Medical Information Retrival System

Shreeram Kumar Singh
M.Tech(CSE)
IIIT Delhi
MT23091
shreeram23091@iiitd.ac.in

Tanmay Parashar
M.Tech(CSE)
IIIT Delhi
MT23100
tanmay23100@iiitd.ac.in

Nitesh Kumar Chaurasia
M.Tech(CSE)
IIIT Delhi
MT23053
nitesh23053@iiitd.ac.in

Dheeraj Pandey
M.Tech(CSE)
IIIT Delhi
MT23034
dheeraj23034@iiitd.ac.in

Shubham Chaudhary
M.Tech(CSE)
IIIT Delhi
MT23093
shubham23093@iiitd.ac.in

Mohit Singh Tanwar
M.Tech(CSE)
IIIT Delhi
MT23127
mohit23127@iiitd.ac.in

1 Problem Statement

The study suggests a model where the user enters unstructured symptoms or chooses from the symptoms the system suggests, and the system returns a list of likely diseases to the user. Additionally, by selecting any of the output diseases, the user can learn more about the illness and their present medical condition, including its causes, associated symptoms, diagnosis, and potential treatments.

In addition, the algorithm recommends other symptoms based on the ones the user has entered. Even someone with little medical background may easily operate the device, which can be helpful for early illness diagnosis and detection. Additionally, it might help those who are initially hesitant to visit hospitals Additionally, it might help those who are initially hesitant to visit hospitals.

2 Motivation

Machine Learning applications in healthcare and biomedical domain has lead to early disease detection and better diagnosis. This has enhanced patient care in recent times. Studies have shown that people take the help of the internet for any possible health-related issues. The problem with this approach is that the search engines provide bulk information in scattered format from which it is difficult to conclude.
There are many disease prediction systems available such as heart disease prediction, neurological disorders prediction, and skin disease prediction. But universal prediction system for diseases based on symptoms is rarely in practice. It is very helpful for doctors or medical experts to diagnose diseases at an early stage based on symptoms. When a query is given, probable diseases are suggested to the user based on the highest probability and scores. With the use of the internet and all resources available to the user, proper diseases are used, and based on that proper medication is done which is very beneficial to all human beings. It is very helpful for doctors and patients to know better about the disease without any medical tests or anything else.

The detection of disease based on disease is a complex game. Being unfamiliar with biological terms, the users feed the symptoms in non-technical or natural terms which add complexity in predicting diseases. The main objective is to develop a novel architecture that could accept and handle such type of user queries by employing techniques like query expansion using a thesaurus, synonym matching, and symptom suggestion that will allow disease prediction with greater accuracy based on user input. We have scraped data from the web and generated dataset which can be used in future research. Query search retrieval and matching are used in such problems to achieve prediction.

3 Literature Review

The research presented introduces WebIRS, lays the foundation for subsequent advancements in medical information retrieval systems. Building upon the concepts of document classification and text summarization, an intelligent medical record retrieval system, extending the capabilities of existing frameworks to enhance symptom-based queries and facilitate more efficient access to patient records. In parallel, we explores innovative approaches to disease diagnosis recommendation systems, leveraging information retrieval techniques to match symptoms against disease datasets. This research complements the efforts of the second paper by focusing on early disease detection and underscores the importance of accurate symptom-based retrieval for timely diagnosis and treatment.Meanwhile, we focus towards empowering individuals in managing their health by developing a symptom tracking system. Inspired by user search behavior on the web, this system bridges the gap between symptom presentation and disease identification, contributing to the broader landscape of healthcare self-management tools. Concurrently, we predictive analytics for heart disease using Naïve Bayes modeling, highlighting the potential of machine learning algorithms to assist medical professionals in making informed decisions. This research builds upon the principles of data-driven healthcare to improve diagnostic accuracy and treatment outcomes. Finally proposes a comprehensive disease prediction system capable of simultaneously predicting multiple diseases, addressing the limitations of single-disease prediction models. By integrating various machine learning techniques and emphasizing early detection, this system aligns with the overarching goal of enhancing healthcare efficiency and patient outcomes.

Together, these interconnected research endeavors form a cohesive narrative, each contributing unique insights and innovations to the broader field of medical informatics and healthcare delivery.

4 Novelty

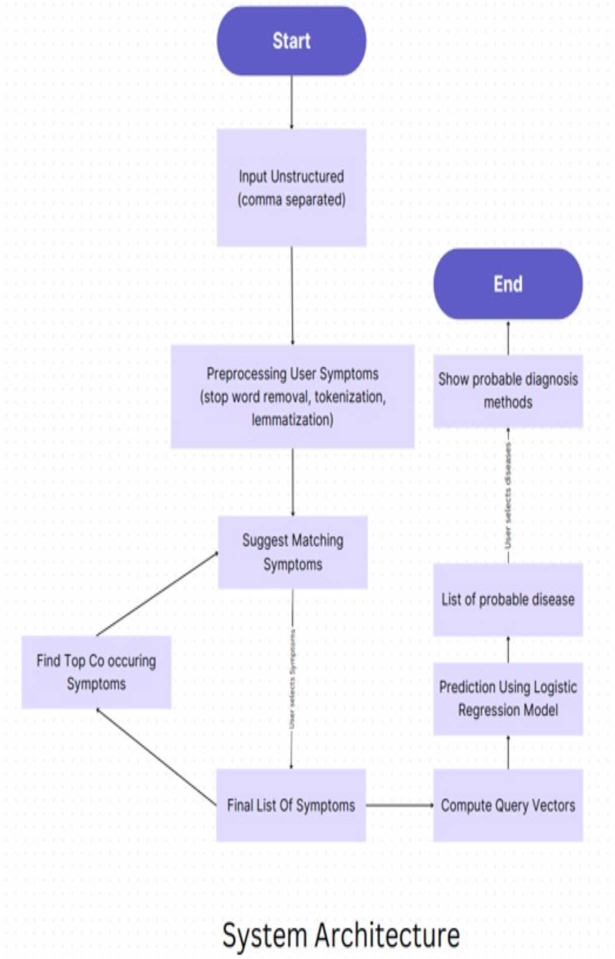
WebMD's Symptom Checker is an interactive application where the users have to select from a list of predefined symptoms. These symptoms are already fed to the data and thus require proper terminology when inputting the symptom, whereas our retrieval system will suggest more symptoms that co-occur along the symptoms provided by the user. Even if the user inputs a symptom that does not follow the proper medical terminology, our system will suggest symptoms that co-occur along with the user input symptom and suggest possible illness. This will help to suggest a more accurate illness as we will consider a wide range of possibilities. The workflow is similar for the Mayoclinic Symptom Checker, however Mayoclinic's system has a very little option for symptoms thus only works for only few symptoms which are already input there.

5 TECHNIQUES OR ALGORITHMS USED

1)ML models:
Support Vector Machine(SVM)
KNN and Decision Tree
Logistic Regression

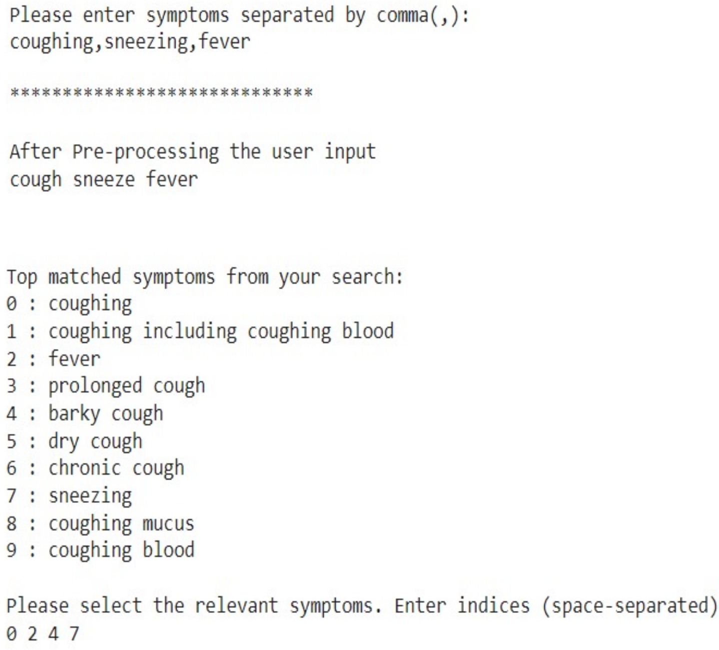
2)Algorithms:
Term Frequency-Inverse Document Frequency
Semantic Search

6 Methodology


Start
Input Unstructured (comma separated)
Preprocessing User Symptoms (stop word removal, tokenization, lemmatization)
Suggest Matching Symptoms
Find Top Co occuring Symptoms
User selects symptoms
Final List Of Symptoms
Compute Query Vectors
Prediction Using Logistic Regression Model
List of probable disease
User selects diseases
Show probable diagnosis methods
End

System Architecture

1.Start: The process begins with user-inputted symptoms. These symptoms are typically provided as an unstructured list, separated by commas.
2. Preprocessing: The symptoms undergo several preprocessing steps: Stop Word Removal: Common words (such as "and," "the," etc.) are removed to focus on relevant terms.
Tokenization: The symptom list is split into individual tokens (words or phrases).
Lemmatization: Words are reduced to their base form (e.g., "running" becomes "run").
3. Suggested Symptoms: The preprocessed symptoms are used to suggest matching symptoms. Additionally, the system identifies top co-occurring symptoms. These steps help create a refined list of symptoms for further analysis.



```
Please enter symptoms separated by comma(,):
coughing,sneezing,fever

*****************************

After Pre-processing the user input
cough sneeze fever


Top matched symptoms from your search:
0 : coughing
1 : coughing including coughing blood
2 : fever
3 : prolonged cough
4 : barky cough
5 : dry cough
6 : chronic cough
7 : sneezing
8 : coughing mucus
9 : coughing blood

Please select the relevant symptoms. Enter indices (space-separated)
0 2 4 7
```

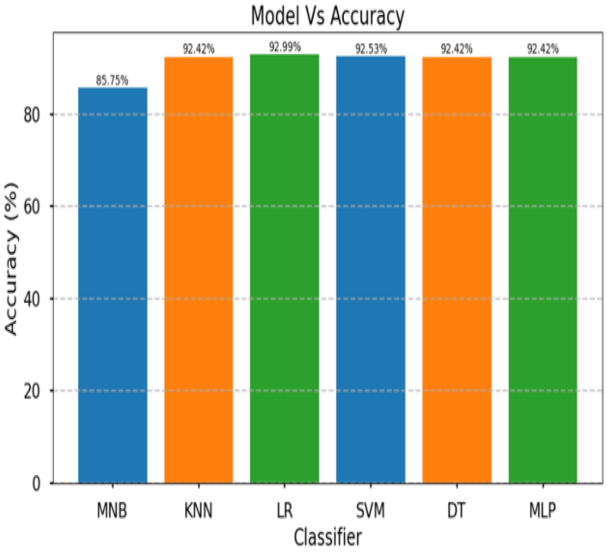
a. Synonym Expansion: User symptoms are expanded with synonyms.
b. Similarity Check: Symptoms in the dataset are compared with expanded user symptoms.
c. Selection Prompt: Top matched symptoms are displayed for user selection.
d. Related Symptom Discovery: Additional symptoms are suggested based on disease co-occurrence.
e. Iterative Suggestion: Commonly co-occurring symptoms are presented iteratively for user selection.
4. Algorithms Used: Logistic Regression Model: The refined symptom list is fed into a logistic regression model. The model predicts probable diseases based on the given symptoms. It computes query vectors to enhance accuracy. Probable Diagnosis Methods: The system provides probable diagnosis methods based on the predicted diseases. These methods could include further tests, consultations, or specific treatments. End: The process concludes with the suggested diagnosis methods.

7 Database

a. For diseases: We scrapped the data from the website of NHS UK to find about different disease that are recognised. Using the values in the HTML tag of the specified website we retrieved a list of diseases. Website link: `https://www.nhs.uk/conditions/` This Python script scrapes a list of medical conditions from the website using the requests and Beautiful Soup libraries and extracts the text of each condition and stores it in a list.

b. For Symptoms: Using the Google package in Python module, the script searched for diseases and fetched their Wikipedia page from the search result obtained. Further the HTML code of the fetched page is processed to retrieve symptoms of the diseases which are available in the ”class”:”infobox” of the Wikipedia page. Attached is a figure demonstrating from where the symptoms have been retrieved. This script iterates through a list of medical conditions extracted from the NHS website. For each condition, it performs a Google search appending ”Wikipedia” to find relevant Wikipedia pages. If a Wikipedia page is found, it retrieves the page content and extracts symptom information from the infobox. The symptoms are then preprocessed and stored in a dictionary with the medical condition as the key. After extracting the diseases and the symptoms, a dictionary has been created mapping diseases with their symptoms. The columns in the dictionary contain the names of diseases and the rows contain the list of symptoms. All symptoms have been specified in the rows treating them as a label to further map whether the symptom persists in any specified disease or not.

8 Code and Evaluation



This project uses novel techniques of Machine learning and IR techniques to detect diseases based on symptoms and provide more details about the top fetched diseases including treatment recommendation. The model which performed best was LR(logistic regression) SVM (Support Vector Machine) with an accuracy of 92.99

References

1. Eric W.K See-To, Y. K Tse, S.L.Ting , “Web Information Retrieval for Health Professionals”10 April 2013, Volume 37, article number 9946, (2013,DOI: `https://doi.org/10.1007/s10916-013-9946-3`

2. Hemant Jain, Cheng Thao & Huimin Zhao” Enhancing electronic medical record retrieval through semantic query expansion” 22 June 2010, Volume 10, pages 165–181, (2012), DOI: `https://doi.org/10.1007/s10257-010-0133-5`

3. Aszani, Hayyu Ilham Wicaksono, Uffi Nadzima, Lukman Heryawan” Information Retrieval for Early Detection of Disease Using Semantic Similarity” IJCCS (Indonesian Journal of Computing and Cybernetics Systems), Vol.17, No.1, January 2023,pp.

45 54, DOI: https://doi.org/10.22146/ijccs.80077

4. Lun-Wei Ku, Wan-Lun Li, Ting-Chih Chang," Disease Detection and Symptom Tracking by Retrieving Information from the Web" LINK: https://cdn.aaai.org/ocs/5722/5722-24520-1-PB

5. Priyanga, Dr. Naveen," Web Analytics Support System for Prediction of Heart Disease Using Naïve Bayes Weighted Approach (NBwa)" DOI 10.1109/AMS.2017.12

6. Dr.R.Shanthakumari, Dr.C.Nalini, Mr.B.Govindaraj" Multi Disease Prediction System using Random Forest Algorithm in Healthcare System" 2022 International Mobile and Embedded Technology Conference (MECON) — 978-1-6654-2020-4/22/$31.00 ©2022 IEEE — DOI: 10.1109/MECON53876.2022.9752432