Disease and Adverse Drug Reaction Prediction using Machine Learning

Reshma Mathai,Ardra K John, Anima M M, Athulya James

U.G. students, Department of Information Technology

Rajagiri School of Engineering and Technology, Kakkanad, Kerala, INDIA

Lakshmi K S

Assistant Professor, Department of Information Technology

Rajagiri School of Engineering and Technology, Kakkanad

Kerala, INDIA

*Abstract*— The aim of the project is to use machine learning techniques for the prediction of diseases and adverse drug reactions. The project is divided into two modules, an android app and a web app. The android app is to predict possible diseases based on the symptoms the person is showing. Along with that the reviews of common drugs from online healthcare forums such as medications.com are extracted and tf-idf is used to find out the possible adverse drug reactions of a given drug. The web app does disease risk prediction based on phenotypic details and lab reports. As an addition to the project, location based medical help and health tips are also implemented.

# Introduction

Machine learning can bring forth a major leap in healthcare by making jobs of clinicians more accurate and easier. Data based on diseases and drugs can be mined and used for predictive and analytic practices. With the growth in medical data, collecting electronic health record is increasingly convenient. Prediction using traditional disease risk models usually involves a machine learning algorithm (e.g., Support Vector Machine, Naïve Bayes, K-Nearest Neighbors etc.). in these models, supervised learning algorithm uses training data with labels to train the model. Using the test set, patients can be classified as either high-risk or low-risk. The idea that we have can be called a “second opinion” as it can be used by patients and doctors to use machines to answer their questions. We also aim in bringing forward a platform where users can find out the potential adverse drug reactions that can be caused by medications that they are using or planning to use. All these features are incorporated into an android application and the risk prediction module is implemented as a web application.

# literature survey

Many studies have been done in the field of prediction of disease prediction and other healthcare services using machine learning. A variety of data mining and machine learning techniques have been applied in the healthcare field to make maximum use of existing clinical data and electronic health records. In this section, we are discussing a few papers that describes a few methodologies and techniques used in this area

## Mining Adverse Drug Reactions from online healthcare forums using Hidden Markov Model

The idea of this research is to extract drug side effects from online healthcare forums. It is done so that they can be used as early indicators in post marketing drug surveillance towards Pharmacovigilance. It is a sequence labeling problem. It presents a Hidden Markov Model (HMM). It is based on Text mining system. Text mining systems can be used to classify as containing side effect details. It then extracts the adverse drug side effects from the online health care forum messages. The online healthcare forum used here for extracting side effects is medications.com. It is then used in the training and validation of the system.

The messages crawled from medications.com were used for creating the train and test data sets used in the evaluation of the HMM classifier. Due to the large number of messages a two phase approach was carried out. Even though there can be multiple occurrences of the same drug name or a side-effect in a forum message, only one such instance is used as a part of the mined information. The prediction from the HMM classifier classifies the messages as either having a drug side-effect relation or not.

From the HMM classifier, a 10 fold cross validation on the dataset produced on an F-score of 0.76 is compared to the 0.575 from the Baseline classifier. The F-score was reduced to 0.378 without the Plain Text Filter component while the HTML Filter component's absence did not have any impact. In addition to this some potential Adverse drug reactions were also found. These mined drug side effects can easily be used as early indicators to improve the efforts in post marketing drug surveillance.

## Heart Disease Prediction using Data Mining Techniques

Healthcare industry has large amounts of data and it requires to be mined properly to discover common trends and relationships in data. The idea behind this paper is to use genetic algorithm with back propagation technique to predict heart disease based on a large number of attributes. The attributes used here include gender, blood pressure etc. to predict the likelihood of patient of getting a heart disease. The classifiers used in their approach towards prediction of heart disease are K-Nearest Neighbor, Naïve Bayes and Decision Trees. In the KNN approach the end result is a class membership. The neighbors are taken from a set of items for which the class is known. But one shortcoming of this approach is that, it is highly dependent on the data’s local structure.

They used data from both local and internet sources. The decision tree used in their approach is j48, which is the most popular method in use. The benefit of using J48 algorithm is that it recursively classifies data until it is classified as accurately as possible and achieving maximum accuracy on test data. The processes involved are data preprocessing, decision tree mining and decision tree mining. The Naïve Bayes algorithm is based on the Bayes theorem. It uses conditional independence. This means that the dependence between attributes is not taken into consideration.

The attributes they used for prediction are gender, pain type, abstinence blood glucose, restack resting electrographic results, exercise induced angina, slope of height exercise ST section, CA variety of major vessels colored by flouropsy, age and so on.

The conclusion of their project was that, KNN is the best classifier amongst all the classification techniques used in their project.

## Survey on Technique for Prediction of Disease in Medical Data

With growing research on disease predicting system, it’s important to discover hidden patterns and relationships from medical database. For classical clinical diagnosis, it require lots of test which could complicate the disease prediction, here data mining technique can be helpful to take a decision about the disease using computer aided decision support system. In this paper various data mining techniques that are used for disease prediction is presented, which are used as classifier to build a cost effective model for disease prediction. It includes various techniques proposed by experts in the field.

Data mining can be used as an approach for extracting knowledge from database. The doctors may find it difficult in taking decisions about the disease as they may not be an expert in all the fields, to solve this problem there is a need for development of decision prediction system that combines knowledge of medical expertise with automated system to achieve best results to serve the society.

Various methods discussed in this paper are: clinical decision support system using weighted fuzzy rules for risk level prediction of heart disease. Here data is preprocessed and carry out generation of weighted fuzzy and a fuzzy rule based decision support system.

Other is a method for predicting intelligent heart diseases, implemented by integrating three models –neural networks, coactive neuro-fuzzy inference system (CANFIS) for discovering nonlinear relationship maps between different attribute models and genetic algorithm.  
A different approach proposed is a data mining application in medical industry for predicting heart attacks. This uses one dependency augmented naïve bayes classifier (ODANB) and naïve creedal classifier2 (NCC2) for data preprocessing.

Data mining technique for predicting acute coronary syndrome uses data reduction technique to reduce the dimensions. Using this technique an observation was made that smoking is the most significant factor of risk for acute coronary syndrome.

The main focus on this paper is to discuss about decision parameter, attribute and features used for predicting the disease. Also discuss the importance of different classification methods for prediction of disease in medical dataset.

*D. Predicting Adverse Drug Events from Personal Health*

*Messages*

An increasing number of people are using internet to search for information about health. Unreported ADEs, ignorance of patient reports by the healthcare professionals and discouraging the reporting of non-serious drug events has led to significant medical consequences. This has led to a large number of people to share and express their medical related issues via online healthcare forums. Within the online healthcare forums the patient describes their experiences of a particular drug, both good and bad. It hypothesizes that drugs that have undergone regulatory actions are talked about are talked about in similar ways particularly regarding sentiment-one’s positive or negative orientation and effect entities. Machine learning classifiers are used to compare messages containing drugs that have undergone regulatory actions.

The online forms consist of public Health Wellness Groups. These consist of unique email ids which are considered as proxy to people. The Health Wellness Groups range from illness based support groups to those supporting home remedies. It is found that people tend to post more negatives than the positives about a particular drug. Considering this, watch list drugs are identified based on the frequency.

The input to the machine learning algorithm are feature vectors consisting of two feature sets based on the words people use to discuss a particular drug. The first feature vector consists of general vocabulary and the second one consists of meta-features and world knowledge in the form of counts over specialized lexicons.

# System Overview

In this section, we are discussing the basic overview of the system we have designed. There are two modules on the system, one being an android application and the other a web application.

Figure 1 shows the overview of the android application. It has four main functionalities which are disease prediction based on symptoms entered by user, adverse drug reaction prediction based on reviews from ‘medications.com’ which uses tf-idf of side effect words, location and contact details of nearby hospitals, clinics and medical shops based on location services and finally health tips.

Fig : System Overview of Android Application

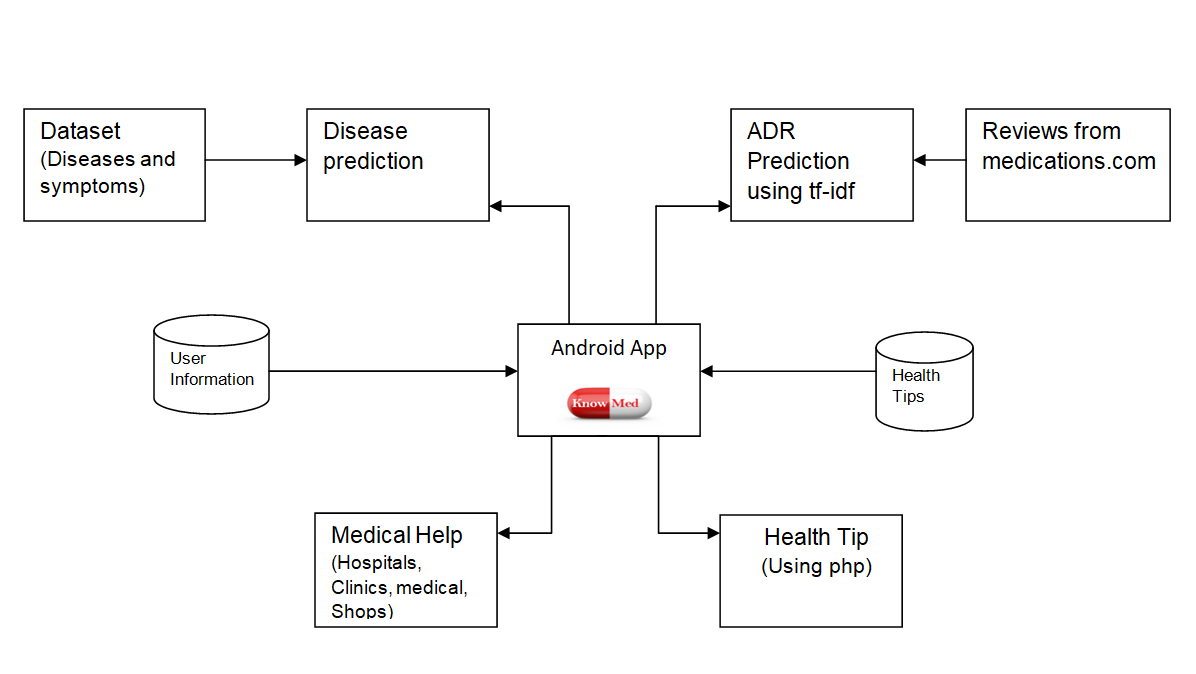


Fig : System Overview of Web Application

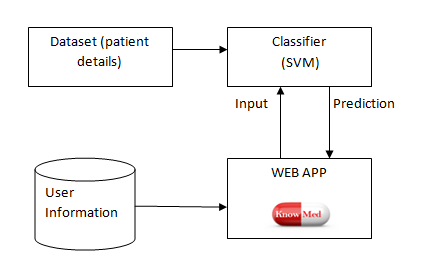


Figure 2 shows the system overview of the web application. The main functionality of the web application is risk prediction based on phenotypic details and lab result details of patients. When this information is entered, the system predicts if this person is at risk of having the disease or not. For this datasets of five diseases were used.

# Dataset Description

In this section, we describe the hospital datasets we use in this study. Furthermore, we provide disease risk prediction model and evaluation methods. We also describe the dataset used for disease prediction. The data collected for predicting adverse drug reaction will also be described here.

## Disease-Symptoms Dataset

This dataset includes a set of diseases and their corresponding symptoms. This is not hospital data and hence it does not raise privacy concerns.

Table : Sample Disease-Symptom Dataset

|  |  |
| --- | --- |
| DISEASE | SYMPTOMS |
| Anemia | fatigue, weakness, pale skin. shortness of breath, brittle nails, headache |
| Amebiasis | abscesses, infections, diarrhea, severe illness |

## Hospital Data

The hospital dataset used in this study contains real-life hospital data, and the data are stored in the database. The dataset includes structured data.  The structured data includes laboratory data and the patient’s basic information such as the patient’s age, gender and life habits, etc. To show risk prediction, we have chosen five main diseases all from different areas of medical science. The diseases we have chosen are Cardio-vascular disease, Liver disease, Psoriasis, Infertility and Diabetes during pregnancy.

Attributes used for cardiovascular disease are: Age, Gender, height, weight, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoking habit, alcohol consumption and active or not.

Attributes used for Psoriasis are erythema, scaling, definite borders, itching, koebner phenomenon, polygonal papules, follicular papules, oral involvement, knee-elbow involvement, scalp invasion, family history and age.

Attributes used for Fertility are season, age, childhood disease, accident or trauma, interventions, high fever last year, frequency of alcohol consumption, smoking habit and number of hours spent sitting.

Attributes used for Diabetes are pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function and age.

Attributes used for Liver Disease are age, gender, total bilirubin, direct bilirubin, alkaline phosphates, alamine aminotransferase, aspartate amino transferase, total proteins, albumin, albumin and globulin ratio.

## Reviews from medications.com

The reviews used for predicting adverse drug reactions are extracted from ‘medications.com’. ‘medications.com’ is a platform where people discuss about medical conditions, medications and raise any queries they have. Reviews about medications are extracted from here and each review is considered as a single document. From this collection of documents, side effect words are identified based on tf-idf.

# Methodology

In this section, we are discussing the methodologies used for disease prediction, risk prediction and adverse drug reaction prediction using the datasets mentioned in the previous section.

## Disease Prediction

* A dataset with around 4000 diseases and their symptoms is used.
* The weights of symptom words are calculated based on their presence in the dataset. The symptom word with the least weight is used for the next step of weight calculation.
* This process is continued until maximum number of symptoms entered by the user is mapped against a disease.
* The possible diseases are listed and their descriptions can also be viewed.

The users enter the symptoms they are experiencing in the android app, and the algorithm mentioned above is done on the dataset to find the potential diseases.

*Algorithm*

1. Weights of all the symptom words are calculated.

Weight = no. of diseases with symptom ‘x’

Total no. of diseases

1. From the set of symptoms entered by user, the symptom word with the least weight is selected. The least weight is chosen in order to make the comparison process faster and easier.
2. Now, the list of diseases under consideration reduces to the ones which have the symptom word with least weight. Repeat step 2 and 3 until maximum number of symptom words are included.
3. Display the diseases found along with the number of symptoms entered by the user which are actual symptoms of the disease.

## Risk Prediction

By risk prediction, what we intend to do is to predict if there is a chance for a person to have a particular disease. For prediction purpose, an ensemble classifier is used. The machine learning algorithms used in the ensemble classifier are SVM, Naïve Bayes and K-Nearest Neighbor. The dataset is split into training set and test set. After training the classifiers using the training set, they are tested using the test set. Maximum accuracy was found for the ensemble classifier as compared to the individual machine learning algorithms on our dataset and hence it is chosen as the backend algorithm for our web application.

Naive Bayes is used as it converges quickly as compared to other discriminative models. This is because Naïve Bayes shows conditional independence. Even if the conditional independence assumption does not hold it performs well.

SVM has high theoretical guarantees regarding over fitting. Another major advantage of SVM is that, once the boundary is set, most of the training data is redundant. Hence, all it needs it a core set of points which will help to set the boundary.

KNN does not use assumptions for classification purpose. This is also memory based and constantly evolves. It is also robust to noise and does not require large amount of training data.

*Algorithm*

1. The dataset is divided into training set and test set.
2. Each classifier has a counter c initialized to 0.
3. Test set with ten instances is given as input to the three classifiers to train the model.
4. The classifiers are tested using test dataset.
5. For each instance of the test dataset, if the output of the classifier is positive, increment c.
6. Repeat step 5 for all three classifiers
7. Set vote=0.
8. If c>5 for each classifier then output is positive else negative.
9. Increment vote for each positive prediction.
10. If vote>2 output is positive, else negative.

Table : Accuracy and Precision of classifiers

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | Accuracy | Precision | Recall |
| SVM |  |  |  |
| Naive Bayes |  |  |  |
| KNN |  |  |  |
| Ensemble Classifier |  |  |  |

* The five diseases chosen for risk prediction are Cardio-vascular disease, Liver disease, Psoriasis, Infertility and Diabetes during pregnancy.
* For each of these diseases, a risk prediction can be done based on the details entered by the user.

## Adverse Drug Reaction Prediction

The reviews from ‘medications.com’ are processed based on tf-idf and the end results are decided after comparison with a dictionary with side effect words. This is done in order to obtain maximum accuracy by avoiding words which are out of context. These side effect words corresponding to a particular medication is stored in a dictionary and when the user types in the name of the medication, possible side effects are displayed.

# Results

In this section, we are including the results that we obtained for the three main prediction components of our project which are disease prediction, risk prediction and adverse drug reaction prediction.

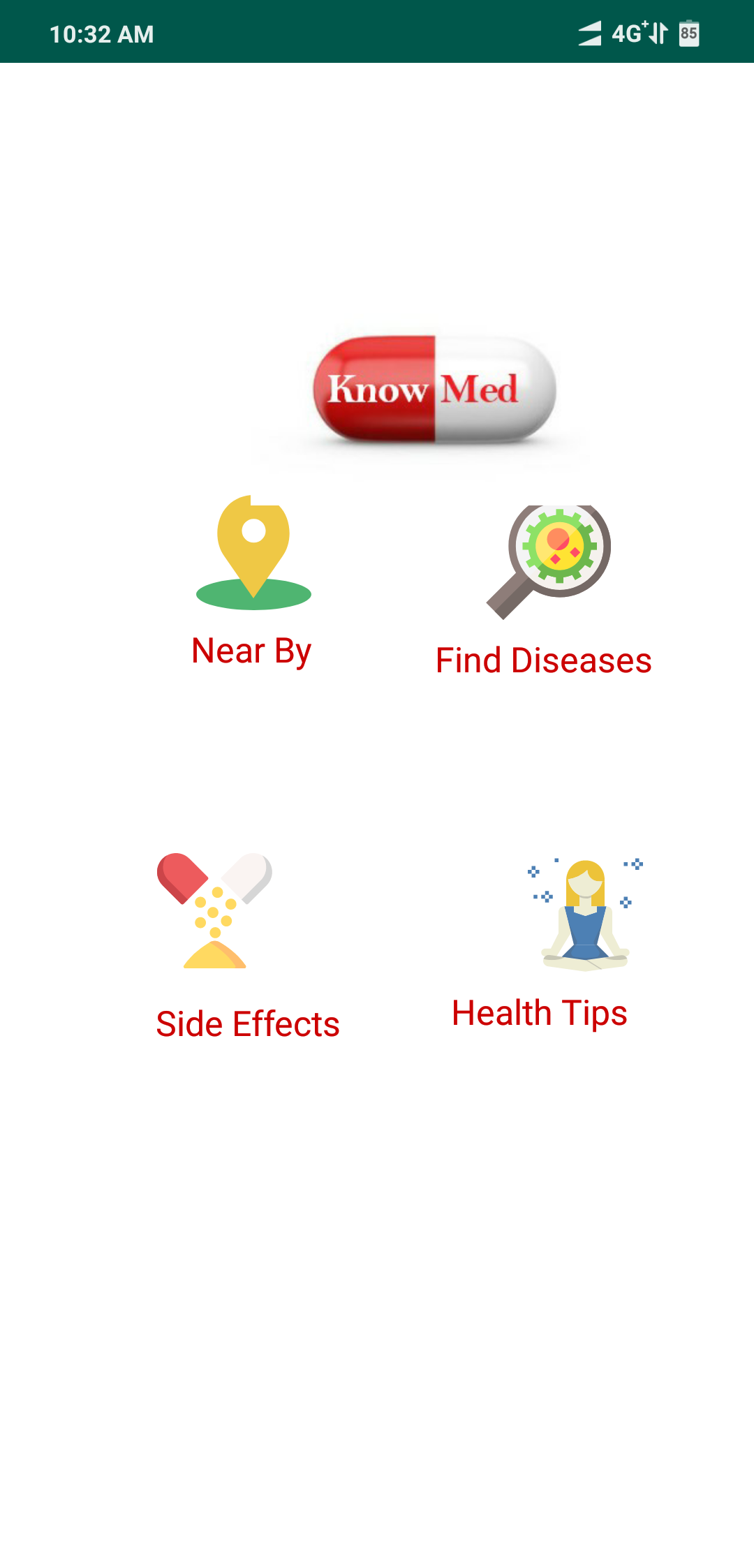


Figure : Android App Opening Page

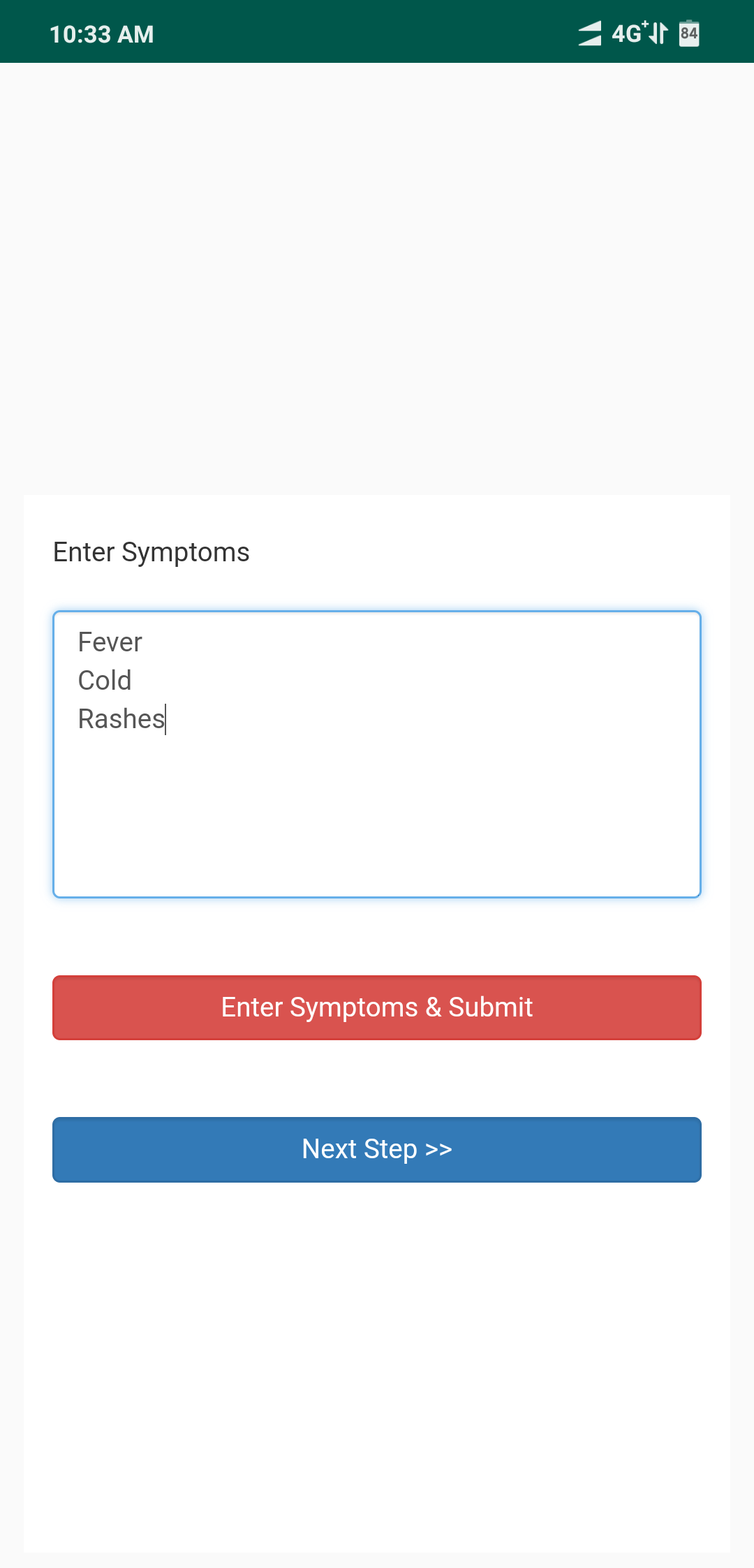


Figure : Entering Symptoms

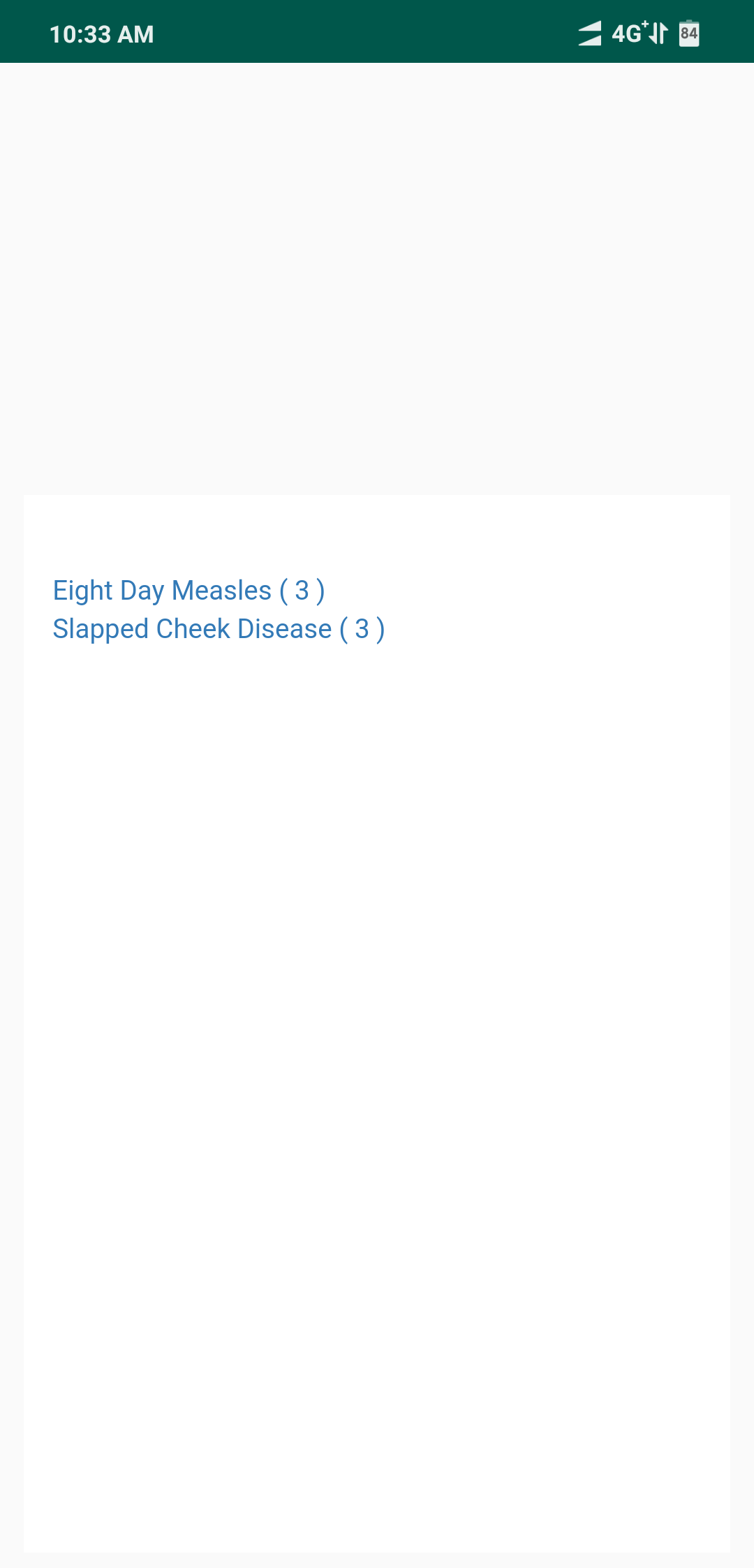


Figure : Displaying Predicted Diseases

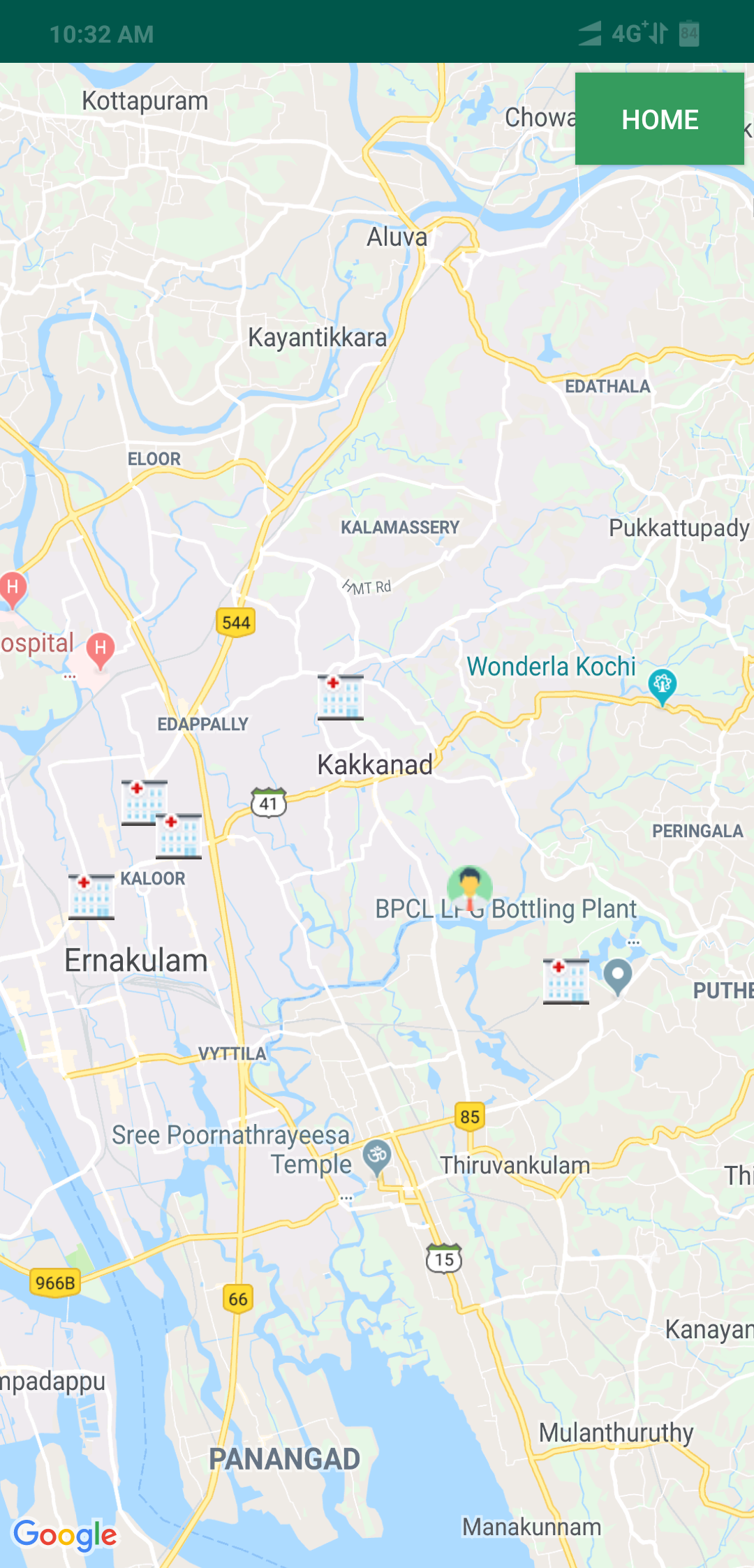


Figure : Location Based Medical Help

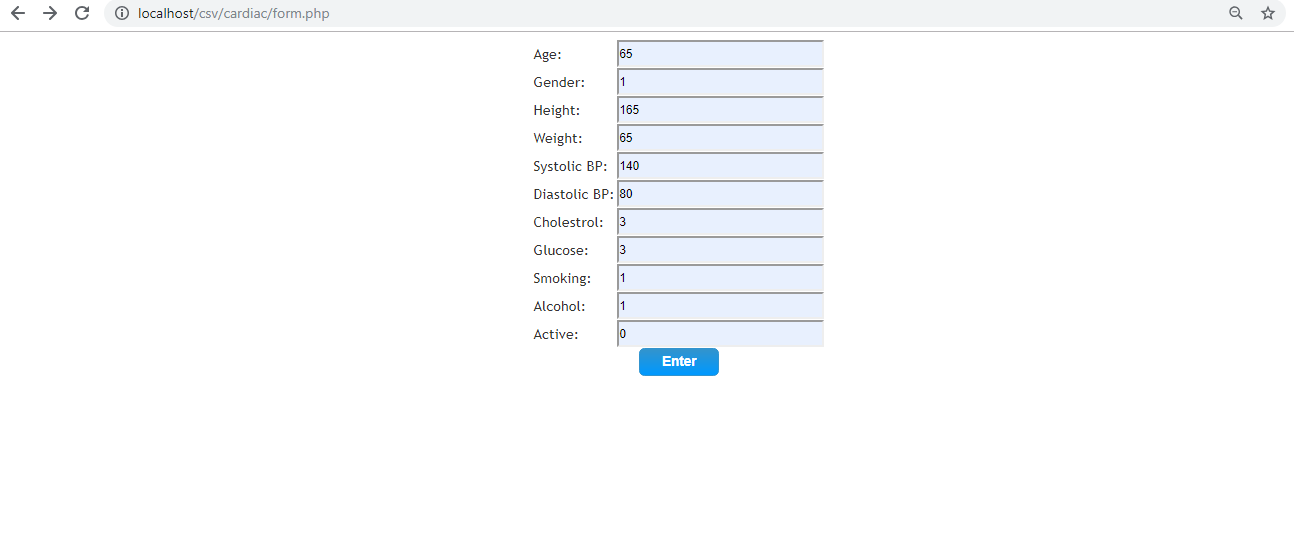


Figure : Web App - Entering Patient Details

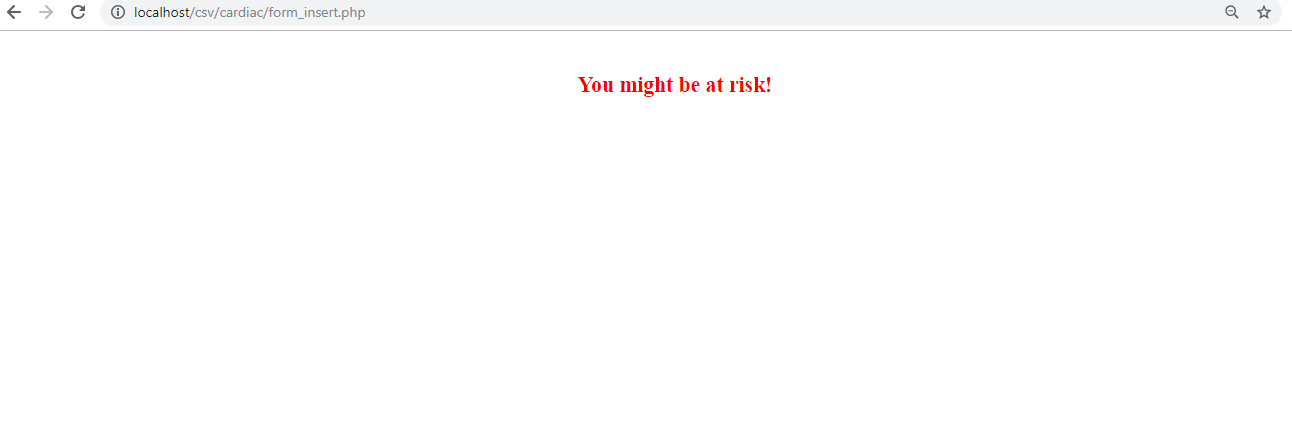


Figure : Risk Prediction

# Conclusion

The overall objective of our project is to predict diseases more accurately based on symptoms and also to predict the adverse drug reactions caused by medications. Our project also has a module that can predict if a person is at risk for the fives diseases included in our system which are Cardio-vascular disease, Liver disease, Psoriasis, Infertility and Diabetes during pregnancy.

In the disease prediction module, from among a set of 4000 diseases, a weight based algorithm is used to predict the disease a person might be having based on symptoms entered.

The adverse drug reaction prediction module extracts reviews from ‘medications.com’ which is an online healthcare forum. Using tf-idf the side effect words are found from these reviews. For this purpose, each review is considered as a single document.

The Risk Prediction module uses SVM classifier. KNN, Naïve Bayes and SVM were applied to the five datasets but maximum accuracy was obtained for SVM as mentioned in Table 2.

Along with all these features, additional functionalities such as location based medical help and health tips have also been included in the android application.

Another advantage is that the system is user friendly with well equipped user interface which enables users to enter their details easily.

##### References

1. Heart Disease Prediction using Data Mining Techniques-Abhishek Rairikar, Vedant Kulkarni et. Al at 2017 International Conference on Intelligent Computing and Control (I2C2)
2. A survey on Technique for Prediction of Disease in Medical Data- Ahelam Tikotikar and Mallikarjun Kodabagi at 2017 International Conference on Smart Technology for Smart Nation
3. Mining Adverse Drug Reactions from online healthcare forums using Hidden Markov Model- Sampathkumar et al. - BMC Medical Informatics and Decision Making , 2014
4. Predicting Adverse Effects of Drugs – Sofia Alexandra –Master of Biomedical Engineering, University of Porto- July 7, 2017
5. Mining Twitter for Adverse Drug Reaction Mentions: A Corpus and Classification Benchmark – Rachel Ginn- Regis University, Denver
6. Post-Marketing Surveillance for Adverse Drug Reaction using Clinical Data- Sujay Bankar, Suyog Gaikwad, Sidhesh Gawas, Mayur Shirkare, Mrs.Deepa Abin- International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5 Issue 1, January 2016
7. Prediction of Adverse Drug Reactions Using Decision Tree Modelling- F Hamman et al.- clinical Pharmacology and Therapeutics- march 10, 2010