Research Article

Disease and Adverse Drug Reaction Prediction using Machine Learning

Reshma Mathai[1], Ardra K John[1], Anima M M[1], Lakshmi K S*[2], Athulya James[1]

1. Former students, Department of Information Technology, Rajagiri School of Engineering and Technology, Kakkanad, Kerala, India.
2. Assistant Professor, Department of Information Technology, Rajagiri School of Engineering and Technology, Kakkanad, Kerala, India.

***Corresponding author's E-mail:** lekshmy.shalu@gmail.com

Received: 18-04-2021; **Revised:** 20-06-2021; **Accepted:** 26-06-2021; **Published on:** 15-07-2021.

ABSTRACT

The aim of the project is to use machine learning techniques for disease prediction, risk prediction and prediction of 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 the drugs may have. 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.

Keywords: Disease, adverse reactions, machine learning.

QUICK RESPONSE CODE →

DOI: 10.47583/ijpsrr.2021.v69i01.026

DOI link: http://dx.doi.org/10.47583/ijpsrr.2021.v69i01.026

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. Collecting electronic health record is increasingly convenient because of the growth in medical data. Prediction usually involves a machine learning algorithm (e.g. Support Vector Machine, Naïve Bayes etc.). It also includes a supervised learning algorithm by using training data along with labels to train the model. Patients can be classified accordingly as high- risk or low- risk in the test set. The idea that we have can be called as a "second opinion" since 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 describe a few methodologies and techniques used in this area.

Mining Adverse Drug Reactions from online healthcare forums using Hidden Markov Model[1]

The idea presented by Sampathkumar et.al is to extract drug side effects from online healthcare forums. They have used Hidden Markov Model (HMM) for accomplishing this task. Adverse drug side effects are extracted from the online health care forum messages and using these messages a prediction system was developed. The online healthcare forum used here for extracting side effects is medications.com. These healthcare messages are then used in the training and validation of the system.

The system consist of 3 phases: The first phase is an information retrieval module. In this phase, messages are crawled from medications.com. It is then fed to the second phase, which is a text processing module. After completing the necessary preprocessing, the data is fed to the third phase, Information Extraction Module which in turn consists of a Named Entity Recognition Module and Relationship Extraction Module.

Using the HMM classifier, a 10 fold cross validation on the dataset produced an F-score of 0.76. 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. The 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[2]

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 work 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 whether a patient gets a heart disease or not. 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 whose 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 classifies the 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 soon. Their major finding 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 [3]

It is important to discover hidden patterns and relationships from medical database. For classical clinical diagnosis, it requires 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, which are used as classifier to build a cost-effective model for disease prediction is presented. 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. 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.

The main focus of 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.

Predicting Adverse Drug Events from Personal Health Messages [4-7]

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.

Fig.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.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.

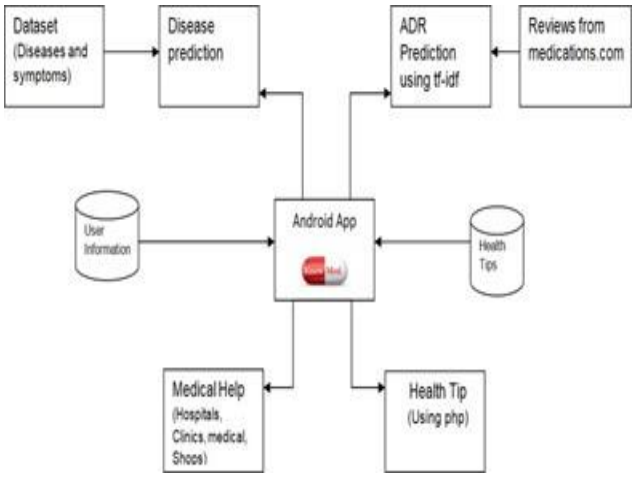

Figure 1: System Overview of Android Application

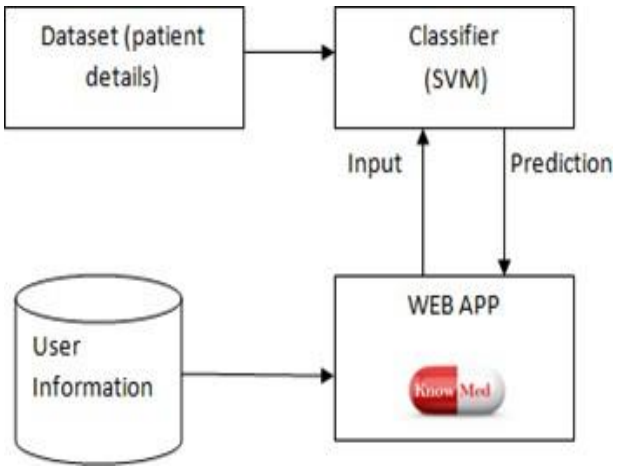

Figure 2: System Overview of Web Application

Dataset Description

Hospital datasets which are used in the study are described in this section. In addition to it, disease risk prediction model and evaluation methods are provided. We also describe the dataset used for disease prediction. The data collected for predicting adverse drug reaction will also be described here.

A. 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 1 shows two samples taken from Disease-Symptom Dataset.

Table 1: Sample Disease-Symptom Dataset

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

B. Hospital Data

Real-life hospital data sets are used in this study. This hospital data are stored in the database. The dataset includes structured data. The structured data consists of laboratory data and the patient's basic information such as the patient's age, gender and life habits. 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 scaling, itching, erythema, oral involvement, knee-elbow involvement, definite borders, polygonal papules, follicular papules, family history, age, scalp invasion and koebner phenomenon. Attributes used for Fertility are season, age, childhood disease, accident or trauma, interventions, high fever last year, smoking habits, alcohol consumption and number of hours spent sitting. Attributes used for Diabetes are insulin, age, BMI, blood pressure, glucose, diabetes pedigree function, skin thickness, and pregnancies. Attributes used for Liver Disease are age, gender, total bilirubin, direct bilirubin, alkaline phosphates, alamine amino transferase, aspartate amino transferase, total proteins, albumin, albumin and globulin ratio.

C. 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.

$$\text{Weight} = \frac{\text{no. of diseases with symptom 'x'}}{\text{Total no. of diseases}}$$

2. 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.
3. Now, the list of diseases under consideration reduces to the ones which has the symptom word with least weight. Repeat step 2 and 3 until maximum number of symptom words are included.
4. 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. Here, machine learning algorithms such as SVM, Naive Bayes and K-Nearest Neighbor are used. The dataset is divided into test set and training set. The classifiers are tested using the test set after they are trained using the training set. Table II shows the performance of various classifiers. Maximum accuracy was found for SVM on our dataset and hence it is chosen as the backend algorithm for our web application.

Table 2: Accuracy and Precision of classifiers

Classifier	Accuracy	Precision	Recall
SVM	93.55	95.56	83.72
Naive Bayes	88.87	93.73	78.35
KNN	84.75	92.18	72.97

- 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. Fig.3 shows the opening page of Android app. Fig.4 shows the screenshot of the window for entering symptoms.

Figure 3: Android App Opening Page

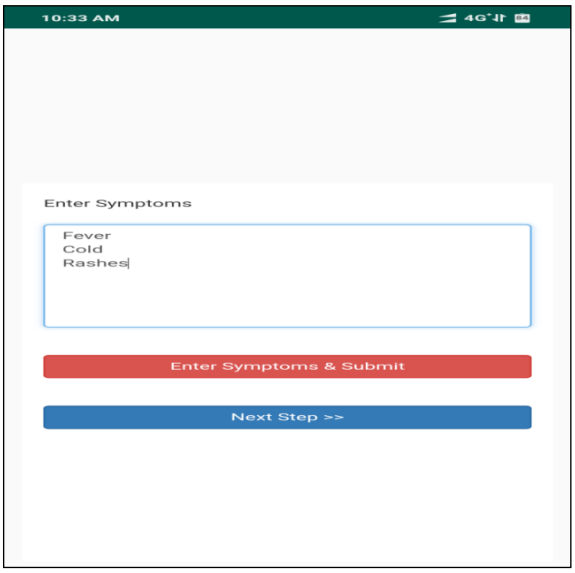

Figure 4: Entering Symptoms

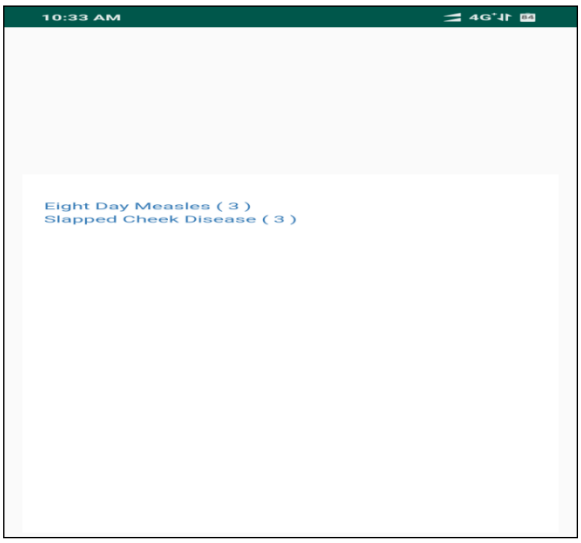

Figure 5 : Displaying predicted Diseases

Once the symptoms are entered, the predicted diseases will be displayed as shown in Fig.5. The user can then go for medical help which would be displayed as in Fig.6.

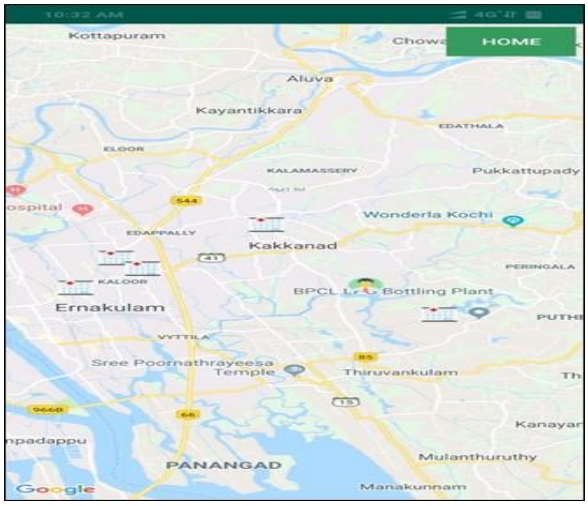

Figure 6: Location Based Medical Help

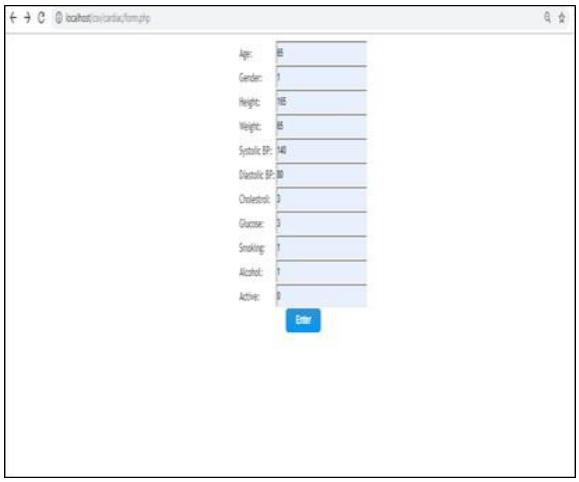

Figure 7: Web App - Entering Patient Details

Fig.7 shows the window for entering patient details in the web app and based on the values entered the app will predict whether the patient is at risk or no risk as shown in Fig.8. Fig.9. shows the performance comparison of various algorithms.

Figure 8: Risk Prediction

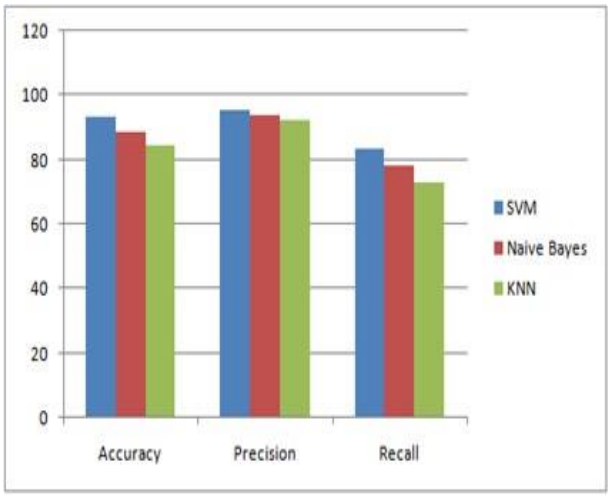

Figure 9: Performance analysis of various algorithms

CONCLUSION

The objective of our project is to predict disease 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 five diseases included in our system which are Cardio-vascular diseases, 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 'medication.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 and Naïve Bayes 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. Sampathkumar H, Chen X, Luo B. Mining Adverse Drug Reactions from online healthcare forums using Hidden Markov Model. BMC Medical Informatics and Decision Making, 2014.
2. Rairikar A, Kulkarni V, Sabale V, Kale H. Heart Disease Prediction using Data Mining Techniques. Proc. of International Conference on Intelligent Computing and Control(I2C2), 2017.
3. Tikotikar A and Kodabagi M. A survey on Technique for Prediction of Disease in Medical Data. Proc. of International Conference on Smart Technology for Smart Nation, 2017.
4. Huang LC, Wu X, Chen JY. Predicting adverse side effects of drugs. BMC Genomics. 2011 Dec 23; 12 Suppl 5(Suppl 5): S11. doi: 10.1186/1471-2164-12-S5-S11. Epub 2011 Dec 23. PMID: 22369493; PMCID: PMC3287493.
5. O'Connor K, Pimpalkhute P, Nikfarjam A, et al. Pharmacovigilance on twitter? Mining tweets for adverse drug reactions. AMIA. Annual Symposium proceedings. AMIA Symposium. 2014; 18: 924-933.
6. Alomar M, Tawfiq AM, Hassan N, Palaian S. Post marketing surveillance of suspected adverse drug reactions through spontaneous reporting: current status, challenges and the future. Ther Adv Drug Saf. 2020; 11: 2042098620938595. Published 2020 Aug 10. doi:10.1177/2042098620938595.
7. Hammann F, Gutmann H, Vogt N, Helma C, Drewe J. Prediction of adverse drug reactions using decision tree modeling. Clin Pharmacol Ther. 2010 Jul; 88(1):52-9. doi: 10.1038/clpt.2009.248. Epub 2010 Mar 10. PMID: 20220749

Source of Support: The author(s) received no financial support for the research, authorship, and/or publication of this article.

Conflict of Interest: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.