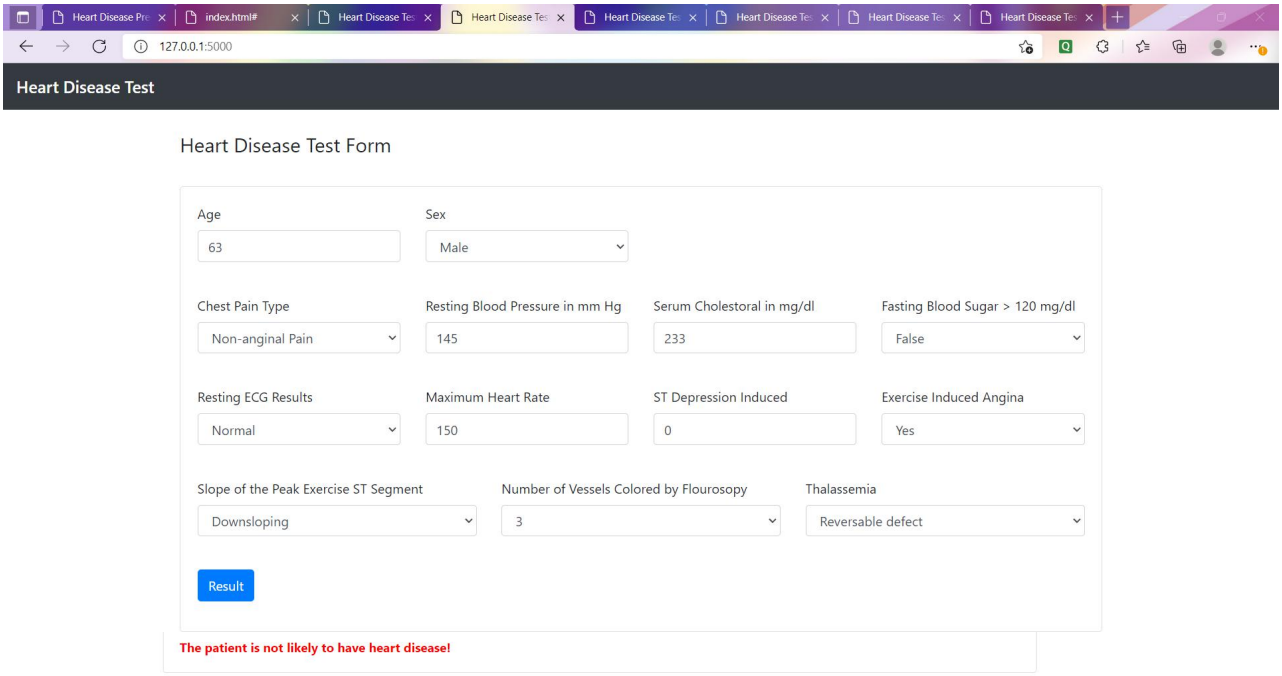
•Finally, the predicted result is shown in the webpage (Step-5).

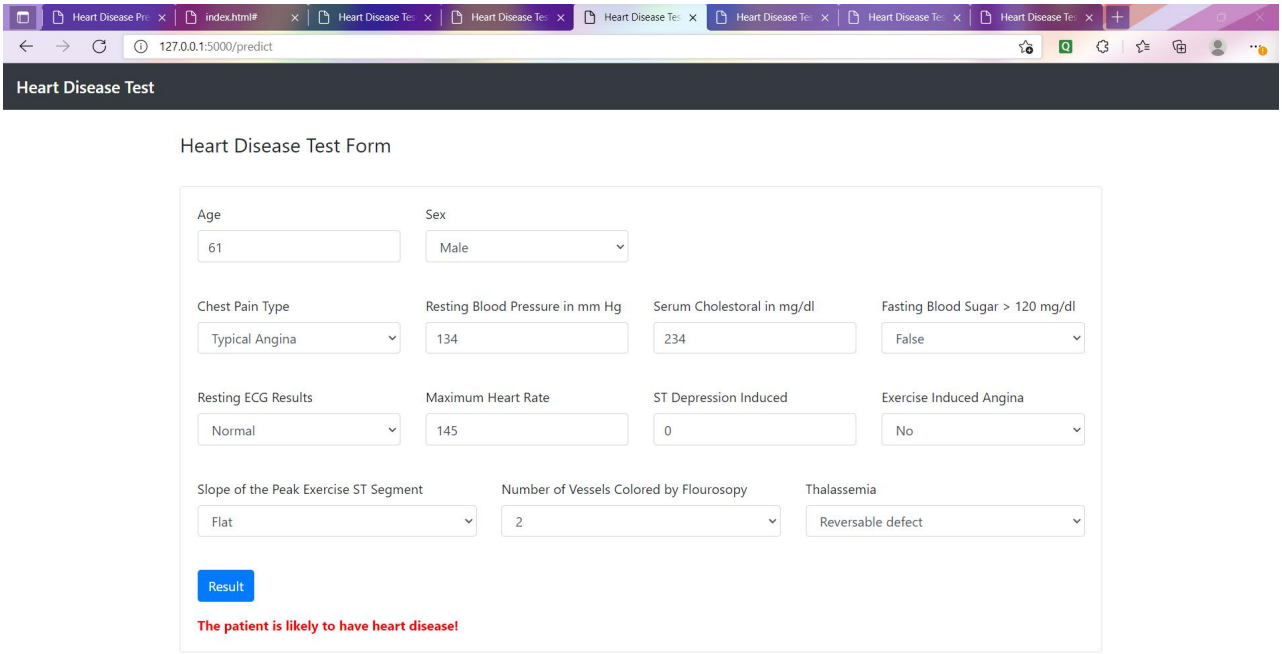
Prediction results of web application

When a user runs the application, a page will appear as shown in Fig. 13. \

The application can check for the valid input for every fields. If the user enters an invalid value for any of the parameters, a warning message is displayed. If the user provides valid information, the application will predict whether the user has diabetes or not, as illustrated in Fig. 14.

Results:





Methods

This section describes the methods used to develop a prediction model to forecast the occurrence of T2D in the following year. To generate the model, data preprocessing, feature selection, hyperparameter tuning, training, testing, and model evaluation procedures were performed

Missing-Data Handling

Data preprocessing is one of the significant steps in ML and data mining. It improves the quality of data and performance of ML models. The technique refers to cleaning and transforming the raw data to make it more suitable to train and evaluate prediction models.

Data preprocessing includes data preparation, cleaning, feature selection, missing values handling, and transformation of data. The result expected after data preprocessing is a final dataset, that can be considered correct and useful for further data mining algorithms [36]. The collected EHRs were a high dimensional dataset. It is unlikely that all the features were obtained during the medical check because the required measurements were dependent on the subjects.

To address the missing-values problem, several solutions were considered, including omission of the row with null values and replacing the missing values by mean, median, or mode values of the feature values [37]. Considering the large size of the dataset, records with null feature values were excluded from the dataset.

Prediction Model

Multiple classifiers are generated using a different combination of feature

sets and aggregated to form the final predictor. Since the ensembled methods (CIM, ST, and SV) use all available classifiers information, their performance is better and/or more robust in most applications [51]. In this study, we utilized the classifier integration model with a confusion table [52], soft voting [18], and stacking classifier models [19].

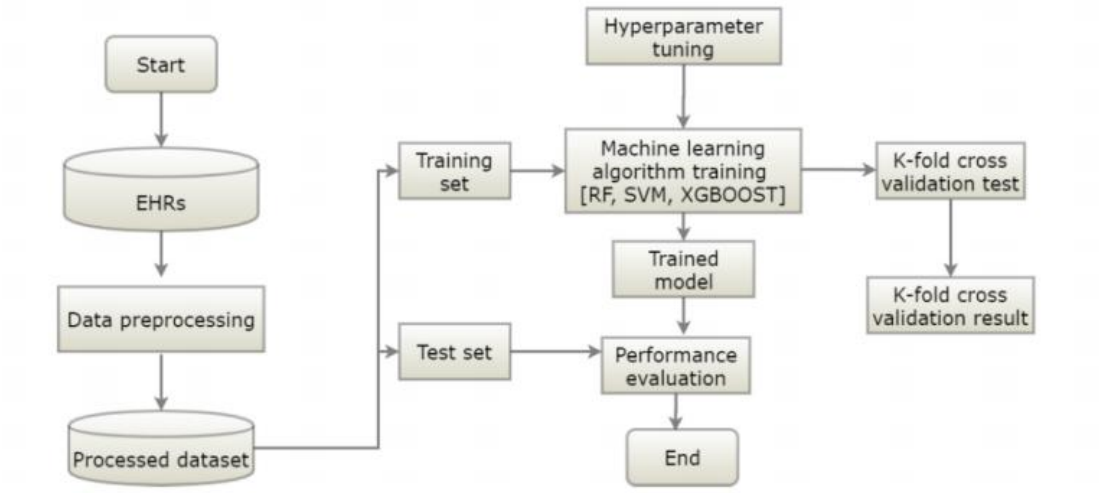


Figure 3. The architecture of the prediction model (RF = random forest, XGB = XGBoost, SVM = support vector machine).

Three sets of experiments were conducted to investigate the performance of the proposed prediction model. The first set of experiments dealt with the evaluation of the models using the test dataset and the ten-fold cross-validation (CV) technique. The CV technique randomly divided the ataset into ten subsets, and the experiments were conducted ten times iteratively.

On each iteration, one of the ten subsets was used as test data, and the emaining nine subsets were used as a training set. The second set of experiments were performed to investigate the performance of the prediction model in comparison with the number of medical follow-up years used to train the prediction model. The training dataset for the experiments was generated by concatenating the medical records over the years. The number of years used to train the dataset ranged from two to four.

The last set of experiments presented the cross-validation performance comparison between the selected 12-feature set and the well-known traditional predictors of T2D. The detailed results of the experiments are presented in Section 4.

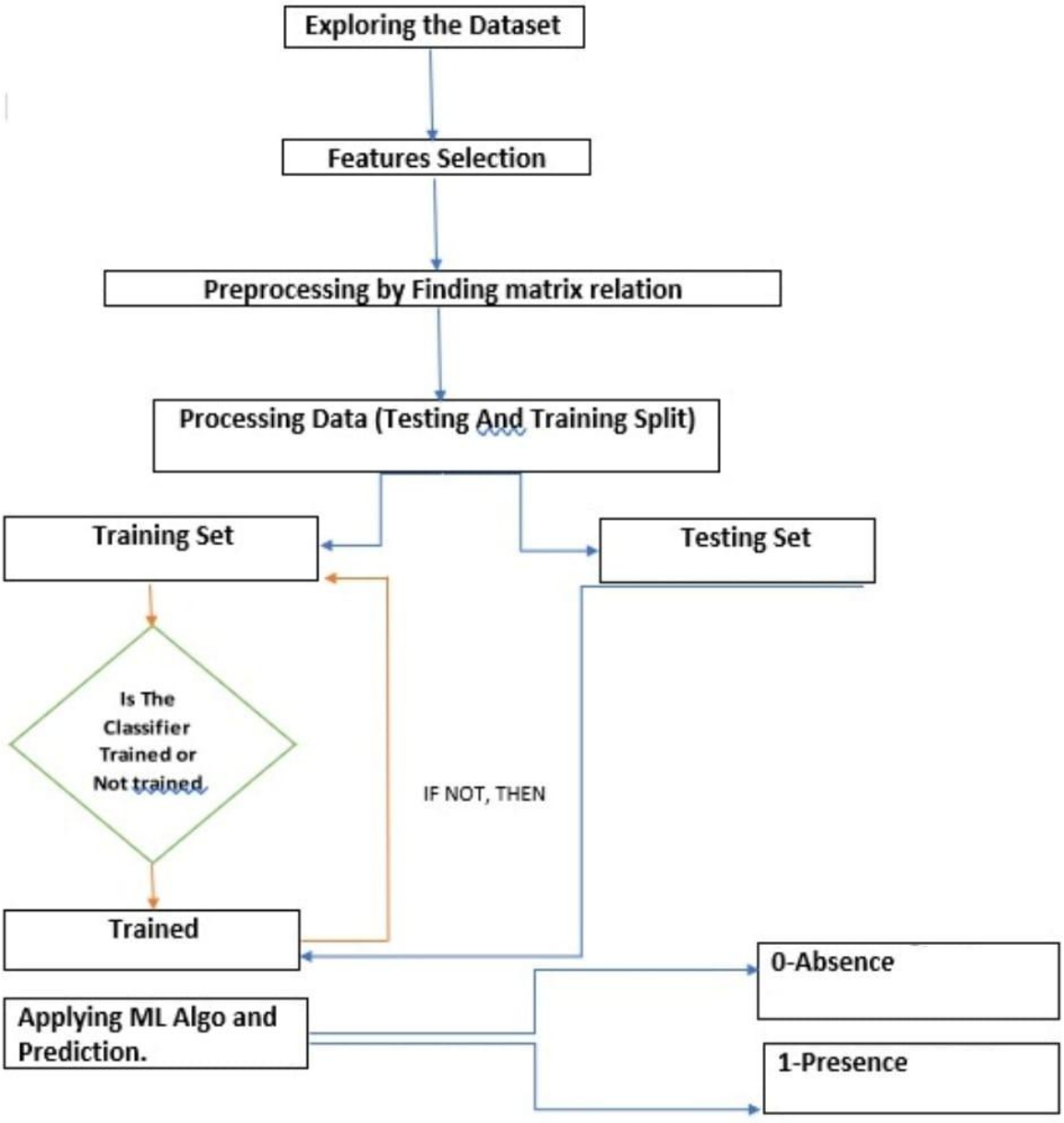


Figure: - Problem Formulation with FLOW CHART

REQUIRED TOOLS

Software and Hardware Requirements: Software Requirements:

1. Python 3.8+
2. Anaconda, Jupyter Notebook
3. Visual Studio Code, Chrome
4. Flask, HTML, CSS, BootStrap, Jquery
5. Scipy, Numpy, Pandas, Matplotlib, Sklearn etc.

Hardware Requirements:

OS : Windows/Linux/Mac

Processor : intel i5

RAM :4 GB

ROM :500 GB

Graphic card : Good but not necessary

Complete work plan layout

This project is implemented in 4 modules as described below each one after another where we will be performing the whole project.

Modules:

1. Heart Disease feature Analysis
2. Machine Learning
3. Base Template
4. Integrating Machinne Learning Pipeline Model to Flask App

Here we will be working with conventional support vector machines along with skcit learn and various algorithms to build this diabetes prediction model on the basis of concepts behind the algorithms and datasets working with conventional support vector machines. To build this conventional support vector machines we will discover different concepts behind the conventional support vector machines support vector machines and the dependencies:

Libraries used:

1. NumPy
2. SciPy
3. Matplotlib (pyplot, rcparams, matshow)
4. Statsmodels
5. Pandas
6. Tkinter
7. Sklearn

Starting from Gathering the data, Data Understanding Data Preprocessing Data Analysis Predictive Modelling

We will work on image processing techniques unconventional support vector machines. We will also do the necessary data analysis and required preprocessing steps for Datas.

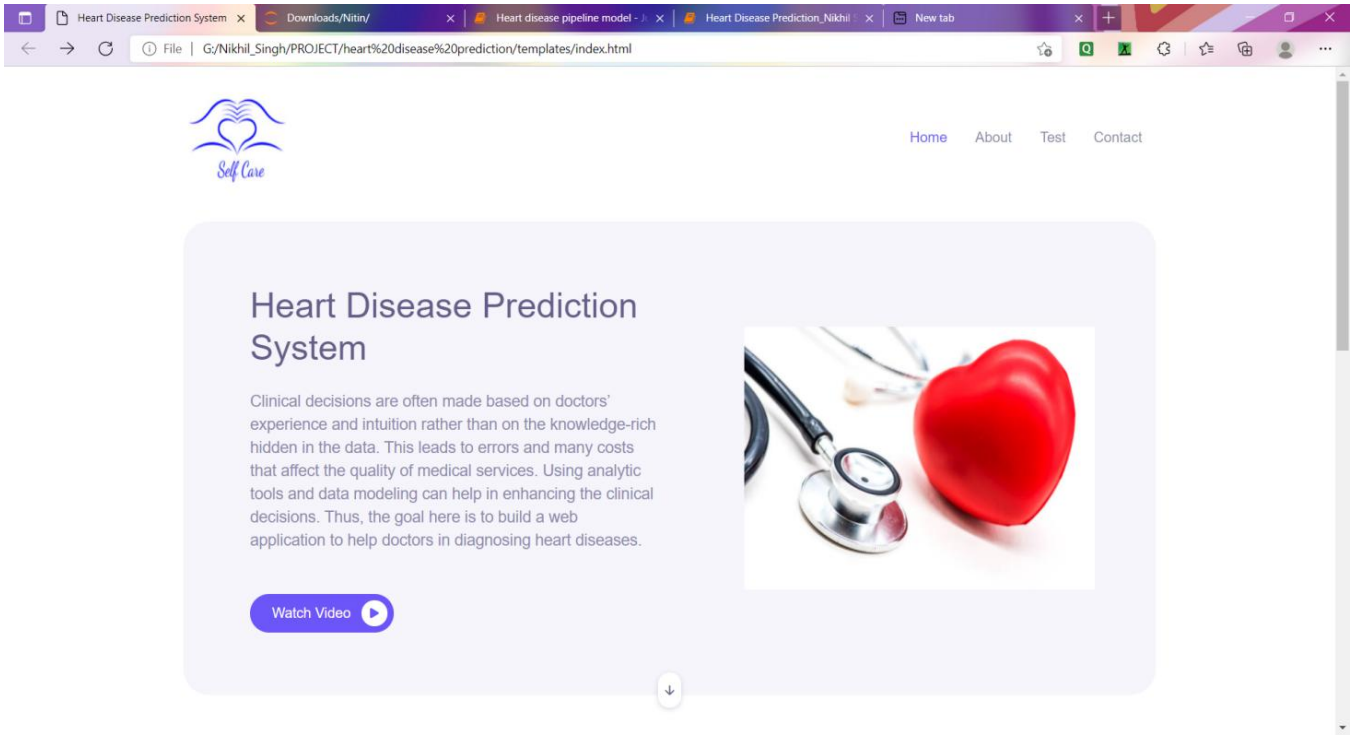
Once our machine learning model is ready, will we learn and develop web server gateway interphase in flask by rendering HTML CSS and bootstrap in the frontend and in the backend written in Python.

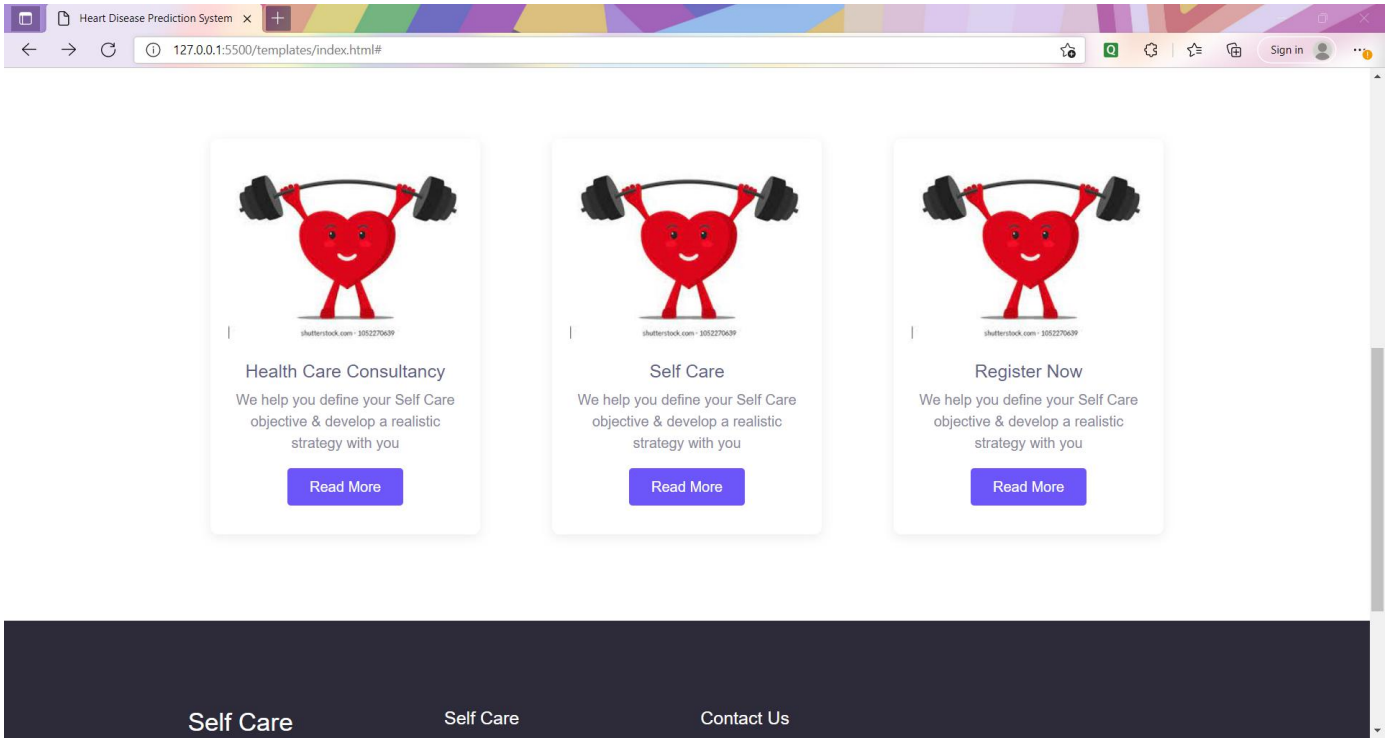
Finally, we will create the project on the Heart disease prediction by integrating the machine learning model

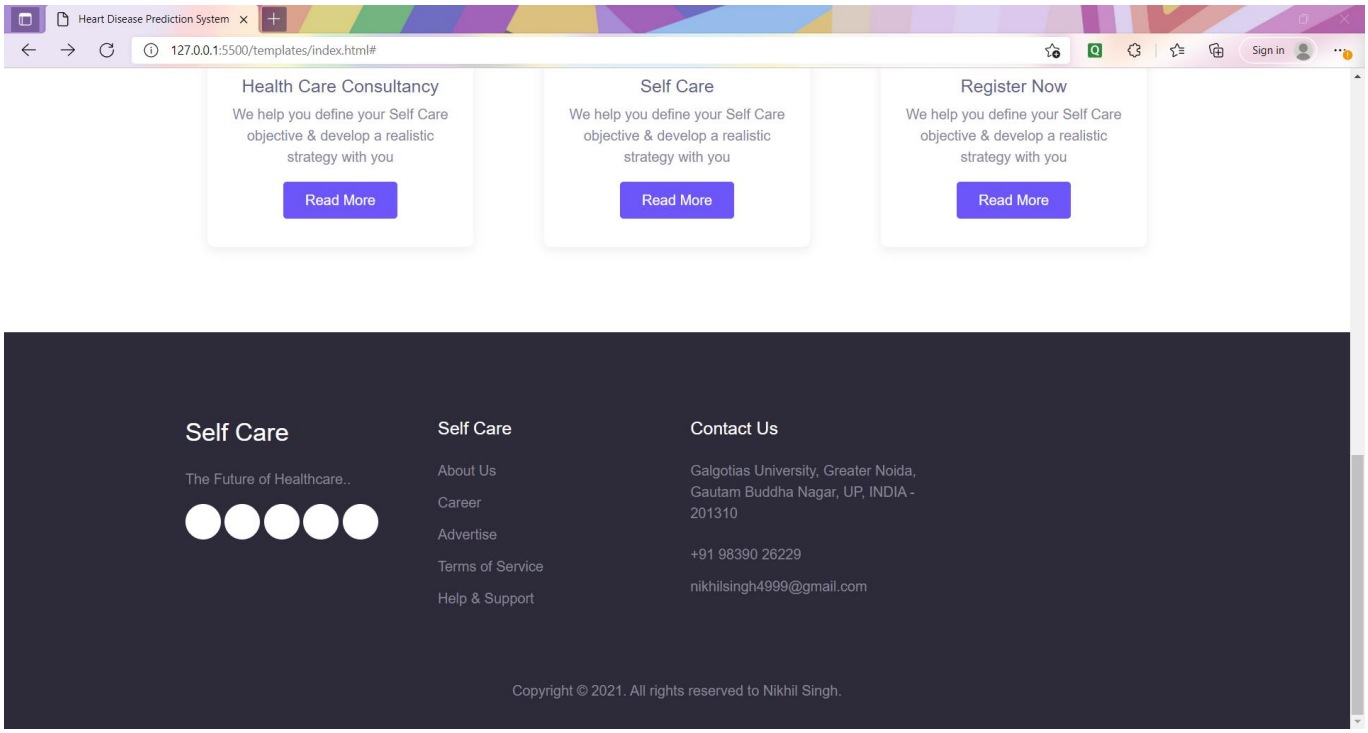
The coding portion were carried out to prepare the data, visualize it, pre-process it, building the model and then evaluating it. The code has been written in Python programming language using Jupyter Notebook as IDE. The experiments and all the models building are done based on python libraries.

Base Templates:

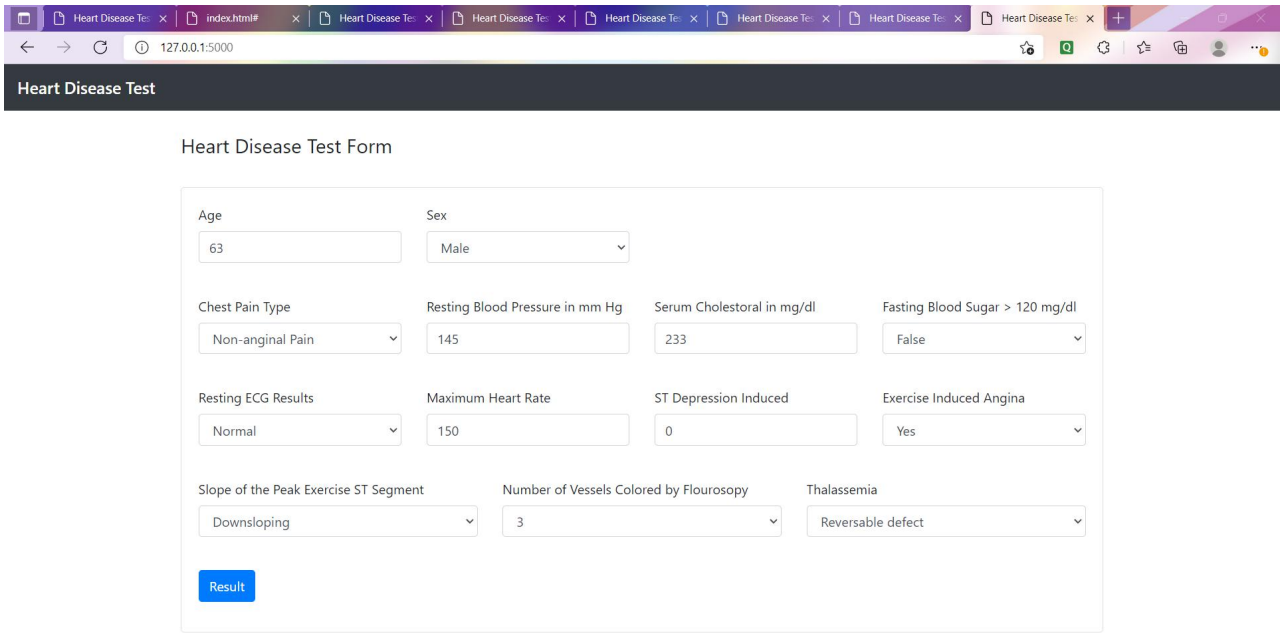
Home:



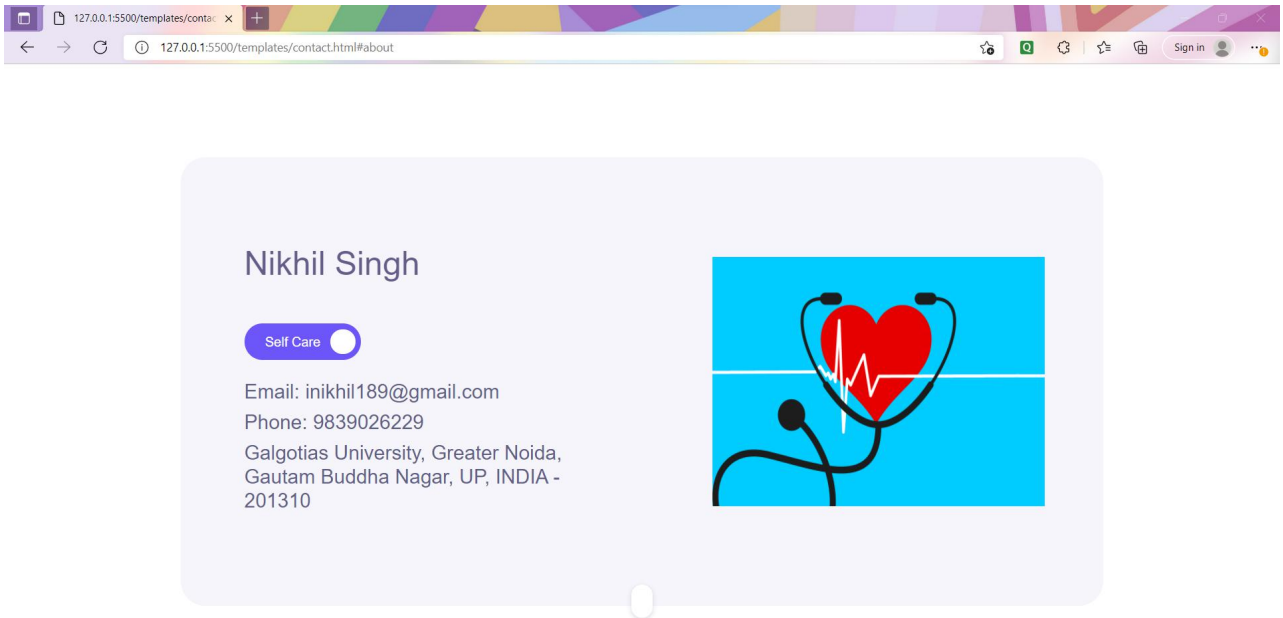



Heart Disease Prediction System
127.0.0.1:5500/templates/index.html#
Sign in
Health Care Consultancy
We help you define your Self Care objective & develop a realistic strategy with you
Read More
Self Care
We help you define your Self Care objective & develop a realistic strategy with you
Read More
Register Now
We help you define your Self Care objective & develop a realistic strategy with you
Read More
Self Care
The Future of Healthcare..
Self Care
About Us
Career
Advertise
Terms of Service
Help & Support
Contact Us
Galgotias University, Greater Noida, Gautam Buddha Nagar, UP, INDIA - 201310
+91 98390 26229
nikhilsingh4999@gmail.com
Copyright © 2021. All rights reserved to Nikhil Singh.

Prediction Page:



Contact Page:



PROJECT SUMMARY

We have started our project on Jupyter Notebook platform of python 3 language. The aim was to predict by the available database of a patient or client that whether they have a heart disease or not.Our project has been divided into different stages of completion.

These stages may be classified as

- Identification of problem.
    - Importing and learning the modules as well as different libraries.
    - Finding and exploring the dataset to be operated
    - Implementation of the solution
    - Processing the data
    - Predicting and analysing the result and so on.
- Feature Extraction
- Creating Model and Training and Testing the data
- Creating the Base Template
- Making Pipeline Model
- Creating Flask App Embedding the Pipeline Model into Web Application

REFRENCES

1. A. H. M. S. U. Marjia Sultana, "Analysis of Data Mining Techniques for Heart Disease Prediction," 2018.

2. M. I. K. ,. A. I. ,. S. Musfiq Ali, "Heart Disease Prediction Using Machine Learning Algorithms".

3. K. Bhanot, "towarddatascience.com," 13 Feb 2019. [Online]. Available: https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning-36f00f3edb2c. [Accessed 2 March 2020].

4. Online]. Available: https://www.kaggle.com/ronitf/heart-disease-uci#heart.csv.. [Accessed 05 December 2019].

5. M. A. K. S. H. K. M. a. V. P. M Marimuthu, "A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach".

6. A. Dantcheva, A, A. S., & Naik, C. (2016). Different Data Mining Approaches for Predicting Heart Disease, 277– 281. https://doi.org/10.15680/IJIRSET.2016.0505545

7. Beyene, C., &Kamat, P. (2018). Survey on prediction and analysis the occurrence of heart disease using data mining techniques. International Journal of Pure and Applied Mathematics, 118(Special Issue 8), 165–173. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-85041895038&partnerID=40&md5=2f0b0c5191a82bc0c3f0daf67d73bc81

8. Brownlee, J. (2016). Naive Bayes for Machine Learning. Retrieved March 4, 2019, from https://machinelearningmastery.com/naive-bayes-for-machine-learning/

9. Kirmani, M. (2017). Cardiovascular Disease Prediction using Data Mining Techniques. Oriental Journal of Computer Science and Technology, 10(2), 520–528. https://doi.org/10.13005/ojcst/10.02.38

10. Polaraju, K., Durga Prasad, D., & Tech Scholar, M. (2017). Prediction of Heart Disease using Multiple Linear Regression Model. International Journal of Engineering Development and

11. Purushottam, Saxena, K., & Sharma, R. (2016). Efficient Heart Disease Prediction System. In Procedia Computer Science (Vol. 85, pp. 962–969).

12. Purushottam, Saxena, K., & Sharma, R. (2016). Efficient Heart Disease Prediction System. In Procedia Computer Science (Vol. 85, pp. 962–969). https://doi.org/10.1016/j.procs.2016.05.288