



# **Disease Diagnosis based on Symptoms**

---

---

**Imankalyan Sarkar  
Pallab Chakraborty  
Sakshi Kaushik  
Abhinav Saurabh**

# Problem Formulation and Motivation

- Disease diagnosis is the primary task performed by a physician.
- Matching of symptoms to disease is often a complicated task because several diseases share common symptoms
- If a particular disease show only shared symptoms it's very hard to diagnose.
- We seek to help the process by utilising the data available for the diseases and use different techniques such as machine learning.
- Doctors focus on the symptoms which are more apparent to the patient since those are the symptoms causing the problem



# Continued.

- A system like this is not acting as a doctor but rather an assistant to the doctor who can fine tune his diagnosis.
- This system can help a patient learn about more medical terms to communicate ailments.
- There are sophisticated AI systems for heart, skin diseases but general physiology has not been explored much.

Conditions that match your symptoms

UNDERSTANDING YOUR RESULTS ⓘ

Coronavirus

Moderate match

>

Chronic Sinusitis

Moderate match

>

Bacterial Pneumonia

Moderate match

>

Adenoidal Hypertrophy

Fair match

>

Influenza (Flu) Adults

Fair match

>

Gender Male Age 23 Edit

My Symptoms Edit

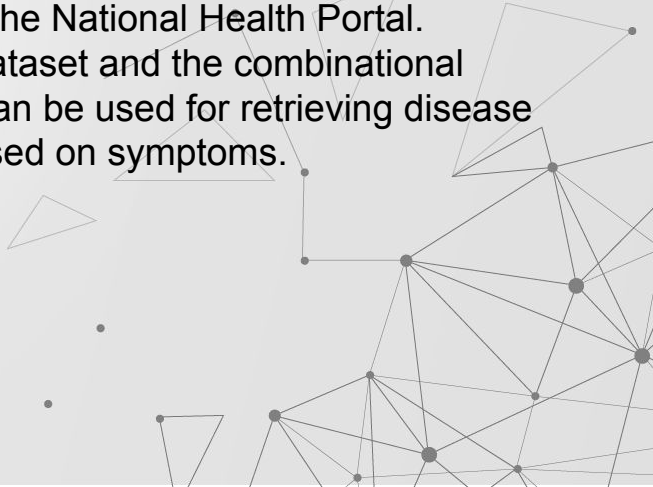
cough , can't taste food , fever below 100.4f

↻

Start Over

# Dataset

- **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',

- There are no central databases for disease and symptoms list.
  - We explore disease prevalent in the Indian subcontinent through National Health Portal.
  - The data is scraped and then preprocessed into a matrix format.
  - We propose 2 datasets based on our scraped data from the National Health Portal.
  - The raw dataset and the combinational datasets can be used for retrieving disease names based on symptoms.
- 

# Dataset Continued

- The dataset will not be useful for diagnosing a disease in the real world.
- Patients show only a subset of the symptoms of a particular disease therefore we need to perform data augmentation in the form of symptoms subset to consider cases like that.
- The combinational dataset will provide more data for machine learning algorithm.

	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	Amaurosis Fugax	0	0	0	0	0
14	Amblyopia	0	0	0	0	0
15	Amebiasis	0	1	0	0	0

	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

# RESULTS

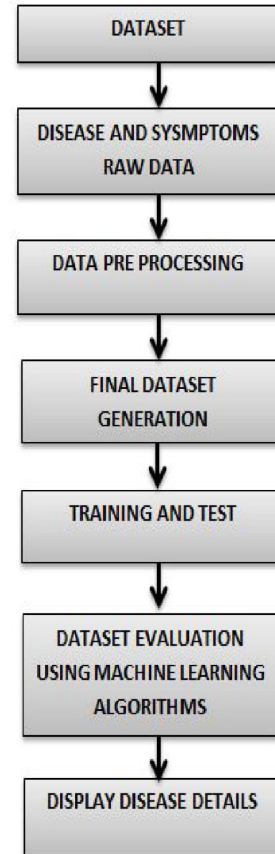
Different techniques can be used to predict diseases based on the symptoms entered the user. We approach the prediction problem using 3 techniques:

- Machine Learning
- Cosine Similarity
- TF-IDF



# Machine Learning Algorithms

- 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.



# Machine Learning Model Results

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



# 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.

$$\cos(A, B) = (A \cdot B) / (|A| * |B|)$$

Top 10 disease based on Cosine Similarity Matching :

0. Disease : Bang's disease	Score : 0.65
1. Disease : Lassa hemorrhagic fever	Score : 0.54
2. Disease : Flu	Score : 0.31
3. Disease : Middle East respiratory syndrome coronavirus (MERS-CoV)	Score : 0.31
4. Disease : Asthma	Score : 0.31
5. Disease : Cough	Score : 0.28
6. Disease : Nipah virus infection	Score : 0.27
7. Disease : Ebola Virus Disease (EVD)	Score : 0.26
8. Disease : Leukemia	Score : 0.26
9. Disease : Legionnaire's pneumonia	Score : 0.24

# 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.

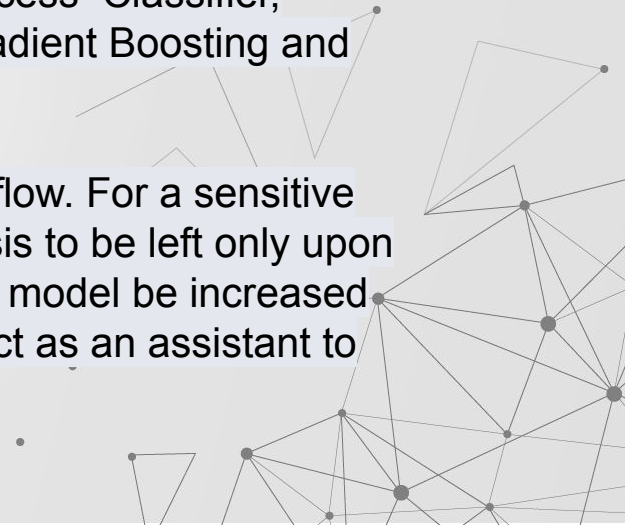
$$\text{tf-idf}(t, d) = \text{tf}(t, d) * \log(N / \text{df} + 1)$$

Top 10 diseases predicted based on TF\_IDF Matching :

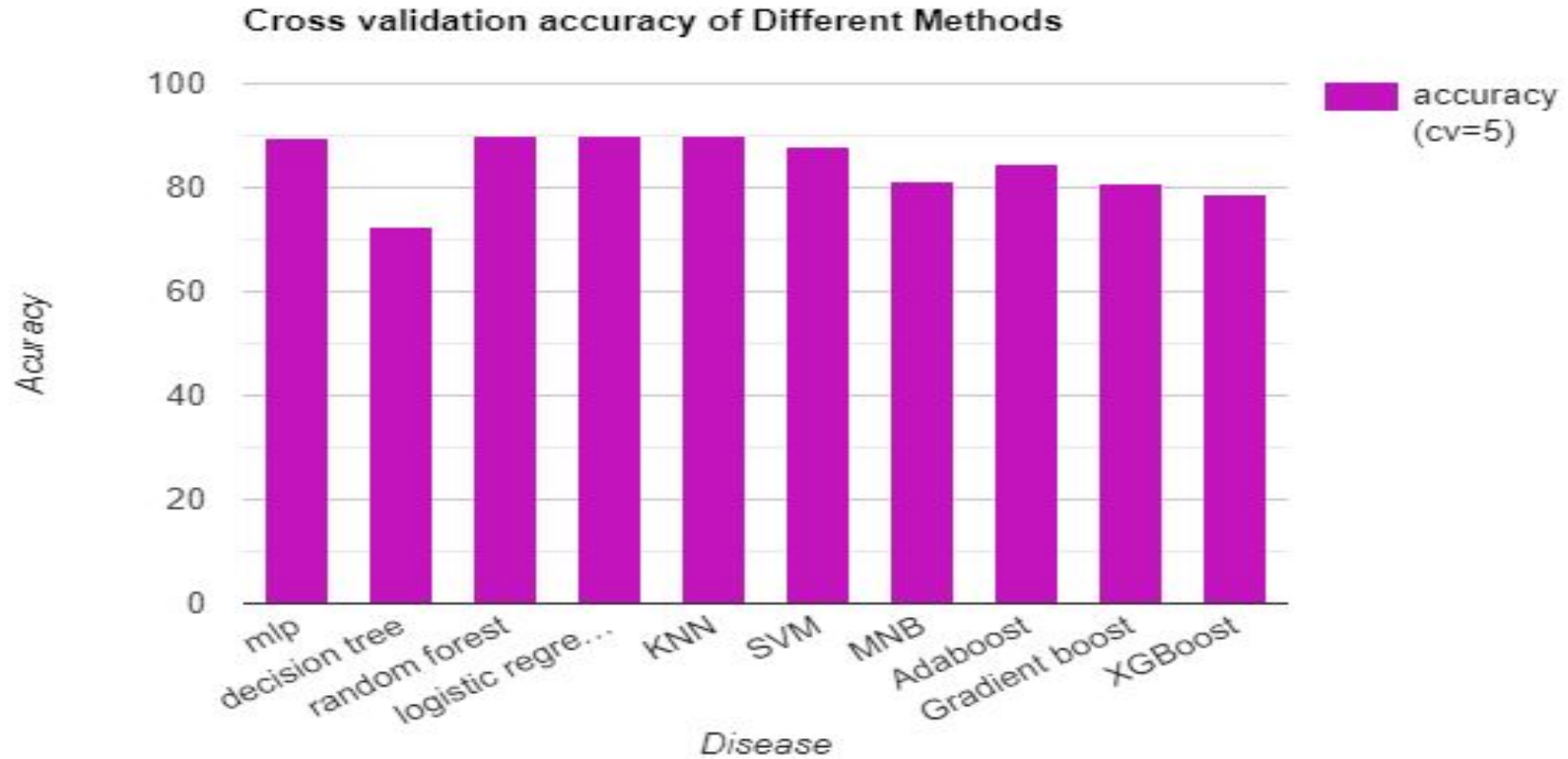
0.	Disease : Asthma	Score : 4.31	
1.	Disease : Bang's disease	Score : 4.31	
2.	Disease : Flu	Score : 4.31	
3.	Disease : Disseminated Intravascular Coagulation		Score : 4.19
4.	Disease : Ebola Virus Disease (EVD)	Score : 4.19	
5.	Disease : Lassa hemorrhagic fever	Score : 4.19	
6.	Disease : Leukemia	Score : 4.19	
7.	Disease : Asbestos-related diseases	Score : 3.07	
8.	Disease : Bronchiolitis	Score : 3.07	
9.	Disease : Common Cold	Score : 3.07	

## PROPOSED METHOD and FUTURE WORK

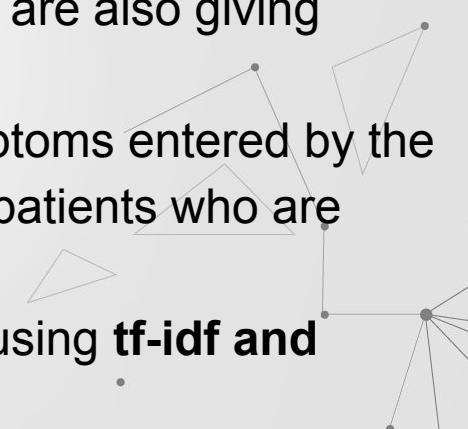
- Most of the work till now has been on sourcing data. Medical Data is very hard to get and the aim of the project is to predict diseases. For bringing a sense of completion in the domain we stick with the diseases list at National Health Portal and we scrape the symptoms related with it.
- In the future we will be fitting more Machine Learning Algorithms like Support Vector Machine, Multinomial Naive Bayes, Gaussian Process Classifier, AdaBoost, Light Gradient Boosting Machine, Extreme Gradient Boosting and will try to integrate some UI interaction with the System.
- This system can be used by a doctor in his diagnosis workflow. For a sensitive field like healthcare it is impossible for such critical diagnosis to be left only upon AI, so only after multiple revisions could the usability of the model be increased to doctors all over. For now a system like this could only act as an assistant to doctors and not actually guide treatment for patients



# EVALUATION



# EVALUATION

- The dataset proposed is a scraped data with disease names from the National Health Portal and the symptoms are scraped from Wikipedia
  - We observe that **Random Forest and KNN** gives the highest results i.e 89.97 CV accuracy.
  - Logistic Regression and Multi Layer Perceptron are also giving comparable results.
  - The user interface is able to take input the symptoms entered by the user and provide synonyms which will help the patients who are unaware of the names of the symptoms.
  - Top 10 predicted list of diseases are displayed using **tf-idf and cosine similarity**.
- 

# Algorithm Accuracy

Multi Layer Perceptron

93.70

Decision Tree

92.42

Random Forest

92.04

Logistic Regression

91.85

K nearest Neighbour

89.96



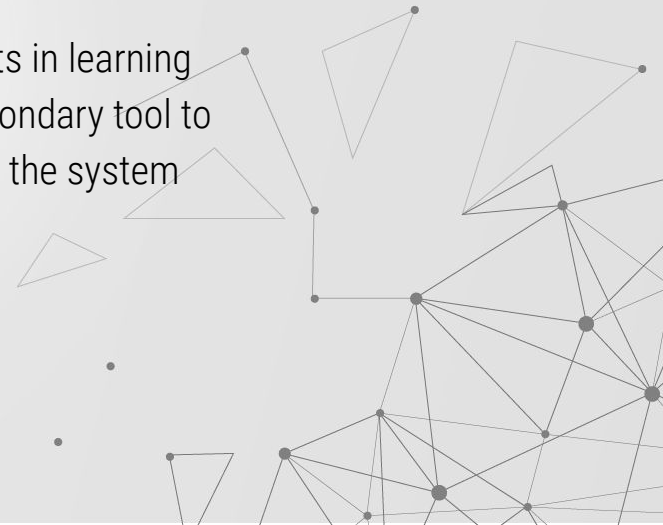
# Limitations and Challenges

- Unavailability of a central database
- There were many diseases which had nearly same symptoms.
- Unavailability of baselines , as there is no such central database for general diseases.



# Contribution

- We have created two new datasets and employed various machine learning methods to predict diseases on given input of symptoms.
- We find some machine learning algorithms which are able to correctly predict the diseases and we also compare it with traditional information retrieval methods like TF-IDF and cosine similarity.
- A fully functional disease prediction system will help both patients in learning more about their symptoms. For a doctor it would serve as a secondary tool to assess the prognosis. Human in the loop feedback would benefit the system for disease prediction.





# Future Work

- The system could be integrated onto a chatbot or a healthcare application in order to assess the situation.
- The models could use be used in an patient-doctor session where the model will be able to pickup the symptoms as spoken by the patient.
- The integrated healthcare application would need to take care a lot of aspects of medicine which is extremely important because the medical field is a sensitive area to work with on technology.
- Privacy, efficacy, inclusiveness, Legality, Accountability, Transparency, Fairness e.t.c factors would need to be considered before the application could be rendered to the users.



# THANK YOU

