A

Major Project

On

HEART DISEASE IDENTIFICATION METHOD USING MACHINE LEARNING

(Submitted in partial fulfillment of the requirements for the award of Degree)

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

BY

K. Rakshitha	**(177R1A05F2)**
V. Bhargavi	**(177R1A05H9)**
G. Vijaya	**(177R1A05E4)**
G. Gopi Krishna	**(177R1A05E5)**

Under the Guidance of

V. NARESH KUMAR

(Associate Professor)

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

CMR TECHNICAL CAMPUS

UGC AUTONOMOUS

(Accredited by NACC, NBA, Permanently Affiliated to JNTUH, Approved by AICTE, New Delhi)Recognized Under Section 2(f) & 12(B) of the UGC Act 1956, Kandlakoya(V), Medchal Road, Hyderabad-501401

2017-2021

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the project entitled "**HEART DISEASE IDENTIFICATION METHOD USING MACHINE LEARNING**" being submitted **by K. RAKSHITHA(177R1A05F2),V. BHARGAVI(177R1A05H9), G.VIJAYA(177R1A05E4),** G. **GOPI KRISHNA(177R1A05E5)** in partial fulfillment of the requirements for the award of the degree of B.Tech in Computer Science and Engineering of the Jawaharlal Nehru Technological University Hyderabad, during the year 2020-2021.It is certified that they have completed the project successfully.

INTERNAL GUIDE

V. Naresh Kumar

DIRECTOR

Dr. A. Raji Reddy

HOD

Dr. K. Srujan Raju

EXTERNAL EXAMINER

Submitted for viva voice Examination held on _____

CMRTC

ACKNOWLEDGEMENT

Apart from the efforts of us, the success of any project depends largely on the encouragement and guidelines of many others. We take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of this project.

We take this opportunity to express my profound gratitude and deep regard to my guide **V. NARESH KUMAR,** Associate Professor for her exemplary guidance, monitoring and constant encouragement throughout the project work. The blessing, help and guidance given by her shall carry us a long way in the journey of life on which we are about to embark.

We also take this opportunity to express a deep sense of gratitude to Project Review Committee(PRC)Coordinators: **Mr.J.NarasimhaRao, Mr.B.P.Deepak Kumar, Dr.Suwarna Gothane** and **Mr.B. Ramji** for their cordial support, valuable information and guidance, which helped us in completing this task through various stages.

We are also thankful to the Head of the Department **Dr. K. Srujan Raju** for providing excellent infrastructure and a nice atmosphere for completing this project successfully.

We are obliged to our Director **Dr. A. Raji Reddy** for being cooperative throughout the course of this project. We would like to express our sincere gratitude to our Chairman Sri. **Ch. Gopal Reddy** for his encouragement throughout the course of this project.

The guidance and support received from all the members of **CMR TECHNICAL CAMPUS** who contributed and who are contributing to this project, was vital for the success of the project. We are grateful for their constant support and help.

Finally, we would like to take this opportunity thank our family for their constant encouragement without which this assignment would not be possible. We sincerely acknowledge and thank all those who gave support directly and indirectly in completion of this project.

K. RAKSHITHA(177R1A05F2)

V. BHARGAVI (177R1A05H9)

G. VIJAYA (177R1A05E4)

G. GOPI KRISHNA(177R1A05E5)

ABSTRACT

Heart disease is one of the complex diseases and globally many people suffered from this disease. On time and efficient identification of heart disease plays a key role in healthcare, particularly in the field of cardiology. In this project, we proposed an efficient and accurate system to diagnosis heart disease and the system is based on machine learning techniques. The system is developed based on classification algorithms includes Support vector machine, Artificial neural network, K-nearest neighbor while standard features selection algorithms have been used such as Relief, Minimal redundancy maximal relevance, Least absolute shrinkage selection operator and Local learning for removing irrelevant and redundant features. We also proposed novel fast conditional mutual information feature selection algorithm to solve feature selection problem.

The features selection algorithms are used for features selection to increase the classification accuracy and reduce the execution time of classification system. Furthermore, the leave one subject out cross-validation method has been used for learning the best practices of model assessment and for hyper parameter tuning. The performance measuring metrics are used for assessment of the performances of the classifiers. The performances of the classifiers have been checked on the selected features as selected by features selection algorithms. The experimental results show that the proposed feature selection algorithm (FCMIM) is feasible with classifier support vector machine for designing a high-level intelligent system to identify heart disease. The suggested diagnosis system (FCMIM-SVM) achieved good accuracy as compared to previously proposed methods. Additionally, the proposed system can easily be implemented in healthcare for the identification of heart disease.

LIST OF FIGURES

LIST OF SCREENSHOTS

TABLE OF CONTENTS

1. INTRODUCTION

1. INTRODUCTION

Heart disease (HD) is the critical health issue and numerous people have been suffered by this disease around the world. The HD occurs with common symptoms of breath shortness, physical body weakness and, feet are swollen. Researchers try to come across an efficient technique for the detection of heart disease, as the current diagnosis techniques of heart disease are not much effective in early time identification due to several reasons, such as accuracy and execution time. The diagnosis and treatment of heart disease is extremely difficult when modern technology and medical experts are not available. The effective diagnosis The associate editor coordinating the review of this manuscript and approving it for publication was Navanietha Krishnaraj Krishnaraj Rathinam. and proper treatment can save the lives of many people.

According to the European Society of Cardiology, 26 million approximately people of HD were diagnosed and diagnosed 3.6 million annually. Most of the people in the United States are suffering from heart disease. Diagnosis of HD is traditionally done by the analysis of the medical history of the patient, physical examination report and analysis of concerned symptoms by a physician. But the results obtained from this diagnosis method are not accurate in identifying the patient of HD. Moreover, it is expensive and computationally difficult to analyze. Thus, to develop a noninvasive diagnosis system based on classifiers of machine learning (ML) to resolve these issues.

The Cleveland heart disease data set was used by various researchers for the identification problem of HD. The machine learning predictive models need proper data for training and testing. The performance of machine learning model can be increased if balanced dataset is use for training and testing of the model. Furthermore, the model predictive capabilities can improved by using proper and related features from the data. Therefore, data balancing and feature selection is significantly important for model performance improvement.

In literature various diagnosis techniques have been proposed by various researchers, however these techniques are not effectively diagnosis HD. In order to improve the predictive capability of machine learning model data preprocessing is important for data standardization. Various Preprocessing techniques such removal of missing feature value instances from the dataset, Standard Scalar (SS), Min-Max Scalar etc. The feature extraction and selection techniques are also improve model performance.

Various feature selection techniques are mostly used for important feature selection such as, Least-absolute-shrinkage-selection-operator (LASSO), Relief, Minimal-Redundancy-Maximal-Relevance (MRMR), Local-learning-based-features-selection (LLBFS), Principle component Analysis (PCA), Greedy Algorithm (GA), and optimization methods, such as Anty Conley Optimization (ACO), fruit fly optimization (FFO), Bacterial Foraging Optimization (BFO) etc.

Similarly Yun et al. presented different techniques for different type of feature selection, such as feature selection for high- dimensional small sample size data, large-scale data, and secure feature selection. They also discussed some important topics for feature selection have emerged, such as stable feature selection, multi-view feature selection, distributed feature selection, multi-label feature selection, online feature selection, and adversarial feature selection. Jundong et al. discussed the challenges of feature selection (FS) for big data. It is necessary to decrease the dimensionality of data for various learning tasks due to the curse of dimensionality. Feature selection has great influence in numerous applications such as building simpler, increasing learning performance, creating clean and understandable data.

The feature selection from big data is challenging job and create big problems because big data has many dimensions. Further, challenges of feature selection for structured, heterogeneous and streaming data as well as its scalability and stability issues. For big data analytics challenges of feature selection is very important to resolved. In designed unsupervised hashing scheme, called topic hyper graph hashing, to report the limitations. The filter based method measures the relevance of a feature by correlation with the dependent variable while the wrapper feature selection algorithm measure the usefulness of a subset of features by actually training the classifier on it. The filter method is less computationally complex than wrapper method. The feature set selected by the filter is general and can be applied to any model and it is independent of a specific model. In feature selection global relevance is of greater importance. On another hand suitable machine learning model is necessary for good results. Obviously, a good machine learning model is a model that not only performs well on data seen during training (else a machine learning model could simply learn the training data), but also on unseen data.

To evaluate all classifiers on data and find that they get, on average, 50% of the cases right. Furthermore, appropriate cross validation techniques and performance

evaluation metrics are critical necessary for a model when model is train and test on dataset. We proposed a machine learning based diagnosis method for the identification of HD in this research work. Machine learning predictive models include ANN, K-NN, SVM, are used for the identification of HD. The standard state of the art features selection algorithms, such as Relief, mRMR, LASSO and Local-learning-based features-selection (LLBFS) have been used to select the features. We also proposed fast conditional mutual information (FCMIM) features selection algorithm for features selection. Leave-one-subject-out cross-validation (LOSO) technique has been applied to select the best hyper-parameters for best model selection.

Apart from this, different performance assessment metrics have been used for classifiers performances evaluation. The proposed method has been tested on Cleveland HD dataset. Furthermore, the performance of the proposed technique have been compared with state of the art existing methods in the literature, such as NB, Three phase ANN (Artificial neural Network) diagnosis system, Neural network ensembles (NNE), ANN-Fuzzy-AHP diagnosis system (AFP), Adaptive- weighted-Fuzzy-system-ensemble (AWFSE).

The research study has the following contributions. Firstly, the authors try to address the problem of features selection by employing pre-processing techniques and standard state of the art four features selection algorithms such as Relief, mRMR, LASSO, and LLBFS for appropriate subset of features and then applied these features for effective training and testing of the classifiers that identify which feature selection algorithm and classifier gives good results in term of accuracy and computation time. Secondly, the authors proposed a fast conditional mutual information (FCMIM) FS algorithm for feature selection and then these features are input to classifiers for improving prediction accuracy and reducing computation time. The classifiers performances have been compared on features selected by the standard state VOLUME 8, 2020 107563 J. P. Li et al.: HD Identification Method Using ML Classification in E-Healthcare of the art FS algorithms with the selected features of the proposed FS algorithm. Thirdly, identify weak features from the dataset which affect the performance of the classifiers. Finally, suggests that heart disease identification system (FCMIM-SVM) effectively identify the HD.

1.1 PROJECT SCOPE

In order to improve the predictive capability of machine learning model data preprocessing is important for data standardization. Various Preprocessing techniques such removal of missing feature value instances from the dataset, Standard Scalar (SS), Min-Max Scalar etc. The feature extraction and selection techniques are also improve model performance. Various feature selection techniques are mostly used for important feature selection such as, Least-absolute-shrinkage-selection-operator (LASSO), Relief, Minimal-Redundancy-Maximal-Relevance (MRMR), Local-learning-based-features-selection (LLBFS), Principle component Analysis (PCA), Greedy Algorithm (GA), and optimization methods, such as Anty Conley Optimization (ACO), fruit fly optimization (FFO), Bacterial Foraging Optimization (BFO)

1.2 LITERATURE SURVEY

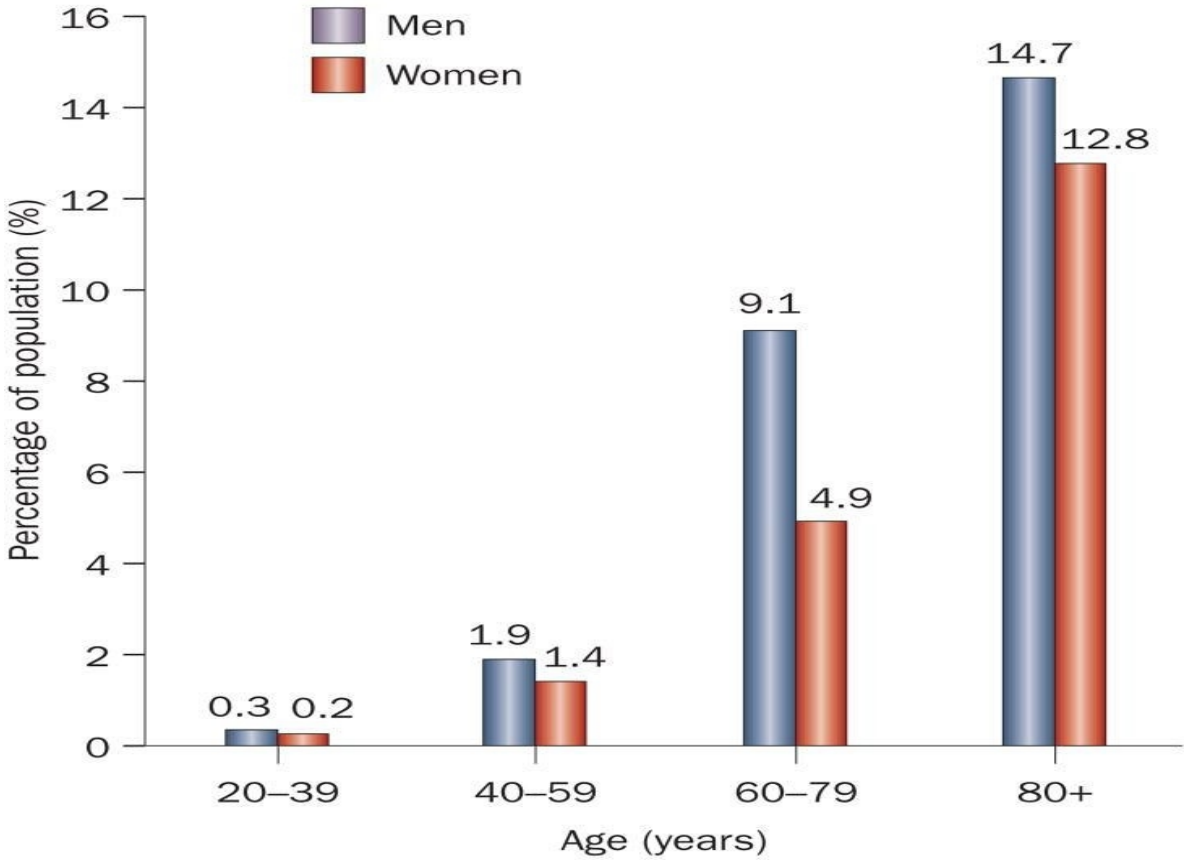
In literature various machine learning based diagnosis techniques have been proposed by researchers to diagnosis HD. This project present some existing machine learning based diagnosis techniques in order to explain the important of the proposed work. Detrano et al. developed HD classification system by using machine learning classification techniques and the performance of the system was 77% in terms of accuracy. Cleveland dataset was utilized with the method of global evolutionary and with features selection method. In another study Gudadhe et al. developed a diagnosis system using multi-layer Perceptron and support vector machine (SVM) algorithms for HD classification and achieved accuracy 80.41%. Humar et al. designed HD classification system by utilizing a neural network with the integration of Fuzzy logic.

The classification system achieved 87.4% accuracy. Resul et al. developed an ANN ensemble based diagnosis system for HD along with statistical measuring system enterprise miner (5.2) and obtained the accuracy of 89.01%, sensitivity 80.09%, and specificity 95.91%. Akil et al. designed a ML based HD diagnosis system. ANN-DBP algorithm along with FS algorithm and performance was good. Palaniappan et al. proposed an expert medical diagnosis system for HD identification. In development of the system the predictive model of machine learning, such as navies bays (NB), and Artificial Neural Network were used. The 86.12% accuracy was achieved by ANN accuracy 88.12% and DT classifier achieved 80.4% accuracy.

Olaniyi et al. developed a three-phase technique based on the artificial neural network technique for HD prediction in angina and achieved 88.89% accuracy. Samuel et al. developed an integrated medical decision support system based on artificial neural network and Fuzzy AHP for diagnosis of HD. The performance of the proposed method in terms of accuracy was 91.10%. Liu et al. proposed a HD classification system using relief and rough set techniques. The proposed method achieved 92.32% classification accuracy.

In proposed a HD identification method using feature selection and classification algorithms. Sequential Backward Selection Algorithm (SBS FS) for Features Selection. The classifier K-Nearest Neighbor (K-NN) performance has been checked on full and on selected features set. The proposed method obtained high accuracy. In another study MOHAN et al. designed a HD prediction method by using hybrid machine learning techniques. He also proposed a new method for significant feature selection from the data for effective training and testing of machine learning classifier. They have been recorded 88.07% classification accuracy. Geweid et al. designed HD identification techniques by using improved SVM based duality optimization technique.

All these existing techniques used numerous methods to identify the HD at early stages. However, all these techniques have lack of prediction accuracy and high computation time for prediction of HD. Thus, the major issues in these previous approaches are low accuracy and high computation time and these might be due the use of irrelevant features in dataset. In order to tackle these problems new methods are needed to detect HD correctly. The improvement in prediction accuracy is a big challenge and research gap.



SURVEY

1.3 EXISTING SYSTEM

This project present some existing machine learning based diagnosis techniques in order to explain the important of the proposed work. Detrano et al. developed an HD classification system by using machine learning classification techniques and the performance of the system was 77% in terms of accuracy. Cleveland dataset was utilized with the method of global evolutionary and with features selection method.

1.3.1 LIMITATIONS OF EXISTING SYSTEM

→ This system was developed on HD classification system.

1.4 PROPOSED SYSTEM

→ The system has been designed for the identification of heart disease. The performances of various machine learning classifiers for HD identification have been checked on selected features.

→ We proposed a machine learning based diagnosis method for the identification of HD in this project. Machine learning predictive models include ANN, K-NN, SVM, are used for the identification of HD.

→ The standard state of the art features selection algorithms, such as Relief, mRMR, LASSO and Local-learning-based features-selection (LLBFS) have been used to select the features.

→ We also proposed fast conditional mutual information (FCMIM) features selection algorithm for features selection. Leave -One subject-out cross validation technique has been applied to select the best hyper parameters or best model selection.

1.4.1 ADVANTAGES OF PROPOSED SYSTEM

→ System consists of three different process model so high accuracy in prediction.
→ Cost effective for patients.
→ Increased accuracy for effective heart disease diagnosis.

1.5 OBJECTIVES

The objective of project : Identification of Heart Disease Using Machine Learning Classification in E-Healthcare.

Applications : It can be used in hospitals.

2. SYSTEM ANALYSIS

SYSTEM ANALYSIS

System Analysis is the important phase in the system development process. The System is studied to the minute details and analysed. The system analyst plays an important role of an interrogator and dwells deep into the working of the present system. In analysis, a detailed study of these operations performed by the system and their relationships within and outside the system is done. A key question considered here is, "what must be done to solve the problem?" The system is viewed as a whole and the inputs to the system are identified. Once analysis is completed the analyst has a firm understanding of what is to be done.

2.1 PROBLEM DEFINITION

→ Heart disease is one of the complex diseases and globally many people suffered from this disease. On time and efficient identification of heart disease plays a key role in healthcare, particularly in the field of cardiology.

→ In this project, we proposed an efficient and accurate system to diagnosis heart disease and the system is based on machine learning techniques.

→ The system is developed based on classification algorithms includes

- Support vector machine
- Artificial neural network,
- K-nearest neighbour

→ While standard features selection algorithms have been used such as Relief, Minimal redundancy maximal relevance, Least absolute shrinkage selection operator and Local learning for removing irrelevant and redundant features.

2.2 SOFTWARE DEVELOPMENT LIFE CYCLE

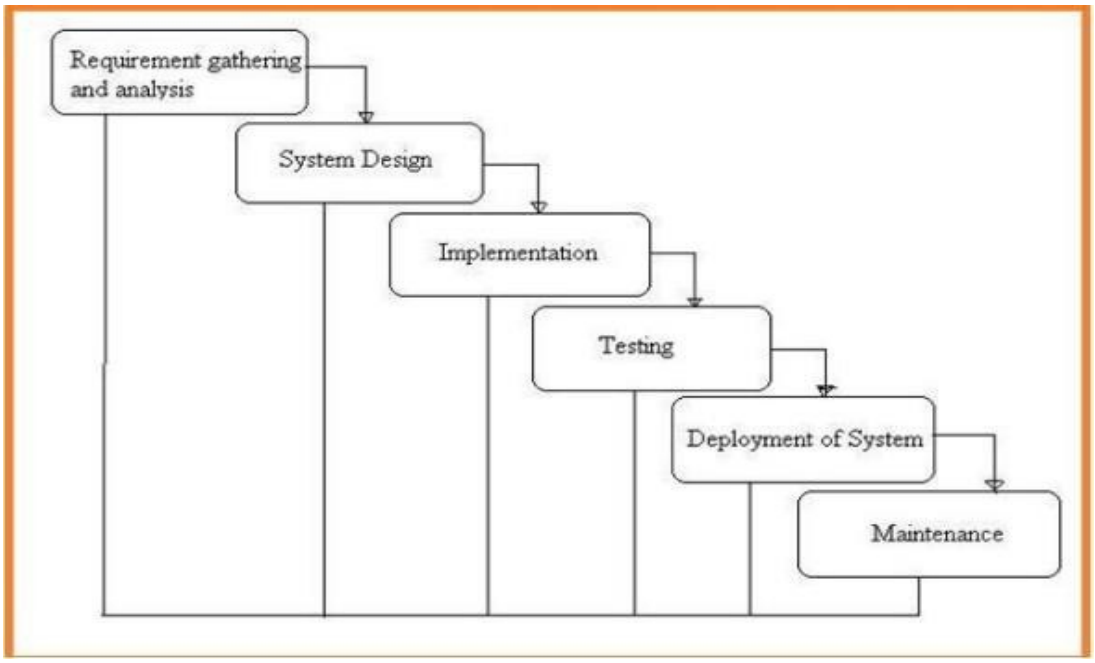


fig-2.2: Structure of SDLC

Software development life cycle consists of six stages. They are:

- ➔ Project Requisites Accumulating and Analysis
- ➔ System Design
- ➔ Implementation
- ➔ Testing
- ➔ Deployment of System
- ➔ Maintenance of the Project

2.2.1 PROJECT REQUISITES ACCUMULATING ANDANALYSIS

It's the first and foremost stage of the any project as our is a an academic leave for requisites amassing we followed of IEEE Journals and Amassed so many IEEE Relegated papers and final culled a Paper designated "Individual web revisitation by setting and substance importance input and for analysis stage we took referees from the paper and did literature survey of some papers and amassed all the Requisites of the project in this stage

2.2.2 SYSTEM DESIGN

In System Design has divided into three types like GUI Designing, UML Designing with avails in development of project in facile way with different actor and its utilizer case by utilizer case diagram, flow of the project utilizing sequence, Class diagram gives information about different class in the project with methods that have to be utilized in the project if comes to our project our UML Will utilizable in this way The third and post import for the project in system design is Data base design where we endeavor to design data base predicated on the number of modules in our project.

2.2.3 IMPLEMENTATION

The Implementation is Phase where we endeavor to give the practical output of the work done in designing stage and most of Coding in Business logic lay coms into action in this stage its main and crucial part of the project.

2.2.4 TESTING

UNIT TESTING : It is done by the developer itself in every stage of the project and fine-tuning the bug and module predicated additionally done by the developer only here we are going to solve all the runtime errors

MANUAL TESTING : As our Project is academic Leave, we can do any automatic testing so we follow manual testing by endeavor and error methods.

2.2.5 DEPLOYMENT OF SYSTEM AND MAINTENANCE

Once the project is total yare, we will come to deployment of client system in genuinely world as its academic leave we did deployment in our college lab only with all need Software's with having Windows OS . The Maintenance of our Project is one-time process only.

2.3 FEASIBILITY STUDY

The feasibility of the project is analysed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. Three key considerations involved in the feasibility analysis are

- ➔ Economic Feasibility
- ➔ Technical Feasibility
- ➔ Social Feasibility

2.3.1 ECONOMIC FEASIBILITY

The developing system must be justified by cost and benefit. Criteria to ensure that effort is concentrated on project, which will give best, return at the earliest. One of the factors, which affect the development of a new system, is the cost it would require. The following are some of the important financial questions asked during preliminary investigation:

- ➔ The costs conduct a full system investigation.
- ➔ The cost of the hardware and software.
- ➔ The benefits in the form of reduced costs or fewer costly errors.

Since the system is developed as part of project work, there is no manual cost to spend for the proposed system. Also, all the resources are already available, it give an indication of the system is economically possible for development.

2.3.2 TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

2.3.3 SOCIAL FEASIBILITY

This includes the following questions:

- ➔ Is there sufficient support for the users?
- ➔ Will the proposed system cause harm?

The project would be beneficial because it satisfies the objectives when developed and installed. All behavioural aspects are considered carefully and conclude that the project is behaviourally feasible.

2.4 HARDWARE AND SOFTWARE REQUIREMENTS

2.4.1 HARDWARE REQUIREMENTS

Hardware interfaces specifies the logical characteristics of each interface between the software product and the hardware components of the system.

- ➔ Processor : Minimum Intel i3
- ➔ Ram : Minimum 4gb
- ➔ Hard disk : Minimum 250gb

2.4.2 SOFTWARE REQUIREMENTS

Software Requirements specifies the logical characteristics of each interface and software components of the system.

- Python : Python idle 3.7 version
- Operating system : Microsoft Windows, Linux

2.5 About Dataset

- Heart disease
- Dataset is stored in static file in our project file. here we collected datasets are breast data, diabetes data and heart disease data. In breast data contains 569 records and each record contains 6 columns

2.5.1 Algorithms

SVM- Support vector Machine
ANN- Artificial neural Network
KNN- K-Nearest Neighbor

3. ARCHITECTURE

3.1 PROJECT ARCHITECTURE

This project architecture shows the procedure followed for heart disease detection using machine learning classification, starting from input to final prediction.

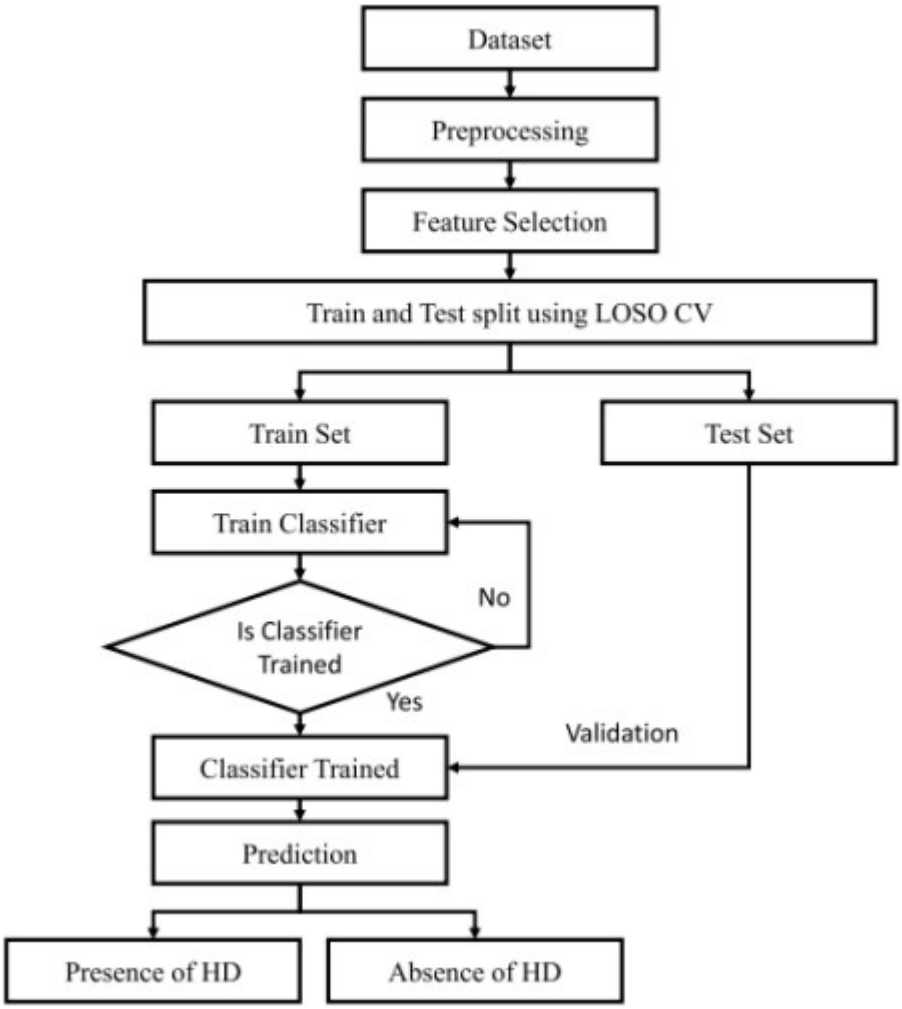


Fig-3.1: Project architecture

3.2 DESCRIPTION

Input data: In this project, we are using Cleveland dataset which is in form of .csv format. This data is imported and uploaded in the idle.

Preprocessing the data: The pre-processing of the dataset is required for good representation. Techniques of pre-processing such as removing attribute missing values, Standard Scalar (SS), Min-Max Scalar have been applied to the dataset.

Feature Selection: After data pre-processing, the selection of features is required for the process. In general, FS is a significant step in constructing a classification model. It works by

reducing the number of input features in a classifier, to have good predictive and short computationally complex models. We have used the LOSO FS algorithm in this study.

Train and Test the Data: Training data is passed to the classifier to train the model and test data is used to test the trained model to check whether it is making correct predictions or not.

Algorithms: In this project we used three algorithms i.e.,K-nearest neighbour (KNN), Artificial Neural Network (ANN), Support Vector Machine (SVM) for classifying the data. We also used a feature selection technique called Leave-One-Subject-Out(LOSO) cross validation technique for more accurate prediction.

Output: The output of this project gives the accuracy values of the three algorithms applied and it also shows the comparison between algorithms in the form of graphs, from the obtained values we can analyze which algorithm gives highest accuracy in prediction of heart disease.

3.3 FLOW CHART

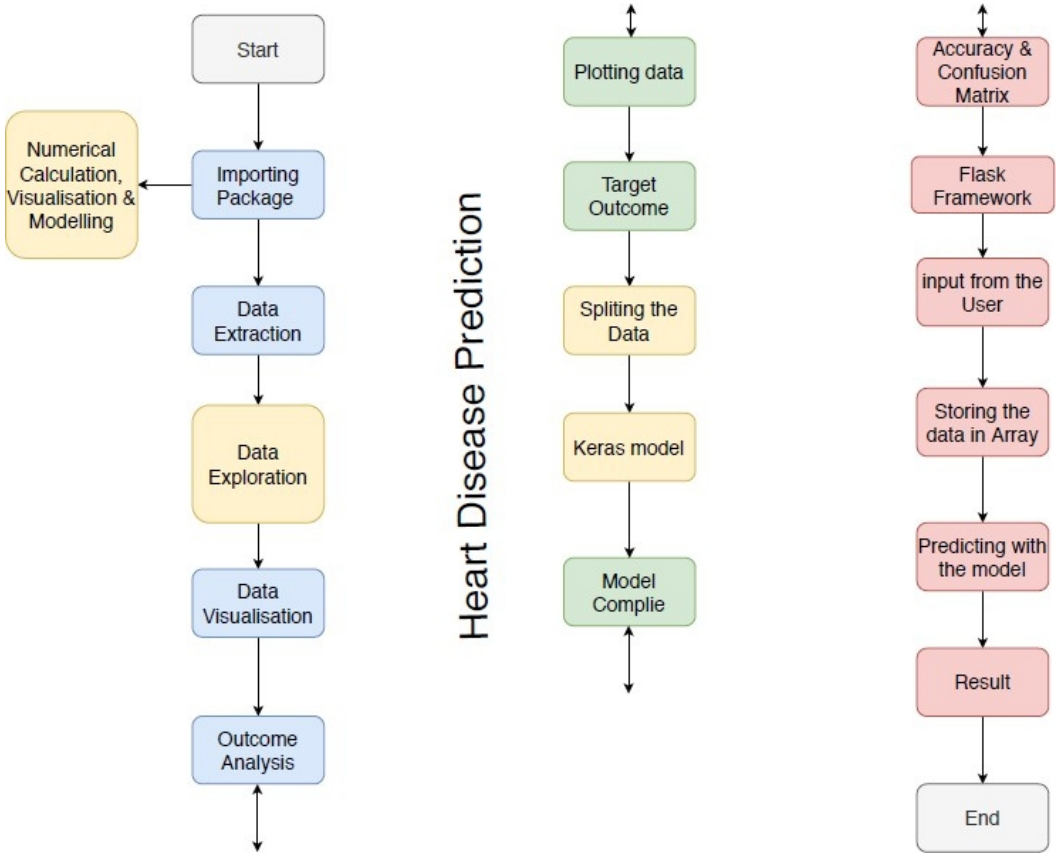


Fig-3.3:Flow chart

3.4 USE CASE DIAGRAM

In the use case diagram we have basically two actors who are the user and the administrator. The user has the rights to login, import the data and view the output. The administrator can upload dataset, preprocess the data and train and test the dataset.

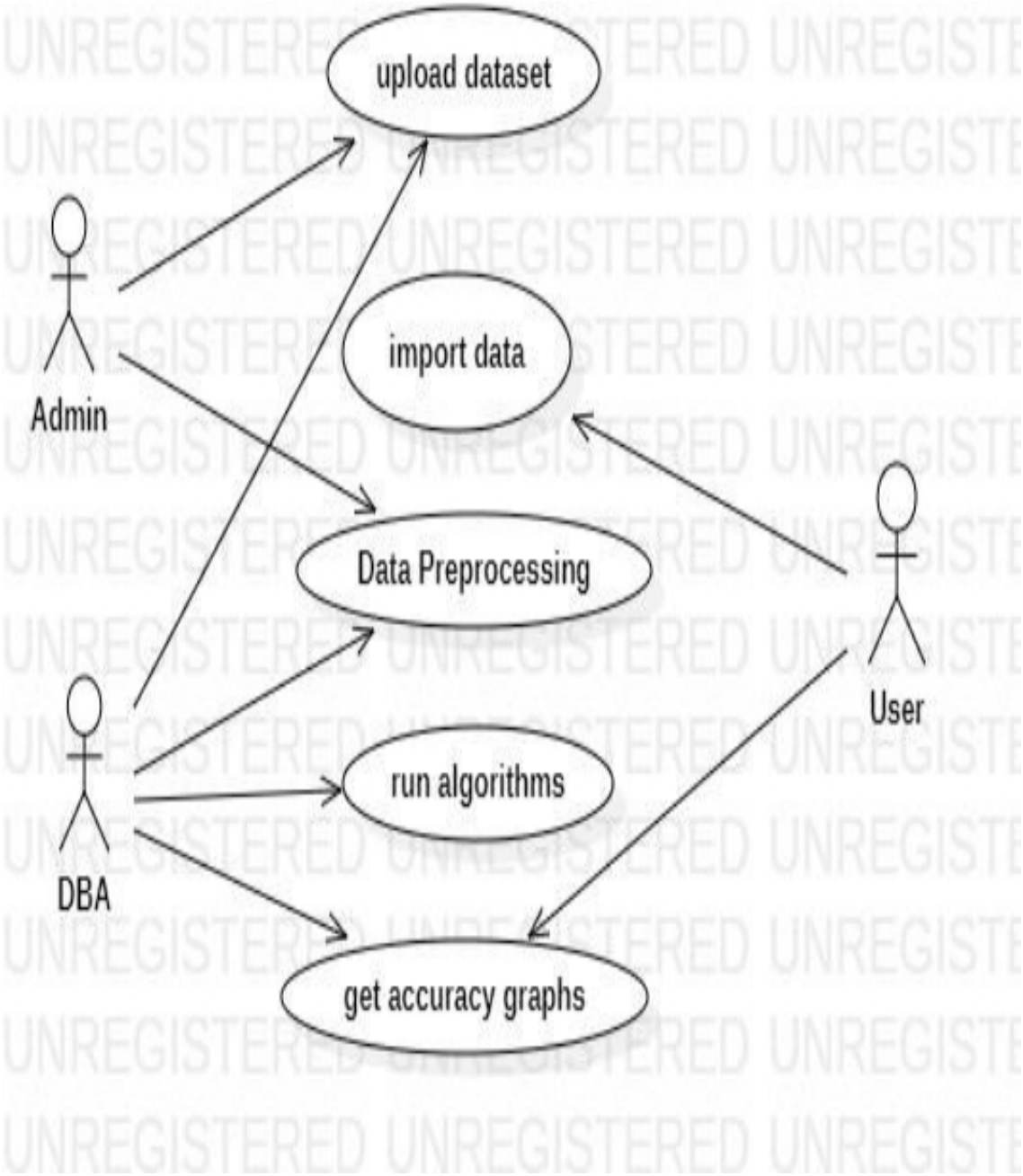


Fig-3.4: Use case diagram

3.5 CLASS DIAGRAM

Class Diagram is a collection of classes and objects. Class diagrams to describe the structure of the system. Classes are abstraction that specify the common structure and behavior of a set of objects. Class diagrams describe the system in terms of objects, classes, attributes, operations and their associations.

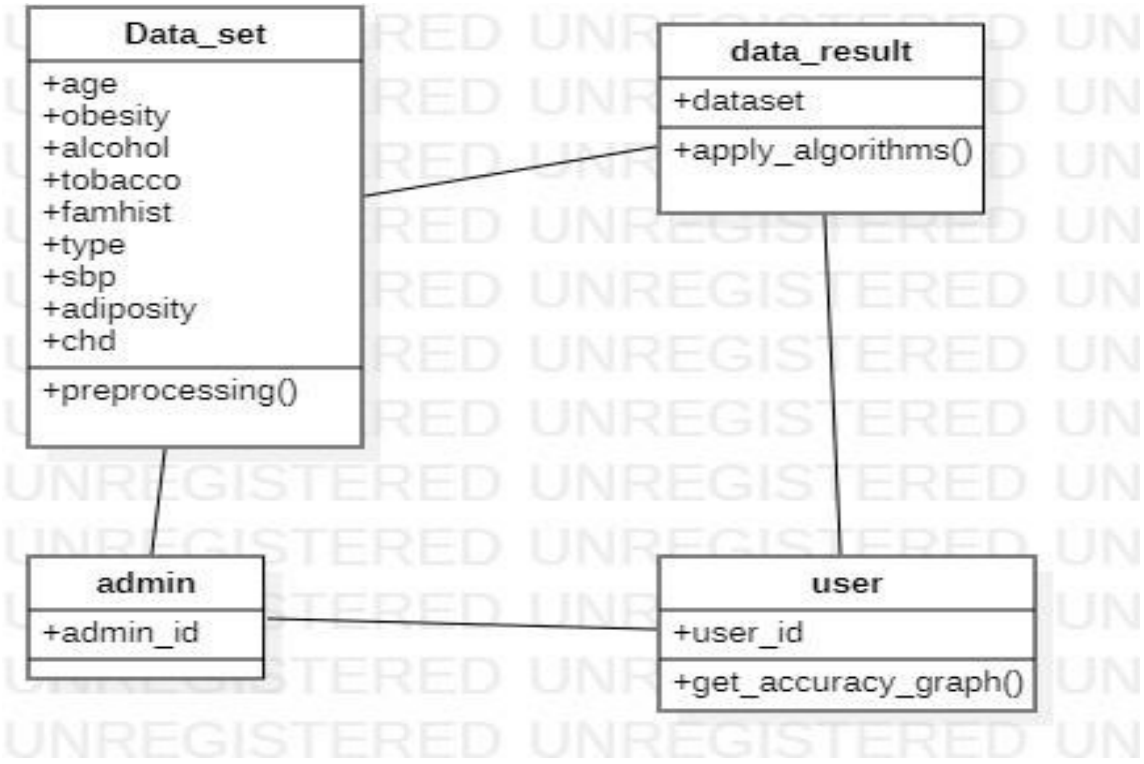


Fig-3.5: Class Diagram

3.6 SEQUENCE DIAGRAM

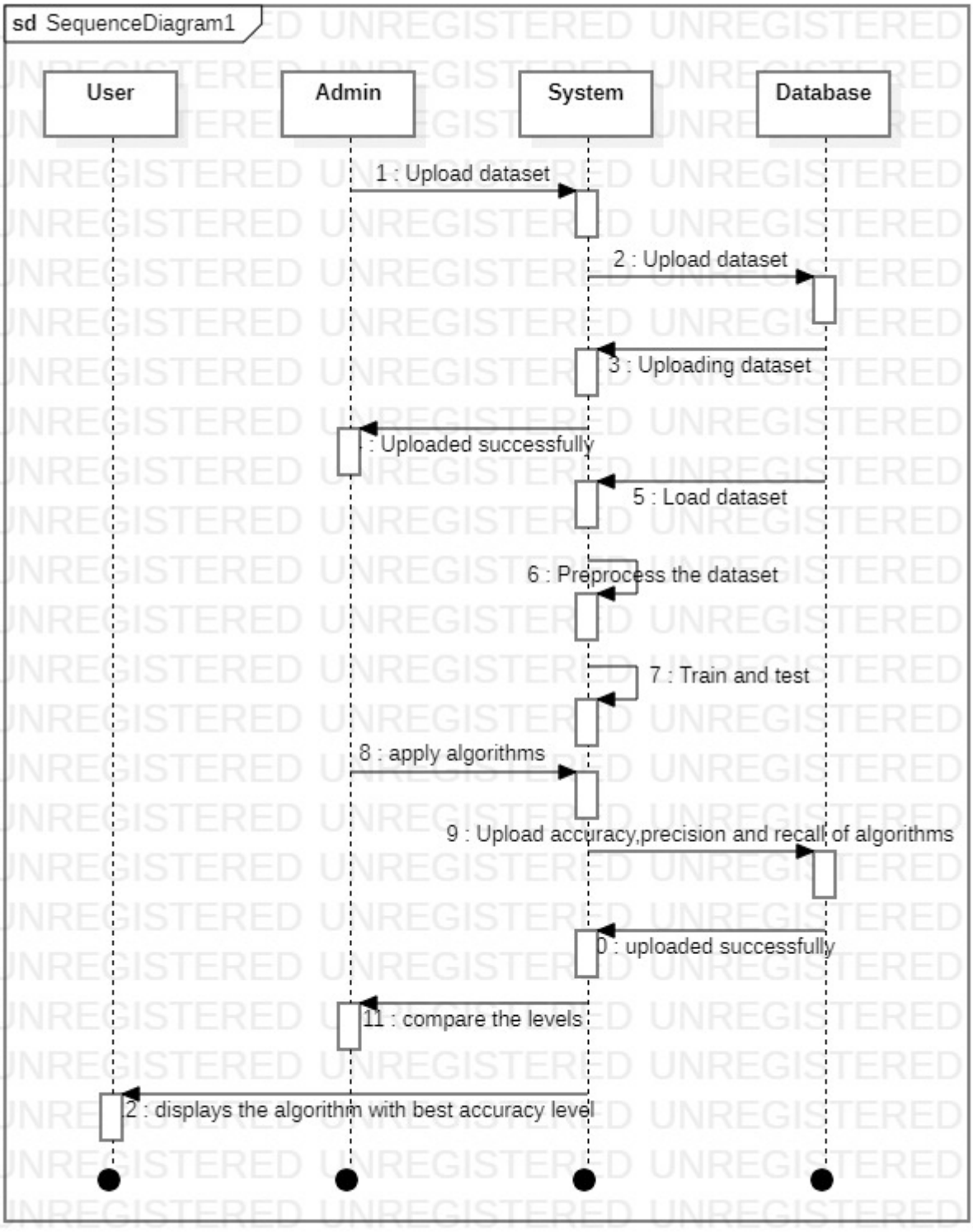


Fig-3.6: Sequence diagram

3.7 ACTIVITY DIAGRAM

It describes the flow of activity states.

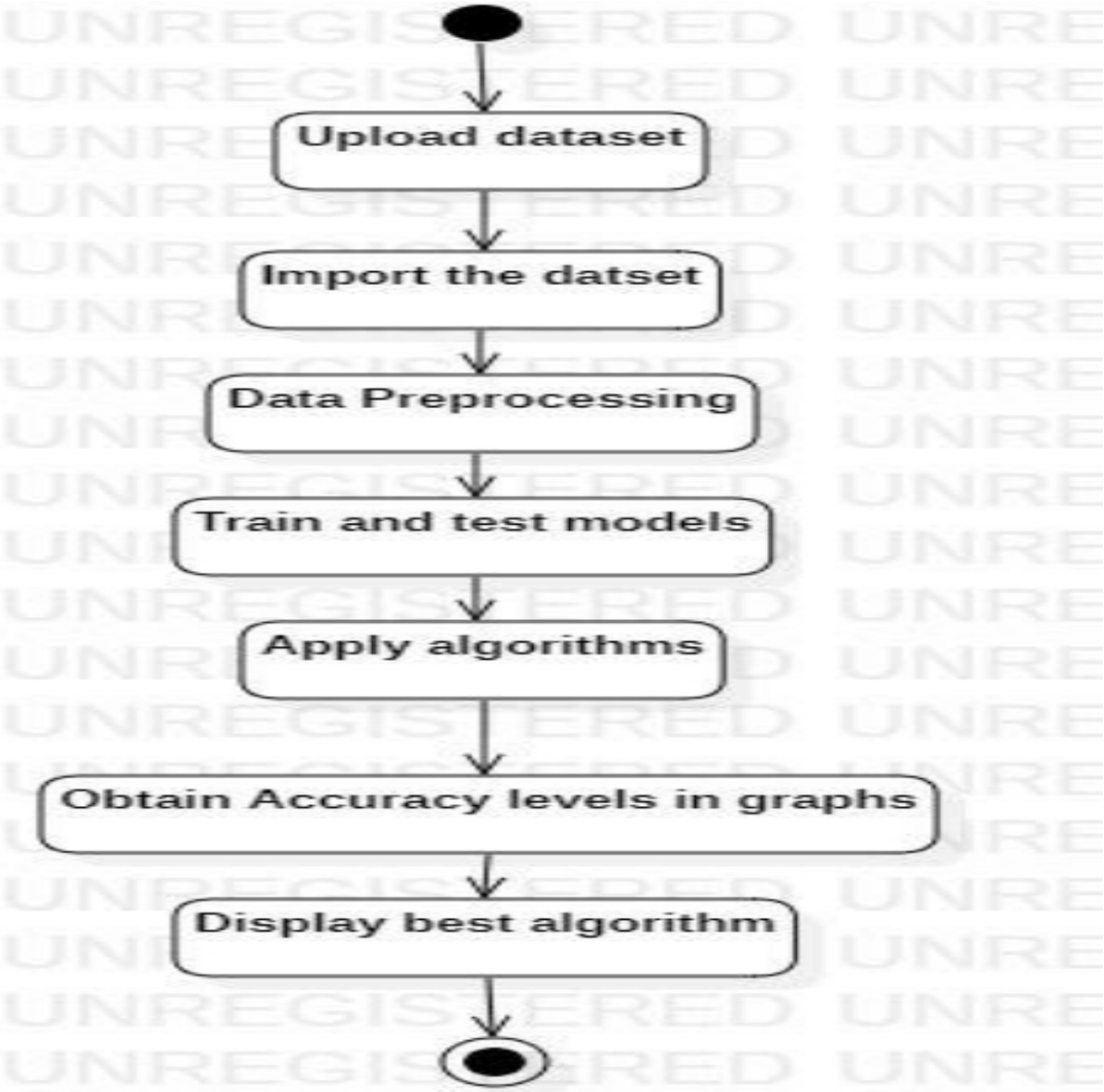


Fig-3.7: Activity diagram

4.IMPLEMENTATION

4.IMPLEMENTATION

4.1 SAMPLE CODE

main.py

```
from tkinter import messagebox
from tkinter import *
from tkinter import simpledialog
import tkinter
from tkinter import filedialog
from tkinter.filedialog import askopenfilename
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import MinMaxScaler
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import confusion_matrix
from sklearn import svm
from sklearn.metrics import accuracy_score
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import make_scorer, accuracy_score,precision_score,recall_score
from sklearn.model_selection import GridSearchCV
from sklearn.neural_network import MLPClassifier
import keras
from keras.models import Sequential
from keras.layers import Dense


main = tkinter.Tk()
main.title("Heart Disease")
main.geometry("1300x1200")
column = ['sbp','tobacco','ldl','adiposity','famhist','type','obesity','alcohol','age','chd']
```

```
def upload():
    global filename
    global data
    text.delete('1.0', END)
    filename = askopenfilename(initialdir = ".")
    pathlabel.config(text=filename)
    text.insert(END,"Dataset loaded\n\n")
def importdata():
    global filename
    global df
    df = pd.read_csv(filename)
    df.columns=column
    text.insert(END,"Data Information:\n"+str(df.head())+"\n")
    text.insert(END,"Columns Information:\n"+str(df.columns)+"\n")
def preprocess():
    global df
    encoder = LabelEncoder()
    df['famhist']=encoder.fit_transform(df['famhist'])
    df['chd']=encoder.fit_transform(df['chd'])
    scale = MinMaxScaler(feature_range =(0,100))
    text.insert(END,"Preprocess Done\n")
    #setting scale of max min value for sbp in range of 0-100, normalise
    df['sbp'] = scale.fit_transform(df['sbp'].values.reshape(-1,1))
    df.head(50).plot(kind='area',figsize=(10,5))
    plt.figure(0)
    df.plot(x='age',y='obesity',kind='scatter',figsize =(10,5))
    plt.figure(1)
    df.plot(x='age',y='tobacco',kind='scatter',figsize =(10,5))
    plt.figure(2)
```

```
    df.plot(x='age',y='alcohol',kind='scatter',figsize =(10,5))
    plt.figure(3)
    df.plot(kind = 'hist',figsize =(10,5))
    plt.figure(4)
    color = dict(boxes='DarkGreen', whiskers='DarkOrange',medians='DarkBlue', caps='Gray')
    df.plot(kind='box',figsize=(10,6),color=color,ylim=[-10,90])
    plt.show()
def ttmodel():
    global df
    global X_train, X_test, y_train, y_test
    # splitting the data into test and train having a test size of 20% and 80% train size
    from sklearn.model_selection import train_test_split
    col = ['sbp','tobacco','ldl','adiposity','famhist','type','obesity','alcohol','age']
    X_train, X_test, y_train, y_test = train_test_split(df[col], df['chd'], test_size=0.2,
random_state=1234)
    text.insert(END,"Shape of Train Data: "+str(X_train.shape)+"\n")
    text.insert(END,"Shape of Test Data: "+str(X_test.shape)+"\n")
    sns.set()
    sns.heatmap(X_train.head(10),robust = True)
    plt.show()
def models():
    global X_train, X_test, y_train, y_test
    global svm_result,knn_result,ann_result
    global recall_svm,recall_knn,recall_ann
    global precision_svm,precision_knn,precision_ann
    svm_clf = svm.SVC(kernel ='linear')
    svm_clf.fit(X_train,y_train)
    y_pred_svm =svm_clf.predict(X_test)
    svm_result = accuracy_score(y_test,y_pred_svm)
    print("Accuracy :",svm_result)
    recall_svm = recall_score(y_test,y_pred_svm)
```

```
    precision_svm = precision_score(y_test,y_pred_svm)
    text.insert(END,"Accuracy of SVM: "+str(svm_result)+"\n")
    text.insert(END,"Recall of SVM: "+str(recall_svm)+"\n")
    text.insert(END,"Precision of SVM: "+str(precision_svm)+"\n")
    knn_clf = KNeighborsClassifier(n_neighbors =5,n_jobs = -1,leaf_size =
60,algorithm='brute')
    knn_clf.fit(X_train,y_train)
    y_pred_knn = knn_clf.predict(X_test)
    knn_result = accuracy_score(y_test,y_pred_knn)
    recall_knn = recall_score(y_test,y_pred_knn)
    precision_knn = precision_score(y_test,y_pred_knn)
    text.insert(END,"Accuracy of KNN: "+str(knn_result)+"\n")
    text.insert(END,"Recall of KNN: "+str(recall_knn)+"\n")
    text.insert(END,"Precision of KNN: "+str(precision_knn)+"\n")
    ann_clf = MLPClassifier()
    parameters = {'solver': ['lbfgs'], 'alpha':[1e-4], 'hidden_layer_sizes':(9,14,14,2),
'random_state': [1]}
    acc_scorer = make_scorer(accuracy_score)
    grid_obj = GridSearchCV(ann_clf, parameters, scoring=acc_scorer)
    grid_obj = grid_obj.fit(X_train, y_train)
    ann_clf = grid_obj.best_estimator_
    # Fit the best algorithm to the data
    ann_clf.fit(X_train, y_train)
    y_pred_ann = ann_clf.predict(X_test)
    ann_result = accuracy_score(y_test,y_pred_ann)
    classifier = Sequential()
    classifier.add(Dense(output_dim = 6, init = 'uniform', activation = 'relu', input_dim = 9))
    classifier.add(Dense(output_dim = 1, init = 'uniform', activation = 'sigmoid'))
    classifier.compile(optimizer = 'adam', loss = 'binary_crossentropy', metrics = ['accuracy'])
    classifier.fit(X_train, y_train, batch_size = 10, nb_epoch = 100)
 # Predicting the Test set results
```

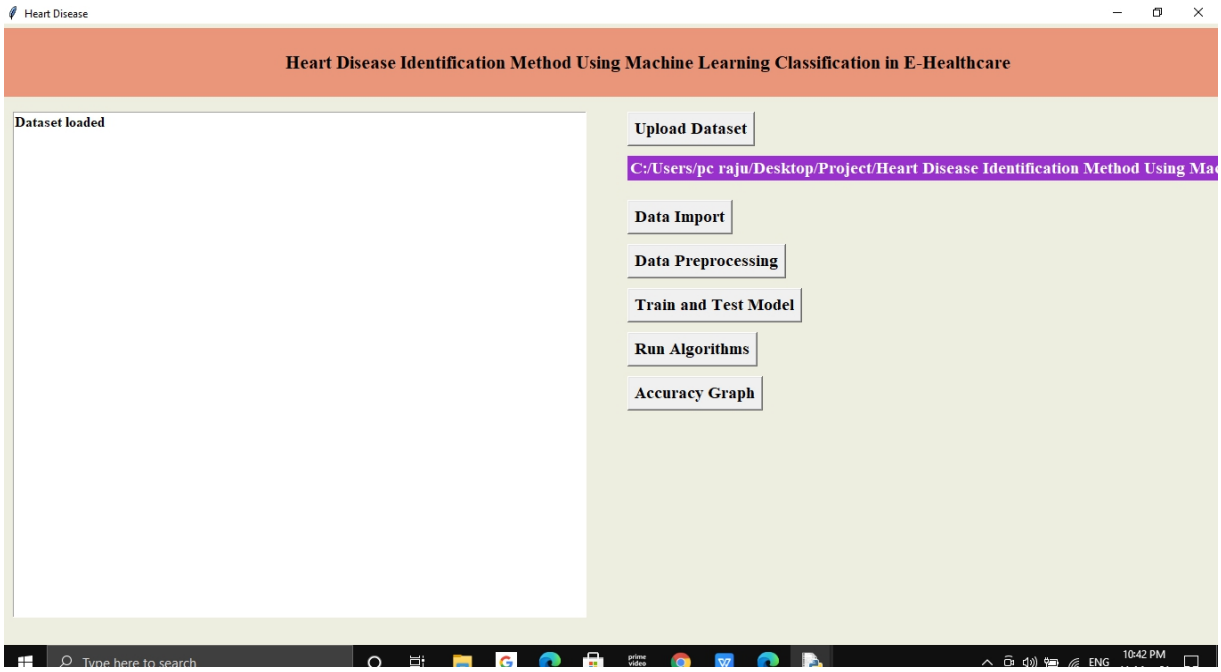
```
    y_pred = classifier.predict(X_test)
    y_pred = (y_pred > 0.5)
    recall_ann = recall_score(y_test,y_pred_ann)
    precision_ann = precision_score(y_test,y_pred_ann)
    text.insert(END,"Accuracy of ANN: "+str(ann_result)+"\n")
    text.insert(END,"Recall of ANN: "+str(recall_ann)+"\n")
    text.insert(END,"Precision of ANN: "+str(precision_ann)+"\n")
def graph():
    global svm_result,knn_result,ann_result
    global recall_svm,recall_knn,recall_ann
    global precision_svm,precision_knn,precision_ann
    results ={'Accuracy': [svm_result*100,knn_result*100,ann_result*100],
            'Recall': [recall_svm*100,recall_knn*100,recall_ann*100],
            'Precision': [precision_svm*100,precision_knn*100,precision_ann*100]}
    index = ['SVM','KNN','ANN']
    results =pd.DataFrame(results,index=index)
    fig =results.plot(kind='bar',title='Comparison of models',figsize =(9,9)).get_figure()
    fig.savefig('Final Result.png')
   fig =results.plot(kind='bar',title='Comparison of models'
,figsize=(6,6),ylim=[50,100]).get_figure()
    fig.savefig('image.png')
    results.plot(subplots=True,kind ='bar',figsize=(4,10))
    plt.show()
font = ('times', 16, 'bold')
title = Label(main, text='Heart Disease Identification Method Using Machine Learning
Classification in E-Healthcare')
title.config(bg='dark salmon', fg='black')
title.config(font=font)
title.config(height=3, width=120)
title.place(x=0,y=5)
font1 = ('times', 14, 'bold')
```

```
upload = Button(main, text="Upload Dataset", command=upload)
upload.place(x=700,y=100)
upload.config(font=font1)
pathlabel = Label(main)
pathlabel.config(bg='dark orchid', fg='white')
pathlabel.config(font=font1)
pathlabel.place(x=700,y=150)
ip = Button(main, text="Data Import", command=importdata)
ip.place(x=700,y=200)
ip.config(font=font1)
pp = Button(main, text="Data Preprocessing", command=preprocess)
pp.place(x=700,y=250)
pp.config(font=font1)
tt = Button(main, text="Train and Test Model", command=ttmodel)
tt.place(x=700,y=300)
tt.config(font=font1)
ml = Button(main, text="Run Algorithms", command=models)
ml.place(x=700,y=350)
ml.config(font=font1)
gph = Button(main, text="Accuracy Graph", command=graph) gph.place(x=700,y=400)
gph.config(font=font1)
font1 = ('times', 12, 'bold')
text=Text(main,height=30,width=80)
scroll=Scrollbar(text)
text.configure(yscrollcommand=scroll.set)
text.place(x=10,y=100)
text.config(font=font1)
main.config(bg='ivory2')main.mainloop()
```

5. SCREENSHOTS

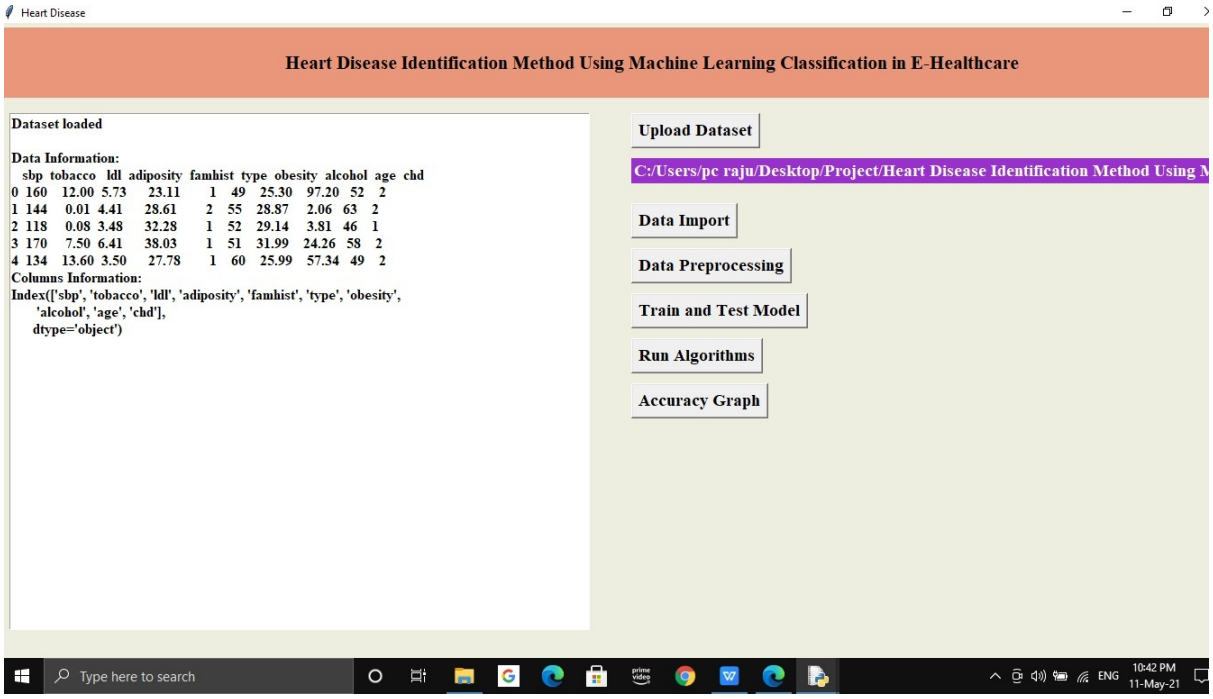
5. SCREENSHOTS

5.1 Dataset uploaded



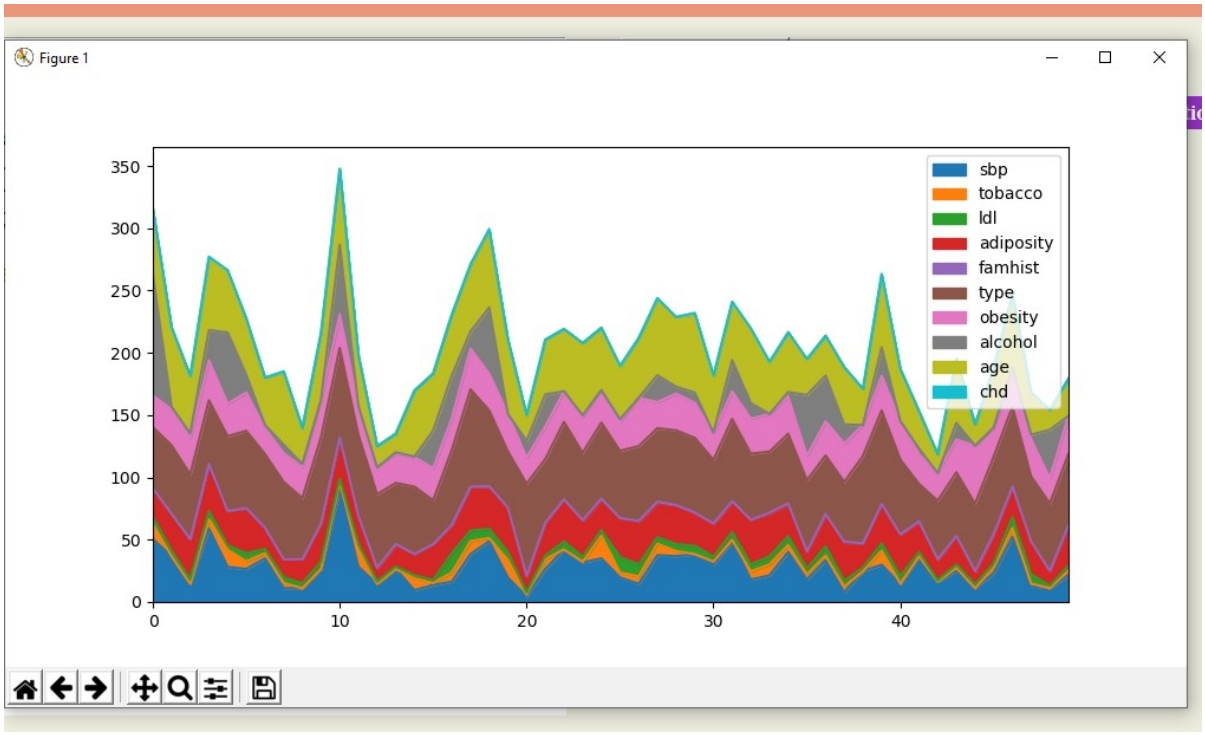
Screenshot 5.1: Dataset uploaded

5.2 Data Import



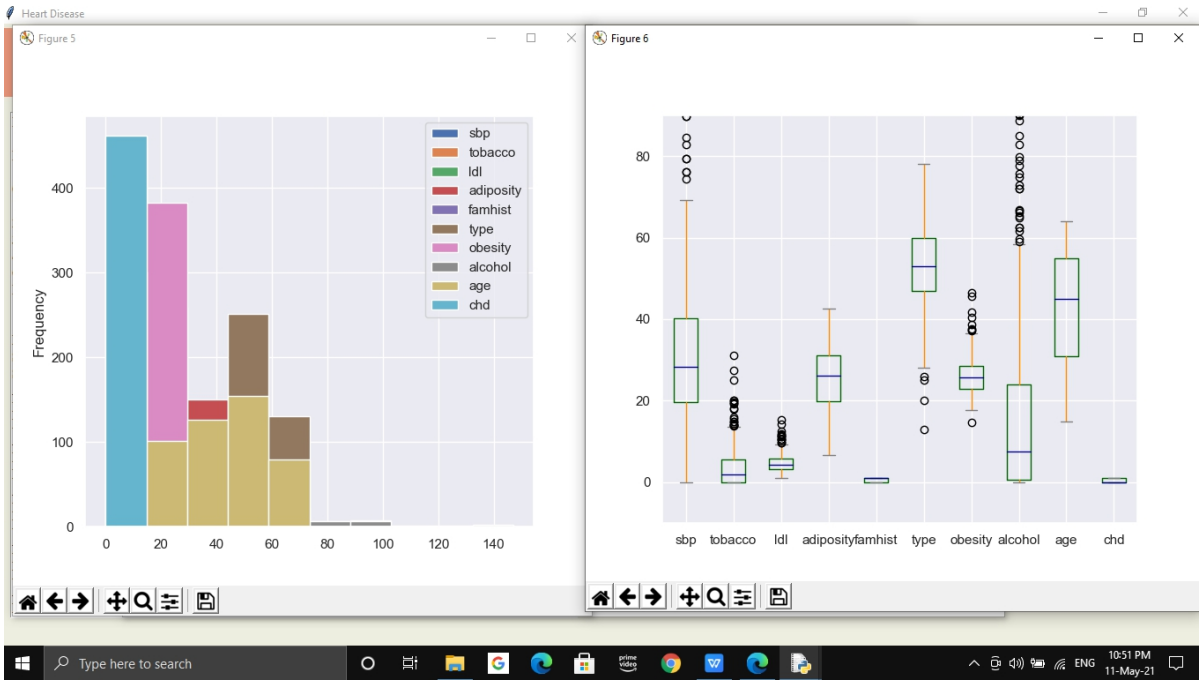
Screenshot 5.2: Data Import

5.3 Data preprocessing



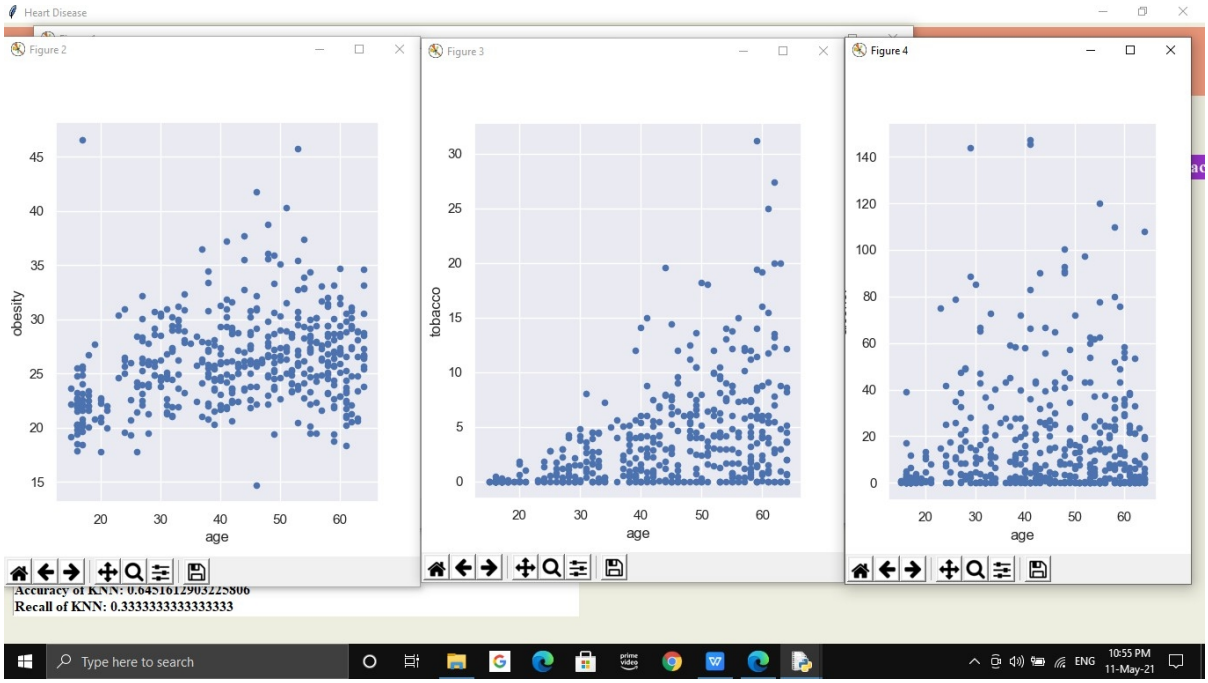
Screenshot 5.3: Data preprocessing

5.4 Frequency of attributes



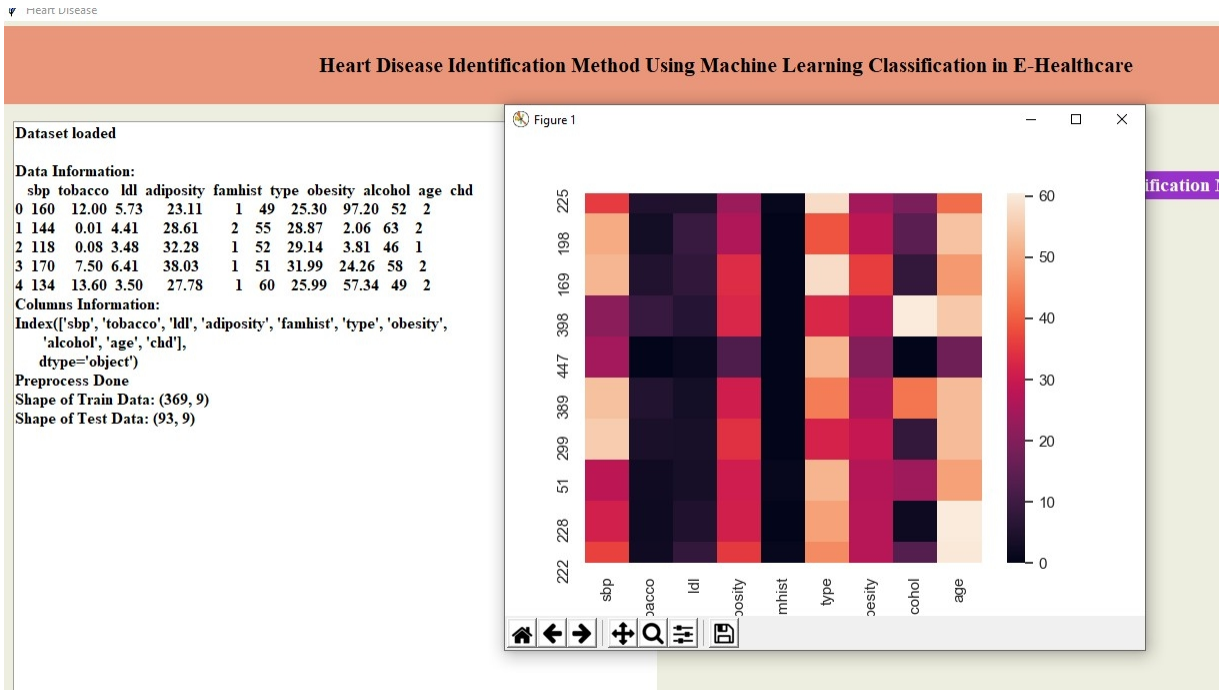
Screenshot 5.4: Frequency of attributes

5.5 Levels of Obesity, Alcohol and Tobacco



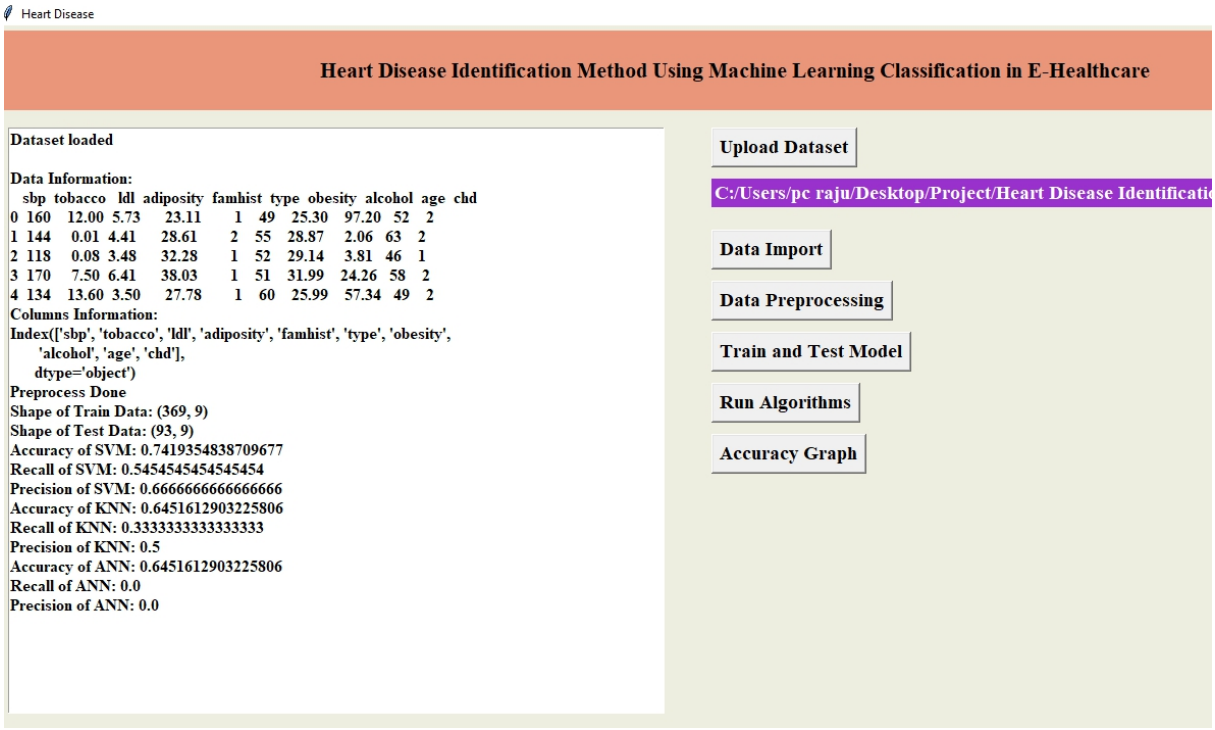
Screenshot 5.5: Levels of Obesity, Alcohol and Tobacco

5.6 Train and test Data



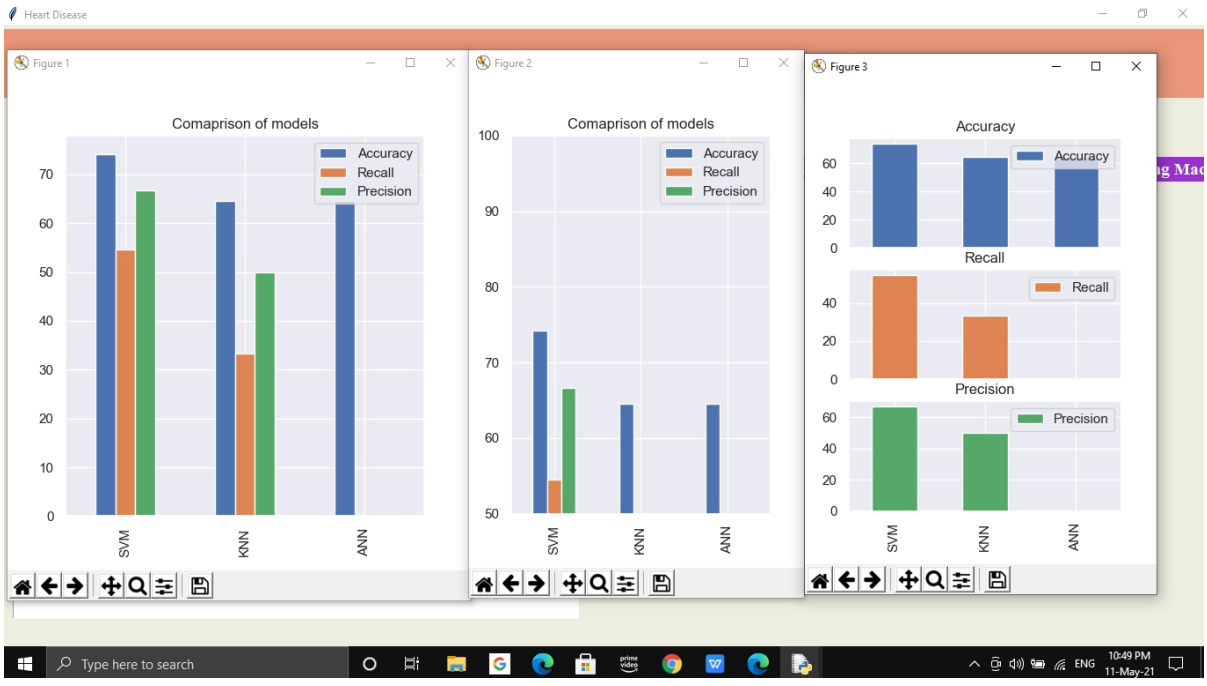
Screenshot 5.6: Train and test Data

5.7 Accuracy levels of three algorithms



Screenshot 5.7: Accuracy levels of three algorithms

5.8 Comparison in form of Graphs



Screenshot 5.8: Comparison in form of Graphs

6. TESTING

6. SOFTWARE TESTING

6.1 INTRODUCTION TO TESTING

Testing is a process of executing a program with the aim of finding error. To make our software perform well it should be error free. If testing is done successfully it will remove all the errors from the software.

6.2 TYPES OF TESTING

6.2.1 Unit Testing

Software verification and validation method in which a programmer tests if individual units of source code are fit for use. It is usually conducted by the development team. Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

6.2.2 Integration Testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

6.2.3 Functional Testing

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures: interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes.

6.2.4 Black Box Testing

Black box testing is testing the functionality of an application without knowing the details of its implementation including internal program structure, data structures etc. Test cases for black box testing are created based on the requirement specifications. Therefore, it is also called as specification-based testing.

When applied to machine learning models, black box testing would mean testing machine learning models without knowing the internal details such as features of the machine learning model, the algorithm used to create the model etc. The challenge, however, is to verify the test outcome against the expected values that are known beforehand.

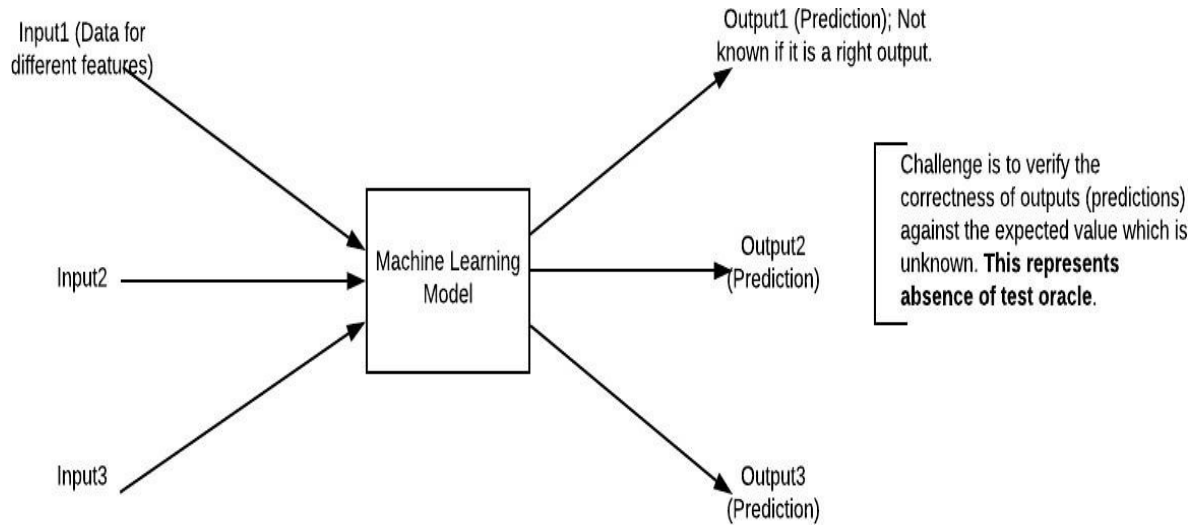


Fig 6.2.4 Black box testing

Input	Actual Output	Predicted Output
[16,6,324,0,0,0,22,0,0,0,0,0,0]	0	0
[16,7,263,7,0,2,700,9,10,1153,832,9,2]	1	1

Table 6.2.4 Black box testing

The model gives out the correct output when different inputs are given which are mentioned in Table 4.1. Therefore the program is said to be executed as expected or correct program.

6.3 TEST CASES

Test Case Id	Test Case Name	Test Case Description	Test Steps			Test Case Status	Test Priority
			Step	Expected	Actual		
01	Start the Application	Host the application and test if it starts making sure the required software is available	If it doesn't Start	We cannot run the applicati on.	The application hosts success.	High	High
02	Home Page	Check the deployment environmen t for properly loading the application.	If it doesn’t load.	We cannot access the applicati on.	The application is running successfully .	High	High
03	User Mode	Verify the working of the application in freestyle mode	If it doesn’t Respond	We cannot use the Freestyle mode.	The application displays the Freestyle Page	High	High
04	Data Input	Verify if the application takes input and updates	If it fails to take the input or store in the database	We cannot proceed further	The application updates the input to application	High	High

7. CONCLUSION AND FUTURE SCOPE

7. CONCLUSION AND FUTURE SCOPE

7.1 PROJECT CONCLUSION

→ In this project, an efficient machine learning based diagnosis system has been developed for the diagnosis of heart disease.

→ Machine learning classifiers include K-NN, ANN, SVM are used in the designing of the system. Four standard feature selection algorithms including Relief, MRMR, LASSO, LLBFS, and proposed a novel feature selection algorithm FCMIM used to solve feature selection problem. LOSO cross-validation method is used in the system for the best hyper parameters selection.

→ The system is tested on Cleveland heart disease dataset. Furthermore, performance evaluation metrics are used to check the performance of the identification system.

→ As we think that developing a decision support system through machine learning algorithms it will be more suitable for the diagnosis of heart disease. Furthermore, we know that irrelevant features also degrade the performance of the diagnosis system and increased computation time. Thus another innovative touch of our study to used features selection algorithms to selects the appropriate features that improve the classification accuracy as well as reduce the processing time of the diagnosis system.

7.2 FUTURE SCOPE

Thus, in our future work we will focus on applying some more advanced algorithms. In the future, we will use other features selection algorithms, optimization methods to further increase the performance of a predictive system for HD diagnosis. The controlling and treatment of disease is significance after diagnosis, therefore, i will work on treatment and recovery of diseases in future also for critical disease such as heart, breast, Parkinson, diabetes

8.BIBILOGRAPHY

8. BIBILOGRAPHY

8.1 REFERENCES

1. X. Liu, X. Wang, Q. Su, M. Zhang, Y. Zhu, Q. Wang, and Q. Wang, ‘‘A hybrid classification system for heart disease diagnosis based on the RFRS method,’’ Comput. Math. Methods Med., vol. 2017, pp. 1–11, Jan. 2017.
2. P. A. Heidenreich, J. G. Trogdon, O. A. Khavjou, J. Butler, K. Dracup, M. D. Ezekowitz, E. A. Finkelstein, Y. Hong, S. C. Johnston, A. Khera, D. M. Lloyd- Jones, S. A. Nelson, G. Nichol, D. Orenstein, P. W. F. Wilson, and Y. J. Woo, ‘‘Forecasting the future of cardiovascular disease in the united states: A policy statement from the American heart association,’’ Circulation, vol. 123, no. 8, pp. 933–944, 2011.
3. A. L. Bui, T. B. Horwich, and G. C. Fonarow, ‘‘Epidemiology and risk profile of heart failure,’’ Nature Rev. Cardiol., vol. 8, no. 1, p. 30, 2011.
4. S. Palaniappan and R. Awang, ‘‘Intelligent heart disease prediction system using data mining techniques,’’ in Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl., Mar. 2008, pp. 108–115.
5. E. O. Olaniyi, O. K. Oyedotun, and K. Adnan, ‘‘Heart diseases diagnosis using neural networks arbitration,’’ Int. J. Intell. Syst. Appl., vol. 7, no. 12, p. 72, 2015.
6. R. Das, I. Turkoglu, and A. Sengur, ‘‘Effective diagnosis of heart disease through neural networks ensembles,’’ Expert Syst. Appl., vol. 36, no. 4, pp. 7675–7680, May 2009.
7. X. Liu, X. Wang, Q. Su, M. Zhang, Y. Zhu, Q. Wang, and Q. Wang, ‘‘A hybrid classification system for heart disease diagnosis based on the RFRS method,’’ Comput. Math. Methods Med., vol. 2017, pp. 1–11, Jan. 2017.

8.2 WEBSITES

- https://llerlymo.com/news-heart-disease-couples-study/
- http://arxiv.org/abs/1811.12808