Illness Diagnosis System

Sanchita Ochani, Yashas Bysani Muralidhar, Nirav Shah, Saumil Patel, Gurjinder Singh Oberoi

*CS691*

*Seidenberg School of Computer Science and Information Systems*

*Pace University, New York, NY, USA*

***Abstract* — The effectiveness of machine learning depends on how much data is available – the more, the better – and the field of medicine is a goldmine. Applications of machine learning in the field of pharmacy and medicine is vast. Record keeping, prediction of outbreaks, and research are some of the areas where machine learning can truly help the field of medicine. This paper focuses on illness identification and personalized treatment. The idea stems from the need for accurate diagnosis of diseases and a more reliable way to treat them. With so many diseases showing similar symptoms, a well-trained system can certainly aid not only patients, but also doctors.**

***Keywords***— **machine learning, illness identification, support vector machine, predictive analytics, healthcare**

1. Introduction

The advancement in the field of technology has helped doctors and scientists all over the world. In a field like medicine, it’s crucial to apply every bit of emerging technology that can help save lives. Machine learning is the ability of a machine to learn things like humans do, based on rigorous training from a given data set. Algorithms are written to recognize complex patterns dynamically and make intelligent decisions based on insight generated from training. Training is the process of building a mathematical model that depends on a data set. It includes querying the relevant dataset and training the classification model. Based on this trained model, the system should be able to make precise predictions of new data.

Disease identification and diagnosis is at the forefront of machine learning applications and research. Many companies have been using machine learning to diagnose cancerous tissues just as accurately as a trained oncologist. This is a breakthrough in the field of medicine. Machine vision and deep learning have been applied to diagnosis of blood stream and genetic diseases as well. In the field of oncology, the first step towards cancer diagnosis is always a visual examination done using a dermatoscope. Based on the evaluation of the lesion, the dermatologist will come to a suitable conclusion. An algorithm developed by a team of researchers at Stanford University managed to produce the same results based on images as a team of highly skilled dermatologists. The results are promising but arduous training is required before the algorithm can be deployed.

Proposed system is based on supervised learning – which focuses on a given data set, and using that can identify correctly the relationship between the input and output. The algorithms which will be used for this system are a combination of Support Vector Machine (SVM) and Forest. The database consists of diseases and symptoms for which an association needs to be found by the system, for correct identification of an illness based on user input, i.e. symptoms. The system will be able to predict illnesses such as common flu, migraines, food poisoning, fractures, vitamin deficiencies, alcohol poisoning, allergies, eating disorders, dehydration, commonly occurring infections among other illnesses. The proposed system can be used not only by scientists and doctors, but also by patients.

1. Literature Review

With artificial intelligence booming in today’s world, many systems have been using deep learning and machine learning techniques wherever possible. Naturally, the bigger companies have started using these techniques to advance the field of medicine and pharma. IBM’s Watson for Genomics [1] is working on tumor detection and personalized treatment plans. A Boston based company is also working on integrating artificial intelligence with oncology, endocrinology, as well as neurology.

Data mining has been engaged in various works in providing illness or disease prediction like hepatitis, cancer, diabetes, and heart diseases. The most frequent dataset undergoes mining in order to find the item sets for user. Association rules that define the methods such as Apriority and FPGrowth are commonly used. To reduce the size of data, genetic algorithms are used and an optimal subset of attributes was used for heart disease prediction. To extract models describing classes of data, classification and supervised learning is required. We discuss in this paper, the exploration of illness prediction and the treatment model to methodically close all the insufficiencies currently found in applying data mining techniques for treatment of illnesses.

With previous research [2] on implementing illness or disease prediction system using data mining, we are clear that the usage of such datasets could boost the medical domain to a great extent. With tools in data mining such as Orange, Weka, Rapidminer the system can analyze the data set and predict the future behavior of diseases. This research focuses on comparison of tools to identify their accuracy.

With more research [3] on the techniques in data mining that provide better performance and accuracy, techniques such as Naive Bayes, J48 and Decision Tree. This research provides us with more facts for implementing prediction for diseases like diabetes and breast cancer. Currently existing and developing systems focus on prediction of diseases like cancers, tumors, heart diseases but not on common illnesses.

1. Current Solutions

A U.K. based startup came up with the mobile and web application Babylon Health whose goal is to predict diseases based on symptoms by dynamically asking questions to the user. The core functionality of our system will be similar to Babylon’s. The users of our proposed system will also be required to answer questions and give symptoms as an input.

The existing system Babylon, is restricted in terms of illnesses that can be predicted. It cannot predict skin issues, or aid in children’s health or pregnancy related issues. The application is also restricted entirely within certain zones of London and does not work for the rest of the city or even the country. The application also requires very simple language and focuses entirely on keywords. It rejects any words not existing in their database. The application also asks too many questions which can be considered frustrating by the user. The application also doesn’t allow you to book appointments.

We can see from Figure 1 the existing system Babylon Health which is similar to what our system will function like.

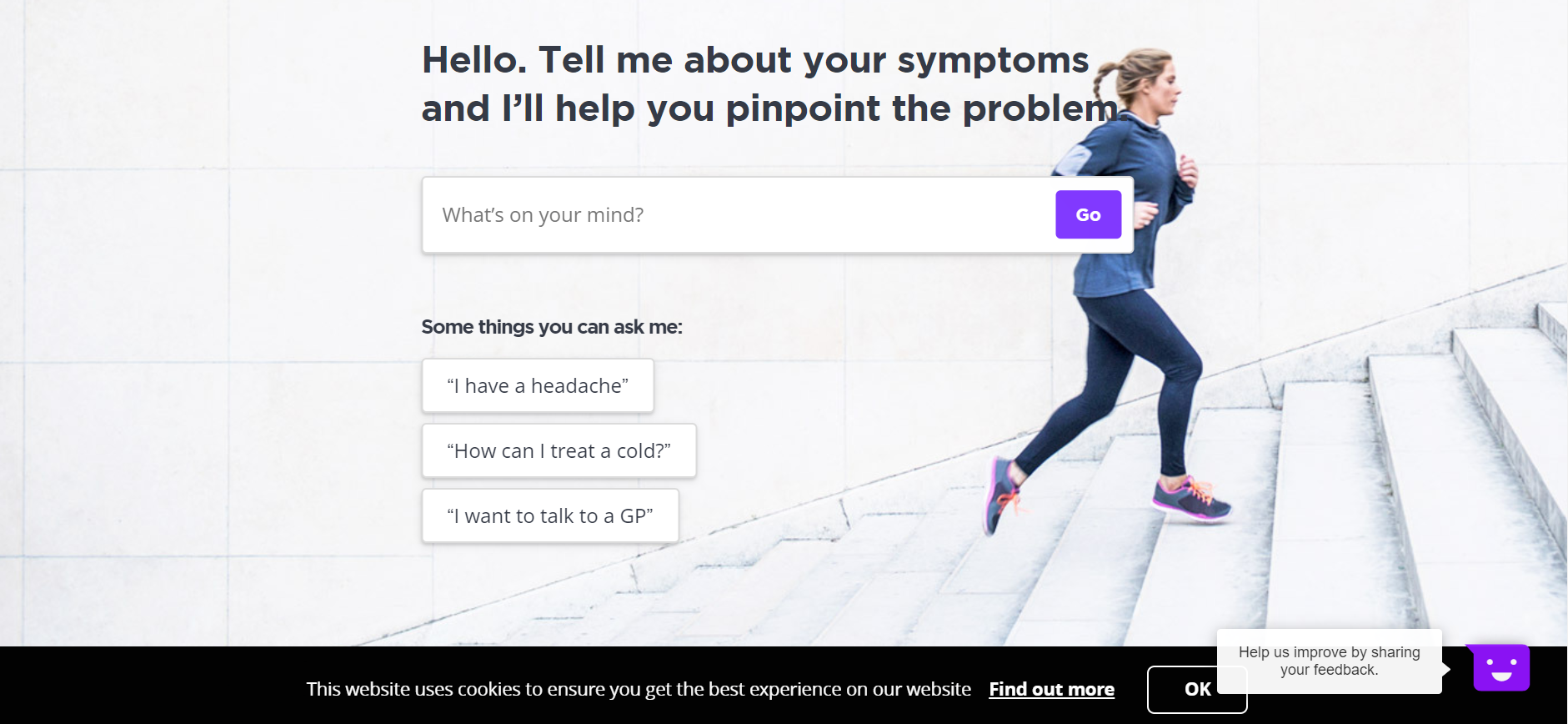


Figure 1. Screenshot of Babylon Health

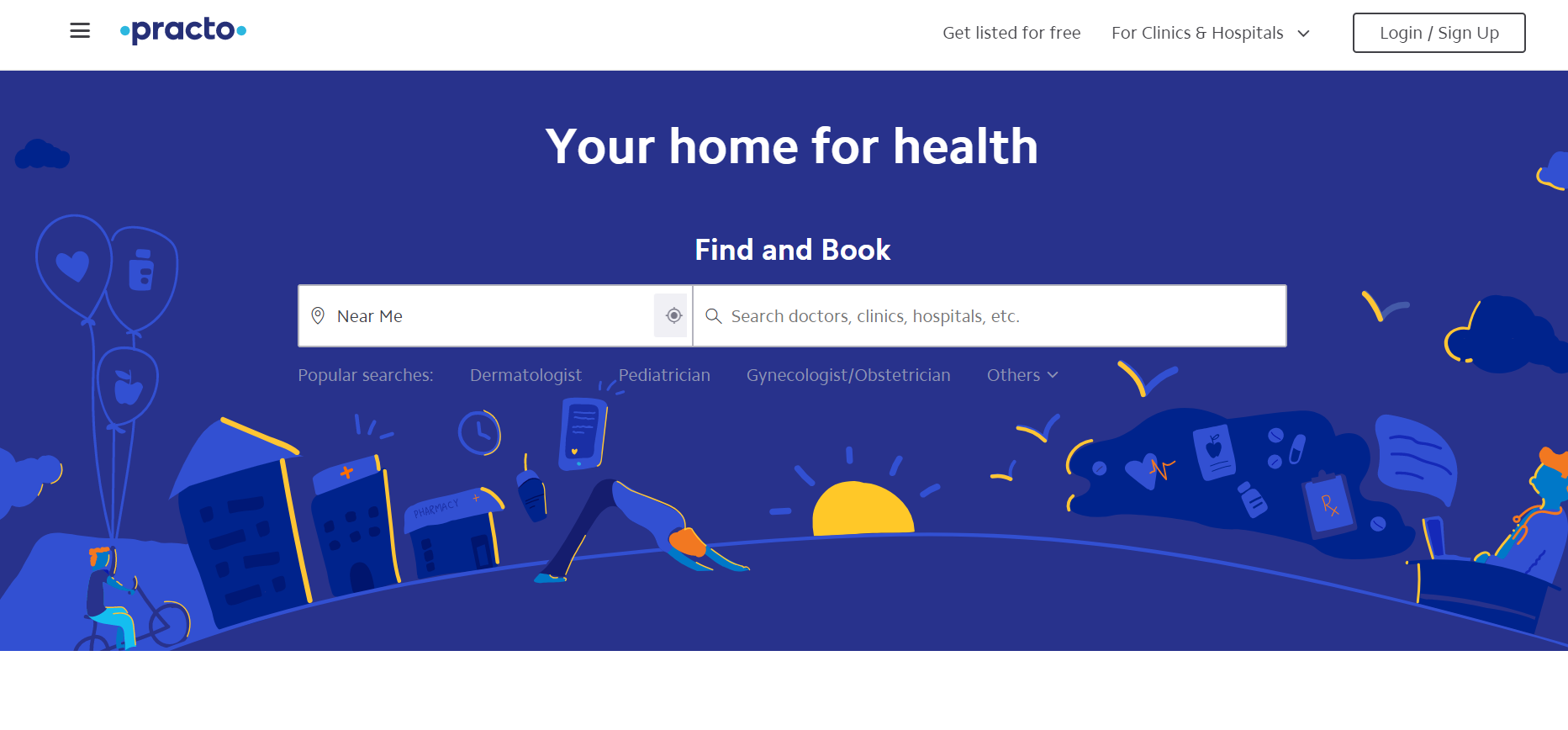


Figure 2. Screenshot of Practo

Practo, another existing system, as shown in Figure 2 does not predict diseases or illnesses but allows you to communicate directly with doctors through the application. This application is also restricted to few Asian countries.

Our application closely resembles a symptom checker tool provided by WebMD, a website used extensively for checking symptoms to certain illnesses. The tool is currently in beta stage and does not provide an interactive way to communicate with the system. It also gives a huge list of possible illnesses and diseases, few of which can be irrelevant. The tool is simple and easy to use, does not store any user information about previous illnesses and medications. Figure shows a screenshot of the tool.

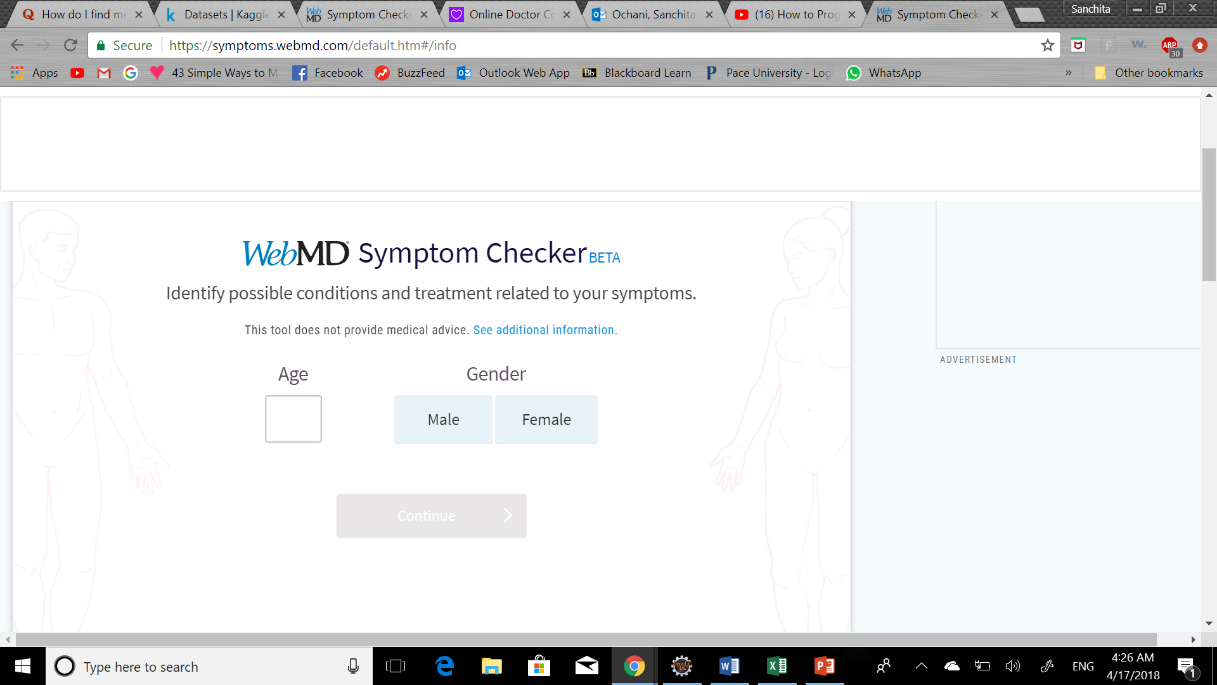


Figure 3. WebMD Symptom Checker

Table 1 provides a comparison of our system with existing systems.

TABLE I  
An overview of existing systems

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Babylon Health | Practo | WebMD symptom checker tool | Illness Diagnosis System |
| Disease Prediction | Yes | No | Yes | Yes |
| Age restriction | 18+ | 18+ | All ages | All ages |
| Free? | In-app payment | In-app payment | Yes | Yes |
| Prescription | Yes | Yes | No. | No. Homemade remedies only. |

1. Project Requirements

Our aim is to develop a machine so intelligent that more people can get correctly diagnosed and treated eliminating entirely the need to rely on underqualified doctors, long lines in the waiting room, and heavy charges that come with physical examinations of minor illnesses. With more and more things getting convenient and at the tip of our fingers, i.e. our phones, the need for getting diagnosed from the comfort of your bed is vast.

The basic functionality of our application is similar to few existing systems, although not limited to. A few key requirements can be derived from existing systems:

1. *User management:* User should be able to log in to his account and access his records and private data. User should be able to log out when he/she needs to and also be able to delete existing data if he/she wishes to delete their profile entirely. The user also controls what data they wish to provide to the system.
2. *Algorithms and testing:* As a machine learning application, our application depends heavily on choosing the correct algorithms and proper training of them.
3. *Obtaining the right datasets:* One of the most challenging steps in our application is finding datasets that resonate well with our requirements.
4. *Additional features:* Our application also should be able to provide additional features such as suggesting nearby doctors, and clinics as well as the ability to book tests and integrate the res
5. *User interface:* A good user interface is essential for any application. UI needs to be simple, friendly and aesthetically pleasing.

1. Steps

A. Obtaining Dataset

B. UML diagrams

C. Decide Algorithms

D. Training Dataset

E. Testing

F. Documentation

[A] Obtaining Dataset - we gather all the symptoms focusing on certain illness, stored in different tables, which later will be used to match the user symptoms to illnesses and make a prediction.

[B] Creating UML and DFD diagrams helps us visualize the model of our system and the interaction between objects using sequence and data flow of the system.

[C] Deciding algorithms is an important step of our project. We are required to choose between multiple machine learning algorithms to pick the most appropriate one

[D] Training dataset is an essential part in machine learning, where the system is rigorously trained resulting in accurate prediction.

[E] Testing stage is necessary to improve the performance and efficiency of the system.

[F] Documentation records all the official information of the system, like working, data flow, source code and much more.

1. Architecture

Machine learning problems are usually divided into classification and regression. In regression based problems, the goal of the algorithm is to predict a continuous valued output. In classification based problems, the goal is to predict a discrete value output. Our system follows a standard way of implementing machine learning by using these classification algorithms. The algorithms we will be using are:

1. *Support Vector Machine (SVM)*
2. *Random Forest*

SVM is a supervised machine learning algorithm which can be used for both classification and regression. Supervised learning has labeled data. We are going to use SVM to classify two sets of data and we are going to optimize it by using gradient descent method. If we have two classes, we will plot both classes on a 2-dimensional graph and draw a decision boundary that best separates the two classes. This line is called hyperplane and SVM creates this hyperplane for us.

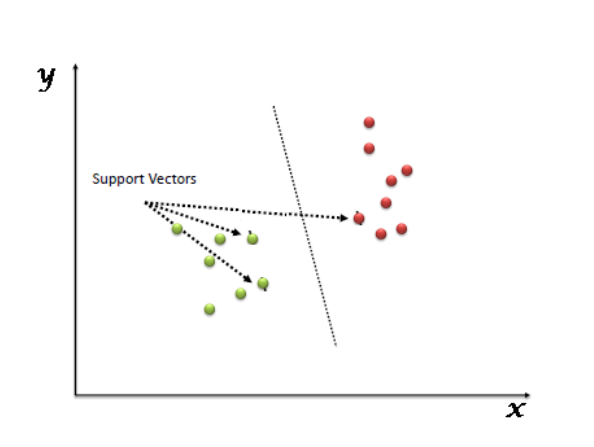


Figure 4. SVM

SVM is great at classification if you have small data sets of thousand rows. Other algorithms like Random Forest or deep neural network requires more data, but almost always come up with very robust model. The decision of which classifier to use depends on the problem and data.

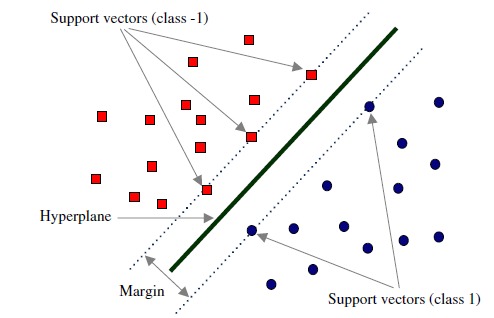


Figure 5. SVM and hyperplanes

Given 2 or more labeled classes of data, SVM acts as discriminative classifier, formally called as a hyperplane that separates all the classes. New data are then mapped into same space which can then be categorized based on which side they fall on. We can build hyperplane or line with decision boundary between the classes by maximizing the margin, that is the space between the line and the points in each of those classes that are closest to decision boundary line. These points which are closest to the hyperplane are called support vectors. They are the data points that support the creation of a hyperplane, so we maximize the margin as we want to draw a hyperplane that is in perfect spot right in the middle between these sets of data such that when we plot new data points it has the maximum likelihood of falling on the correct side of decision boundary. We are trying to draw a line which is the decision boundary and the line that best classifies the two sections.

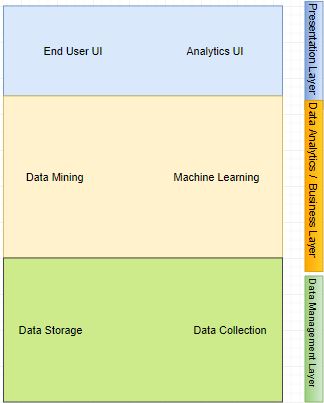


Figure 6. Functional Diagram

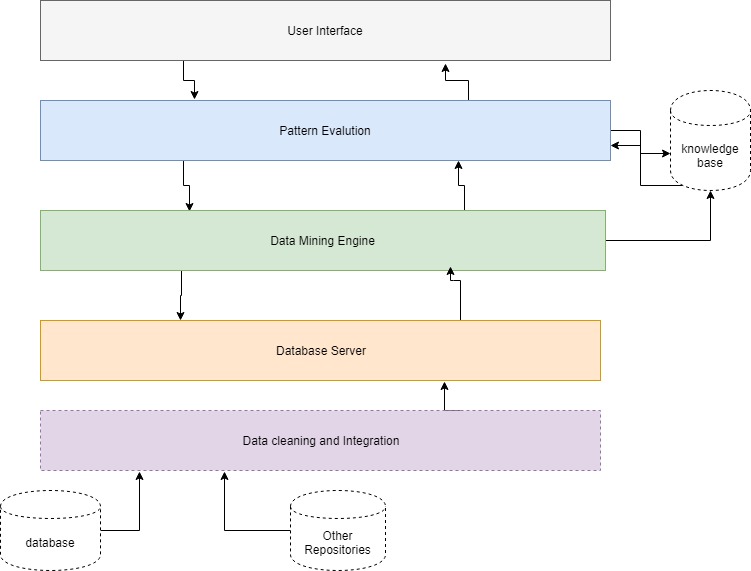


Figure 7. Data Mining architecture

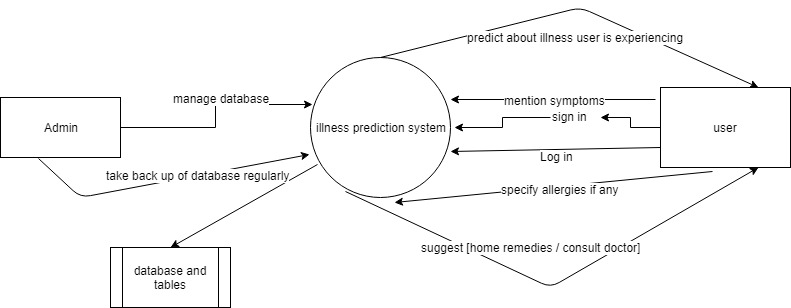


Figure 8. Data Flow diagram

The above diagrams show the flow and functionality of our system.

Our system also uses a database to store the user information. The database consists of login information along with some factors that can affect health of the user, like previous medical conditions, disabilities, country of residence, etc.

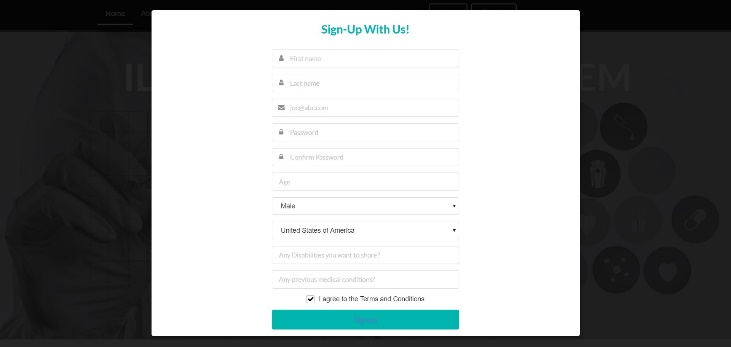


Figure 9. Signup page for user

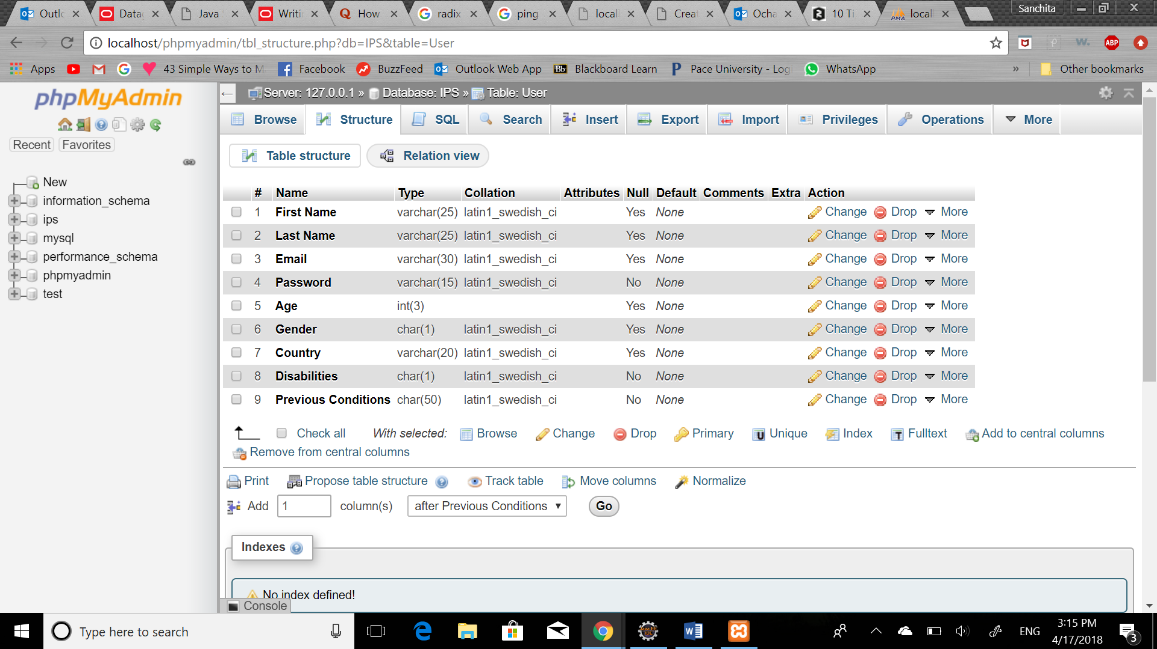


Figure 10. User table

1. Advantages
2. Through this system the patients would get informed about their diseases and in some situations get basic treatments. In particular instances one could get instant relief by system suggestions for remedies.
3. Predictive analysis helps patients diagnose an illness based on unidentified symptoms. With early intercession, numerous sicknesses can be anticipated. In the domain of genomics, it will help users to foresee a potential danger. Patients can choose to change their way of life to prevent certain diseases.
4. Users will also get information about what precautions they can take based on current location and weather conditions.
5. They can also be informed about things to avoid consuming for certain allergies.
6. Future Scope

Patients will have more information about what illnesses they are going through without having to consult doctors and spend a lot of money. There will be a significant change in the field of pharma. We will also collaborate with medical suppliers to prescribe medicine for the predicted illness. Using machine learning and data mining techniques we will conduct experiments on real time datasets to predict even more diseases, in number and seriousness.

1. Product results

Here are the screenshots of the website implemented so far.



Figure . Landing page

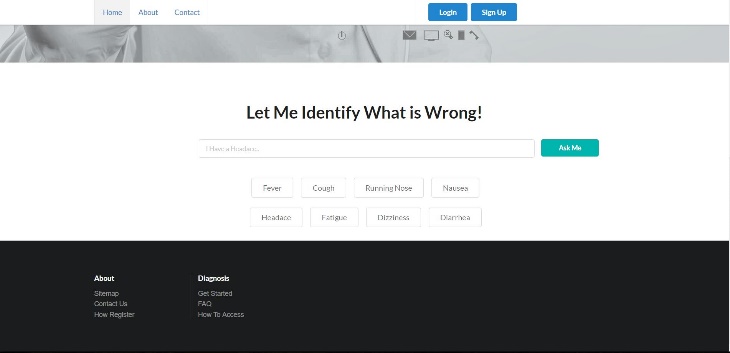


Figure . Homepage

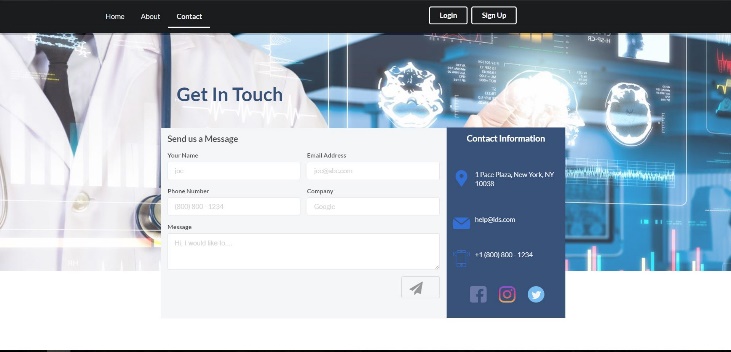
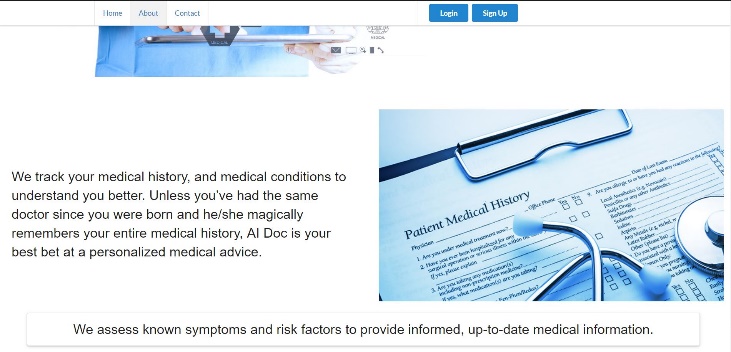
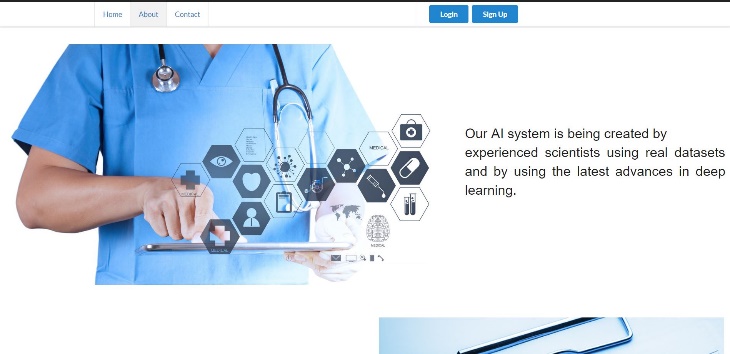
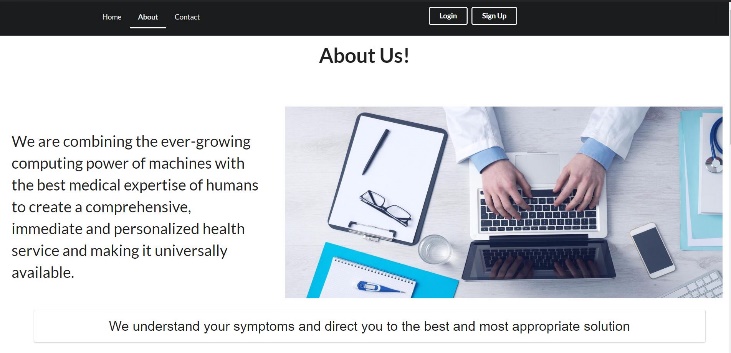


Figure 13. Contact page

1. Conclusions

With the current scenario, we are expected to rely completely on a doctor’s memory for illness diagnosis. While doctors are well trained and smart, they are not necessarily up to date and can’t possibly remember everything. An algorithm, on the other hand, relies entirely on a dataset even after thorough training. Given a massive dataset and access to it, a human brain cannot be expected to analyze and memorize it in its entirety and then also integrate it with the medical records of every single patient. This is where machine learning has an advantage. The use of predictive analytics is not only useful for predicting diseases but it may also reveal surprising associations that a human brain can easily miss. Learning models can only improve with time and more training and the same cannot be said about a human mind.

References

1. IBM Watson health “internet: https://www.ibm.com/watson/health/oncology-and-genomics/genomics/”
2. Kauser Ahmed p (2017). *Analysis of data mining tools for disease prediction*
3. K.Gomathi\*,Dr. D. Shanmuga Priyaa (2016, December). *Multi disease prediction using data mining techniques*
4. Back to biology for a healthier tomorrow “internet: http://berghealth.com/pipeline/”
5. 7 applications of machine learning in pharma and medicine “internet: https://www.techemergence.com/machine-learning-in-pharma-medicine/”
6. Seven ways predictive analytics can improve healthcare
7. “inter/net: https://www.elsevier.com/connect/seven-ways-predictive-analytics-can-improve-healthcare”
8. Understanding support vector machine algorithm from examples “internet: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/”
9. Sathyabama Balasubramanian, Balaji Subramani (2014, February). *Symptoms based diseases prediction in medical system by using k-means algorithm*
10. Md. Tahmid Rahman Laskar, Md. Tahmid Hossain, Abu Raihan Mostofa Kamal, Nafiul Rashid (2016, January). *Automated disease prediction system (ADPS): a user input-based reliable architecture for disease prediction*
11. Machine learning for medical diagnostics – 4 current applications “internet: https://www.techemergence.com/machine-learning-medical-diagnostics-4-current-applications/”