ACADEMIC CITY UNIVERSITY COLLEGE

SELF DIAGNOSIS OF AILMENTS USING PREDICTIVE MODELLING MACHINE LEARNING

HARRIET FIAGBOR

2022

ACADEMIC CITY UNIVERSITY COLLEGE

SELF DIAGNOSIS OF AILMENTS USING PREDICTIVE MODELLING USING MACHINE LEARNING

BY

HARRIET FIAGBOR

10180200241

PROJECT SUBMITTED TO THE SCHOOL OF ENGINEERING, ACADEMIC CITY UNIVERSITY COLLEGE, IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE AWARD OF BACHELOR OF ENGINEERING DEGREE IN COMPUTER ENGINEERING.

MAY 2022

DECLARATION

This is to declare that, the research work documented in this Thesis has been carried out by the under-mentioned student under the supervision of the under-mentioned supervisor. The student and supervisor certify that the work documented in this Thesis is the output of the research conducted by the student as part of final year project work in partial fulfillment of the requirements for the Bachelor of Engineering (BSc) in Electrical and Electronics Engineering degree.

………………………………………………… …………………………

Harriet Fiagbor **(Date)**
(Student)

………………………………………………….. ……………………….

Mr. Julius Amegadzie **(Date)**
(Supervisor)

ABSTRACT

Although there are disadvantages to self-diagnosis, it is a habit that cannot be completely abandoned. Hospitals would be overcrowded if everyone went to a clinic for every ailment, resulting in high costs and long travel time. As a result, others, make it a habit to self-diagnose and avoid going to clinics. If the practice of self-diagnosis becomes common, it can lead to addiction, overdose, and other health complications.

Self-diagnosis enabled by Predictive Modeling is a health-prediction system that serves as an end-user support and online consultation project to address the issue of health facility dependency and avoid sole reliance on over-the-counter (OTC) stores. Here is a proposed system that allows users to get real-time advice on their health issues via an online intelligent health care system. Various symptoms associated with each disease/illness are fed into the system. It then examines the user's symptoms to predict the corresponding disease/illness that may be present. In determining the best prediction model, four data mining techniques or classifiers (Naive Bayes, Random Forest, Decision Trees, and Multi-layer Perceptron Neural Network) were employed. The Random Forest classifier was chosen due to its robustness in terms of performance, that is it reduces overfitting of the data.

ACKNOWLEDGEMENT

I would like to acknowledge the inspirational and instructional guidance, selfless devotion, leadership and coaching skills of my supervisor, Julius Amegadzie of Academic City University College, Ghana. There is no doubt in my mind that without his continued support and counsel I could not have completed this project. I would also like to express my utmost thanks to Pius Gadosey of Academic City College for assisting with his skills in Machine Learning and also Dr. Akosua K. Nyame-Kusi of International SOS for her medical knowledge.

I would also like to thank my families and friends for their continuous encouragement and moral support.

DEDICATION

I dedicate this project to my mother, Ms. Vida Fiagbor for her love, sacrifice and support for me throughout this program. Finally, to God for all the things he has done.

TABLE OF CONTENTS

COVER PAGE ____ i
TITLE PAGE ____ ii
DECLARATION ____ iii
ABSTRACT ____ iv
ACKNOWLEDGEMENT ____ v
DEDICATION ____ vi
TABLE OF CONTENTS ____ vii
LIST OF FIGURES ____ ix
LIST OF ABBREVIATIONS AND SYMBOLS USED ____ xi

CHAPTER ONE - INTRODUCTION ____ 1

1.1 Background ____ 1
1.2 Statement of Problem ____ 2
1.3 Objectives of the study ____ 3
1.3.1 General Objective ____ 3
1.3.2 Specific Objectives ____ 3
1.4 Scope of the Study ____ 3
1.5 Significance of the Study ____ 5
1.6 Limitation ____ 5
1.7 Organization of Study ____ 5
1.8 Technology Used ____ 6

CHAPTER TWO - LITERATURE REVIEW ____ 7

2.1 Introduction ____ 7
2.2 Predictive Modelling ____ 10
2.3 Review of Related Work ____ 11

CHAPTER THREE - MODEL BUILDING ____ 14

3.1 Data Gathering ____ 14
3.2 Data Preprocessing ____ 14
3.3 Exploratory Data Analysis (EDA) ____ 15
3.4 Choosing the right Estimator ____ 18
3.5 Classifiers ____ 19
3.5.1 Multilayer Perceptron Neural Network ____ 19
3.5.2 Decision Tree ____ 20
3.5.3 Random Forest ____ 20
3.5.4 Naïve Bayes ____ 22
3.6 Model Evaluation ____ 24

CHAPTER FOUR - IMPLEMENTATION AND INTERFACE BUILDING ____ 26

4.1 Software Architecture ____ 26

4.2 Design Architecture ____ 27
4.3 User Interface ____ 30

CHAPTER FIVE - RESULTS, CONCLUSION AND RECOMMENDATIONS ____ 40

5.1 Results ____ 40
5.2 Conclusion ____ 48
5.3 Recommendations ____ 48

REFERENCES ____ 49

References ____ 49

LIST OF FIGURES

FIGURE	TITLE	PAGE
2.1:	The General Machine Learning cycle	8
3.2:	Diagram showing proportionality of features in the dataset	15
3.3:	Correlation between symptoms of the dataset	16
3.4:	Table showing generic or common symptoms in our data	17
3.5:	Table showing telling symptoms from our data	18
3.6:	Diagram showing data modelling stages with the four classifiers	24
3.7:	Performance Accuracies of classifiers chosen	25
4.1:	Diagram showing the software architecture of the system	27
4.2:	Diagram showing the Design Architecture of the system	28
4.3:	Diagram showing system user's registration into the system	30
4.4:	Home page of the application	31
4.5:	Users Login	31
4.6:	Pop-up Login Modal	32
4.7:	Patient login	32
4.8:	Patient user Interface	33
4.9:	Entering Symptoms	33
4.10:	Predictions of the system from symptoms	34
4.11:	Feedback modal	34
4.12:	Consultation User Interface	35
4.13:	Doctor Profile	35
4.14:	Consultation History Interface	36

4.15: Consultation and chat Interface…………………………………………..36

4.16: Admin Sign-in Interface………………………………………………….37

4.17: Admin User Interface……………………………………………………..37

4.18: User Table from database………………………………………………...38

4.19: Patient Table from database……………………………………………...38

4.20: Consultation Table from database………………………………………..39

5.1: GERD common symptoms excluding a very telling symptom…………….40

5.2: GERD symptoms including acid reflux as a telling symptom……………..41

5.3: Piles telling symptom bloody stool………………………………………42

5.4: Increasing confidence score of piles by adding telling symptom…………..43

5.5: Prediction of Malaria……………………………………………………..44

5.6: Prediction of Malaria using a telling symptom……………………………..45

5.7: Typhoid fever and a telling symptom……………………………………….46

5.8: Dengue fever with a telling symptom……………………………………….47

LIST OF ABBREVIATIONS AND SYMBOLS USED

DT..Decision Tree

RF..Random Forest

NB..Naive Bayes

MLP...Multilayer Perceptron

MLPNN...Multilayer Perceptron Neural Network

EDA……………………………………………Exploratory Data Analysis

CHAPTER ONE

INTRODUCTION

1.1 Background and Justification of the Study

Self-medication is a human practice in which a person administers medications based on the ailments they may be experiencing. This usually does not necessitate the use of a medical prescription or instruction [1]. These are known as "over-the-counter" or "non-prescription" drugs. Self-medication can also be used to treat stress, anxiety, depression, and psychological trauma [2]. Self-medication is as a result of self-diagnosis.

For both males and females, it is age-independent [3]. Age, gender, income and expenditure, self-care attitude, education level, medical knowledge, satisfaction, and perception of ailments are all characteristics that influence self-medication trends in diverse groups [4]. This is a popular technique for so-called "minor diseases," such as headaches, coughs, flu, and other mild maladies. In these instances, self-medication is resorted to when the discomfort is not so great as to require an appointment with the doctor.

This is common practice in many parts of Africa. When people are unwell, they tend to take their medication and dismiss it as a minor problem. Most of the time, there are no medical specialists available to provide guidance. People do it because it saves them money, time and travelling costs. Some folks do it incorrectly. These drugs may have unfavorable effects on the body because there are no rules or proper prescriptions for them. One of the biggest issues with self-medication is the intake of wrong dosage and underestimating its effects.

It is tough to eliminate self-medication due to the numerous benefits it provides. In actuality, self-medication has been historically used in different cultures to provide a means of healing. In recent times, the associated problems of self-medication can be overcome by technology. Through technology and research, a lot of progress has been made in the medical field. Using emerging technologies such as Artificial intelligence to alleviate societal challenges including self-diagnosis and medication would be a good counter-argument [5]–[7]. It would be a safer idea to have a guide that helps people to self-diagnose their symptoms and provide appropriate prescription for whatever ailments they are experiencing. Although it is not the only possible solution that could put an end to these challenges, it could be a step in the right direction.

1.2 Problem Statement

Self-diagnosis has drawbacks, but it is a habit that cannot be completely abandoned. If every ailment was taken to a clinic or hospital, it would be overcrowded, resulting in expensive costs and travel time. If a person goes to an emergency room for a hypochondriac's concerns, it can be a long wait for little treatment ("Visit the doctor? He will just listen to my breathing and send me away"). A person seeking help for a self-medicated condition faces the risk of an annoying assessment by physicians and filling out unnecessary paperwork [8]. On the other hand, there are those who make a habit of self-medication and do not go to clinics because of the high cost. If the practice is widespread, it can lead to addiction, over dosage, and other problems. This circumstance creates a tough social environment.

There is also the danger of mistaking major illnesses for "minor" ones, which can progress to more serious problems if not handled properly [9], [10]. There are chronic illnesses that appear to be trivial at first. It's not a headache just because it's a headache. Some "minor" illnesses are stepping stones to more significant medical problems. For example, strep throat may seem like a run-of-the-mill cold, but strep throat can also lead to rheumatic fever [11].

1.3 Objectives of the study

1.3.1 Aim – General Objective

The goal of this project is to create an end-user health prediction system that serves as a guide to self-diagnose and also provide online consultation for better medical advice.

1.3.2 Specific Objectives of the Study

The specific objectives of the study are;

- To create a smart health prediction system that serves as a guide in diagnosing and consulting doctors after diagnosis is performed.
- To enable people have access to medical advice in the fastest way possible.
- To reduce over-dependence on healthcare facilities.
- To reduce the risks generated through self-diagnosis.

1.4 Scope of Study

- The health prediction system has three users; doctor, patient and admin.
- Each user of the system is authenticated by the system.
- There is a role-based access to the system.

- The system allows the patient to select symptoms which they are experiencing for disease prediction.
- The system suggests doctors for predicted diseases.
- The system allows online consultation for patients.
- The system helps the patients to consult the doctor at their convenience by sitting at home.

The following are the stages the project will go through;

1. **Data Preparation** is the process of collecting, combining, structuring, and organizing data in order to make it usable.
2. **Data Transformation** is the process of changing the format, structure, or values of data is known as data transformation.
3. **Feature Extraction** is to reduce the number of resources needed to describe a large set of data. One of the major issues with performing complex data analysis is the large number of variables involved. An analysis with a large number of variables usually necessitates a lot of memory and processing power. It may also lead to an overfitting of a classification algorithm to training samples and poor generalization to new samples. Feature extraction is a broad term that refers to methods for constructing combinations of variables to get around these issues while still adequately describing the data [12].
4. **Machine Learning Algorithms Implementation** involves implementing classification algorithms.

5. **Model Creation** is an iterative process in which data is continuously trained and machine learning models are tested to find the best one for the job.

The system preprocesses the data using data mining techniques before applying machine learning algorithms to the results. As a result, an accurate prediction of the likelihood (frequency) of a specific disease or condition can be made.

1.5 Significance of Study

Machine learning is a relatively new digital technology for solving problems in today's world [13]. This technique has grown particularly promising in tackling real-world problems due to its capacity to construct an algorithm and teach it data rather than write it down. This project uses machine learning to improve the possibilities of self-diagnosis by creating a health care prediction system.

1.6 Limitation

This study will not prescribe pharmaceutical or medications that will be used in conjunction with the condition. To avoid keeping sensitive patient data, the system's input will exclude a patient's medical history and/or lab test results. In this case, the product will a be a prototype which is a web based app, so a mobile app will be a secondary priority. Prototyping will be done using a Django framework app. Django is a free and open-source web framework written in Python that follows the model–template–views architectural paradigm.

1.7 Organization of Study

The project is briefly introduced in the first chapter. This chapter discusses the project's background, problem definition, objectives, and general scope, all of which are relevant to the development of a self-medication guide utilizing machine learning. The second chapter discusses research relevant to the project's associated topics, which are primarily about different design implementations of disease prediction using machine learning.

Chapter 3 describes the project's methodology, which includes exploratory data analysis, data transformation, feature extraction, and model creation. The software implementation of the project and interface are discussed in Chapter 4. Finally, Chapter 5 summarizes all of the findings and offers some suggestions for further research.

1.8 Technology Used

Front end: HTML, CSS, Bootstrap, Javascript, Jquery

Back end: Django (python-based web framework)

Database: PostgreSQL Tools: PgMyadmin 4

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

Machine learning, as the name implies, is the process of machines learning without being explicitly programmed or without the need for direct human participation. This machine learning method begins with providing them with high-quality data, followed by training them by creating multiple machine learning models based on the data and various algorithms. The algorithms we use are determined by the type of data we have and the task we are attempting to automate.

As for the formal definition of machine learning, we can say that a machine learning algorithm learns from experience E with respect to some type of task T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E [14].

If a machine learning algorithm is used to play chess, the experience E is gathered from playing many games of chess, the task T is to play chess with a large number of players, and the performance measure P is the probability that the algorithm will win the game of chess.

Supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning are the different categories (methods) of machine learning. This is a project based on supervised learning. The method uses a training dataset to create predictions, which are then compared to the actual output values. If the predictions are not accurate, the algorithm is tweaked until it is perfect. The algorithm will continue to learn until it reaches the needed level of performance. Then, for any additional inputs, it can deliver the desired output values.

As indicated in Fig. 2.1, there are several ways that this research will pursue as a general machine learning problem.

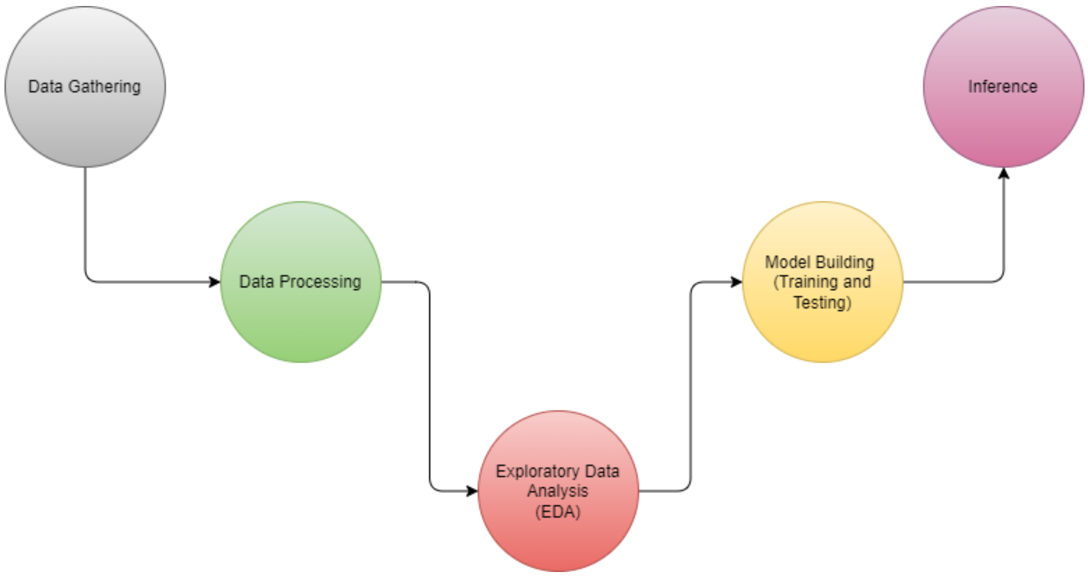

Figure 2.1: The General Machine Learning cycle

1. Gathering of data

The first phase in the machine learning life cycle is data collection. It is one of the most crucial stages of the life cycle. The purpose of this step is to identify and collect all data-related issues. Firstly, identify the numerous data sources, as data can be obtained from a variety of places, including files, databases, the internet, and mobile devices. The output's efficiency will be determined by the quantity and quality of the data collected. The more data you have, the more precise your prediction will be.

2. Data Processing

Data preprocessing is a data mining approach for converting unstructured data into a usable format. Data cleansing, data integration, data transformation, and data reduction are the four primary stages of this process. To ensure quality data and analysis outcomes, data cleaning will filter, detect, and treat unclean data. There may be noise in this scenario due to unrealistic and extreme values, outliers, and missing numbers. Inconsistent data, redundant attributes, and data are examples of errors. Null values in the dataset will be discovered and, if possible, correctly replaced as a first step.

The most crucial phase in a machine learning project is data cleaning. The machine learning model's quality is determined by the quality of the data. As a result, data must always be cleaned before being fed to the model for training.

3. Exploratory Data Analysis

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. The key objectives of EDA include:

- Identifying and extracting variables from the underlying structure.
- Detecting abnormalities and putting assumptions to the test.
- Increasing the value of data sets and selecting the best settings for factors.
- Selecting the tools required for training the data.
- Identifying new data-gathering channels

4. Model Building

The data is ready to be utilized to train a machine learning model after it has been collected and cleaned. The Decision Tree Support Vector Classifier, Multilayer Perceptron Neural Network, Naive Bayes Classifier, and Random Forest Classifier models will be used to train the clean data for this project.

5. Inference

After training the data with the selected four models, the disease for the input symptoms will be predicted by choosing the best performing model. This makes our overall prediction more robust and accurate.

2.2 Predictive Modelling

Predictive modelling, also known as predictive analytics, is a statistical method that attempts to predict future events or outcomes by studying patterns that are likely to tell future occurrences. The purpose of predictive modeling is to answer the following question: "What is most likely to happen in the future based on known past behavior?" [15], [16]

A predictive model isn't static; it's updated or validated on a regular basis to account for changes in the underlying data. Predictive models can also be customized for individual needs by incorporating user input in the form of information or assumptions. That is, this isn't a once-and-done prediction. Predictive models make assumptions based on what has already occurred and what is currently occurring. If new data arrives that reveals changes in the current situation, the impact on the expected future outcome must also be recalculated. These adjustments are made every time new data is acquired.

Machine learning uses predictive modeling to predict a variety of things, such as diseases and weather conditions [17]. Prediction is hard, and it can come roughly right most of the time, but even there, the process is never completely accurate. This project uses predictive modeling to take a patient's symptoms and predict or recognize the illness he or she may be suffering from. Predictive modeling has recently been applied to Parkinson's disease, heart cancer, hypertension, and other conditions as elaborated below.

2.3 Review of Related Work

Much progress has been made in the field of medical diagnosis. Hospitals use database systems to help diagnose patients. These systems consist of doctors entering patients' symptoms. This is basically a "crutch" that helps doctors diagnose patients' ailments.

Much technological research has been done to improve the algorithms of these systems in order to make them autonomous and easy to use [18]. Artificial intelligence has made some of the most significant advances in making these systems more intelligent and assist in the diagnosis of symptoms with greater accuracy.

Machine learning has enabled many advancements in detecting cancerous cells. A typical scenario is a study by Gayathri et al.[19], who discovered that mammogram images occasionally have a risk of false detection, putting the patient's health at risk. They used machine learning algorithms to predict and detect breast cancer in its early stages in order to find alternative and vital approaches to detect cancer. These machine learning methods are easier to implement and used with a variety of data sets, less expensive and safer, and can produce more accurate predictions. Support Vector Machine (SVM) [20], Artificial Neural Network (ANN) [21], K-Nearest Neighbor (KNN) [22], and Decision Tree (DT)

[23] machine learning algorithms were used to predict breast cancer in patients in order to find the machine learning technique that produces the best and most accurate predictions. According to the findings of these researchers, SVM is the most widely used method for cancer detection applications [19]. To improve performance, SVM was used alone or in combination with other methods. The maximum accuracy achieved by SVM (single or hybrid) was 99.8%, which can be improved to 100%.

Another study by Amin Ul Haq et al. [24], used the "Cleveland heart disease dataset 2016" [25] to predict heart disease and discussed the use of machine learning algorithms to detect complex human diseases that can lead to early heart failure. In the system, popular algorithms such as Logistic Regression, K-NN, ANN, SVM, DT, and Naïve Baiyes (NB) were used to distinguish healthy people from those with heart disease. This resulted in the creation of a hybrid intelligent system framework for the prediction of heart disease.

Parkinson's disease is a medical problem that is currently undergoing extensive research, and it is critical to accelerate the development of more accurate models for early detection. Parkinson's disease is a progressive nervous system disease characterized by tremor, muscular rigidity, and slow, imprecise movement that primarily affects the middle aged and elderly. It is linked to degeneration of the brain's basal ganglia and a lack of the neurotransmitter dopamine [26].

The processing of voice signals for detecting Parkinson's disease using machine learning techniques is investigated in a study by Jefferson S. Almeida et al [27]. To classify data obtained from sustained phonation and speech tasks, the method compares the use of eighteen feature extraction techniques and four machine learning methods. Phonation is concerned with the voicing of the vowel /a/, while speech is concerned with the pronunciation of a short sentence in Lithuanian. The audio tasks were recorded using two microphone channels from an acoustic cardioid and a smartphone, allowing researchers to compare the performance of various microphone types. The researchers found that phonation tasks were more effective than speech tasks in detecting disease.

Gopi Battineni et al [28] conducted a study on machine learning (ML) predictive models in the diagnosis of chronic diseases. Because each machine learning method has advantages and disadvantages, their findings suggest that there are no standard methods for determining the best approach in real-time clinical practice. Support vector machines (SVM), logistic regression (LR), and clustering methods were the most commonly used methods among those considered. These models are extremely useful in the classification and diagnosis of chronic diseases, and they are expected to play a larger role in medical practice in the near future. The goal of the study is to use machine learning techniques to detect chronic diseases like diabetes, hepatitis, and other similar conditions early on, preventing medical complications.

These medical advances have aided the medical industry in a lot of aspects.

Machine learning has made major advances to disease diagnosis, detection, and prediction in medicine, according to all of the studies reviewed. These algorithms have mostly enabled scientists to discover new diseases earlier and with greater accuracy.

CHAPTER THREE

MODEL BUILDING

3.1 Data Gathering

The first step in a machine learning task is to gather, measure, and compile the necessary and targeted data from internal or external data sources, then compile it into a pre-existing system. For some, this may be a simple task, but for those with large amounts of data, it is a time-consuming process.

Data cleaning will filter, detect, and handle dirty data to ensure quality data and quality analysis results. In this case, there may be noise due to extreme values, outliers, and missing values. The errors may include inconsistent data and redundant attributes and data. This research makes use of dataset that is acquired on Kaggle (a reputable and reliable data science and machine learning platform). Kaushil268's disease prediction dataset [29] is used in this project. The dataset consists of 42 disorders, including chronic conditions. There are two types of data in the dataset: training data and test data.

3.2 Data Preprocessing

Data preprocessing is a data mining approach for converting unstructured data into a usable format. Data cleaning, data integration, data transformation, and data analysis are the four primary stages of this process. In total, there are 42 different types of illnesses in the training data. There are 42 rows of records and 133 columns of characteristics in this collection of data. To start preprocessing, null values in the dataset will be identified and, if possible, replaced. There was just one column that held null values after the discovery of null values, and this column was eliminated.

3.3 Exploratory Data Analysis (EDA)

Our dataset must be balanced to some extent, as this will have an impact on our model's training. When the dataset is unbalanced, the model will look down upon a certain disease, highlighting its preponderance. It's vital to train models on well-balanced datasets to avoid distribution bias in predicting abilities (unless there is a specific application that requires a given class to be weighted more heavily) [30]. As a result, as shown in Fig 3.1, the dataset was analyzed for balance.

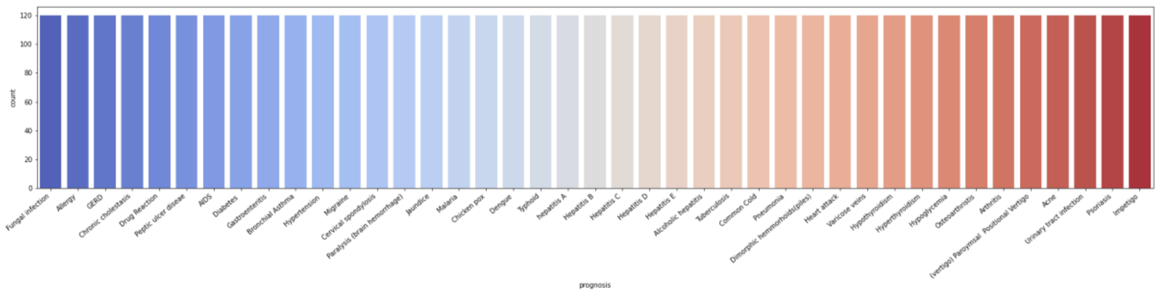

Fig 3.2: Diagram showing proportionality of features in the dataset

A correlation heat map is used to display all of the correlation coefficients in order to detect multi-collinearity, or high inter-correlation between two qualities over an absolute value of 0.5. When there are two multi-collinear attributes, one will be eliminated because it would be redundant to include both of them with nearly mirrored values and hence practically perfect descriptions of each other. Preventing overfitting is another factor.

Through the type of correlation and its strength, the correlation will compare and characterize the linear connection and link between pairs of features. A positive correlation means that both features will change in the same direction, whereas a negative correlation means they will change in opposite directions. The stronger the connection and association, the higher the correlation strength.

Also, when analyzing a correlation, it's important to distinguish between correlation and causality, because while a correlation shows that two variables are related, it doesn't necessarily mean that one causes the other (cause-and-effect relationship).

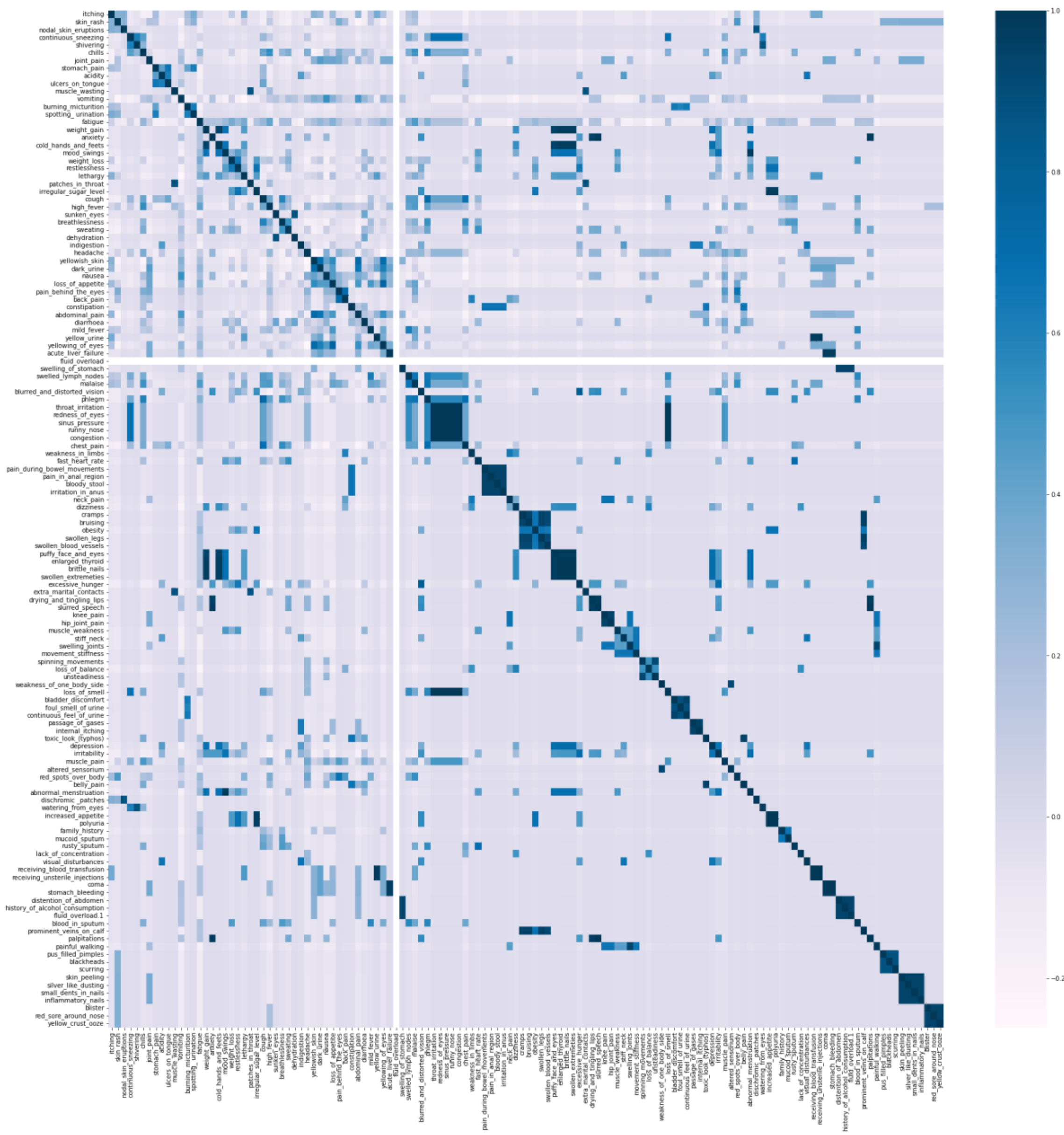

Fig 3.3: Correlation between symptoms of the dataset

Almost all symptoms have weak linear correlations, which is indicative that these symptoms do not come hand-in-hand.

In order to further analyze the data, the dataset was evaluated for common and generic symptoms. Fatigue, vomiting, and high temperature are all common signs of a various ailments, according to our data.

	Symptoms	Prognosis	length
41	fatigue	[Diabetes , Bronchial Asthma, Jaundice, Chicke...	17.0
122	vomiting	[GERD, Chronic cholestasis, Peptic ulcer disea...	17.0
46	high_fever	[AIDS, Bronchial Asthma, Jaundice, Malaria, Ch...	12.0
72	nausea	[Chronic cholestasis, Malaria, Dengue, Typhoid...	10.0
61	loss_of_appetite	[Chronic cholestasis, Peptic ulcer diseae, Chi...	10.0
45	headache	[Hypertension , Migraine, Paralysis (brain hem...	10.0
0	abdominal_pain	[Chronic cholestasis, Peptic ulcer diseae, Jau...	9.0
131	yellowish_skin	[Chronic cholestasis, Jaundice, hepatitis A, H...	8.0
130	yellowing_of_eyes	[Chronic cholestasis, hepatitis A, Hepatitis B...	7.0
101	skin_rash	[Fungal infection, Drug Reaction, Chicken pox,...	7.0

Fig 3.4: Table showing generic or common symptoms in our data

The dataset was analyzed for the most significant or telling symptoms, of which there are several, such as patches in throat, with an AIDS prognosis.

	Symptoms	Prognosis	length
42	fluid_overload	[]	0.0
126	weight_gain	[Hypothyroidism]	1.0
87	pus_filled_pimples	[Acne]	1.0
86	puffy_face_and_eyes	[Hypothyroidism]	1.0
85	prominent_veins_on_calf	[Varicose veins]	1.0
84	polyuria	[Diabetes]	1.0
82	patches_in_throat	[AIDS]	1.0
81	passage_of_gases	[Peptic ulcer diseae]	1.0
80	palpitations	[Hypoglycemia]	1.0
78	pain_in_anal_region	[Dimorphic hemmorhoids(piles)]	1.0

Fig 3.5: Table showing telling symptoms from our data

3.4 Choosing the Right Estimator

The estimators considered in this situation are:

- Decision Tree Classifier (DT)
- Random Forest Classifier (RF)
- Naive Bayes Classifier (NB)
- Multilayer Perceptron Neural Network (MLPNN)

The most appropriate classification models are DT, RF, NB, and MLPNN, which are all immune to multi-collinearity. Because all of the models are non-parametric, the Decision Tree evaluates only one of the features at a time during the splitting process. Similarly, while neural networks (MLPNN) are parallel, they tend to be over-parameterized. Random Forests consider only a subset of all characteristics, reducing the feature space that each tree optimizes over and therefore combating the effects of multi-collinearity.

The Naive Bayes algorithm, on the other hand, employs the Bayes theorem of probability. It assumes that the presence of one feature has no bearing on the presence or absence of another feature, regardless of how linked the features are. For these reasons, and because all predictors in this illness prediction project have Boolean values, no characteristics were eliminated in order to avoid losing relevant information and degrading the overall Exploratory Data Analysis (EDA) and supervised machine learning prediction process.

3.5 Classifiers

3.5.1 Multi Layer Perceptron Neural Network

A feedforward artificial neural network called a multilayer perceptron (MLP) is a type of feedforward artificial neural network. There are at least three levels of nodes in an MLP: an input layer, a hidden layer, and an output layer. Each node, with the exception of the input nodes, is a neuron with a nonlinear activation function. MLP is distinguished from a linear perceptron by its numerous layers and non-linear activation. It can differentiate non-linearly separable data.

- The input layer is supplied with a vector of predictor variable values (x1... xp). The input layer (or the processing that comes before it) standardizes these values so that each variable's range is -1 to 1. The values are distributed to each of the neurons in the hidden layer by the input layer. In addition to the predictor variables, each of the hidden layers receives a constant input of 1.0 called the bias, which is multiplied by a weight and added to the sum going into the neuron.
- When the value from each input neuron reaches a hidden layer neuron, it is multiplied by a weight (wji), and the weighted values are joined together to get a

combined value uj. A transfer function receives the weighted sum (uj) and returns a value hj. The output layer receives the outputs from the hidden layer.

- When the value from each hidden layer neuron reaches a neuron in the output layer, it is multiplied by a weight (wkj), and the weighted values are joined together to get a combined value, vj. The weighted sum (vj) is sent via a transfer function, which generates the value yk. The network's outputs are represented by the y values.

3.5.2 The Decision Tree Classifier

Decision trees are a type of categorization strategy that is both successful and adaptable. It's utilized in image categorization and pattern recognition. Because of its versatility, it is employed for categorization in exceedingly complicated problems. It can also be used to enroll higher-dimensional problems. The root, nodes, and leaf are the three main components. The root represents the attribute that has the greatest impact on the outcome; the leaf represents the attribute that is tested for value, and the leaf is the tree's output.

Advantages:

- The categorization tree that results is simpler to comprehend and interpret. It is simpler to prepare data.
- Multiple data kinds are supported, including numeric, nominal, and categorical.

3.5.3 Random Forest

The Random Forest algorithm is a supervised learning algorithm. In Machine Learning, it solves both classification and regression issues. This method combines numerous classifiers to solve a difficult problem. This improves the model's performance, which is referred to as ensemble learning.

A Random Forest is a technique that combines a variety of decision trees to forecast subsets of a dataset and then averages the results to increase the dataset's predictive accuracy. The random forest algorithm uses many decision trees instead of just one. It takes the predictions from each tree and forecasts the final output based on the majority vote of predictions. The higher the number of trees in the forest, the more accurate it is, avoiding the problem of overfitting. The random forest algorithm's illness prediction technique is as follows:

1. The algorithm selects k symptoms randomly from total symptoms m where it builds a decision tree using these K symptoms.
2. In the next step, it repeats step 1 multiple times to get a total of n decision trees.
3. Pass a random variable to n decisions to predict the disease.
4. Frequent predicted diseases are calculated, and the most frequent one is decided as final

Advantages:

- In comparison to DT, RF has a decreased possibility of variation and overfitting of training data. It calculates the average value from the results of its decision trees.
- This ensemble-based classifier outperforms individual base classifiers, such as DTs, in terms of performance.

- It handles huge datasets nicely. It can calculate which variables or attributes are most important in the classification.

3.5.4 The Nave Bayes algorithm

The Naive Bayesian classifier is based on a theorem known as Bayes' theorem, which presupposes predictor independence. A Naive Bayes model is simple to construct and does not require complicated iterative parameter estimates, making it suitable for huge datasets. Although the Naive Bayesian classifier is a basic method, it frequently performs well and is widely used. It's popular because it frequently outperforms more advanced categorization methods. It obtains results using the Bayes' theorem and strong independence assumptions between the features.

The Bayes theorem works on conditional probability. When something depends on a previous event and is predicted to occur as a result, this is known as conditional probability.

Conditional probability =

P(A|B) = P(B|A).P(A)/P(B)

Where,

P(A): The probability of a hypothesis being true also known as prior probability.

P(B): The probability of the evidence.

P(A|B): The probability of the evidence given that the hypothesis is true.

P(B|A): The probability of the hypothesis given that the evidence is true

Advantages of Naïve Bayes Theorem

- Simple and very useful for large datasets
- It can be used for both binary and multi-class classification problems.
- It requires a lower amount of training data.
- It can make probabilistic predictions and can handle both continuous and discrete data.

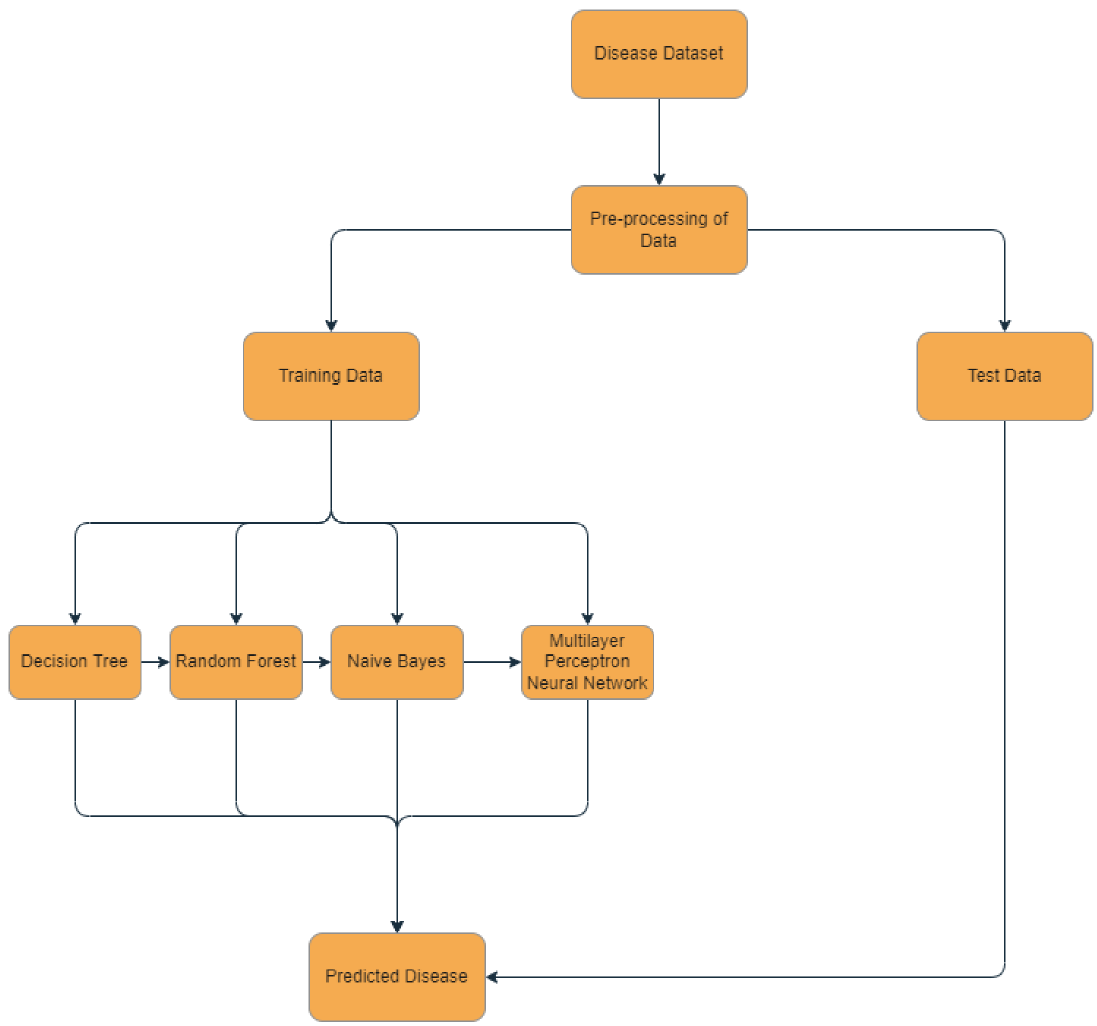

Fig 3.6: Diagram showing data modelling stages with the four classifiers

3.6 Model Evaluation

On our dataset, each of the models chosen performed exceptionally well. As a result, one had to be chosen from the group to be saved as our trained model. The MLPNN gave an accuracy of 97.52% on the test data. This is quite a good score but compared to the rest, it performance is not top notch. DT, RF and NB performed well on the training and test data. There is the need to choose only one of the classifiers to serve as our main model. Because the three classifiers performed well, one had to be selected. The Random Forest Classifier was chosen because it works as an ensemble classifier and reduces data overfitting.

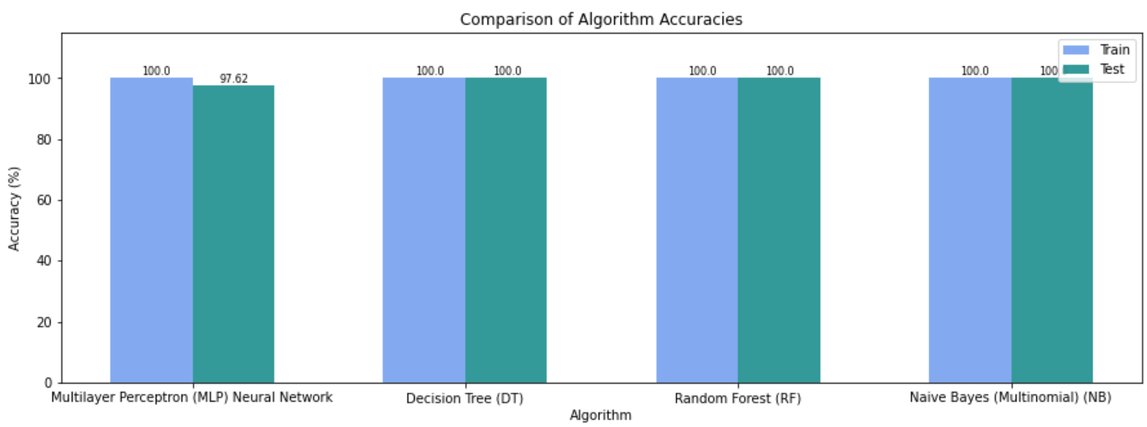

Fig 3.7: Performance Accuracies of classifiers chosen

CHAPTER FOUR

IMPLEMENTATION AND INTERFACE BUILDING

4.1 Software architecture

On the user side, the software architecture follows an encoded path from machine learning model to hosted website, where the user has access to patient portal information and administrator have access to the database for all of the evaluated symptoms and predictions. The doctor has the access to check and discuss with the patient, depending on the seriousness of the situation.

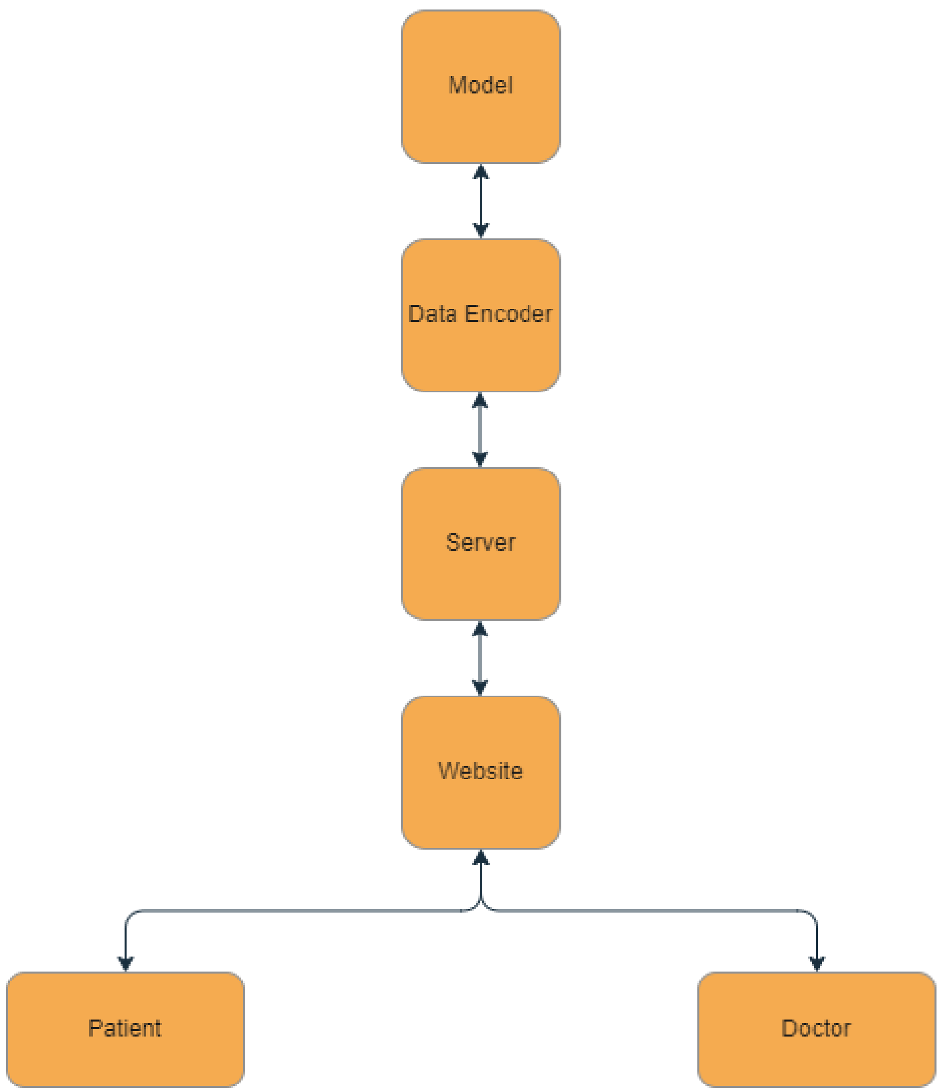

Fig 4.1: Diagram showing the software architecture of the system

4.2 Design Architecture

The system is an online consultation and end-user support project. A proposed method that allows users to receive immediate guidance on their health concerns via an online intelligent health care system. Various symptoms and the disease/illness linked with those symptoms are supplied into the system. Users will be able to share their symptoms and problems with the system. It then examines the user's symptoms to see if there are any illnesses that could be linked to them. Here, the machine learning model trained is used to determine the most likely sickness to be linked to a patient's symptoms.

When a doctor checks in to the system, he or she can access his or her patient's information as well as the report. Doctors can see information regarding the patient search, as well as what the patient searched for based on their predictions. Doctor has access to his own personal information.

Admin can add new disease details to the database by defining the disease type and symptoms. The database contains many diseases and symptoms that the administrator can view. When the user defines the symptoms of his sickness, the system will provide appropriate guidance.

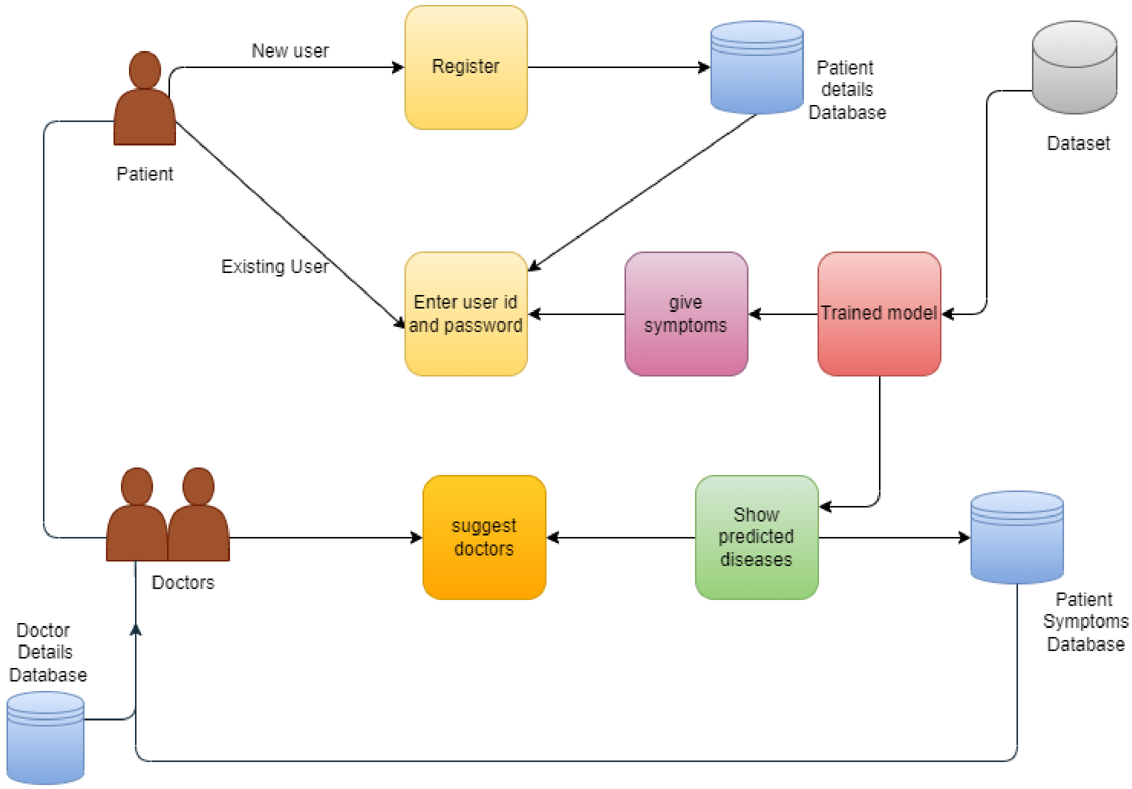

Fig 4.2: Diagram showing the Design Architecture of the system

The overall system is user-friendly and mainly consists of three users:

- Patients
- Doctors
- Admin

The modules and functionalities present in the patient dashboard are:

- Check disease
- Consultation
- Feedback
- View and update your profile.

The user must first register, and then sign in to view the features available. If the user is a patient, the features available include checking the disease by providing symptoms, consultation information, and a feedback option. The patient can access and edit the information they submitted during registration.

The modules and functionalities present in the doctor's dashboard are:

- Consultation
- Feedback
- Update Profile

The modules and functionalities present in the admin's dashboard are:

- Add and change groups
- Add and Change Users
- View Doctors and patients
- View Disease Information
- View ratings and feedbacks
- View chats and consultations

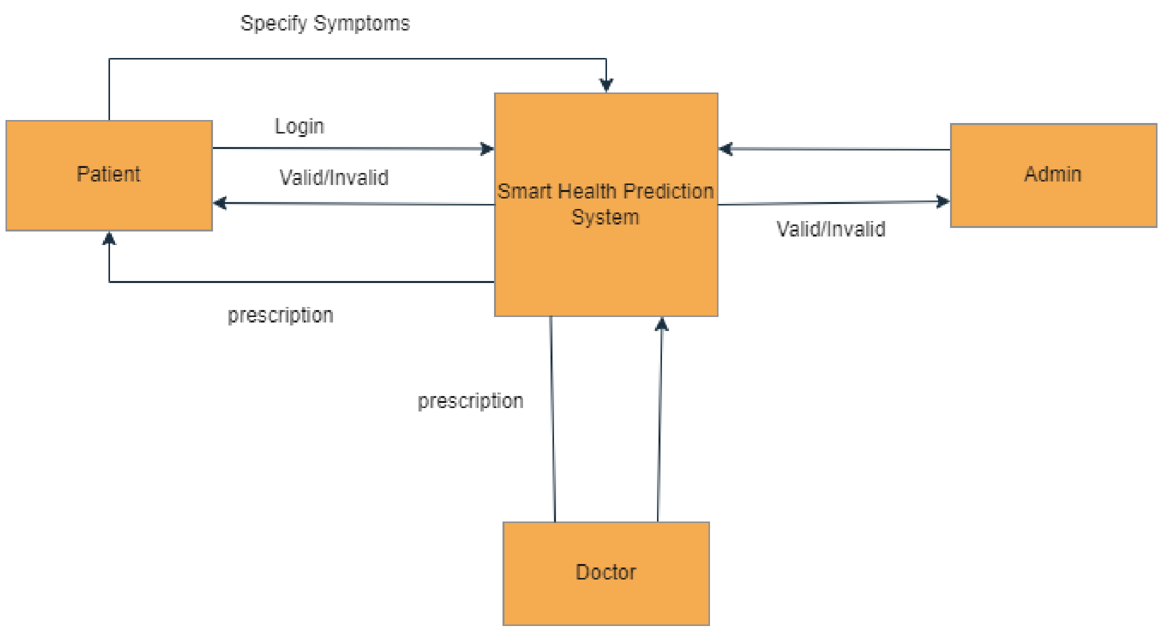

Fig 4.3: Diagram showing system user’s registration into the system

There is also the option of learning more about the ailment. Based on the disease indicated, a list of doctors is also presented. This will allow the patient to receive the essential consultation and assistance sooner. This can be accomplished by chatting or consulting with the doctor.

Patients can view a doctor's profile before deciding whether or not to consult with him. The ability to view the specifics and history of the consultation is also included. Additionally, ratings and feedback can be offered to the system. If the user is a doctor, he can fill up his information, which will be visible to the patients. The doctor can view and respond to the patient's disease-related questions.

4.3 User Interface

After loading the web page, viewers or users of the system would see this home page as shown in Figure 4.4.

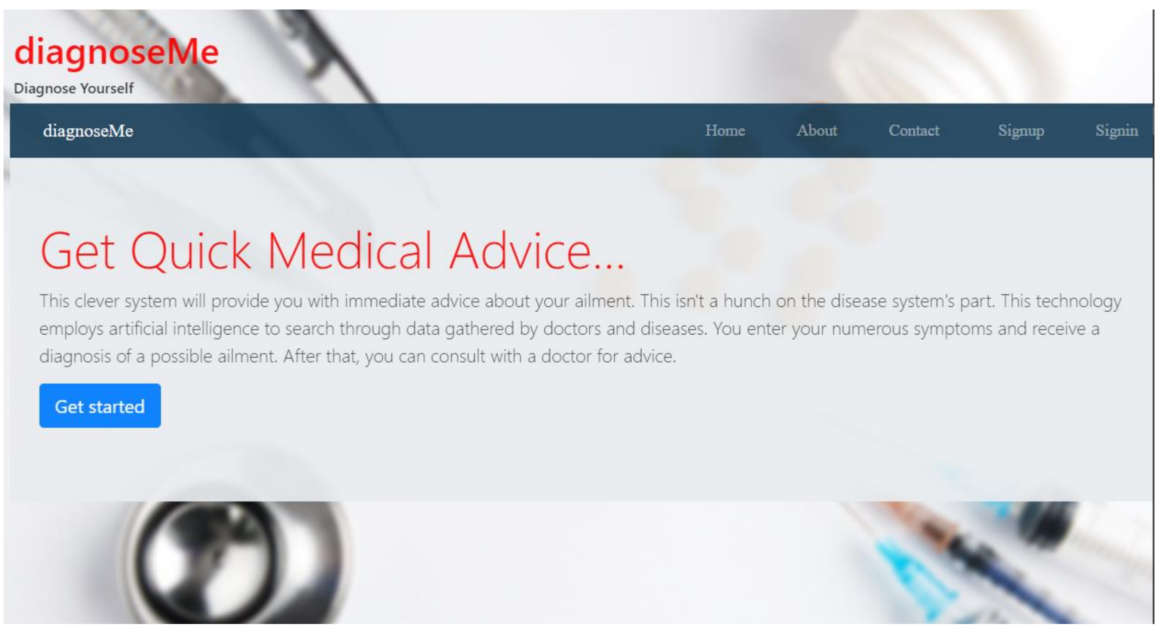

Fig 4.4: Home page of the application

Figure 4.5 shows the login interface for all users of the system

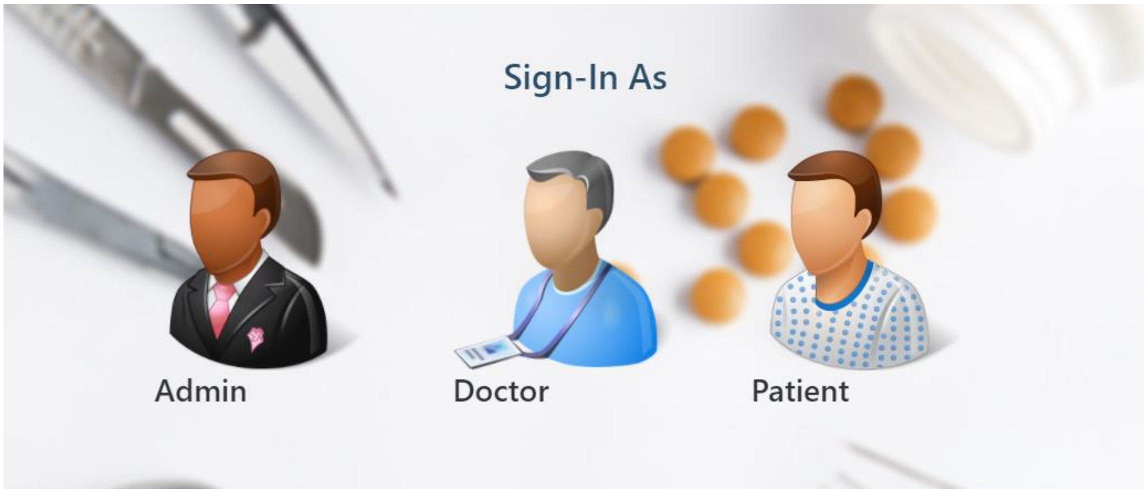

Fig 4.5: Users Login

This login can also be popped up as a model when “Get started” button gets clicked

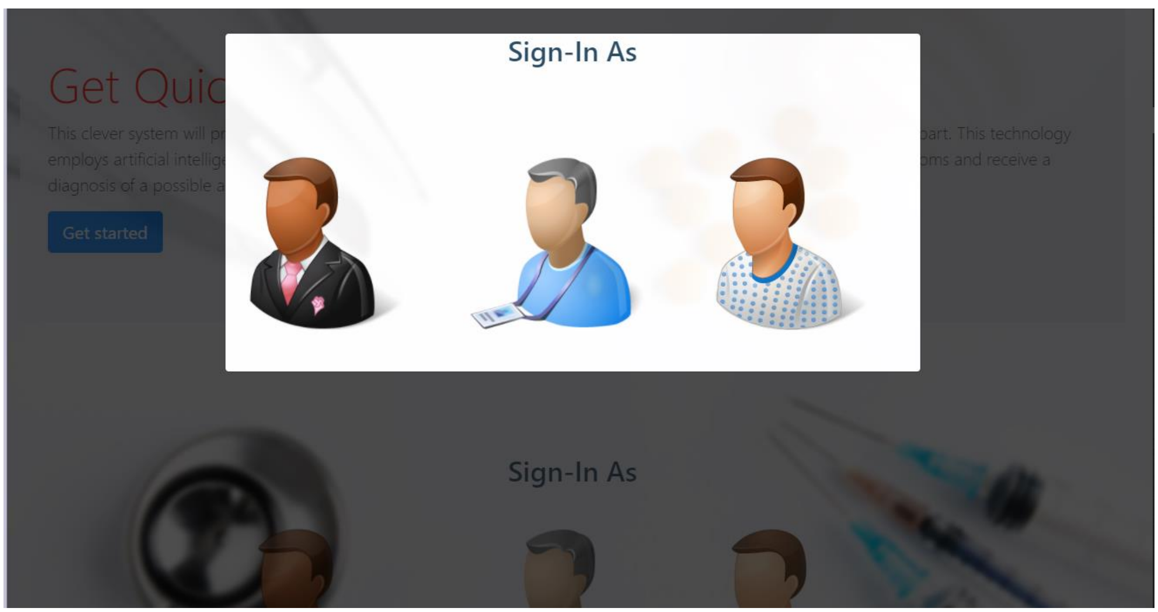

Fig 4.6: Pop-up Login Modal

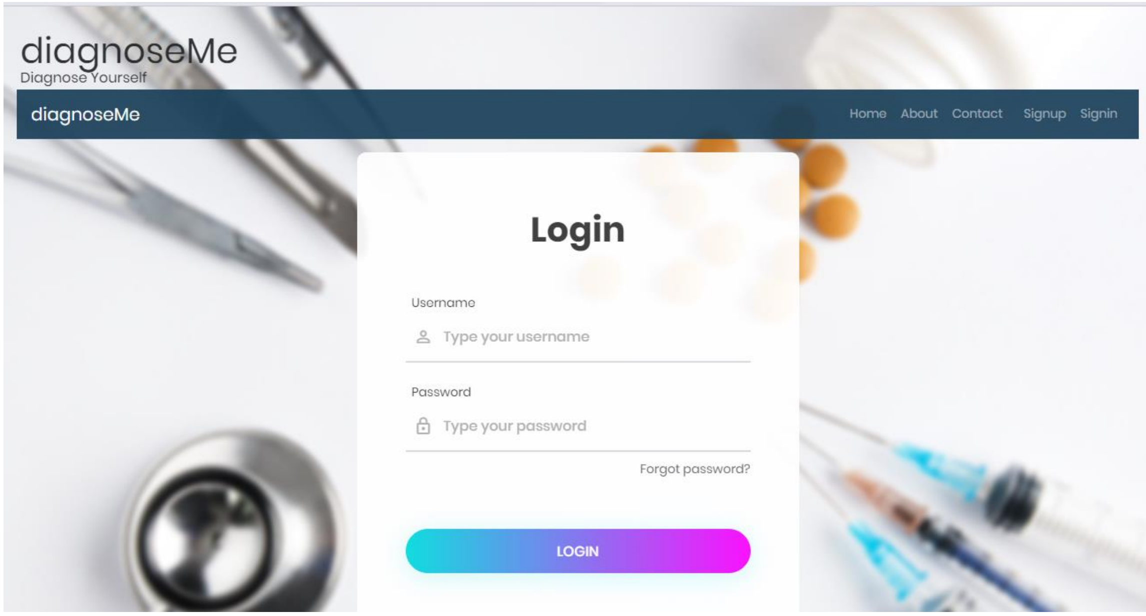

Fig 4.7: Patient login

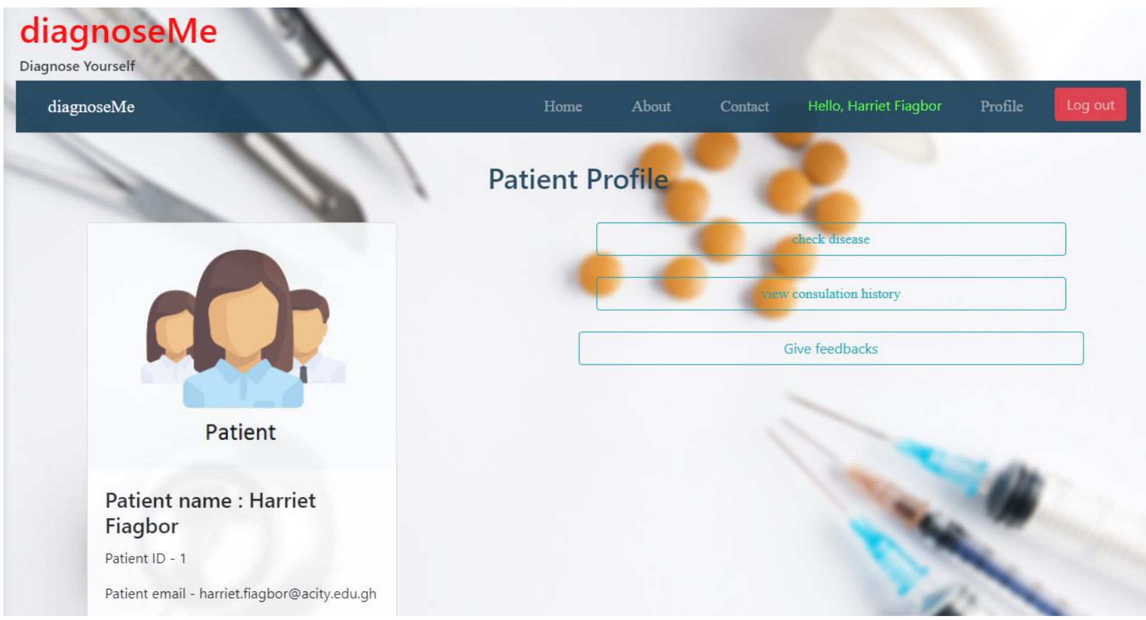

Fig 4.8: Patient user Interface

Patients can enter symptoms to predict or check ailments, allowing the system to provide medical advice.

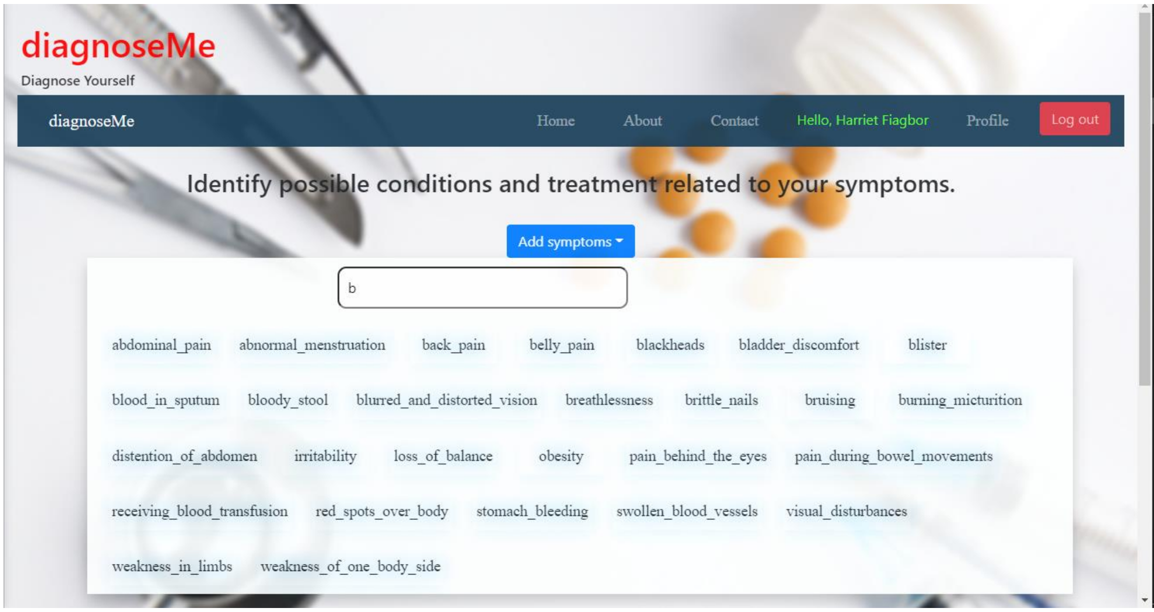

Fig 4.9: Entering Symptoms

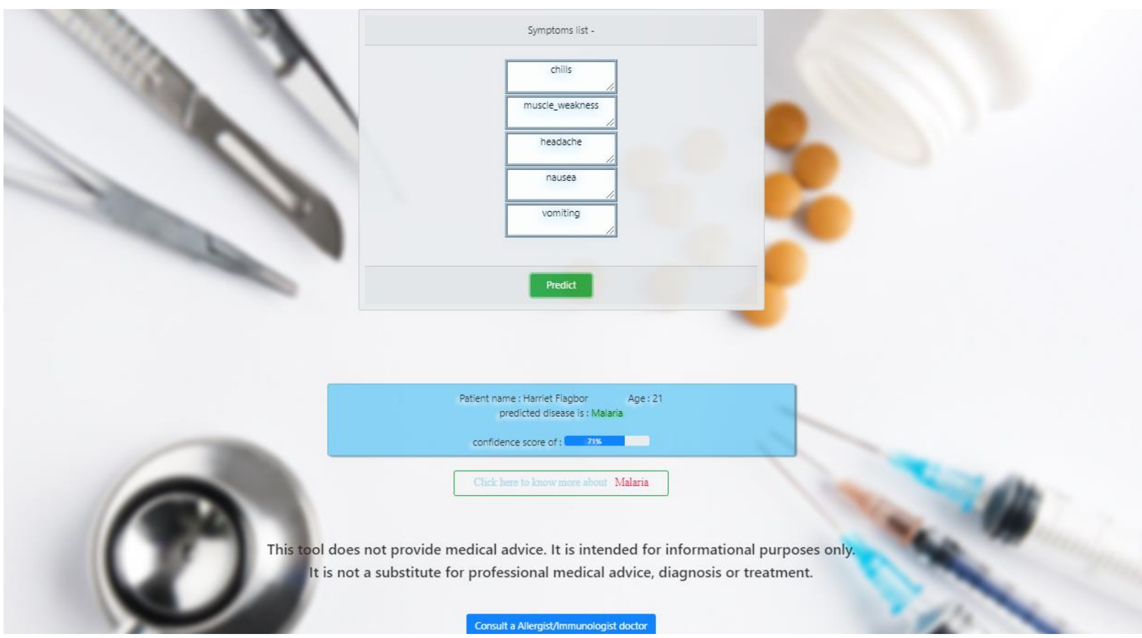

Fig 4.10: Predictions of the system from symptoms

Patients can give feedback to the system

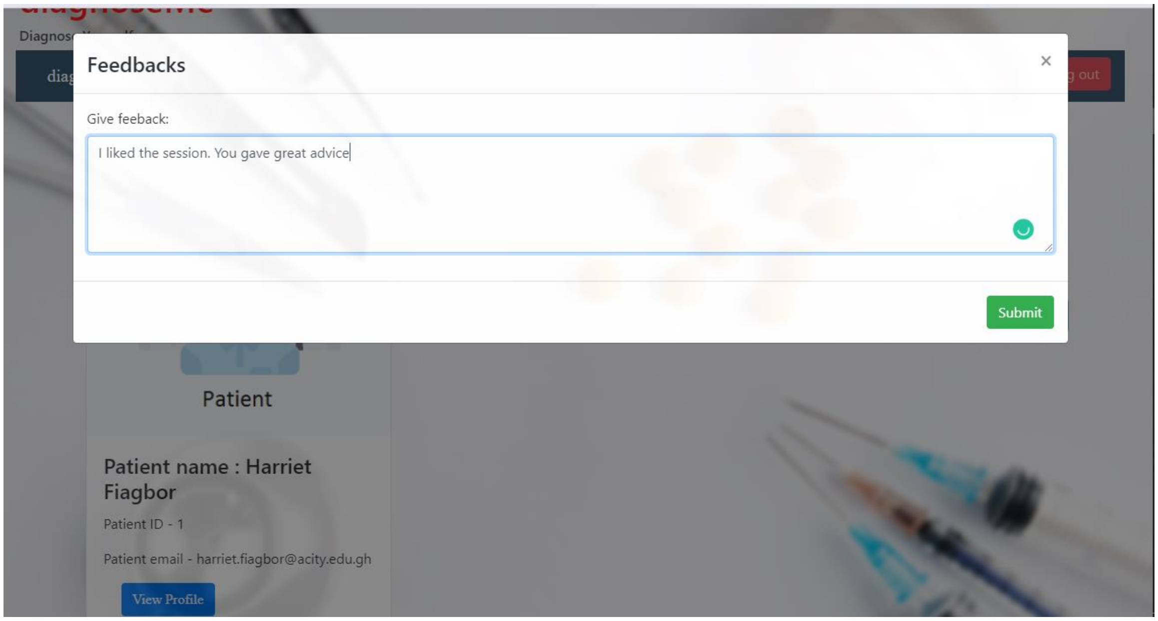

Fig 4.11: Feedback modal

Patients can view the doctors available for consultation and other information about the doctors such as doctors' specializations and ratings.

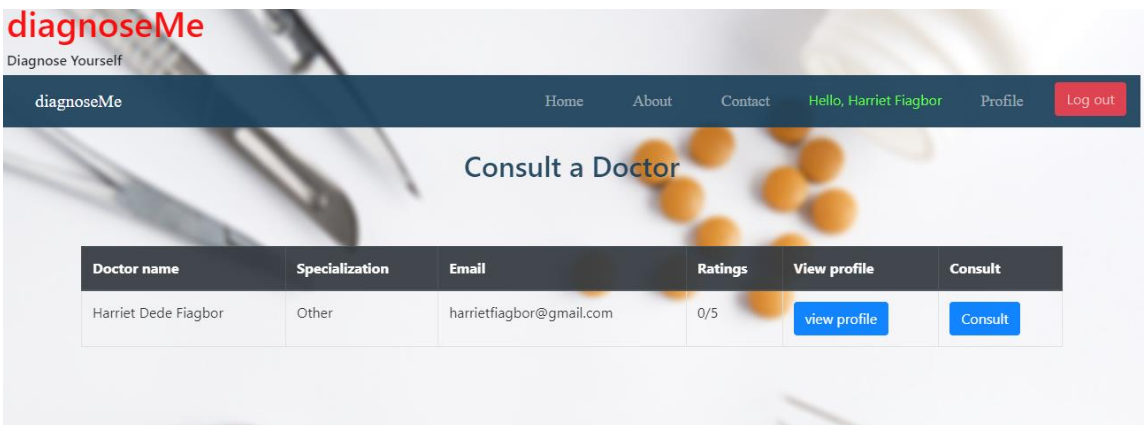

Fig 4.12: Consultation User Interface

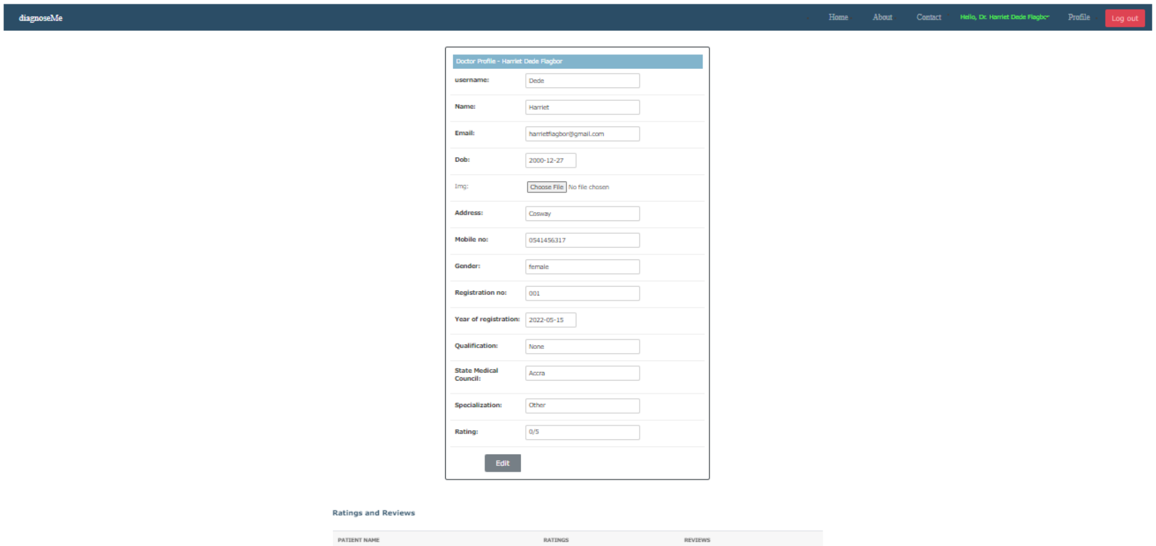

Fig 4.13: Doctor Profile

Doctors can check the number of consultations they have and whether the consultations sessions are still open or closed.

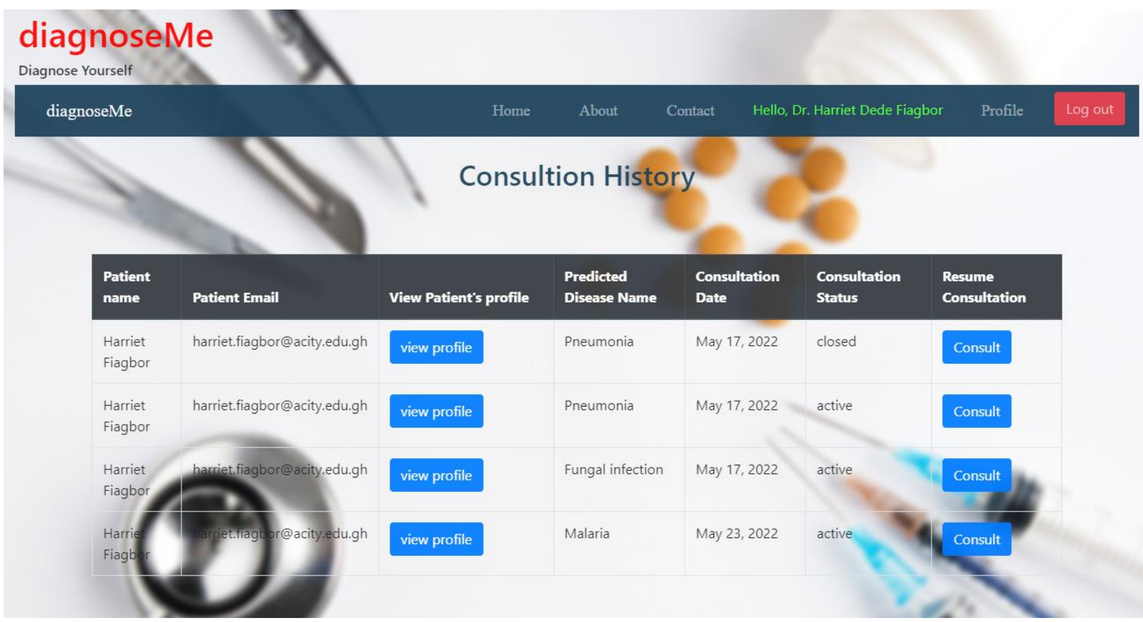

Fig 4.14: Consultation History Interface

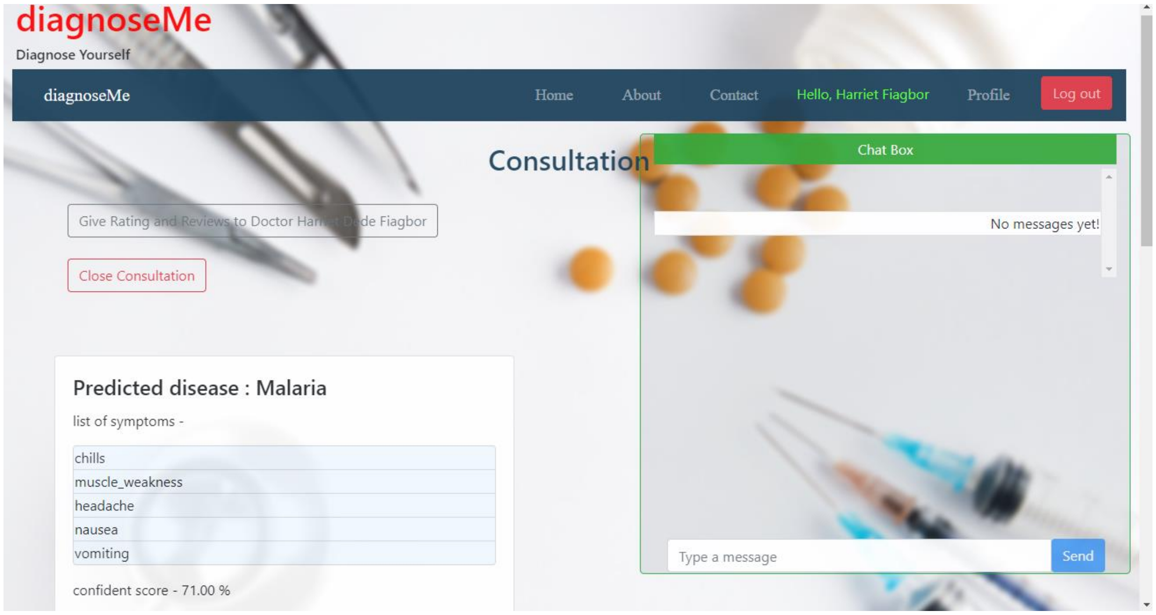

Fig 4.15: Consultation and chat Interface

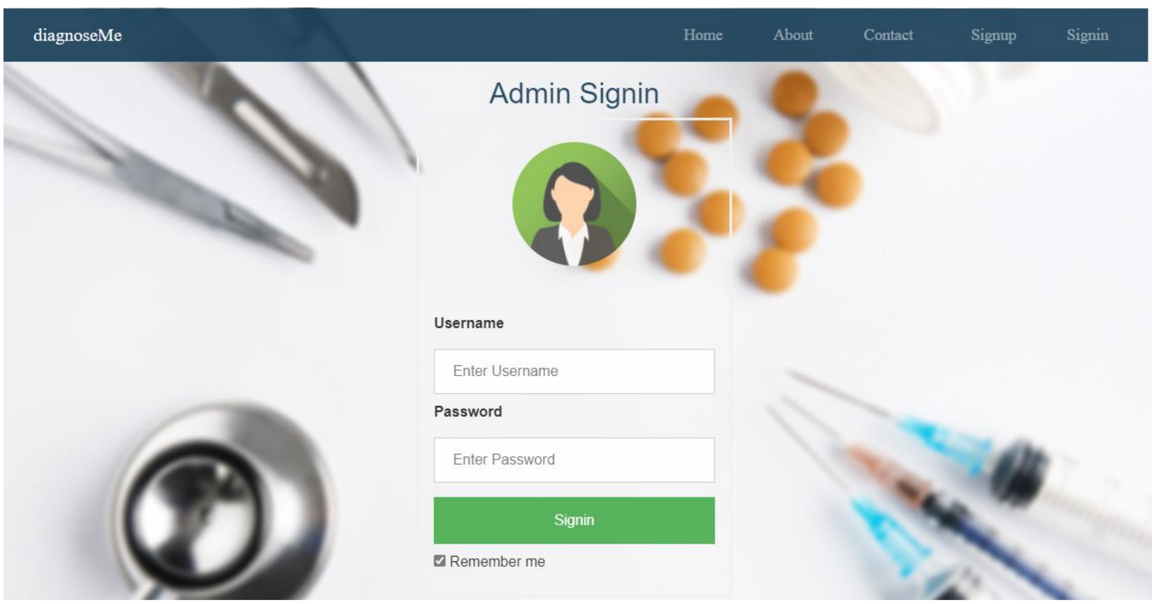

Fig 4.16: Admin Sign-in Interface

The Admin can perform several roles such as add users, check chats, feedbacks, ratings and reviews as shown in Figure 4.16

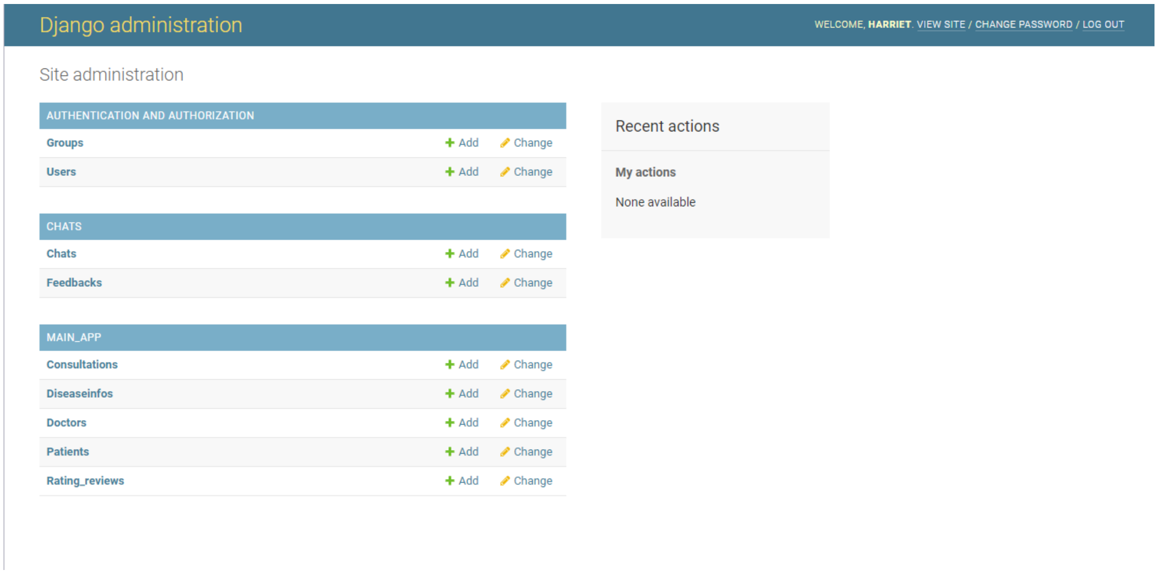

Fig 4.17: Admin User Interface

On the database side, it keeps information on users of the system, such as login credentials, so that users can keep their information and access it at any time.

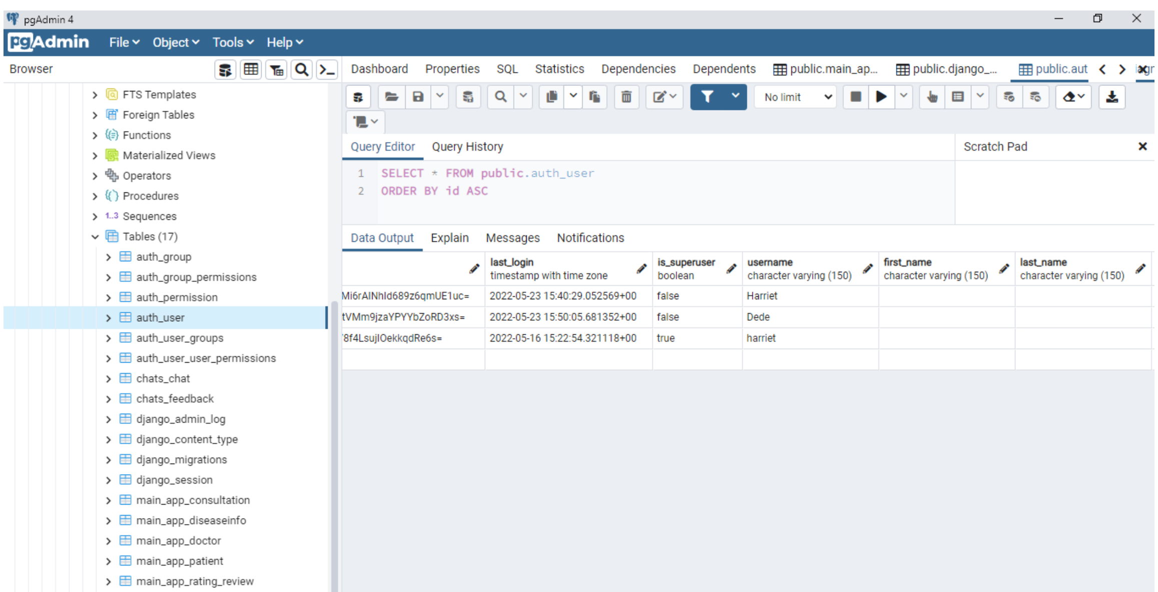

Fig 4.18: User Table from database

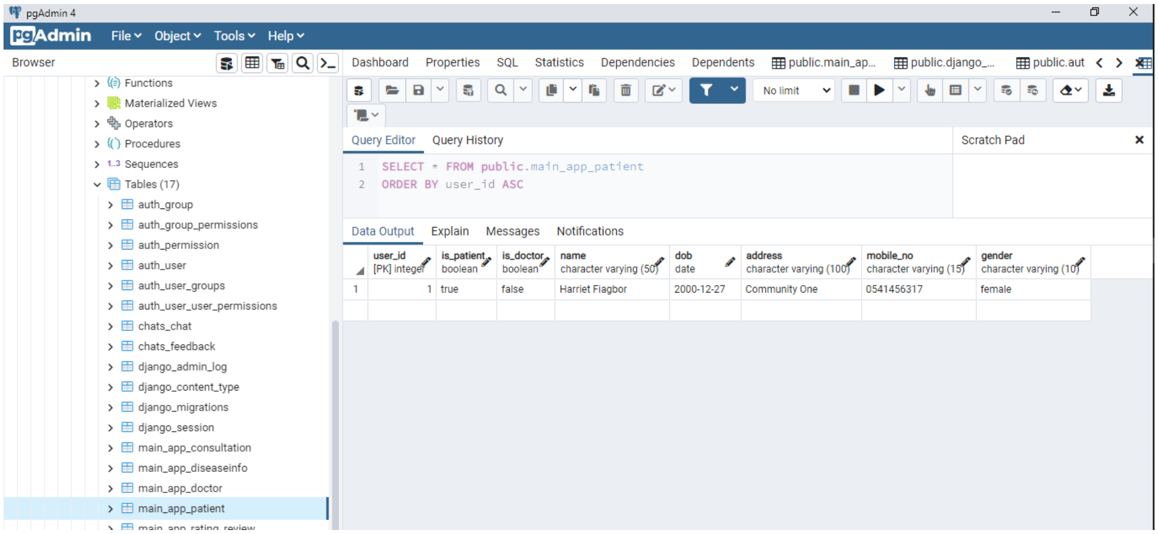

Fig 4.19: Patient Table from database

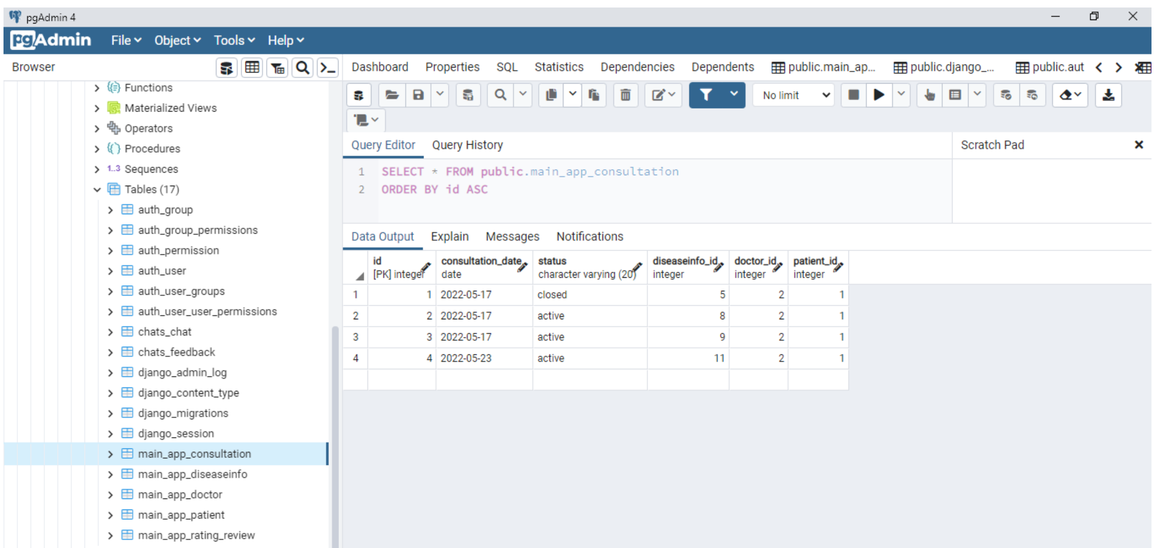

Fig 4.20: Consultation Table from database

CHAPTER FIVE

RESULTS, CONCLUSION AND RECOMMENDATIONS

5.1 Results

In this session, the prediction system is put to the test in order to determine its accuracy in providing medical advice. Symptoms were entered into the system and compared to trusted sources such as the Mayo Clinic, health boards, and reputable doctors.

To commence, Gastroesophageal reflux disease (GERD) symptoms [31] gotten from the Mayo Clinic were entered into the system. The GERD symptoms included chest pain, breathlessness, cough. These symptoms are common to pneumonia also. Our model predicted pneumonia because of symptoms such as chest pain and cough.

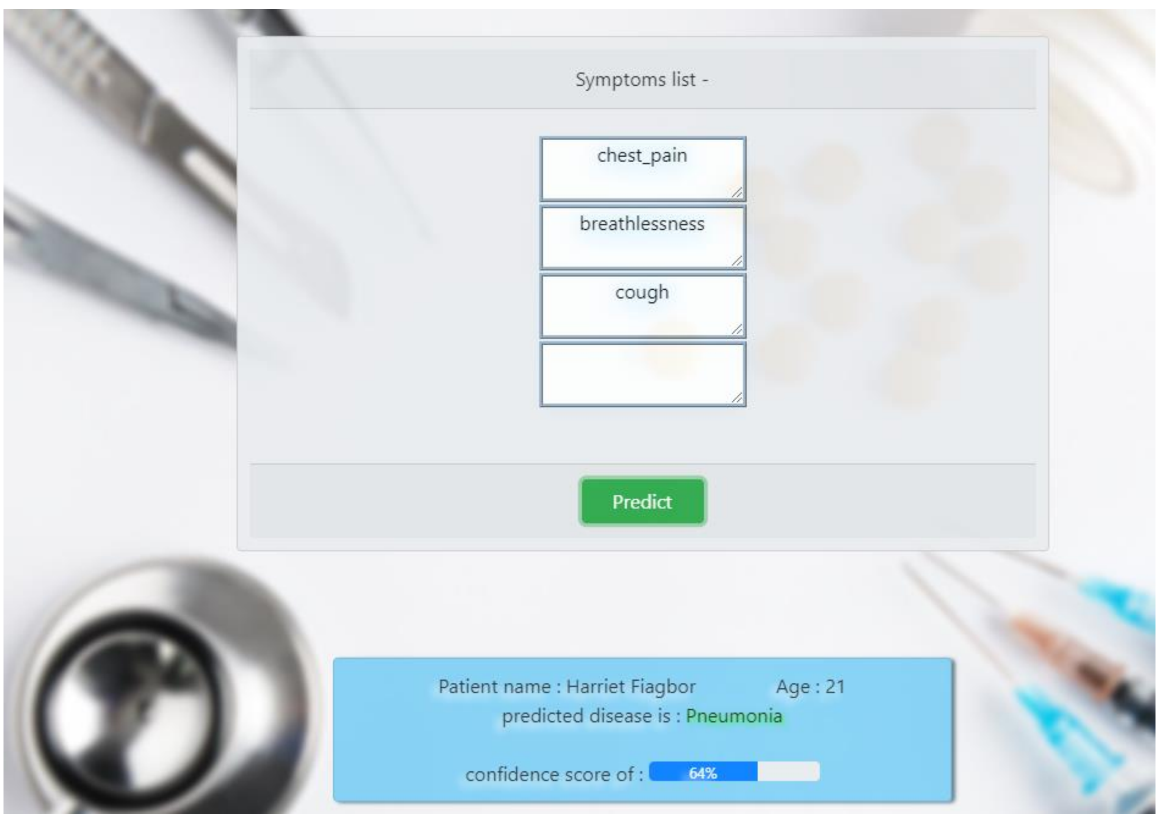

Figure 5.1: GERD common symptoms excluding a very telling symptom

When a significant symptom that is very telling of GERD, acid reflux was added, our model quickly picked it up as a GERD disease with a confidence of 80%.

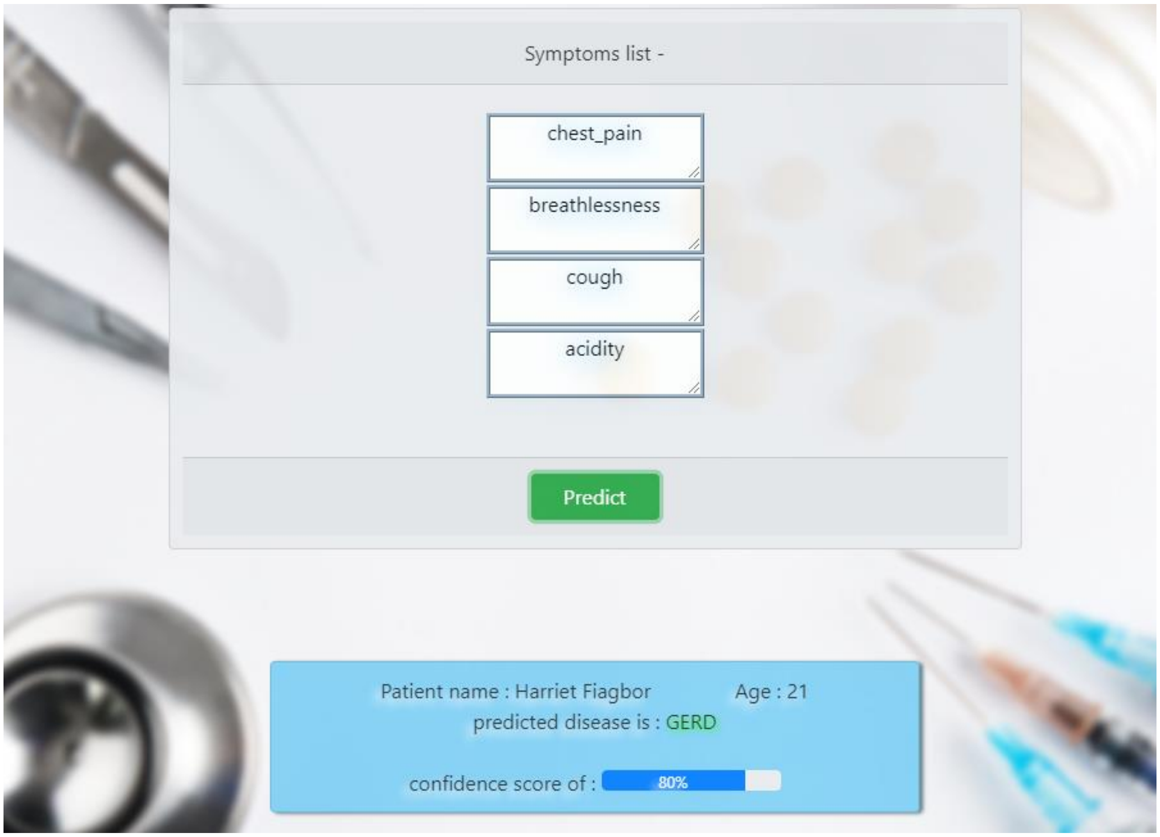

Figure 5.2: GERD symptoms including acid reflux as a telling symptom

The next disease tested was Dimorphic hemorrhoids (piles). Sources of the symptoms for this disease were verified on LinkedIn by Dr. Rakshith Raj Bharadwaj, Director at Dr. Belaku Private Healthcare Limited [32]

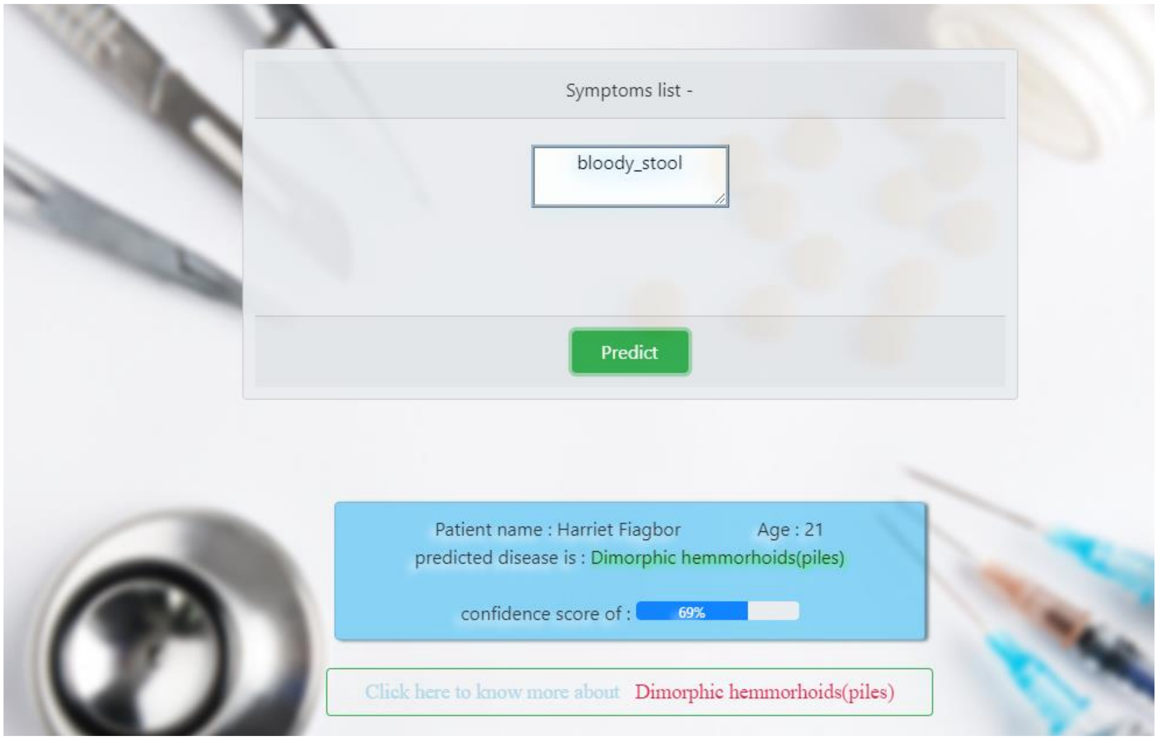

Figure 5.3: Piles telling symptom bloody stool

A very telling symptom of piles is bloody stool. Our model did extremely well by picking that up on the first go with a confidence of 60%. When another telling symptom such as pain in the anal region was included, the confidence score increased to 90% as in Figure 5.4

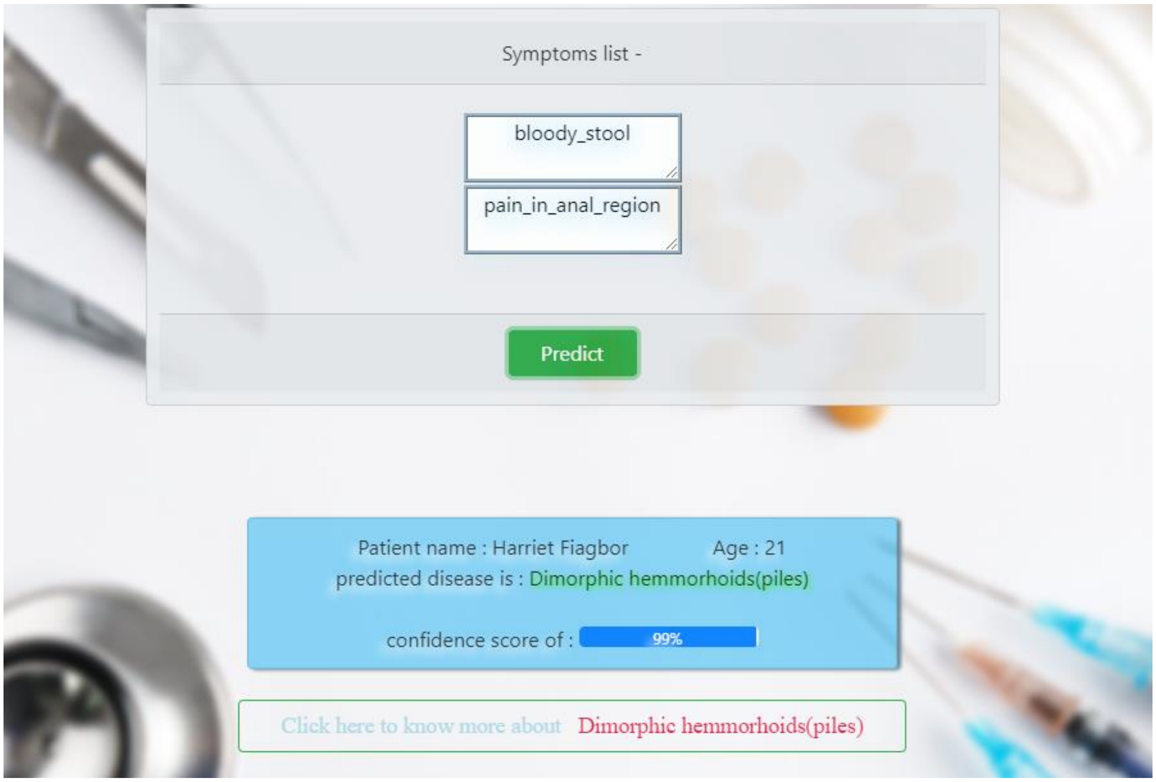

Figure 5.4: Increasing confidence score of piles disease by adding another telling symptom

The analysis of these diseases shows that a telling symptoms leads to a higher correct diagnose. Let us consider situations where symptoms seem very closely related to each other such as Typhoid fever and Malaria. Symptoms of Malaria is very common to a lot of diseases. Typhoid fever and Malaria, although caused by different organisms both typhoid and malaria share a similar symptomatology and epidemiology [33]. In even some situations symptoms of both these can be similar to dengue fever. But Dengue fever has a telling symptom since it is known as a “bone-breaking” disease [34]. Entering similar symptoms like high fever, chills and headache predicts Malaria with a confidence of 56% which is quite low.

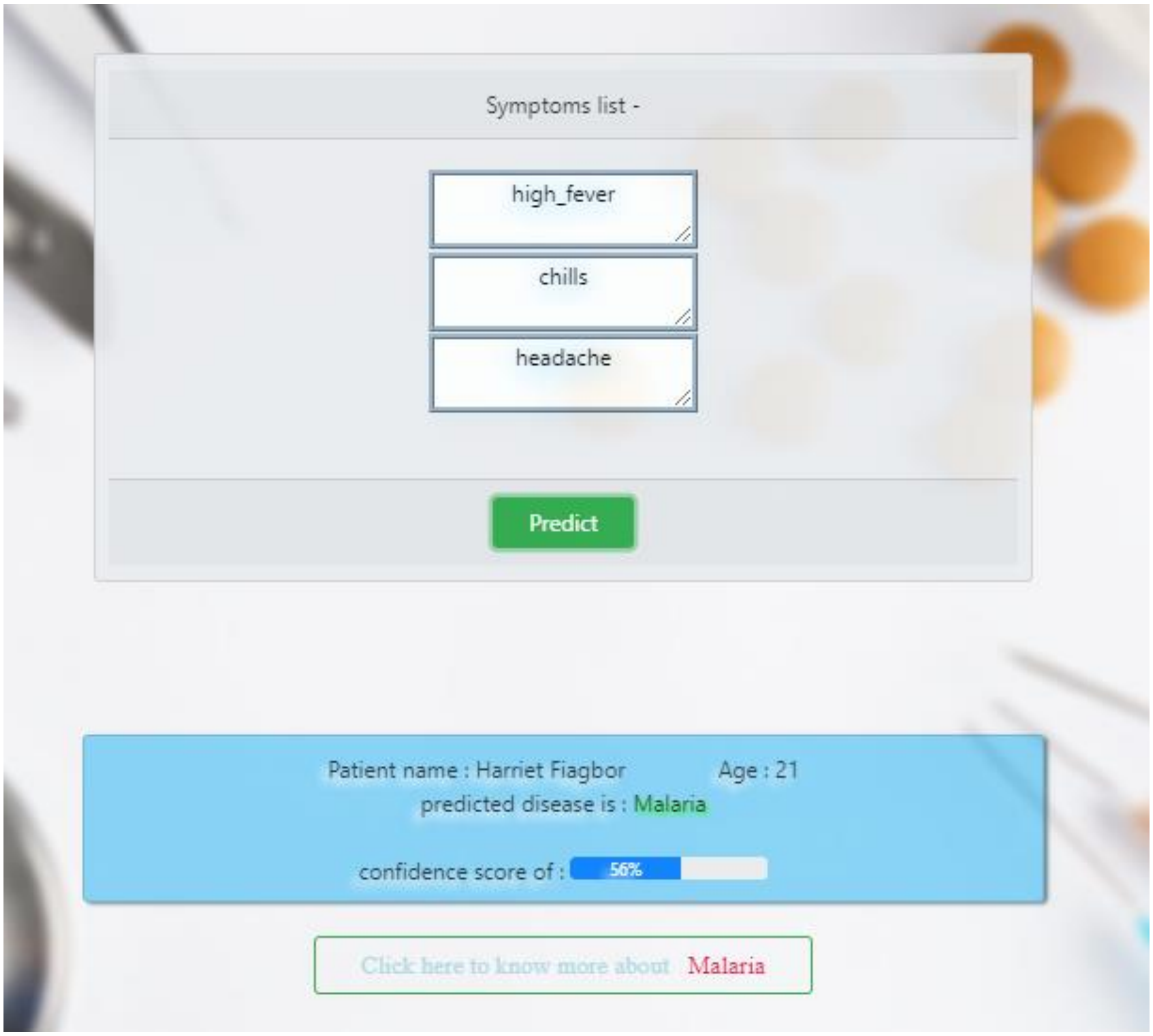

Fig 5.5: Prediction of Malaria

According to the Mayo Clinic, Malaria has a telling symptom which is sweating [35]. Sweating was entered to see if the system prediction accuracy increases.

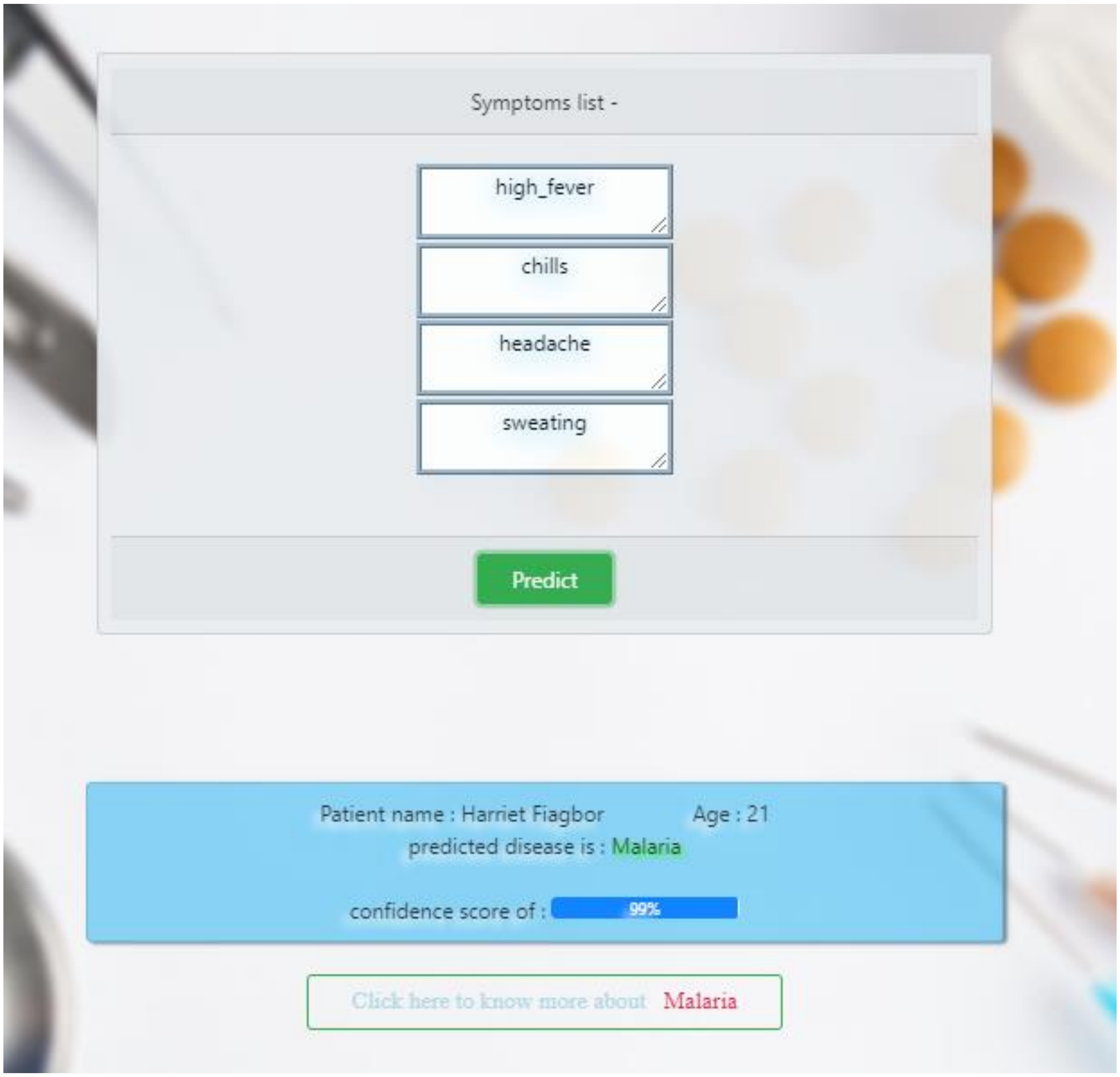

Fig 5.6: Prediction of Malaria using a telling symptom

By adding sweating to the common symptoms, our confidence score increased showing the presence of Malaria clearly.

Another case is with Typhoid fever. Although symptoms may seem similar to Malaria, an added symptom such as constipation draws a clear line [36] as suggested by the Mayo clinic. By testing that out, we realize from figure 5.7 that our results correspond with the research.

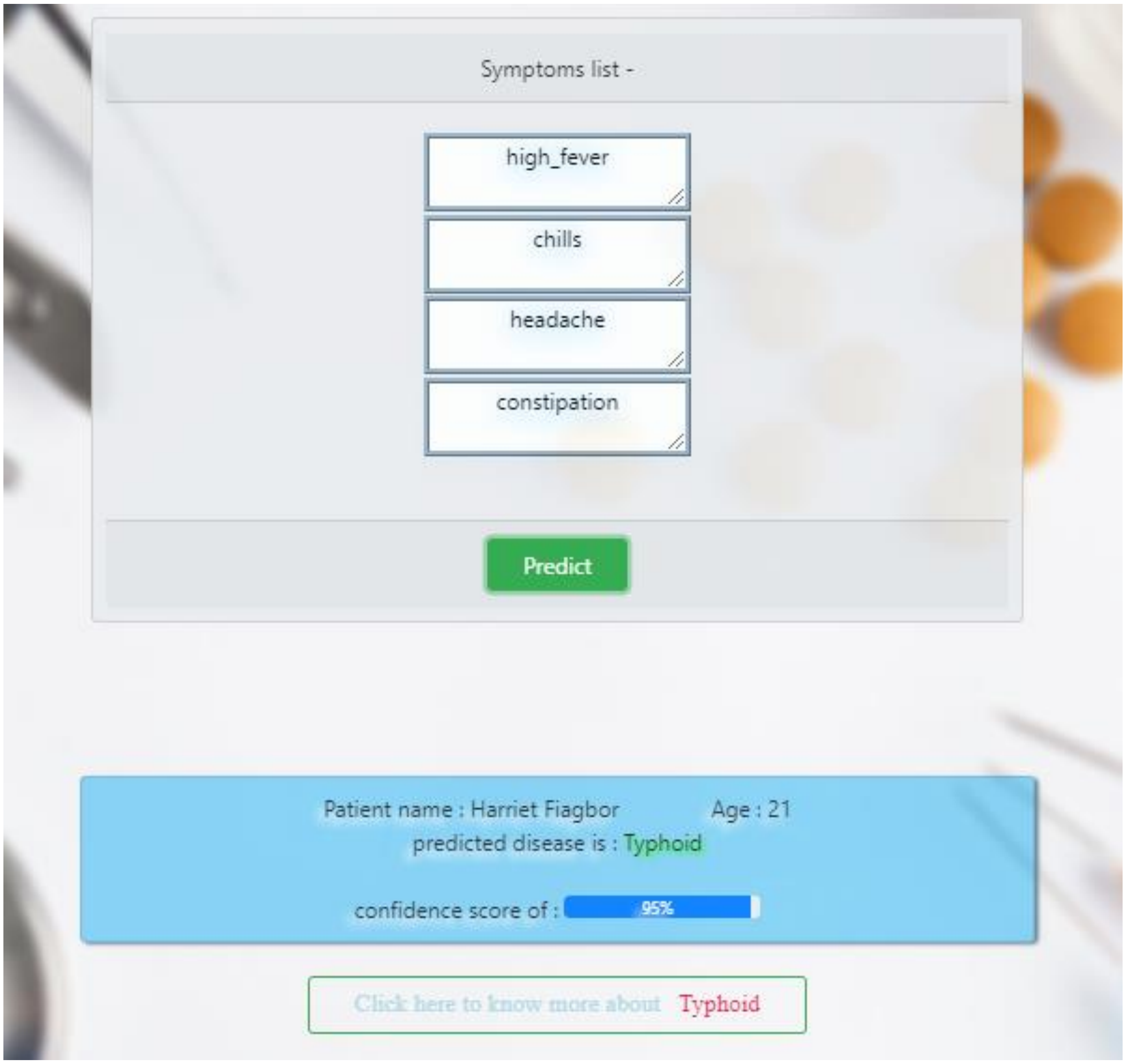

Fig 5.7: Typhoid fever and a telling symptom.

By adding a “bone breaking” symptom such as joint pain or muscle pain, we expect the results to prove dengue fever as can be seen from Figure 5.8

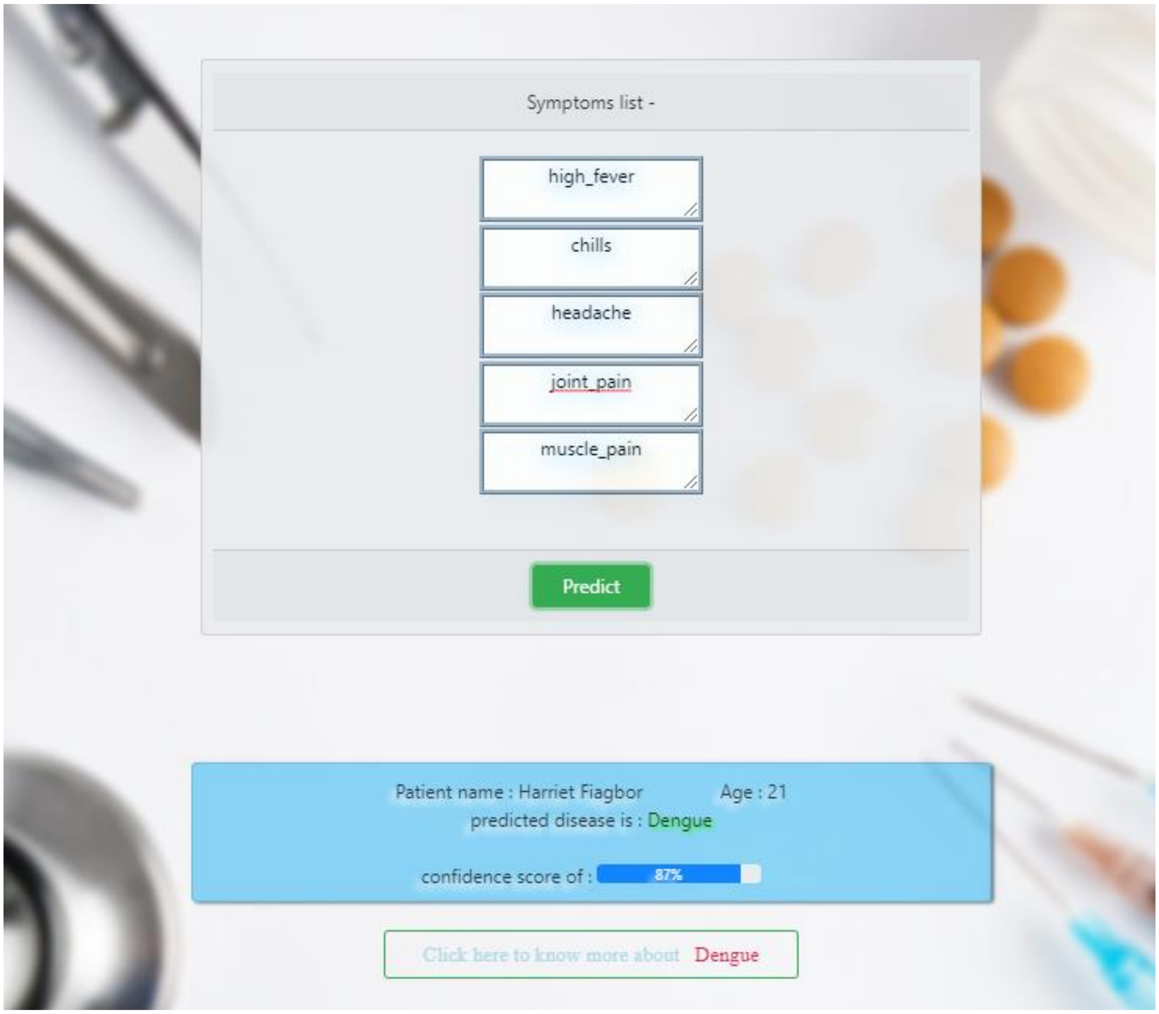

Figure 5.8: Dengue fever with a telling symptom

5.2 Conclusion

Based on the findings of the aforementioned study, it can be stated that some illnesses have a particularly revealing symptom that indicates the presence of a specific condition. As a result, entering a very telling symptom into the system increases the likelihood of receiving a proper diagnosis. More precise symptoms of that particular condition might aid the system discern better when it comes to ailments that appear similar or very perplexing. As with Typhoid, Malaria, and Dengue, the addition of a fourth symptom differentiated the diagnoses and allowed the system to identify more accurately.

5.3 Recommendations

- The system has a list of drop down symptoms that user has to select from. It would be really useful to add a feature such as a text box to insert preferred symptoms.
- Natural Language Processing [37] can be used to extract the correct symptoms from the text to overcome the problem of inaccurate wordings in the text box.
- A larger data set for training the system would almost certainly result in better results.
- A flag or emphasis on ailments that are not minor to be reported to the nearest hospital once system detects one.

REFERENCES

[1] L. A. F. Souza, C. D. da Silva, G. C. Ferraz, F. A. E. F. Sousa, and L. V. Pereira, “The prevalence and characterization of self-medication for obtaining pain relief among undergraduate nursing students,” *Rev. Lat. Am. Enfermagem*, vol. 19, pp. 245–251, Apr. 2011, doi: 10.1590/S0104-11692011000200004.
[2] K. A. Bashrahil and A. S. Baruzaig, “Self-medication: concept, prevalence & risks in Mukalla City (Yemen) 2004–2005,” *Andal. Stud. Res*, vol. 2, pp. 1–3, 2008.
[3] A. Befekadu, “Self Medication”, Accessed: Jan. 17, 2022. [Online]. Available: https://www.academia.edu/4228902/Self_Medication
[4] A. Sharrif, D. Almasdy, and A. Sherrif, “Self-Medication Practice with Nonprescription Medication among University Students: a review of the literature,” *Arch. Pharm. Pract.*, vol. 2, no. 3, p. 7, 2011.
[5] L. Thomas, *The technology of medicine*, vol. 285, no. 24. Mass Medical Soc, 1971, pp. 1366–1368.
[6] K. J. Petrie and S. Wessely, *Modern worries, new technology, and medicine*, vol. 324, no. 7339. British Medical Journal Publishing Group, 2002, pp. 690–691.
[7] S. J. Reiser and S. J. Reiser, *Medicine and the Reign of Technology*. Cambridge University Press, 1981.
[8] H. N. Sherwin, M. McKeown, M. F. Evans, and O. K. Bhattacharyya, “The waiting room ‘wait’: From annoyance to opportunity,” *Can. Fam. Physician*, vol. 59, no. 5, pp. 479–481, 2013.
[9] S. B. Agampodi, N. D. Wickramasinghe, J. Horton, and T. C. Agampodi, “Minor ailments in pregnancy are not a minor concern for pregnant women: a morbidity assessment survey in rural Sri Lanka,” *PloS One*, vol. 8, no. 5, p. e64214, 2013.
[10] W. C. Cockerham, M. C. Creditor, U. K. Creditor, and P. B. Imrey, “Minor ailments and illness behavior among physicians,” *Med. Care*, pp. 164–173, 1980.
[11] “Strep Throat: All You Need to Know | CDC.” https://www.cdc.gov/groupastrep/diseases-public/strep-throat.html (accessed Jun. 27, 2022).
[12] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature extraction: foundations and applications*, vol. 207. Springer, 2008.
[13] K. Siau and W. Wang, “Building trust in artificial intelligence, machine learning, and robotics,” *Cut. Bus. Technol. J.*, vol. 31, no. 2, pp. 47–53, 2018.
[14] J. Alzubi, A. Nayyar, and A. Kumar, “Machine learning from theory to algorithms: an overview,” in *Journal of physics: conference series*, 2018, vol. 1142, no. 1, p. 012012.
[15] C. Brooks and C. Thompson, “Predictive modelling in teaching and learning,” *Handb. Learn. Anal.*, pp. 61–68, 2017.
[16] L. Swani and P. Tyagi, “Predictive Modelling Analytics through Data Mining,” *Int. Res. J. Eng. Technol. IRJET*, vol. 4, no. 09, 2017.
[17] C. Mair *et al.*, “An investigation of machine learning based prediction systems,” *J. Syst. Softw.*, vol. 53, no. 1, pp. 23–29, 2000.
[18] P. Szolovits, R. S. Patil, and W. B. Schwartz, “Artificial intelligence in medical diagnosis,” *Ann. Intern. Med.*, vol. 108, no. 1, pp. 80–87, 1988.

[19] B. M. Gayathri, C. P. Sumathi, and T. Santhanam, “Breast cancer diagnosis using machine learning algorithms-a survey,” *Int. J. Distrib. Parallel Syst.*, vol. 4, no. 3, p. 105, 2013.
[20] W. S. Noble, “What is a support vector machine?,” *Nat. Biotechnol.*, vol. 24, no. 12, pp. 1565–1567, 2006.
[21] B. Yegnanarayana, *Artificial neural networks*. PHI Learning Pvt. Ltd., 2009.
[22] M.-L. Zhang and Z.-H. Zhou, “ML-KNN: A lazy learning approach to multi-label learning,” *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, 2007.
[23] S. R. Safavian and D. Landgrebe, “A survey of decision tree classifier methodology,” *IEEE Trans. Syst. Man Cybern.*, vol. 21, no. 3, pp. 660–674, 1991.
[24] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, “A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms,” *Mob. Inf. Syst.*, vol. 2018, p. e3860146, Dec. 2018, doi: 10.1155/2018/3860146.
[25] “UCI Machine Learning Repository: Heart Disease Data Set.” https://archive.ics.uci.edu/ml/datasets/heart+disease (accessed Jan. 24, 2022).
[26] A. E. Lang and A. M. Lozano, “Parkinson’s disease,” *N. Engl. J. Med.*, vol. 339, no. 16, pp. 1130–1143, 1998.
[27] J. S. Almeida *et al.*, “Detecting Parkinson’s disease with sustained phonation and speech signals using machine learning techniques,” *Pattern Recognit. Lett.*, vol. 125, pp. 55–62, 2019.
[28] G. Battineni, G. G. Sagaro, N. Chinatalapudi, and F. Amenta, “Applications of Machine Learning Predictive Models in the Chronic Disease Diagnosis,” *J. Pers. Med.*, vol. 10, no. 2, Art. no. 2, Jun. 2020, doi: 10.3390/jpm10020021.
[29] “Disease Prediction Using Machine Learning.” https://www.kaggle.com/kaushil268/disease-prediction-using-machine-learning (accessed Mar. 28, 2022).
[30] A. D’Angela, “Why Weight? The Importance of Training on Balanced Datasets,” *Medium*, Feb. 04, 2021. https://towardsdatascience.com/why-weight-the-importance-of-training-on-balanced-datasets-f1e54688e7df (accessed Mar. 28, 2022).
[31] “Gastroesophageal reflux disease (GERD) - Symptoms and causes,” *Mayo Clinic*. https://www.mayoclinic.org/diseases-conditions/gerd/symptoms-causes/syc-20361940 (accessed May 25, 2022).
[32] “(99+) Dr. Rakshith Raj Bharadwaj | LinkedIn.” https://www.linkedin.com/in/drrakshith/ (accessed May 26, 2022).
[33] B. C. M. Keong and W. Sulaiman, “Typhoid and Malaria Co-Infection – An Interesting Finding in the Investigation of a Tropical Fever,” *Malays. J. Med. Sci. MJMS*, vol. 13, no. 1, pp. 74–75, Jan. 2006.
[34] M. H, “Typhoid, Malaria, Dengue Fever,” *Grabbing Life By the Lapels*, Jan. 27, 2016. https://penthesilearising.wordpress.com/2016/01/27/typhoid-malaria-dengue-fever/ (accessed May 27, 2022).
[35] “Malaria - Symptoms and causes - Mayo Clinic.” https://www.mayoclinic.org/diseases-conditions/malaria/symptoms-causes/syc-20351184 (accessed May 27, 2022).

[36] "Typhoid fever - Symptoms and causes," *Mayo Clinic*. https://www.mayoclinic.org/diseases-conditions/typhoid-fever/symptoms-causes/syc-20378661 (accessed May 27, 2022).
[37] "What is Natural Language Processing? | IBM." https://www.ibm.com/cloud/learn/natural-language-processing (accessed May 27, 2022).