

# Disease Diagnosis based on Symptoms

Imankalyan Sarkar  
imankalyan20010@iiitd.ac.in

Pallab Chakraborty  
pallab20063@iiitd.ac.in

Sakshi Kaushik  
sakshi20094@iiitd.ac.in

Abhinav Saurabh  
abhinav20127@iiitd.ac.in

## 1. PROBLEM FORMULATION

The first diagnosis to any sort of ailment is usually performed by a physician. The patient gets recommended some tests or drugs and medication as a cure. The diagnosis is performed based on symptoms that are apparent to the patient. This process is often complex since the doctor has to narrow down the list of diseases the patient might be suffering from. This matching of symptoms to disease is a highly confusing since often diseases have common symptoms. Some cases where patients might have shown early signs are often missed by doctors as some symptoms were more prevalent in the patient. We seek to aid in this process through the use of machine learning, utilising the data available for diseases and it's respective symptoms with respect to diseases which are prevalent in India. This sort of disease detection would serve as an aid in the diagnosis flow for the Doctor. In no way is this system acting as a Doctor but with more medical diagnosis data the system would get better and serve as a integrated healthcare solution embedded into the methods of diagnosis as we know today for doctors.

## 2. MOTIVATION

As a first response to any sort of ailment we approach a physician who can diagnose the problem we're having. The doctor asks for symptoms and tries his best knowledge to narrow down to some diseases we might be suffering from. This process is not simple since often diseases have common symptoms. Also doctors focus on the symptoms which are more apparent to the patient since those are the symptoms causing the problem. The availability of information at one's fingertips also enables a patient to search for symptoms. Some popular domains like WebMD offer symptom checker [4]. It takes symptoms from the user and tries to match the most likely disease with it. The more distinct symptoms weigh in the most and gives a confidence to the prediction of what the user might be suffering from. Such systems are not comparable to the diagnosis of a doctor but might provide some insight.

Therefore, detection of disease is a difficult task which is very much essential for any patient who is suffering from an unknown ailment. Early detection of such diseases would result in better treatment. Additionally, human error can creep in while diagnosing and a second opinion isn't always an option due to financial limitations. A system for diagnosing diseases which also helps the patients learn about

medical terms which helps them describe the symptoms will benefit the doctors as well.

There exist sophisticated systems which can detect heart diseases, skin diseases much more accurately than cardiologists [1] and dermatologists. However, such expertise doesn't yet exist in the general physiology domain. Such a system would be hugely beneficial to doctors and patients alike.

## 3. DATASET

There are no central databases for diseases complete with symptoms and other information. Heart diseases, skin diseases have several datasets which are available but normal physiological diseases lists are not complete. Due to this limit we only target diseases which are obtained from National Health Portal of India[3].

Using data from National Health Portal gives a sense of completion since we cover the diseases which are prevalent in India. The symptoms data are scrapped either from the website itself or Wikipedia.

The data was scrapped from National Health Portal of India [3] and also from Neal Chamberlain, Ph.D., A. T. Still University of Health Sciences/Kirksville College of Osteopathic Medicine [2]. The data was obtained in the form of a dictionary with key being the disease in question and the value being the symptoms associated with the disease. Some diseases along with their symptoms are shown below :-

- **Acute encephalitis syndrome:** 'Headache, fever, confusion, stiff neck, vomiting'
- **Ascariasis:** 'Abdominal swelling, abdominal pain, diarrhea, shortness of breath'
- **Bronchitis:** 'Coughing up mucus, wheezing, shortness of breath, chest discomfort'
- **Dementia:** 'Decreased ability to think and remember, emotional problems, problems with language, decreased motivation'
- **Gangrene:** 'Change in skin color to red or black, numbness, pain, skin breakdown, coolness',

The number of diseases obtained in this fashion is 480. Next task is to structure the data such that it could be used by various algorithms to process diseases along with their symptoms. Some preprocessing needs to be done on the raw data before we can structure it according to our needs. The steps are:-

1. Lowercase all the text.
2. Remove Stop words.
3. Lemmatize the words.

So we make a dataset in the form of term incidence matrices:-  
The are several problems with using a dataset like this for

	A	B	C	D	E	F
1	label_dis	abdominal cramp	abdominal pain	abdominal paininjection form fever	abdominal swelling	abnormal bleeding
2	AIDS	0	0	0	0	0
3	Acne	0	0	0	0	0
4	Acquired Capillary Haemangioma of Eyelid	0	0	0	0	0
5	Acute flaccid myelitis	0	0	0	0	0
6	Acute hemorrhagic cystitis	0	0	0	0	0
7	Adult T	0	0	0	0	0
8	Airbag Eye Injury	0	0	0	0	0
9	Alcohol Abuse and Alcoholism	0	0	0	0	0
10	Alopecia (hair loss)	0	0	0	0	0
11	Alveolar hydatid	0	0	0	0	0
12	Alzheimer's Disease	0	0	0	0	0
13	Amatoxins Fugax	0	0	0	0	0
14	Amblyopia	0	0	0	0	0
15	Ameliasis	0	1	0	0	0

Figure 1: Dataset

our use case. **Often times all the symptoms are not clearly expressed in a patient. To handle such cases we have to modify our dataset to accommodate the diseases with a combination of the symptoms.** This is an extremely important step since machine learning models function well with large amounts of data and there could be different combination of symptoms for each disease. Once this is done our data is ready for use in various algorithms and techniques. After generating combination the dataset would look like this :

	A	B	C	D	E	F
1	label_dis	abdominal cramp	abdominal pain	abdominal paininjection form fever	abdominal swelling	abnormal bleeding
2	AIDS	0	0	0	0	0
3	Acne	0	0	0	0	0
4	Acne	0	0	0	0	0
5	Acne	0	0	0	0	0
6	Acne	0	0	0	0	0
7	Acne	0	0	0	0	0
8	Acne	0	0	0	0	0
9	Acne	0	0	0	0	0
10	Acne	0	0	0	0	0
11	Acne	0	0	0	0	0
12	Acne	0	0	0	0	0
13	Acne	0	0	0	0	0
14	Acne	0	0	0	0	0
15	Acne	0	0	0	0	0

Figure 2: Dataset after applying Symptom Combination

## 4. LITERATURE REVIEW

In this paper[10], they developed automated methods for acquisition and discovery of medical knowledge embedded in clinical narrative reports. The paper focuses on two types of entities, disease and symptom. Evaluation based on a random sample of disease-symptom associations indicates an overall recall of 90% and a precision of 92%. In the IEEE paper[8], the implementation of Personalized Medical assistant has been done that heavily relies on AI algorithms. The system can predict the diseases based on the symptoms and give the list of available treatments. It can also give the composition of the medicines and their prescribed uses

which helps them to take the correct treatment.

In the paper by Yi Zhang and Bing Liu [11], Traditional text classification is based on a particular class of topic like sports, politics or sciences. However, many real-world text classification problems need more refinement based on the semantic characteristics of the sentences. The conventional "bag of words" model is no longer enough. Here the author reports that sentence semantic and structure features are beneficial in the refined classification problems. Five varieties of semantic features are pulled out from the sentence, i.e. centre noun, center verb, adjective, modifiers of the center noun, and the center verb's modifiers. Center noun is the noun of noun phrase containing an infectious disease name while Center verb is the verb governing the center noun. Adjective word is used jointly with the center verb. Negative modifiers to centre noun are like "no", "zero". Negative modifiers to centre verb are "not" "never". The subjective mood in modifiers to centre verb are "could", "would". If any of the above features are found in the dependency tree, it is added to the semantic tree. The dataset has been created manually by extracting sentences from documents of ProMED-mail and labelling the sentences. The dataset contains 604 EDR and 1533 Non-EDR sentences. The experiment is carried out on six random run of algorithms, with 90% as training data and rest as test data. SVM and Naive Bayes on sentence features gave f1 score of 0.590 and 0.662, respectively. SVM and Naive Bayes on semantic features gave f1 score of 0.500 and 0.557, respectively. SVM and Naive Bayes on (semantic + sentence) features gave an f1 score of 0.659 and 0.702. Naive Bayes gave the best results when used, combined with semantic features and sentence features. Using semantic features and sentence features alone produce much lower f1 score. In conclusion, results show that sentence semantic and structure features are useful in getting better accuracies.

In the IEEE paper [9], 'Chatbot for Disease Prediction and Treatment Recommendation using Machine Learning', the user sends messages and as a response the system gives appropriate messages. For this to happen smoothly, the system is trained with some possible questions and predefined answers, that the user can ask. Text processing is done on the input using NLP. When a user asks a question/symptoms, it will undergo a series of text operations and converted to a vector. Then the vector will be given to the model which will produce the index of the answer which will be later mapped to find the disease. The system will predict the disease to the user and will also provide a link where the user can search about the treatment needed for the disease predicted.

In IRJET paper[7], 'Disease Prediction and Doctor recommendation system', the authors takes in the symptoms as input and predicts the disease as well as a near by physician based on rating. In this paper the authors gives a novel approach based on data mining techniques coupled with a naive bayes classifier for disease prediction. Data mining is done by using WEKA tool. The naive bayes classifier finds the similarities of the diseases with the input symptoms and gives accuracy above 80 percent. The recommendation systems finds a nearby doctor on the basis of a dataset of user reviews and location. CoreNLP is used to power the recommendation system.

In [6] by Dhiraj Dahiade et. al. discusses two approaches using KNN and CNN for prediction of the disease. Their

approach is limited with the size of the dataset which is obtained from UCI. The accuracy obtained is 84.5% which is not satisfactory for deployment in a sensitive field like medicine.

In [5] by J. Chen, K. Li, H. Rong, K. Bilal, N. Yang, and K. Li, it provides a method to diagnose disease and gives treatment recommendation. It will help medical treatments by diagnosing reports of the patients. The system proposed here first inspects the medical reports and obtain disease symptoms as clustering centres. Then it tries to find the strongly related links between diseases and its treatments by implementing it with the association analysis algorithm. From the survey of the above papers we infer that there is much more to be inferred in the field of general physiology. This area deserves more attention since a sophisticated AI physician could act as the go-to Doctor or a second opinion for a patient. It is also beneficial for doctors since an AI assistant with the knowledge of large medical disease database could assist the doctor.

## 5. METHODOLOGY

The scraped data has been reshaped into a symptom disease data base. Each of the rows in the datasets corresponds to a disease and the symptoms associated with it. The two datasets obtained after scraping the websites for symptoms were used further for the prediction of diseases. Different methods used have been discussed in detail in the following section.

## 6. RESULTS

Different techniques can be used to predict diseases based on the symptoms entered. We approach the prediction problem using 3 techniques:

- Machine Learning
- Cosine Similarity
- TF-IDF

### 6.1 Machine Learning

Machine Learning can be applied in this case since there can be multiple combination of symptoms. Only matching symptoms would not result in accurate prediction because of overlapping symptoms between diseases. Also human errors or inability to identify symptoms will always result in a subset of symptoms for a particular disease. With Machine Learning we can vectorize the symptoms and give a better matching for the disease.

All Machine Learning Models share a similar pipeline which is shown in the diagram alongside.

The following pipeline was applied to different Machine Learning algorithms and the results are tabulated :

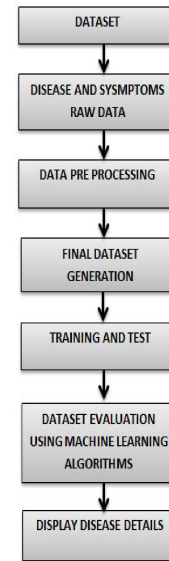


Figure 3: Workflow of the Machine Learning System for disease prediction

Algorithm	5 Fold CV
Multi Layer Perceptron	89.42
Decision Tree	72.62
<b>Random Forest</b>	<b>89.97</b>
Logistic Regression	89.88
<b>K nearest Neighbour</b>	<b>89.97</b>
Support Vector Machine	87.99
Multinomial Naive Bayes	81.30
Adaboost	84.69
Gradient Boosting Machine	80.94
Extreme Gradient Boosting Machine	78.83

Accuracy Table using Machine Learning

Almost all algorithms have provided a decent enough accuracy. This would not have been possible if the original dataset of diseases was only used. As we can see from the table we obtain the highest accuracy with KNN and Random Forest. Since this is a Disease detection we actually need to train the model with the entire dataset that has been generated using the combination of symptoms. The accuracy here obtained therefore doesn't actually indicate the effectiveness of the model that has been created using various algorithms. The only way to test these models would be to actually involve a human in the loop who can verify and test the accuracy of the model under real circumstances. However, we can conclude that Machine Learning provides the best results in this case.

### 6.2 Cosine Similarity

Cosine Similarity metric considers angle between two vectors to determine the similarity score. Higher the cosine similarity more chance is that the disease and queried symptoms have higher similarity. Then further we can sort the scores in decreasing order to get the First K disease.

$$\cos(A, B) = \frac{A \cdot B}{|A| * |B|}$$

Top 10 disease based on Cosine Similarity Matching :

0. Disease : Bang's disease	Score : 0.51	
1. Disease : Asthma	Score : 0.35	
2. Disease : Babesiosis	Score : 0.31	
3. Disease : Coronavirus Disease	Score : 0.3	
4. Disease : Nasal Polyps	Score : 0.3	
5. Disease : Middle East respiratory syndrome coronavirus (MERS-CoV)	Score : 0.27	
6. Disease : Flu	Score : 0.26	
7. Disease : Coronary Heart Disease	Score : 0.22	
8. Disease : Legionnaire's pneumonia	Score : 0.21	
9. Disease : Leukemia	Score : 0.2	

Figure 4: Top 10 results using Cosine Similarity

The results obtained through cosine similarity are not at par with the results obtained from Machine Learning models. So we further go for better machine learning models like logistic regression, Multi Layer Perceptron, Decision Tree, K nearest Neighbour, Random Forest. In the later part of the project we explore more options for disease prediction which have been explained in section 6.

### 6.3 TF-IDF

TF-IDF (term frequency-inverse document frequency) is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. This can be done by multiplying two metrics: how many times a word appears in a document, and the inverse document frequency of the word across a set of documents. The higher the score, the more relevant that word is in that particular document. The simplest way of calculating term frequency is a raw count of instances a word appears in a document. The inverse document frequency means, how common or rare a word is in the entire document set. The closer it is to 0, the more common a word is. This metric can be calculated by taking the total number of documents, dividing it by the number of documents that contain a word, and calculating the logarithm.

**TF-IDF = Term Frequency \* Inverse Doc Frequency**

$$tf - idf(t, d) = tf(t, d) * \log\left(\frac{N}{df + 1}\right)$$

Top 10 diseases predicted based on TF\_IDF Matching :

0. Disease : Coronavirus Disease	Score : 8.29	
1. Disease : Asthma	Score : 7.55	
2. Disease : Flu	Score : 5.76	
3. Disease : Nasal Polyps	Score : 5.61	
4. Disease : Babesiosis	Score : 5.25	
5. Disease : Leukemia	Score : 5.25	
6. Disease : Bang's disease	Score : 4.51	
7. Disease : Hantavirus Pulmonary Syndrome	Score : 4.29	
8. Disease : Legionnaire's pneumonia	Score : 4.29	
9. Disease : Middle East respiratory syndrome coronavirus (MERS-CoV)	Score : 4.29	

Figure 5: Top 10 results using tf-idf

## 7. USER INTERFACE

We have developed an user interface using cloud platform which could be accessed by public internet. This system will be able to input their symptoms and the system will match it against the known symptoms from the database. Synonym matching has been implemented in order to consider variations of a word implicating the same meaning. After that the system would find the co related symptoms and allow the user to chose if the correlated occur in them. After the user have input the symptoms the system would predict the top 10 disease based on the symptoms given as input. The users can also know more about the predicted disease if they want. The following steps will showcase how to use to system in order to get the predicted disease with confidence values.

- Open the website and enter the symptoms in the given field separated by comma.

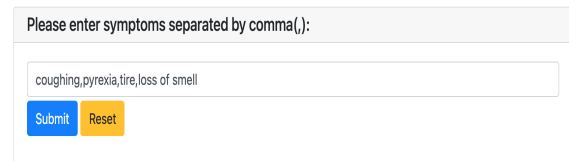


Figure 6: Enter Symptoms

- Then select the symptoms that are relevant to the symptoms entered before.

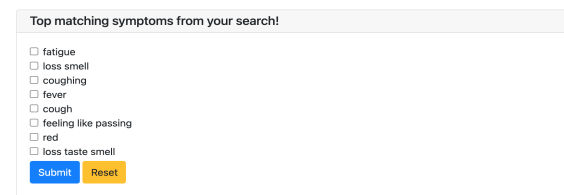


Figure 7: Select the symptoms

- Then select symptoms that is possibly relevant to you i.e co-occurring symptoms else select stop to skip the step.

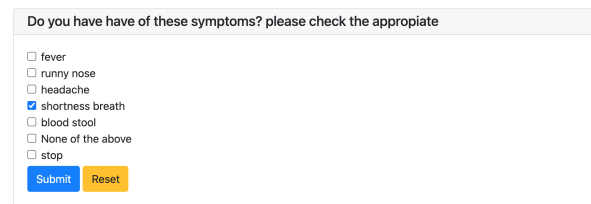


Figure 8: Select co-occurring symptoms

- Select the method to predict possible set of disease.

do you want tf-idf or cosine similarity to predict ?

☐ tf-idf
 ☐ cosine

Figure 9: Select method

- Then look at results, most probable disease is at the top.

Top 10 diseases predicted based on cosine similarity :

select the disease you want to know more of check none to end

☐ 0. Disease : Coronavirus Disease Score : 0.55  
☐ 1. Disease : Bang's disease Score : 0.43  
☐ 2. Disease : Asthma Score : 0.3  
☐ 3. Disease : Nasal Polyps Score : 0.25  
☐ 4. Disease : Babesiosis Score : 0.24  
☐ 5. Disease : Flu Score : 0.21  
☐ 6. Disease : Middle East respiratory syndrome coronavirus (MERS-CoV) Score : 0.2  
☐ 7. Disease : Coronary Heart Disease Score : 0.19  
☐ 8. Disease : Leukemia Score : 0.15  
☐ 9. Disease : Legionnaire's pneumonia Score : 0.15  
☐ none

Figure 10: Results

## 8. EVALUATION

The results of 5 fold cross validation have already been shown before. We observe that Random Forest and KNN gives the highest results. Logistic Regression and Multi Layer Perceptron are also giving comparable results. The dataset proposed is a scraped data with disease names from the National Health Portal and the symptoms are scraped from Wikipedia. For the use of the system in the real world the entire dataset should be used for training as it'll make sure we don't omit any diseases.

The user interface is able to take input the symptoms entered by the user and provide synonyms which will help the patients who're unaware of the names of the symptoms. The system is thus able to handle the synonyms it is unaware of by making the patient aware of what they mean. The system will handle new data even those that are unseen by the model because of the presence of the synonyms module.

## 9. CONTRIBUTION

We have created two new datasets and employed various machine learning methods to predict diseases on given input of symptoms. The first dataset is a raw dataset of the associated symptoms with a disease and the second dataset is a combinational dataset with a subset of symptoms for each disease as all symptoms might not be expressed for every patient naturally. We find some machine learning algorithms which are able to correctly predict the diseases and we also compare it with traditional information retrieval methods

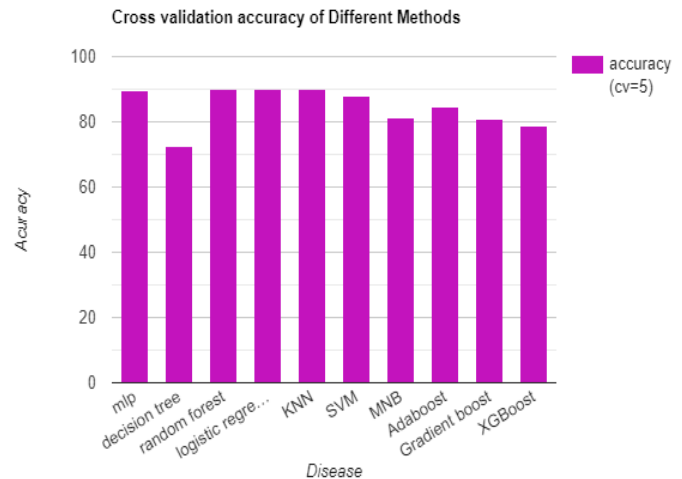


Figure 11: Results

like TF-IDF and cosine similarity. The machine learning model in this task. All of the members were actively involved in the project. All of the work was done in collaboration. The contribution of each of the members is as follows:-

- Imankalyan Sarkar: Responsible for implementing MLP, Decision Tree, Random Forest and Logistic Regression algorithms. Implemented the UI for interaction with the TF-IDF, cosine similarity, ML models to predict diseases.
- Pallab Chakraborty: Responsible for the data collection through the use of web scraping and generating the raw and combinational datasets used for training.
- Sakshi Kaushik: Responsible for implementing Machine learning models like SVM, Light GBM, XGB, Naive Bayes, ADABOOST algorithms.
- Abhinav Saurabh: Responsible for implementing the TFIDF and Cosine similarity disease matching.

## 10. ACKNOWLEDGEMENTS

We would like to thank Dr Rajiv Ratn Shah for providing us the opportunity to work on this project in our MTech Computer Science Course on Information Retrieval, 2021 and Anmol Singhal for providing us his valuable guidance, who is our mentor for this project.

## 11. CONCLUSION

The primary objective of the project is to predict the disease on the basis of the symptoms entered by the user. The designed system aims at bridging gap between Doctors and Patients. The project is designed in such a way that the system takes symptoms from the user as input and produces output i.e. predict disease. The system uses a GUI to provide support for disease prediction using different Machine Learning algorithms, tf-idf and cosine similarity.

## 12. REFERENCES

- [1] Harvard business review-ai can outperform doctors. so why don't patients trust it? <https://hbr.org/2019/10/ai-can-outperform-doctors-so-why-dont-patients-trust-it>: :text=MedicalAccessed: 2021-02-18.
- [2] Infectious disease names and their etiologies. <https://www.atsu.edu/faculty/chamberlain/Website/diseases.htm>. Accessed: 2021-02-18.
- [3] National health portal of india. <https://www.nhp.gov.in/disease-a-z>. Accessed: 2021-02-18.
- [4] Webmd symptom checker(identify possible conditions and treatment related to your symptoms.). <https://symptoms.webmd.com/>. Accessed: 2021-02-18.
- [5] J. Chen, K. Li, H. Rong, K. Bilal, N. Yang, and K. Li. A disease diagnosis and treatment recommendation system based on big data mining and cloud computing. *Information Sciences*, 435:124–149, 2018.
- [6] D. Dahiwade, G. Patle, and E. Meshram. Designing disease prediction model using machine learning approach. In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, pages 1211–1215, 2019.
- [7] D. Gujar, R. Biyani, T. Bramhane, S. Bhosale, and T. P. Vaidya. Disease prediction and doctor recommendation system. *International Research Journal of Engineering and Technology (IRJET)*, 5:3207–3209, 2018.
- [8] R. B. Mathew, S. Varghese, S. E. Joy, and S. S. Alex. Chatbot for disease prediction and treatment recommendation using machine learning. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 851–856. IEEE, 2019.
- [9] R. B. Mathew, S. Varghese, S. E. Joy, and S. S. Alex. Chatbot for disease prediction and treatment recommendation using machine learning. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 851–856, 2019.
- [10] X. Wang, A. Chused, N. Elhadad, C. Friedman, and M. Markatou. Automated knowledge acquisition from clinical narrative reports. In *AMIA Annual Symposium Proceedings*, volume 2008, page 783. American Medical Informatics Association, 2008.
- [11] Y. Zhang and B. Liu. Semantic text classification of disease reporting. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 747–748, 2007.