



**A**  
**PROJECT REPORT**  
**FOR**  
**SUBJECT: LAB II- PROJECT PHASE II**  
**ON**  
**“HUMAN DISEASE PREDICTION USING**  
**MACHINE LEARNING”**

Submitted in partial fulfilment of the requirement for the award of

**Bachelor of Engineering**  
**In**  
**Computer Science and Engineering**  
**Punyashlok Ahilyadevi Holkar Solapur University**

By

Name	Roll. No.	Exam Seat No.
Mr. Shrinivas Satyanarayan Kolpyak	62	622826
Mr. Sandeep Shriniwas Manglaram	63	622967
Mr. Swapnaj Dhananjay Nandgaonkar	64	622652
Mr. Pratik Umakant Tarkasband	66	622619

Under the Guidance of  
**Prof. Mr. Shivappa. M. Metagar**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING  
WALCHAND INSTITUTE OF TECHNOLOGY  
SOLAPUR - 413006  
(2019-2020)**



## Certificate

This is to certify that the project entitled

**“Human Disease Prediction Using Machine Learning”**

Is submitted by

Name	Roll. No.	Exam Seat No.
Mr. Shrinivas Satyanarayan Kolpyak	62	622826
Mr. Sandeep Shriniwas Manglaram	63	622697
Mr. Swapnaj Dhananjay Nandgaonkar	64	622652
Mr. Pratik Umakant Tarkasband	66	622619

(Mr. Shivappa. M. Metagar)

*Project Guide*

(Mrs. Anita Kulkarni)

*Head CSE Dept*

**Dr. S. A. Halkude**

*Principal*

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING  
WALCHAND INSTITUTE OF TECHNOLOGY  
SOLAPUR - 413006  
(2019-2020)**

## **Project Approval Sheet**

The Project Entitled

**“Human Disease Prediction Using Machine Learning”**

Submitted by

Name	Roll. No.	Exam Seat No.
Mr. Shrinivas Satyanarayan Kolpyak	62	622826
Mr. Sandeep Shriniwas Manglaram	63	622697
Mr. Swapnaj Dhananjay Nandgaonkar	64	622652
Mr. Pratik Umakant Tarkasband	66	622619

“Is hereby approved in partial fulfilment for the degree of  
Bachelor of Computer Science and Engineering”

**(Mr. Shivappa. M. Metagar)**

*Project Guide*

**(Mrs. Anita Kulkarni)**

*Head CSE Dept*

**Dr. S. A. Halkude**

*Principal*

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING  
WALCHAND INSTITUTE OF TECHNOLOGY  
SOLAPUR - 413006  
(2019-2020)**

## **Acknowledgment**

At the outset, we would like to take this opportunity to express our deep gratitude to our guide **Mr. Shivappa M. Metagar** of CSE Department for his guidance and moral support throughout this successful completion of our project.

We heartily thank **Dr. Mrs. A.M.Pujar**, Head of CSE Dept for her moral support and promoting us through completion of our project.

We would also like to thank our Principal **Dr. S. A. Halkude** and all staff members for their whole hearted co-operation in completing this project.

## UNDERTAKING

We solemnly declare that project work presented in the report titled “**Human Disease Prediction Using Machine Learning**” is solely my project work with no significant contribution from any other person except project guide. Small contribution/help wherever taken has been duly acknowledged and that complete report has been written by the members of the project group.

We understand the zero tolerance policy of the WIT, Solapur and University towards plagiarism. Therefore we as Authors of the above titled report declare that no portion of the report has been plagiarized and any material used as reference is properly referred / cited.

We undertake that if found guilty of any formal plagiarism in the above titled report even after award of the degree, WIT, Solapur and Solapur University reserves the rights to withdraw/revoke the degree granted and that WIT, Solapur and the University has the right to publish our name on the website on which names of students are placed who submitted plagiarized report.

Name	Exam Number	University PRN Number	Signature
Mr. Shrinivas Satyanarayan Kolpyak	622826	2016032500228267	
Mr. Sandeep Shriniwas Manglaram	622697	2016032500229673	
Mr. Swapnaj Dhananjay Nandgaonkar	622652	2016032500226527	
Mr. Pratik Umakant Tarkasband	622619	2016032500226195	

Date: / /

## Abstract

“Disease Prediction System” is based on predictive modeling which predicts the disease of the user on the basis of the symptoms that are provided by the user as an input to the system. The system analyzes the symptoms provided by the user as input and gives the disease as an output. Disease Prediction is done by implementing the Decision Tree Algorithm, Naïve Bayes Algorithm and Random Forest. This algorithms calculates the probability of the disease and average prediction accuracy probability is greater than 80%. In the decision tree we are comparing values from the root attribute i.e the symptoms gives by the user. On the basis of comparison we follow the branch corresponding to the values and jump for the next comparison.

### **Keywords –**

DT: Decision Tree,

UML: Unified Modeling Language,

DFD : Data Flow Diagram.

DM: Data Mining,

ML: Machine Learning,

# Index

<b>Sr. No.</b>	<b>Title</b>	<b>Page No.</b>
<b>1</b>	<b>Introduction</b>	
	1.1 Introduction	8
	1.2 Problem Statement and Objective	9
<b>2</b>	<b>Background</b>	10
<b>3</b>	<b>Proposed Solution</b>	
	3.1 Solution	11
	3.2 Advantages of proposed system	12
<b>4</b>	<b>Working Environment</b>	
	4.1 Hardware Requirements	13
	4.2 Software Requirements	14
<b>5</b>	<b>Methodology</b>	
	5.1 System Architecture	15
	5.2 Work Flow	16
<b>6</b>	<b>Implementation</b>	
	6.1 Code Snippet	16
	6.2 Screenshots & Results	19
<b>7</b>	<b>Flow diagrams</b>	
	7.1 Data Flow Diagrams	21
	7.2 Sequential UML Diagram	22
<b>8</b>	<b>Project Cost (if required, based on project type)</b>	
<b>9</b>	<b>Future Work</b>	23
<b>10</b>	<b>Conclusion</b>	24
<b>11</b>	<b>References</b>	25

# **Chapter 1**

## **INTRODUCTION**

### **1.1 Introduction**

As the use of internet is growing every day, people are always curious to know different new things. People always try to refer to the internet if any problem arises. People are more addicted to the internet. People do not have immediate option when they suffer with particular disease. So, this system can be helpful to the people as they have access to internet 24 hours.

Today many hospitals have installed databases systems to manage their clinical data or patient data. These information systems typically generate large amounts of data which can be in any format like numbers, text, charts and images but unfortunately, this database that contains rich information is rarely used for clinical decision making. There is much information stored in repositories that can be used effectively to support decision making in healthcare. Data mining techniques are widely used in medical field for extracting data from database. In data mining decision tree is a method which is used extensively. Decision trees are non-parametric supervised learning methods used for classification. The main aim is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The structure of the decision tree is in the form of tree and leaf nodes. Decision trees are most commonly used in operations research, mainly in decision analysis. Advantages are that they are easy to understand and interpret. They are robust, performed well with large datasets, able to handle both numerical and categorical data. By providing efficient treatments, it can help to reduce costs of treatment. Using decision tree algorithm techniques it takes less time for the prediction of the disease with more accuracy.

At present, when one suffers from a particular disease, then the person has to visit to the doctor which is time consuming and costly too. Also if the patient is out of reach of doctors and hospitals it may be difficult for the user as the disease can not be identified. So, if the above process can be completed using an automated program which can save time as well as money, it could be easier to the patient which can make the process easier. There are other Heart related Disease Prediction System using data mining techniques that analyzes the risk level of the patient. Disease Prediction System is a web based application that predicts the disease of the user with respect to the symptoms given by the user. With the help



of Disease Predictor the user will be able to know the probability of the disease with the given symptoms.

## 1.2 Problem Statement and Objective

**Problem Statement:** Prediction of disease with which user is suffering from with the help of symptoms user provided.

**Objective:** To train a Machine Learning Model such that with the symptoms as input it should predict disease accurately.

## **Chapter 2**

# **Background**

Today many hospitals have installed databases systems to manage their clinical data or patient data. These information systems typically generate large amounts of data which can be in any format like numbers, a text, charts and images but unfortunately, this database that contains rich information is rarely used for clinical decision making. There is much information stored in repositories that can be used effectively to support decision making in healthcare.

In data mining decision tree is a method which is used extensively. Decision trees are non-parametric supervised learning methods used for classification. So we are displaying the list of symptoms where user can select the symptoms. By the help of the input the model will predict disease with the given symptoms. We are providing the datasets so the machine will analyse it and will perform the required decisions.

The project is technically feasible as it can be built using the existing available technologies. Disease Prediction System is based on client-server architecture where client is user and server is the machine where datasets are stored.

We are also providing additional information to take from the user like Feedback mechanism which will eventually increase predictive accuracy of our model and try to not repeat the query again.

# Chapter 3

## PROPOSED SOLUTION

### 3.1 Solution

User initially opens Web Application, where he can login if he already has an account or he can register if he is a new user, he will provide all necessary user credentials as well as details regarding his health and physic such as weight, height, blood group etc.

Then user will select one of the options for between chronic and heart disease.

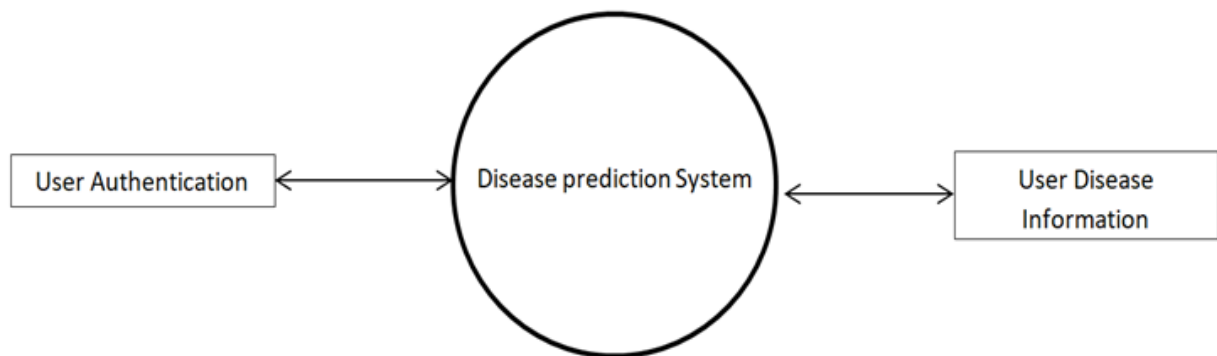


Fig 1. DFD Level 1.

There are n number of symptoms from which that user will select that option that he is suffering from. Based upon the symptoms given by the user on our machine will calculate the probability of the disease so that it will be easier for him to get cured on time.

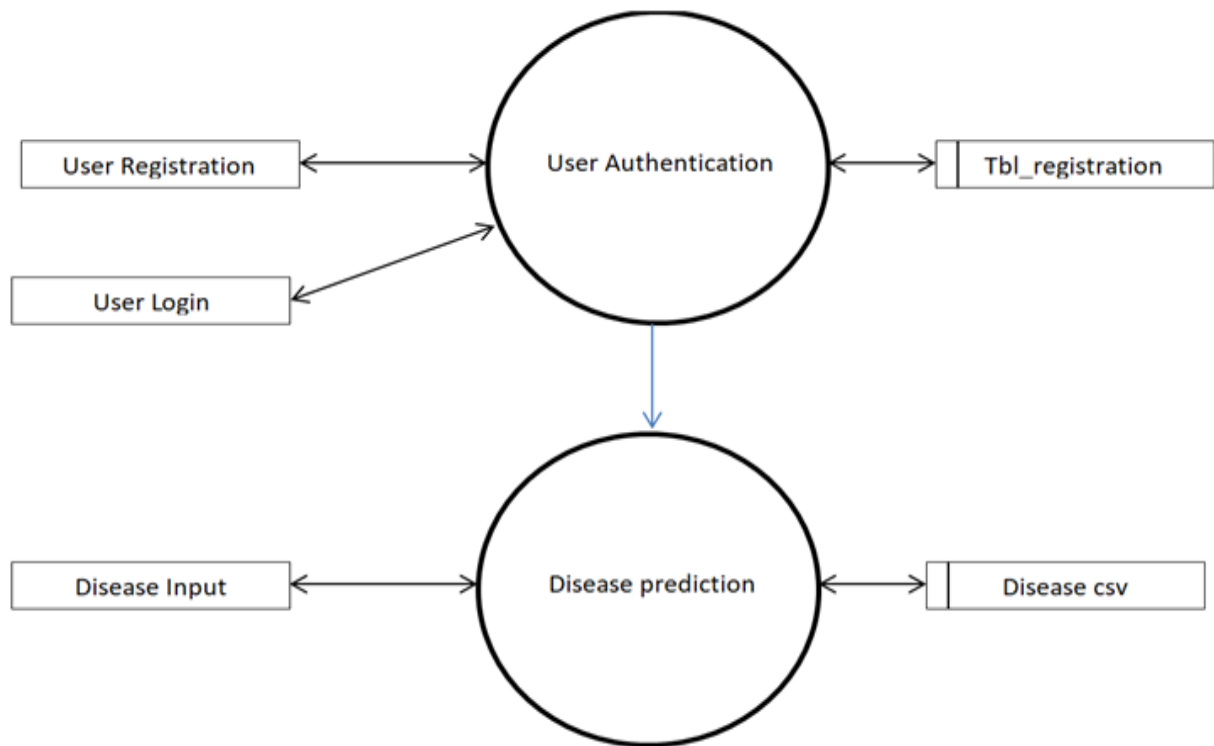


Fig 2. DFD Level 2.

- Dataset of disease and their symptoms is randomly divided into 80-20%.
- 80% of dataset is used to train the data.
- 20% of dataset is used to test the model's working and accuracy.
- This trained model will be used for practical implementation.

### 3.2 Advantages of the proposed system:

#### Random Forest:

It builds multiple decision trees and merges them together to obtain more accurate and stable prediction.

#### Decision Tree:

It is a graphical representation of all the possible solution to a decision based on certain conditions.

#### Naïve Bayes:

It is classification technique based on Bayes' theorem.

All the above algorithms work differently so predictions may differ so to overcome this issue random forest takes mean of all the predictions.

# Chapter 4

## WORKING ENVIRONMENT

### 4.1 Hardware Requirements

- 2 GB RAM
- 20 GB HDD
- 1.8 GHz Processor

### 4.2 Software Requirements

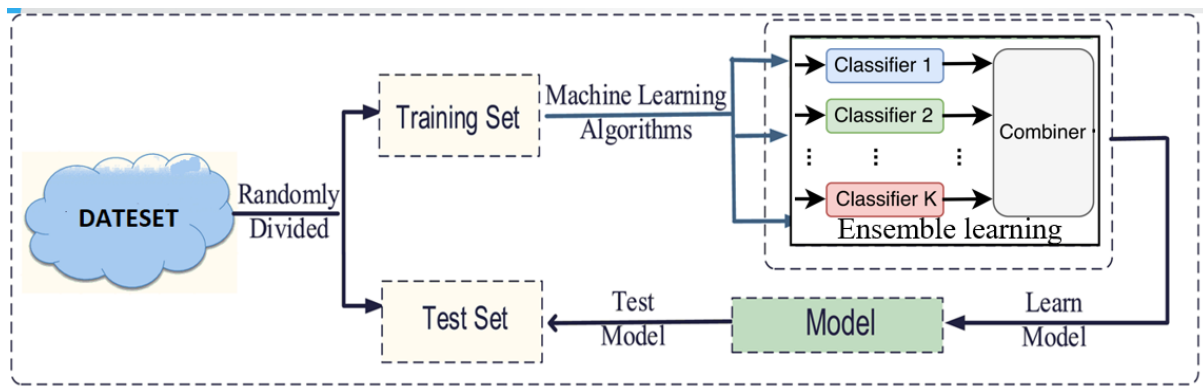
- **Scikit**: Scikit-learn is a free software machine learning library for the python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, K-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.
- **Flask**: Flask is an API of Python that allows to build up web-applications. It was developed by Armin Ronacher. Flask's framework is more explicit than Django's framework and is also easier to learn because it have less base code to implement a simple web-Application.
- **PyCharm**: **PyCharm** is an integrated development environment (IDE) used in computer programming, specifically for the Python language. It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems.
- **HTML**: Hypertext Markup Language is the standard markup language for documents designed to be displayed in a web browser. It can be assisted by technologies such as Cascading Style Sheets and scripting languages such as JavaScript.
- **JavaScript**: JavaScript, often abbreviated as JS, is a programming language that conforms to the ECMAScript specification. JavaScript is high-level, often just-in-time compiled, and multi-paradigm. It has curly-bracket syntax, dynamic typing, prototype-based object-orientation, and first-class functions.

- **CSS** : Cascading Style Sheets is a style sheet language used for describing the presentation of a document written in a markup language like HTML. CSS is a cornerstone technology of the World Wide Web, alongside HTML and JavaScript.

# Chapter 5

## METHODOLOGY

### 5.1 System Architecture:



### 5.2 Work flow:

- Establishing connection between client server web application.
- When compilation is successful web page will appear.
- Provides input of symptoms to web page.
- All the provided symptoms are given as input to Trained Machine Model.
- Disease is predicted from the input and provided to the user.

# Chapter 6

## IMPLEMENTATION

### 6.1 Code Snippet

- **Web Application:**

```
from flask import Flask, request, jsonify, render_template
from flask_cors import CORS
import train
import pickle
import numpy as np

loaded_model = pickle.load(open('decision_tree_classifier.pickle', 'rb'))
a = train.column_headings
app = Flask(__name__)
CORS(app, resource = {r"/symptoms/*" : {"origin" : "*"} })
@app.route("/")
def index():
    return render_template("project.html")
@app.route("/heart")
def heart():
    return render_template("heart.html")
@app.route("/find", methods = ["POST"])
def heartdisease():
    return render_template("heart.html")
@app.route("/symptoms", methods = ["POST"])
def syms():
    vect = np.zeros(len(a)-1)
    symptoms = [ str(s).lower().replace(" ", "_") for s in
request.form.getlist('symptoms[]')]
    for ix in symptoms:
        x = a.index(ix)
        vect[x] = 1
    bimari = loaded_model.predict([vect])[0]
    resp = {
        'desease' : bimari
    }
    return jsonify(resp)
```



```
def main():
    app.run(port = 8000, debug = True)

if __name__ == "__main__":
    main()
```

- **Training Model:**

```
import pandas as pd
data = pd.read_csv("Manual-Data/Training.csv")
data.head()
data.columns
len(data.columns)
len(data['prognosis'].unique())
df = pd.DataFrame(data)
df.head()
cols = df.columns
cols = cols[:-1]
cols
len(cols)
x = df[cols]
y = df['prognosis']

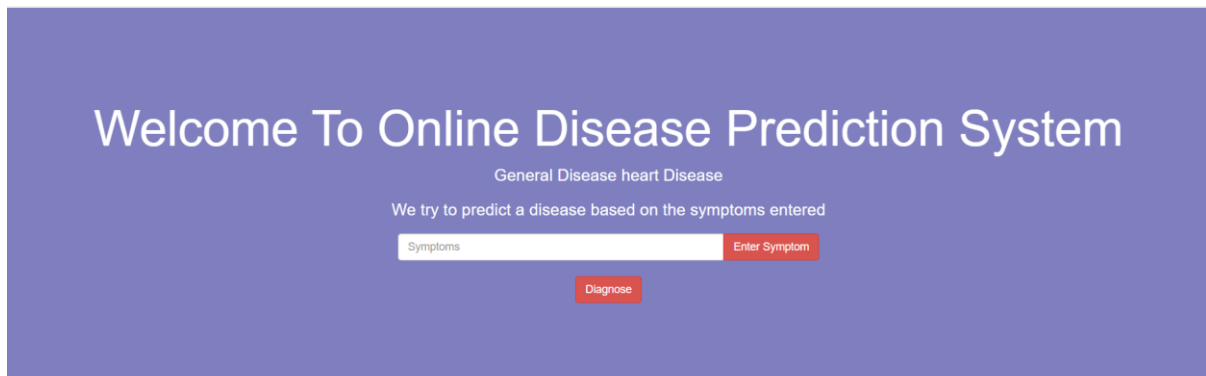
import os
import csv
with open('Manual-Data/Training.csv') as f:
    reader = csv.reader(f)
    i = reader.__next__()
    rest = [row for row in reader]
column_headings = i
for ix in i:
    ix = ix.replace('_', ' ')
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.naive_bayes import MultinomialNB
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.33,
random_state=42)
mnb = MultinomialNB()
mnb = mnb.fit(x_train, y_train)
mnb.score(x_test, y_test)
```

```

from sklearn import model_selection
scores = model_selection.cross_val_score(mnb, x_test, y_test, cv=3)
test_data = pd.read_csv("Manual-Data/Testing.csv")
test_data.head()
testx = test_data[cols]
testy = test_data['prognosis']
mnb.score(testx, testy)
from sklearn.tree import DecisionTreeClassifier, export_graphviz
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.33,
random_state=42)
dt = DecisionTreeClassifier()
clf_dt=dt.fit(x_train,y_train)
from sklearn import model_selection
scores = model_selection.cross_val_score(dt, x_test, y_test, cv=3)
import numpy as np
import matplotlib.pyplot as plt
importances = dt.feature_importances_
indices = np.argsort(importances)[::-1]
features = cols
feature_dict = { }
for i,f in enumerate(features):
    feature_dict[f] = i
feature_dict['hip_joint_pain']
sample_x = [i/79 if i==79 else i*0 for i in range(len(features))]
len(sample_x)
sample_x = np.array(sample_x).reshape(1,len(sample_x))
import pickle
decision_tree_pkl_filename = 'decision_tree_classifier.pickle'
decision_tree_model_pkl = open(decision_tree_pkl_filename, 'wb')
pickle.dump(dt, decision_tree_model_pkl)
decision_tree_model_pkl.close()

```

## 6.2 Screen shots & Results

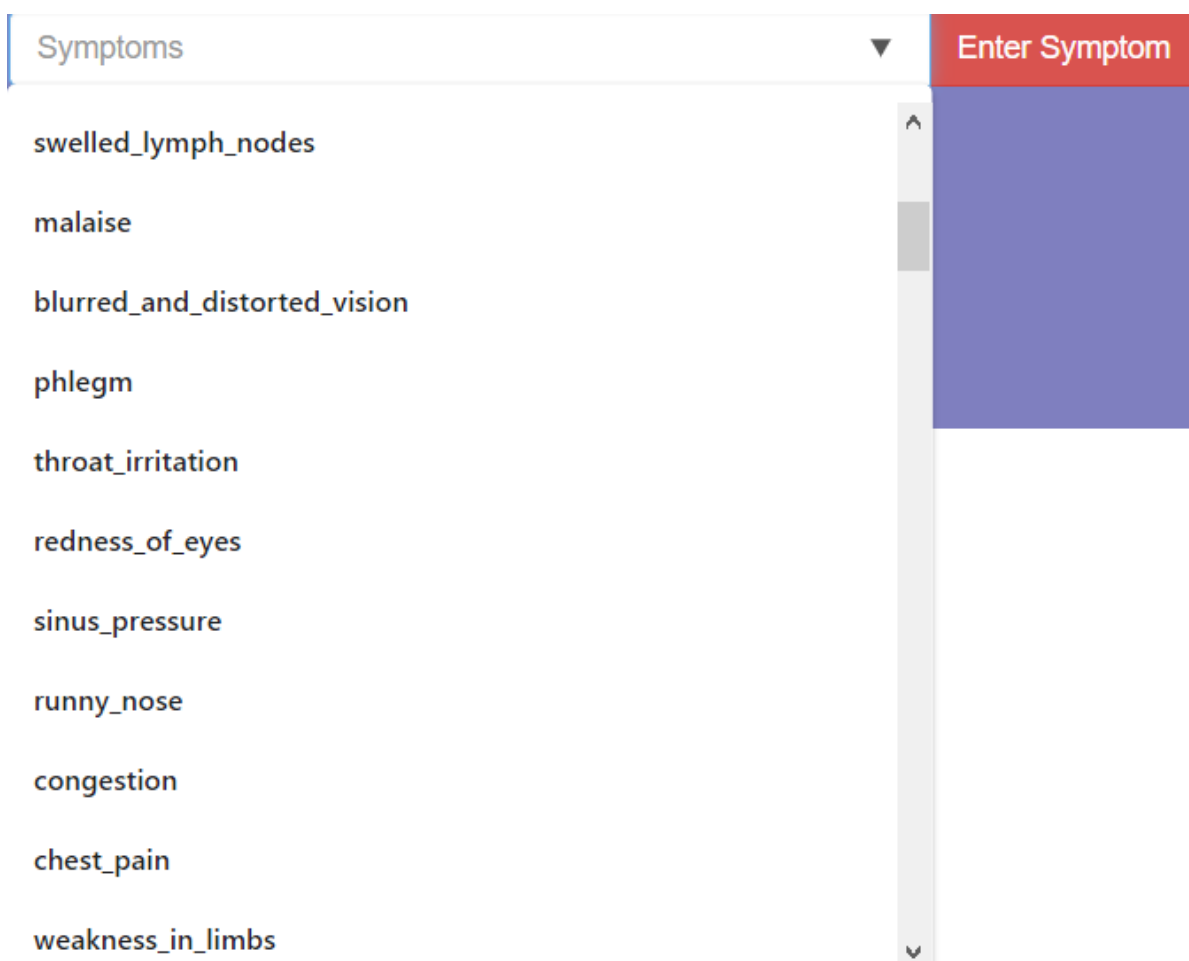


Welcome To Online Disease Prediction System

General Disease heart Disease

We try to predict a disease based on the symptoms entered

Symptoms



Symptoms ▼

- swelled\_lymph\_nodes
- malaise
- blurred\_and\_distorted\_vision
- phlegm
- throat\_irritation
- redness\_of\_eyes
- sinus\_pressure
- runny\_nose
- congestion
- chest\_pain
- weakness\_in\_limbs

# Welcome To Online Disease Prediction System

General Disease heart Disease

We try to predict a disease based on the symptoms entered

Symptoms	Enter Symptom
Diagnose	

- chills
- vomiting
- sweating
- headache
- muscle\_pain

You may have

Malaria  
Malaria  
Malaria

## Heart Disease Predictor

General Disease heart Disease

We try to predict a heart disease based on the Following entered Information

Age	
Sex (1=M,0=F)	
CP (0-4)	
RBP (94-200)	
Serum Chol	
Fasting BP(0-4)	
ECG (0,1,2)	
thalach(71-202)	
exAngina(0/1)	
Old Peak(0-6.2)	
Slope(0,1,2)	
C. A (0-3)	
THAL(0,1,2,3)	
Diagnose	

Apps

127.0.0.1:8000 says  
NO DETECTION OF HEART DISEASES

OK

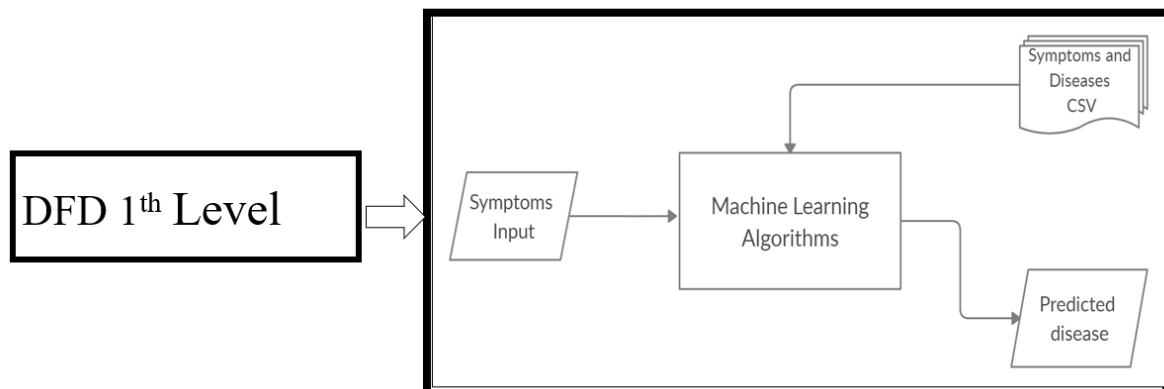
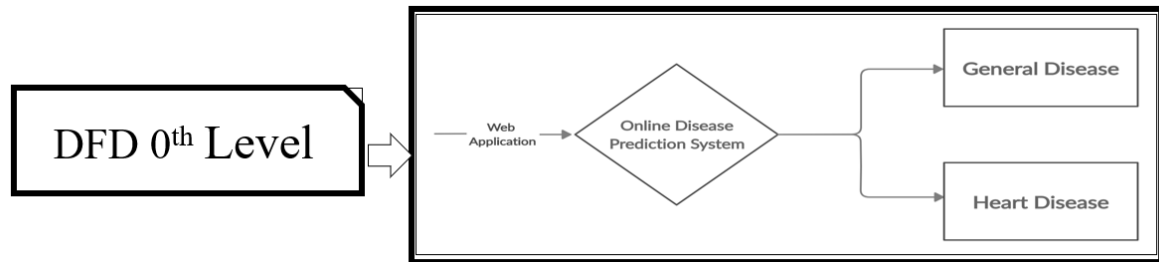
We try to predict a heart disease based on the Following entered Information

30	
0	
2	
140	
289	
1	
1	
128	
1	
4	
3	
3	
7	
Diagnose	

# Chapter 7

## FLOW DIAGRAMS

### 7.1 Data flow diagrams



## 7.2 Sequential UML Diagrams

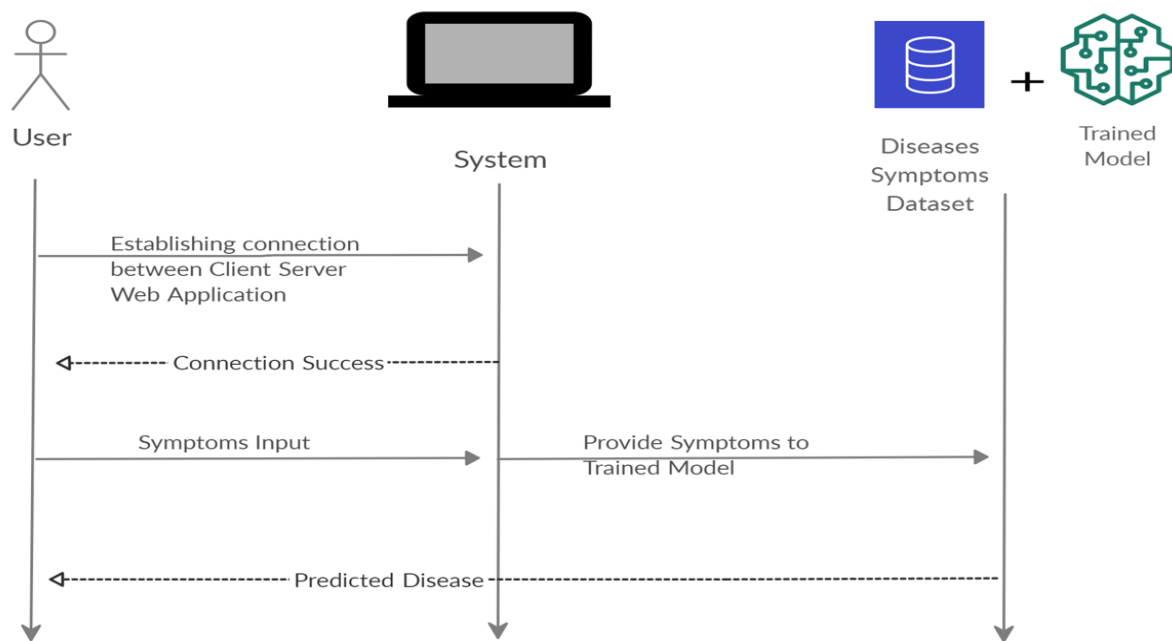


Fig: UML Diagram

## **8. FUTURE WORK**

Training of model and implementation is completed successfully.

## 9. CONCLUSION

Disease Prediction using Machine Learning is successfully implemented with accuracy greater than 80%. Algorithms such as Decision tree, Naïve Bayes, Random Forest , Gaussian works differently so with the help of different algorithms accuracy of output prediction is increased.



# 10. REFERENCES

1. David L. Olson and Dursun, D., Advanced Data Mining Techniques. Springer-Verlag Berlin Heidelberg (2008).
2. Han, J. and Kamber, M., Data Mining: Concepts and Techniques. Second Edition, Morgan Kaufmann Publishers, San Francisco (2006).
3. UCI Machine Learning Databases  
“<https://archive.ics.uci.edu/ml/datasets/heart+Disease>”
4. Sellappan Palaniappan, Rafiah Awang, Intelligent Heart Disease Prediction System Using
  1. David L. Olson and Dursun, D., Advanced Data Mining Techniques. Springer-Verlag Berlin Heidelberg (2008).
  2. Han, J. and Kamber, M., Data Mining: Concepts and Techniques. Second Edition, Morgan Kaufmann Publishers, San Francisco (2006).
  3. UCI Machine Learning Databases  
“<https://archive.ics.uci.edu/ml/datasets/heart+Disease>”
  4. Sellappan Palaniappan, Rafiah Awang, Intelligent Heart Disease Prediction System Using  
Data Mining Techniques, 978-1-4244-1968- 5/08/\$25.00 ©2008 IEEE.