Disease Detection using Machine Learning

Pranav Kulkarni, Shubham Gupta, Animesh Singh
UG scholar

Abstract- **India is home to 1.3 billion people and with such a huge population arises sanitation and cleanliness problems. These problems further nurture deadly diseases. Indians are prone to a variety of disease be it lifestyle related or deficiency related. Indians, majorly constituting of conservative societies often shy away from relaying their symptoms to doctors which further leads to mortal complications. We used machine learning to bridge the gap between the lack of communication. Machine learning is a widely over used topic in medical field but rather than removing the doctor completely and substituting it with an Artificial Intelligence we made them work together to simplify the whole process and make it less time consuming and less prone to errors. We investigated different diseases and their symptoms. Post the analysis we made a model to classify the symptoms and provide an overview of the patient to the doctor while limiting the scope to provide a better diagnostics. It not only will reduce the burden on the doctor but also save the patient from unnecessary medical tests thus reducing the overall costs of medication.**

I. INTRODUCTION

Patients arriving at a primary care service sometimes need to wait for a long time before being advised by a doctor . This is often due to high workload and limited resources at the primary care service To facilitate the process, nurses and other health care staffs usually take the role of patient intake. An incoming patient would be first greeted by a receptionist who carries out an intake inquiry. The receptionist would typically be someone who has a certain level of medical proficiency, and the inquiry involves collecting patient information and understanding the symptoms of the patient. A brief report is generated as outcome of this inquiry to narrow down the causes of the symptoms, so that the doctor may then use minimum effort to perform differential diagnosis. The old-fashioned way of going to see the doctor in the hospital where the patient speaks out his symptoms and the doctor on his personal analysis provides the necessary prescriptions.

Machines to predict the disease using deep learning for certain specified diseases. The main drawback of the current system is that there are not enough doctors for every patient in India. The next is that the people in India are conservative they never discuss their symptoms or overlook them so the doctors have to work with that possibility which is prone to errors due to lack of information. Even the doctors are sometimes prone to errors while predicting the disease or might have something on their mind to no pay heed to an important symptom and as the human mind work, they can't remember every possible disease from each symptom. Overall which leads to an error prone system which reduces the level of health care required and many more problems. Often the doctors prescribe more drugs than what is necessary which leads to overdose or the recent growing concern of antibiotics which is because of over use of antibiotics even if it was unnecessary. The machines to predict only certain disease works at a really limited scope and has a high set up cost and not efficient enough after the labor put in it. The proposed system works using speech recognition. The proposed model would be a pre trained model using a pre-defined set of diseases with their symptoms.

II. LITERATURE SURVEY

The overcrowding is a big issue at the health care centers due to large waiting lines of patients and is a really difficult hurdle to overcome for almost every country. As the number of patients grow so does the discontent felt by the patients, the earlier system of one to one communication between the doctors and patients is now replaced with short encounters often leading to few or none positive outcomes for the patient.

Research on human health is actually the most fundamental part of science for people, as none of us are resistant to physical afflictions as the research progresses more and more data is available to integrate in the already existing models and database to further improve the disease prediction modules using the said database Existing patient diagnosis bolster applications frequently appear as master frameworks. A typical test looked by every one of these applications is the equivocalness furthermore, decent variety of patient answers. Thus, conventional master frameworks ordinarily neglect to convey compelling choice help and does not have the adaptability that suits person needs.

An example on existing system using sparse learning for disease inference from health-related queries the main drawback of the system was it couldn't focus on the distinctive features of each disease. The whole system in general was trying to totally remove the doctor from the process which in-turn might lead to errors or a vague diagnosis as the machine works on the data sets and models and doesn't have experience like a doctor and often works on predictions. This system works on a chatbot type application. A similar application of medical diagnosis using chatbots is Mandy. Mandy uses mapping and an interface for patient to work as a chatbot and give predictions for the given symptoms. The drawback for Mandy is that the prediction accuracy is really low because the dataset is huge but Mandy might serve as the starting base for many such applications to improve on and work to further this idea of medical diagnosis using chatbots. The process is much more complex and unique but the same problem comes in play that is they disregard the importance of the doctor.

There are several disadvantages when a doctor is removed but there are places a doctor can't be reached and doesn't have required skill level to operate an AI medical diagnosis bot hence to reach these remote places where we need the doctors for diagnosis this idea deals with the problem or lack of doctors in certain areas .The suggested idea was to connect the patients with the doctors via a system to provide access even in remote areas. The next module was for disease diagnosis but the main problem is that the man power and the sheer capital needed to implement such a system would be really big and would require a really skilled labor force to operate and maintain it.

This reference serves the base idea for the development of machines for diagnosis using symptoms but was rather held down due to the lack of technology in that era but can be implemented perfectly now. The main cause of concern in this was that the doctor would feed the symptoms rather we could simplify the whole process if we use a machine to fed the symptoms rather than a human who is rather prone to errors and we make the doctor work together with the machine to get better outcomes. The security and prediction can also be improved as the technology used is really ancient. As we improve on this idea, we get a better and more perfect system working on the synergy between humans and machines. This reference help in the future development of the system where it doesn't need a prefeed dataset model to work on the samples. It can also work with undiagnosed samples by improving the computer aided diagnosis. As we go further into the whole spectrum of symptoms and their diseases we get many unknown symptoms that sometimes match but are not in the data set .This improved system will learn and automatically update itself to later on work with these unknown samples although this idea is really convenient the accuracy decreases by many folds hence is not preferred for now but can be improved on. The reference uses deep learning to predict the disease of patients by mapping the symptoms from their medical records if grouped together with and can help sort their respective drawbacks and give a better prediction as the drawback of this system was only that it didn't involve the expertise of doctors in its system and that it need pre fed records and can't manage undiagnosed samples.

So we might cover all the disadvantages of all the different system if we can make them work by adding features of the others but it might be really complex and might lead to more errors but still provide decent results.

III. ARCHITECTURE DIAGRAM

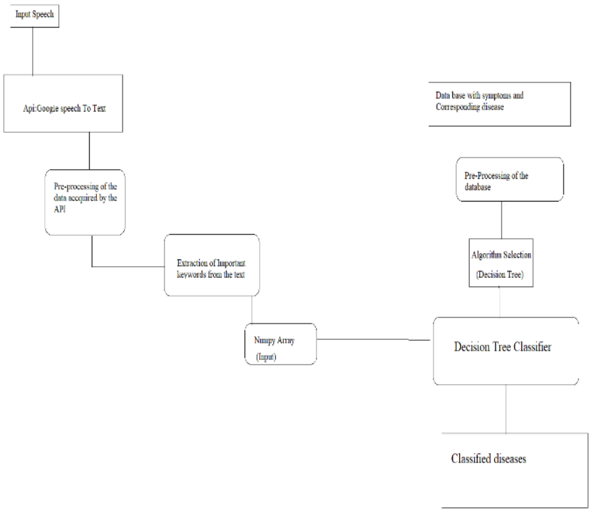

Input Speech
Api:Google speech To Text
Pre-processing of the data accquired by the API
Extraction of Important keywords from the text
Numpy Array (Input)
Data base with symptoms and Corresponding disease
Pre-Processing of the database
Algorithm Selection (Decision Tree)
Decision Tree Classifier
Classified diseases

Fig1.Artictecture of the proposed system

Figure1 is the architecture diagram of the system. The input is taken from the user in the form of speech and then Google text to speech API is used to get the keywords from the text. The keywords are then filled into a NumPy array and then fed into a classifier which is trained on a pre-existing data set. The algorithm analyzes the data and the result is given out by the algorithm.

IV. MODULES

Modules

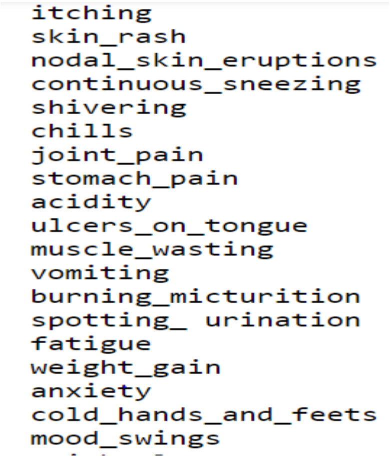
I. Speech recognition

This marks the start of the process for the diagnosis as the patient comes in and starts describing the symptoms to a mic. The mic will recognize the whole thing whatever the patient will speak and save it as a file in his database. We use the Google speech recognition API to get the input from the patient through mic and it will convert it to a text and then we use the python to save it in a file. Now let us denote the main text file with 'S'.

II. Grouping Symptoms by mapping

After the speech is converted into text the text is fed into the mapping algorithm which then separates the symptoms from the useless text using mapping.

Then the module will compare the text file with the symptoms file (say 'X') and it will extract the symptoms and shows to the screen. Then these symptoms will be given to the main algorithm for classification to the disease. To extract the main symptom, we used the simple text comparison from a file using python and we save it into a NumPy array for faster computation for the next module.

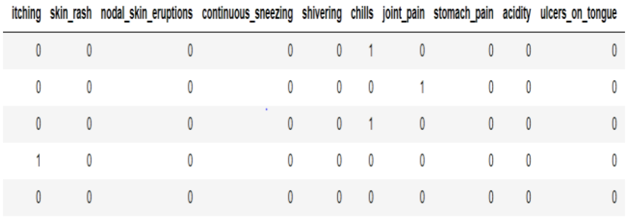


```
itching
skin_rash
nodal_skin_eruptions
continuous_sneezing
shivering
chills
joint_pain
stomach_pain
acidity
ulcers_on_tongue
muscle_wasting
vomiting
burning_micturition
spotting_ urination
fatigue
weight_gain
anxiety
cold_hands_and_feets
mood_swings
```

Fig 2.list of symptoms

III. Model formation

To classify the disease from symptoms we need to make model. To make a model we used different classification algorithm like Logistic regression, Naïve base, random forest, Decision tree and the support vector machine. To train the model we separate the dataset into two different set called Training set and Test set. The training set will be used to get train our classification model and the test set will be used to check the efficiency of the model.



itching	skin_rash	nodal_skin_eruptions	continuous_sneezing	shivering	chills	joint_pain	stomach_pain	acidity	ulcers_on_tongue
0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	1	0	0	0	0
1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

Fig3.database of different diseases

The dataset has 132 differed symptoms which are used to cl assify the disease. example These are some of the symptoms .

{ itching , skin rash , sneezing , shivering , joint pain , stoma ch pain , acidity }

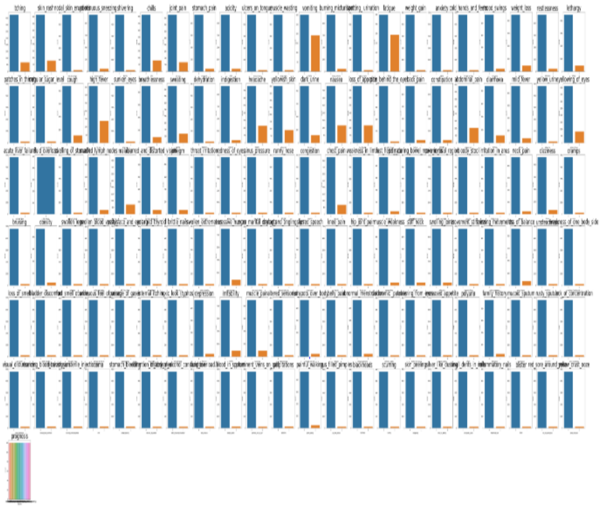


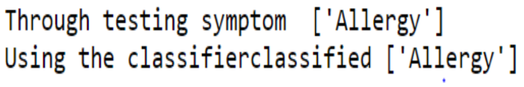
Fig.4 :Disease distribution

Now we try all the algorithm to train our model and we test the model using the test set. Now we encounter the most difficult problem of overfitting of the data set. In every algorithm we got 100% efficiency which is not possible. We decided to go for changing the k fold value for each algorithm and after the hit and trial we got the algorithm which is decision tree which gives the efficiency of 87 % with a k fold value of 2.

IV: Disease prediction

Now our model was ready to classify disease using the symptoms which we have already separated from the text of the patient. Using the NumPy array we gave the model with the symptoms and then it classifies the most probable disease the patient may have. This will be used by the doctor to diagnosis further.

This provides all the predicted disease by the machine as an output through an interface to the doctor who further prioritizes among them using his own experience and expertise and thus provide a close enough diagnosis.



```
Through testing symptom  ['Allergy']
Using the classifierclassified ['Allergy']
```

V.RESULTS

The proposed work successfully suggests deployment of a assistive technology for the doctor to recognize disease and thus will prevent further implications. This will also reduce the tremendous burden on the medical institutions and will also result in proper utilization of medical resources. This in-turn will reduce the overall costs of the medical treatment.

REFERENCES

1. Liqiang Nie, Meng Wang, Luming Zhang, Shuicheng Yan, *Member, IEEE*,Bo Zhang, *Senior Member, IEEE*, Tat-Seng Chua, *Senior Member, IEEE.* "Disease Inference from Health-Related Questions via Sparse Deep Learning",2006

2. Ming Li and Zhi-Hua Zhou, *Senior Member, IEEE* Improve Computer-Aided Diagnosis With Machine Learning Techniques Using Undiagnosed Samples

3. Bernstein, S.L., Aronsky, D., Duseja, R., Epstein, S., Handel, D., Hwang, McCarthy, M., John McConnell, K., Pines, J.M., Rathlev, The effect of emergency department crowding on clinically oriented outcomes. Acad. Med. **16**(1), 1–10 (2009)

4. Lin Ni(B), Chenhao Lu, Niu Liu, and Jiamou Liu Department of Computer Science, The University of Auckland, Auckland, New Zealand
MANDY: Towards a Smart Primary Care Chatbot Application 1998

5. Edward Choi, Mohammad Taha Bahadori, Doctor AI: Predicting Clinical Events via Recurrent Neural Networks,(2004)

6. Amiya Kumar Tripathy , Rebeck Carvalho, Ajit Puthenputhussery, Nikita Chhabhaiya, Bijoy Anthony
MediAssistEdge-Simplifying diagnosis procedure & Improving patient doctor connectivity, Department of Computer Engineering, Don Bosco Institute of Technology, Kurla (W), Mumbai, India School of Computing and Security Sciences, Edith Cowan University, Perth, Australia

7. David Rosenthal and Rachael Sokolowski, Kurzweil Applied Intelligence, a division of Lernout and Hauspie
Voice enabled, structured medical reporting,(2010)

8. Hammond Gerido, MPH School of Information Florida State University Tallahasse, FL, USA Patient-Centered Strategies to Increase Participation in Cancer Clinical Trials Lynette ,(2010)

9. Qiuling Suo, Fenglong Ma, Ye Yuan, Mengdi Huai, Weida Zhong,and Jing Gao, Member,IEEE Aidong Zhang, Fellow,IEEE,Deep Patient Similarity Learning for Personalized Healthcare (2014)

10. Di Somma, S., Paladino, L., Vaughan, L., Lalle, I., Magrini, L., Magnanti, M,Overcrowding in emergency department: an international issue. Intern. Emerg.
Med. **10**(2), 171–175 (2015)

Pranav Kulkarni is a student at SRM institute of science and technology. He is currently pursuing his 3rd year in the institute and is working on various domains like machine learning, Mobile Application development and Web-development.

Shubham Gupta is a student of SRM institute of science and technology. He is currently pursuing his 3rd year in the institute and is working on various domains like machine learning, artificial intelligence etc.

Animesh Singh is a student at SRM institute of science and technology. He is currently pursuing his 3rd year in the institute and is working on various domains like machine learning, Mobile Application development and Web-development.