Disease prediction based on symptoms

Udit Kumar(MT21148)
IIIT DELHI
udit21148@iiitd.ac.in

Sayan Mitra(MT21142)
IIIT DELHI
sayan21142@iiitd.ac.in

Mahvash Fatima(MT21126)
IIIT DELHI
mahvash21126@iiitd.ac.in

Gaurav Agarwal(MT21121)
IIIT DELHI
gaurav21121@iiitd.ac.in

Shivnath Singh Gaur(MT21085)
IIIT DELHI
shivnath21085@iiitd.ac.in

Rajat Pal(MT21138)
IIIT DELHI
rajat21138@iiitd.ac.in

1 PROBLEM DEFINITION :

In today's time, everyone is striving to achieve the best in their careers and lead a successful life; however, minor health issues are often ignored in this race. One recent development is people's mentality due to the ongoing pandemic is taking even small health issues seriously but the lack of time and financial conditions remain an issue, which often leads to delay in diagnosis and thus, worsening peoples' health by the time they are attended by a doctor. Meenakshi Mission Hospital and Research Centre (MMHRC) says that more than 50% of cancer deaths are due to late diagnosis. In the case of Covid, newspaper headlines said 'Late diagnosis, delay in reporting to hospitals caused 75% of fatalities' which clearly shows the importance of early diagnosis. Thus, with this project, we highlight the importance of early disease detection by attempting to detect disease probability based on symptoms given by the user at any time which can be used as a reference by medical experts for further treatment. This has been made possible with several information retrieval techniques used on scraped data to further refine the input and then employing several machine learning algorithms to finally give the diagnosis.

Update : Considering the role of a person's medical history, we have decided to incorporate the user's medical report (optional) along with typed symptoms. And earlier papers focused on specific diseases but our approach is generalized, it focuses on multiple diseases simultaneously.

2 UPDATED LITERATURE REVIEW :

Apart from the 5 research papers that we reviewed in our first submission, we've worked on the coding part and read one new research paper : Intelligent Disease Pre Diagnosis Only Based on Symptom

[1] This paper emphasizes that the extent of a particular disease can vary for every individual and every disease. These three areas are- Heart disease, Coronavirus, and Diabetes. This group of researchers tried to increase the prediction accuracy of particular diseases by pre-processing the existing data by performing cleaning, binarizing, performing feature selection by identifying the strong and weak correlations of the attributes with target attribute and other attributes using heatmaps on each disease data and further applying several machine learning models (Logistic Regression, J48, KNN, SVM, ANN, RF, GB, ANFIS, GANFIS) in comparison to the proposed model where data are entered into an android app, the analysis is then performed in a real-time database using a pre-trained machine learning model which was trained on the same dataset and deployed in firebase. Finally, the disease detection result is shown in the android app. They got F-measure improved by 1.4765 and 1.2782, respectively, for the COVID-19 dataset, by 1.8274 and 1.7264, for the diabetes dataset, and 1.7362 and 1.3821, for the heart disease dataset on the proposed methodology which was evaluated using the F-measure analysis. The proposed method performed better on all evaluation metrics with a low percentage of uncertainty compared to other models. Tools used: Python 3, Java, Android Studio. Datasets:(acquired by corresponding authors on request)

[2] This paper states that the effectiveness of similarity analysis between the user's symptoms contributes to a better prediction of disease.RNN model was applied, but the con of using RNN was its gradient vanished for long text data and couldn't find a strong correlation between words of symptoms.LSTM was used for the above problem, and it showed an outstanding sentiment classification and semantic relatedness prediction. The drawback of LSTM: Although the prediction is good, it takes a lot of time to train LSTM. A better solution was using CNN for better symptom similarity analysis, which takes less time than LSTM. Word2Vec is used to map a symptom to a specific vector, and The Manhattan distance calculates the similarity score.CNN showed a 4

[3] This paper proposed a model in which they have extracted the symptoms from the MIMIC-3 dataset and then used TF-IDF(this model captures the association between the disease and the symptoms) and word2vec(this model captures the association between two symptoms) model to convert symptoms to vector form and used BiLSTM model for prediction. They used 3 different models(TFIDF+BiLSTM, word2vec+BiLSTM, TF-IDF+word2vec+BiLSTM) and used 4 different evolution measuring techniques(Precision, Recall, F1-score, AUC) and got highest score as 0.87 for AUC with the 3rd model.

[4] This paper had created their data set by fetching data from MedlinePlus(which consists of documents related to most diseases.) and Wikipedia. Data is stored in the MongoDB database The technology used for web development is Flask Framework. Documents scored with the help of TF-IDF scoring and cosine similarity between the query and the documents and Documents processed with the help of texting. They used Linear support vector machines, K-nearest neighbor, Bernoulli Naive Bayes, and MultiNomial Naive Bayes classifier and they suggest using some advanced NLP tools rather than using normal modals.

this paper elucidates the importance and methods of spelling correction and has mainly focused on the clinical records that include names of various disease and their symptom. The paper can be beneficial in symptom spelling refining using various IR techniques. Also, the process will lead to stability in medical data and results.

[6] The very two adept methods for the same are:- Metaphone phonetic algorithm: This algorithm is one of the methods for spell refining that sorts the words according to their orthographic and phonetic edit distances. The Metaphone algorithm maps the particular misspelling to a generated code; words with the same or similar code are returned as suggestions. Crowell et al. found this very useful and said that the accuracy achieved with this algorithm was 76.2 Rule-Based Named Entity Recognition: This is used to make the text normalized and detect the named entities in the documents and categorize them into predefined classes. Such methods have ultimately contributed to the spell correction.

[7] This paper.We have studied this paper and found it interesting and thought that it would help implement our project. This paper aims at identifying key trends in different types of supervised machine learning algorithms and their performance and usage for disease risk prediction. It mainly focuses on various kinds of Machine Learning Techniques and Deep Learning techniques (ANN) and elaborates their accuracy scores useful in this domain. Here basically, more than one supervised machine learning algorithm is used for single disease prediction. The supervised machine learning algorithms that were used are SVM, Decision Tree, Random Forest, Logistic Regression, and ANN. Two databases Scopus and PubMed were searched thoroughly for different search items and among them, 48 articles were selected for comparison among different supervised machine learning algorithms for instance SVM, KNN, Naive Bayes in which it was seen that for some articles SVM was showing the best results whereas for some KNN was good. Using this information we were able to analyze what nature of data fits what ML technique. In terms of frequency usage of models, SVM is ahead followed by Naive Bayes and others, but in terms of accuracy performance, random forest is ahead although it is not used that much. In terms of frequently modeled diseases, SVM showed superior accuracy most times for three diseases. For breast cancer, ANN showed superior accuracy at most times. This paper widely elaborates on the ML techniques and ANN available for training and testing purposes. Hence gathering all this knowledge will help us in choosing and analyzing the best model for our dataset.

[8] This paper focuses on two major methods for disease prediction using symptoms and those are support vector machines(S.V.M) and neural network technology. The essence of neural equations has been used and formulated in such a way that the relation between nth hidden and the output layer is the next training process. Before exploring both methods, this paper labels the symptom database in a very adept way such that the final disease code formed gives out the most accurate output.
For this labeling there are three tasks associated:-
a)Identification of the main disease category.
b)Identification of subclass disease type.
c)Identification of specific disease categories.
For the above purposes, the various methods used are:-
1)Leave one Leave out cross-validation:-
Here the neural the support vector machine (SVM) and network disease identification method and disease identification methods are compared. Here it was found that binary classification of the specific disease diagnosis was more accurate as compared to multi-classification modules
2)Diagnosis with weight samples:-
Here some weights are assigned artificially to some samples based on clinical records. Those weights were further used in loss function in machine learning procedures to increase accuracy.
3)Multiple disease diagnosis:-
This module has been designed because a person may have more than one disease and those diseases may refer to different types. Hence results are obtained by putting them into the classification module of disease successively. It was found out that in the identification of concurrence of multiple disease type the accuracy of machine learning algorithms are decreasing and therefore for a more similar number of diseases, let's say three, the accuracy will drop more. "So, in this, the essence of neural network and support vector machine has been used here. Also, each symptom has been used as a feature and the above three categories i.e Identification of the main disease category, Identification of subclass disease type, Identification of specific disease category are used as layers"

[9] The article by Rahul Maheshwari about Disease detection based on symptoms with treatment recommendation(with scraped dataset) uses interaction based disease prediction system.

3 BASELINE RESULTS

After making the data set we have used different machine learning classifiers to predict the results. And the results are measured in terms of accuracy. The reason to choose accuracy is that the data set has an almost equal number of labels for each class. So accuracy is the right measure for this classification. The accuracies of the used classifiers are given in the figure 1:

Accuracies are as follows:
1) Multinomial naive Bayes: 81.20% (naive Bayes are not performing well because here symptoms are correlated but naive Bayes assumed them to be independent).
2) Random Forest: 83.20% (In our case we have used 533 symptoms so in this case, random forest is not performing well because in case of multiple features decision tree and random forest does not perform well).
3) K Nearest Neighbour: 86.99% (KNN generally perform poorly when there are more no of classes, so in our case there are 264 diseases which makes it hard for KNN classifiers to classify).
4) Logistic Regression: 90.51% (here logistic regression is performing well because our dataset is simple (1 represents presence of symptoms and 0 represents absence of symptoms) and it classifies it well).
5) Support Vector Machine: 90.16% (svm is performing well because svm captures multidimensional data very well).
6) Decision Tree: 77.50% (decision tree is not performing well due to multidimensionality).

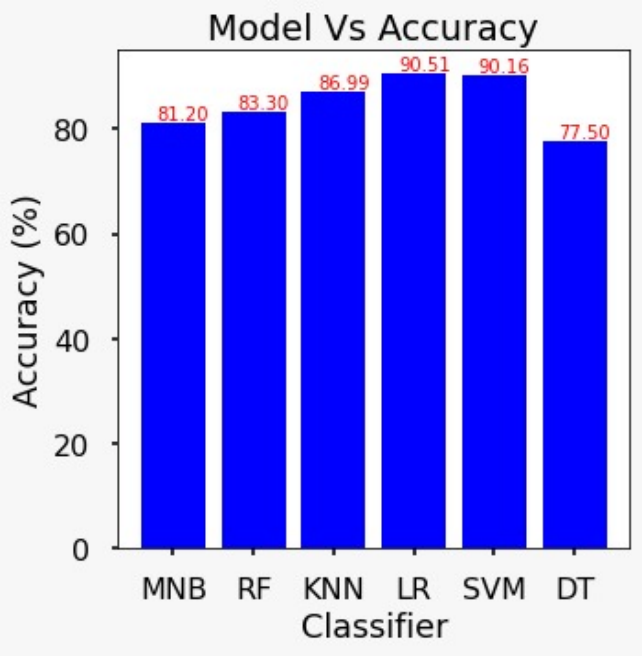

Figure 1: Model vs Accuracy

4 PROPOSED METHOD

There is a lot of work already done in the disease prediction field but no one focused on using the symptoms along with the patient's medical history. so we are planning to solve this problem using machine learning and deep learning techniques. Right now we have used machine learning techniques for prediction.
The whole project is divided into 2 sections:
1) Data scraping and dataset formation.
2) User interaction and prediction

- 1) Dataset scraping and dataset formation:
- We have used the Beautiful Soup library to scrap the disease from the two most reliable government websites. One is National Health Portal https://www.nhp.gov.in/disease-a-z, and the second is from the Centers for Disease Control and Prevention https://www.cdc.gov/diseasesconditions/az/.
- We have scrapped 303 diseases from the nhp portal and 932 from the cdc portal. After taking the union of both, we left with 1156 diseases.
- We have used the most reliable source of information (Wikipedia) for scraping the symptoms of the diseases. We made a dictionary of a particular disease as key and value as its corresponding symptoms. Due to the redundancy of diseases and unavailability of some symptoms on Wikipedia, we are left with 265 keys and values.
- We have made two CSV files, dataset1 and dataset2.
- In dataset 1, we have 264 rows and 553 columns. The first column represents the name of the disease, and the other remaining columns represent the total symptoms in the dataset. In this dataset, each row represents a unique disease. This dataset will help us in prediction.
- In dataset 2, we have the same 553 columns but 5682 rows. For a particular disease, we have tried the combinations of 2 and 3 pairs of symptoms at once so that it helps in more rigid prediction. The reason for making multiple rows for a single disease is that a single symptom can also point to a disease that can happen due to multiple symptoms. datascraping procedure is described in figure 2.

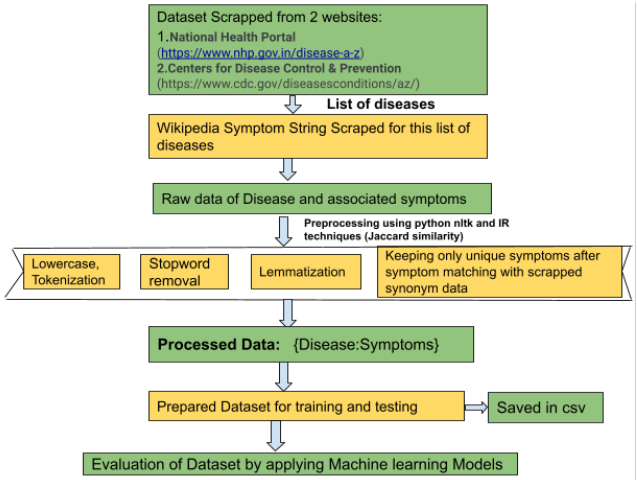

Figure 2: Data scraping and preparation flow

- Processing of scraped data:
  - Tokenization of retrieved data
  - Removal of stopwords
  - Replacement of Special characters.
  - Lemmatization of the prepared data
  - Removal of None values
  - Calculation of Jaccard similarity between each symptom and every other symptom synonyms.
  - Removal of too similar(greater than threshold score of Jaccard=0.75) symptoms to populate our final processed symptoms with unique symptoms
- User interaction and prediction The flow of the application is like that and it is shown in figure 3 and 4:
  - Taking input from the user (in the form of symptoms or medical history).
  - Processing the symptoms, adding synonyms of the symptoms, and asking the user to confirm the symptoms.
  - Processing is for user input data:
    - Tokenization,stemming,lemmatization
    - Stopwords removal
    - Adding synonyms for the input symptoms
    - Selecting symptoms of our dataset by matching with the user symptoms(using Jaccard coefficient score).
    - Adding some more occurring symptoms into the final selected symptom for prediction..
  - Then we have used 2 different approaches for prediction:
    - TF-IDF + Cosine similarity: TF-IDF score is a measure of the importance of the document.it is the product of TF(Term frequency of a symptom, which is 1 in our case,

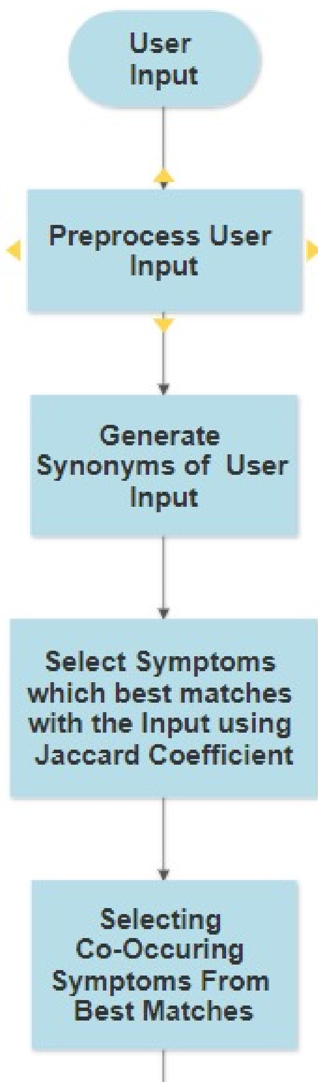

Figure 3: Workflow Diagram part 1

due to the uniqueness of the symptoms) and IDF(inverse document frequency- which shows if a symptom comes in many diseases then it is an irrelevant symptom).

∗ Machine Learning techniques: Machine learning techniques are very effective in prediction and classification tasks. So we have used different machine learning techniques like logistic regression, decision trees, naive bayes, support vector machine, random forest, and KNN.

5 REFERENCE :

[1] Naresh Kumar, Nripendra Narayan Das, Deepali Gupta, Kamali Gupta, and Jatin Bindra, "Efficient automated disease diagnosis using machine learning methods"

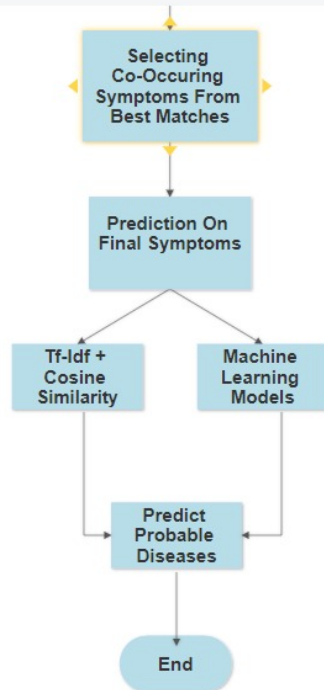

Figure 4: Workflow Diagram part 2

[2] Zhang, P., Huang, X., Li, M., “Disease Prediction and Early Intervention System Based on Symptom Similarity Analysis”.(2019)

[3] D.Guo, G. Duan, Y. Yu, Y. Li, F-X. Wu, M. Li,” A Disease Inference Method Based on Symptom Extraction and Bidirectional Long Short Term Memory networks, Methods (2019)”

[4] Gunjan Dhole1, Nilesh Uke2 “Nlp based retrieval of medical information for the diagnosis of human diseases” (2014) [5] Kenneth H. Lai a, Maxim Topaz a,b, Foster R. Goss c, Li Zhou “Automated misspelling detection and correction in clinical free-text records “(2015).

[6] Vu H. Nguyen, Hien T. Nguyen Vaclav Snasel “Text normalization for named entity recognition in Vietnamese tweets” (2016)

[7] Shahadat Uddin, Arif Khan, Md Ekramul Hossain Mohammad Ali Moni “Comparing different supervised machine learning algorithms for disease prediction”(2019).

[8] Fangfang Luo, and Xu Luo, “ Intelligent Disease Pre Diagnosis Only Based on Symptoms”(2021).

[9] Rahul Maheshwari. Disease detection based on symptoms with treatment recom- mendation(with scraped dataset). 2020.