AI HEALTHCARE CHATBOT WITH DISEASE PREDICTION USING MACHINE LEARNING

DONE BY

Students: Vishal Kumar, Shubham Sharma, Rahul Gupta and Sumeet Patiyal

Under Guidance of : Dr G.P.S Raghava

.

ABSTRACT

People are increasingly concerned about their health in the present international circumstances. Regrettably, today's medical human resource is inferior to that of the patient. The goal is to develop a medical chatbot that can converse with a patient if he is aware of his ailment, deliver relevant information, diagnose the disease using disease prediction, and provide fundamental details about the condition prior to approaching a doctor. A med-bot will minimize healthcare costs and increase access to medical understanding. Chatbots are coded systems that interact with users using natural speech. The answers are kept in the repository so that the bot can recognize the query phrase, make a query choice, and revert to the query. The n-gram, TFIDF, and cosine similarity are a method for classifying and calculating phrase alikeness. An outcome will be calculated for each phrase from the supplied input sentence, and additional comparable penalties will be found for the query. The chatbot has two different parts. One of them is a conversing agent that converses with the user. Another one is disease prediction, which takes input from the user and predicts what the disease is and what precautions can be taken. For the disease prediction part, we have used multiple machine learning classifiers such as logistic Regression, K-NN, SVM, RF, DT, and Convolutional Neural Network [1] on the dataset, and we got the highest accuracy of 91.44 using CNN classifier on the test dataset.

INTRODUCTION

Chatbots are artificial intelligence systems that process natural language and perform the role of artificial conversational agents, simulating human interactions as part of AI devices. As this cohort is still in its early period of evolution, healthcare chatbots are likely to be a growth driver. Better availability to healthcare, better doctor-patient and hospital–patient conveying, or aid in managing the increased request for fitness solutions such as faraway testing, medication obedience followup, and online consultations, these are some of the main purposes of healthcare chatbots [2]. India has an expanding population, rising birth rate, and dropping mortality rates due to medical advancements, while the number of physicians has fallen to match the population's growing need. When we visit cities with public hospitals, we can better understand this scenario, where a lack of physicians is a crucial cause of incorrect

patient treatment and, in some circumstances, death, even doctors can make mistakes that result in a patient's death when they fail to provide adequate care. To cope with such scenarios, a clever chatbot is required to advise physicians and even patients on what to do in such instances, eventually saving the lives of hundreds of people [3]. The scheme's main goal is to close the gap in reporting between patients and healthcare providers by providing elicit replies to queries faced by patients. People these days are more dependent on the internet, yet they are unconcerned about their health. People avoid going to the hospital for minor ailments that might develop into serious illnesses in the future. Creating a query and response forum is becoming a simple approach rather than sifting through a list of potentially relevant materials on the internet [4].

Creating a user interface that allows you to input data and receive replies was used to construct our chatbot application. It is an arrangement that connects with the user by keeping track of the present condition of the interaction and remembering old instructions to provide purpose. AI algorithms are used to create our chatbot that examines user queries, recognizes them, and responds. If a user wants to know about the specific disease he is facing by using the symptoms that he is facing, we can also predict the disease and get relevant information about it.

The study has been divided into four sub-parts

1) Literature Review
2) Proposed Methodology
3) Results and Analysis
4) Conclusion and Future Scope

LITERATURE REVIEW

The study is based on the acceptance of a chatbot in tuberculosis patients through quantitative research, the researchers have first interviewed the patients who have TB. Then they are using the Technology Acceptance Model for generating the results of acceptance of chatbot. The Korea Disease Agency donated information for the dataset. [5]

The study investigates the feasibility, usability, and effectiveness of a ML based bodily activity chatbot. Contributors wore a Fitbit Flex 1 and attached to the bot via Messenger. Fitbit pursuit data were synchronized from the Fitbit platform to the bot platform. The strategy is Dialog Flow , a modern ML platform for creating conversational AI applications. [6]

The paper focuses on the recent trends in Chabot technology. It gives us an overview of the brief history of chatbots in medical science. It uses various chatbots as research examples to find out whether humans are replaceable or not in this arena. Disease Covered: Cancer therapy, Dataset is available on Github. [7,8]

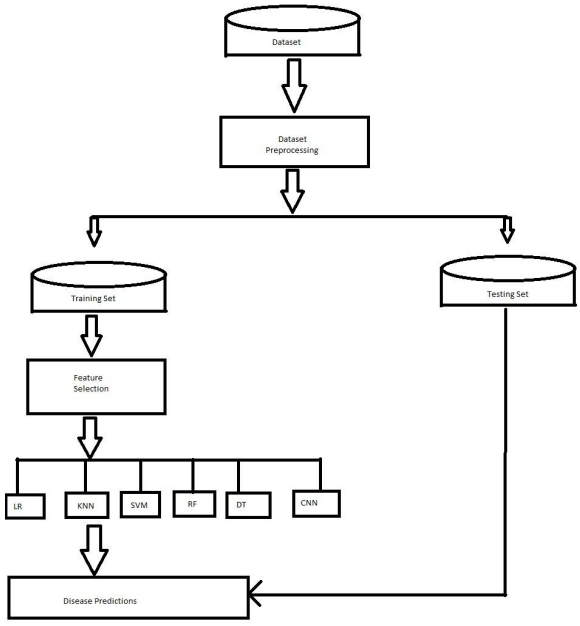
The research focuses on making an artificial intelligence chatbot for identifying social and behavioral changes regarding sexual and reproductive health in India. Sneha bot focuses to provide a secure place for Indian youth to have communication about SRH, debunk sex-related taboos and misconceptions, provide correct insight regarding secure sex and birth-control products, and address mental health problems. They are using NLP, an Azure LUIS app that gets user queries as a conversation and natural language text then handles them as user expressions, operates the data by getting keywords and forecasting user intents, and then works on the JSON data to make conclusions about fulfilling the user request. [9,10]

The paper tells about Safebot, a chatbot that allows people to communicate with it in normal language. Safebot relies on its own trusted users to discriminate against malicious users and then it uses the data obtained to control the action of malicious users. Their analysis shows the viability of Safebot. They plan to situate SafeBot as an augmentation of the intelligent home aide, and permit genuine cooperative learning. Since Safebot completely works on the natural language, It may be used as the brain of a toy like a talking robot. [11,12]

This paper focuses on the disease prediction part by using various machine learning algorithms and using the dataset available on kaggle [14]. The highest accuracy achieved after preprocessing on the dataset using CNN classifier is 84.5%. [13]

PROPOSED METHODOLOGY

We have done two different types of work, one of them is a healthcare chatbot in which we have merged many other datasets [15,16,17,18]. The generated dataset will have question and answer pairs. After that, we put the dataset to work using the chatbot. We used the Disease Symptom dataset in the second part of our work [14]. We preprocessed the dataset, split the dataset into an 80:20 ratio, and applied various machine learning algorithms.

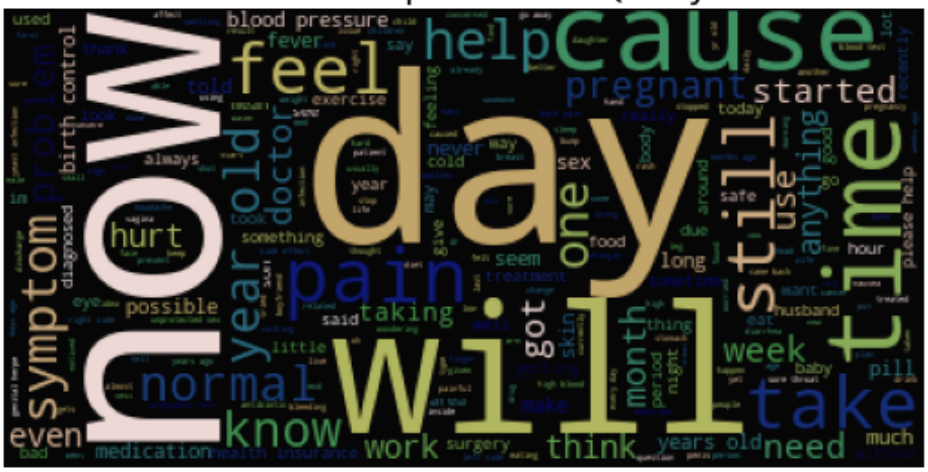


Fig(1): Disease Prediction Methodology

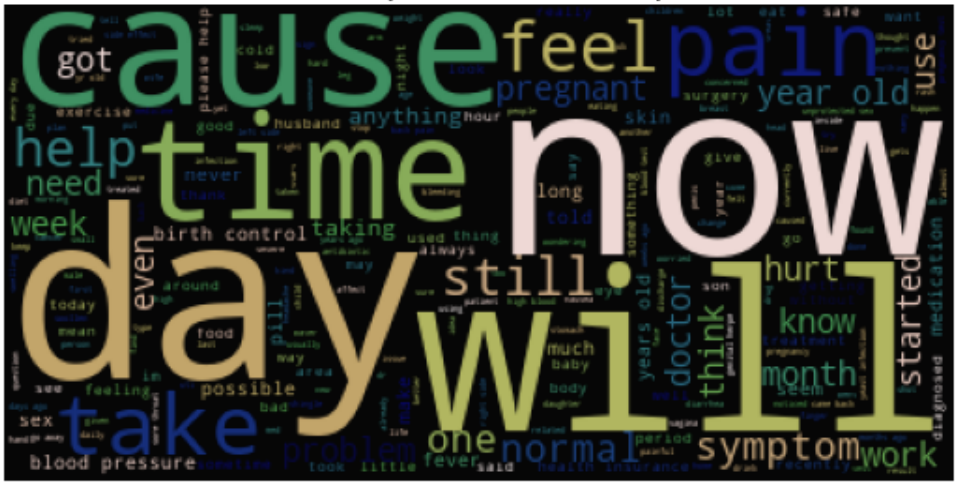
1) **Dataset Description**: The dataset used for the chatbot is a merged dataset that contains 11278 question-answer pairs [15,16,17,18] regarding covid, mental health, depression, and general doctor-patient quotes. The dataset used for disease prediction is the disease symptom dataset [14]. The dataset contains 4961 rows containing 129 features (Disease symptoms) and one target variable (Disease Name).

2) **Dataset Preprocessing**: On the chatbot dataset, Text Normalization is done after that important sentences are marked like top positive,negative and neutral sentences. Then sentence vectorizer is used and vectorized scores are saved in the dataset itself. Then cosine similarity is used upon the dataset.

Words frequented in Query

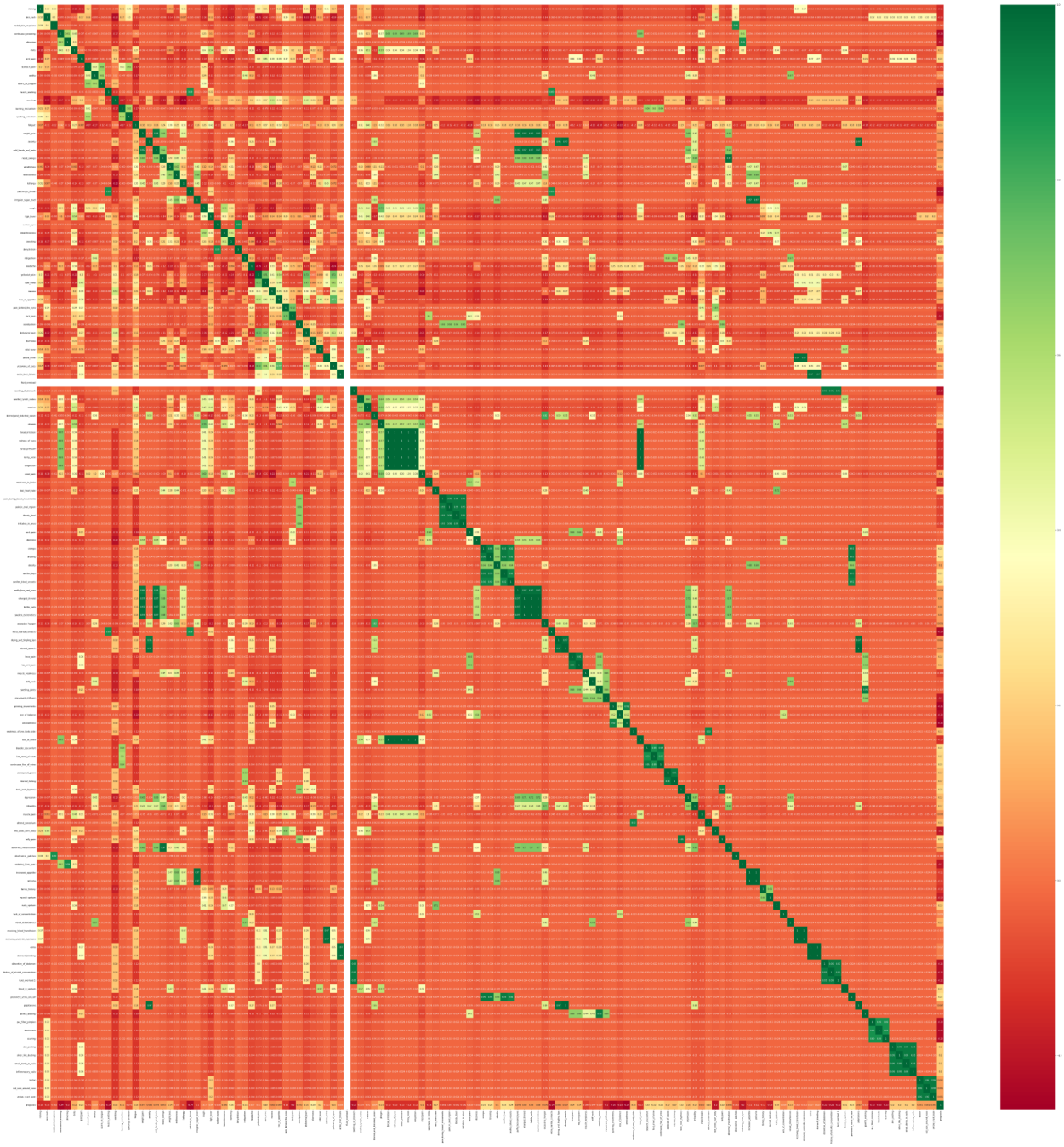


Words frequented in Response



Fig(2): Frequent words in query and response

On the other hand in the disease prediction dataset label encoder is used to encode the target variable “prognosis”. After that features are extracted based on the kendall correlation among the different variables.



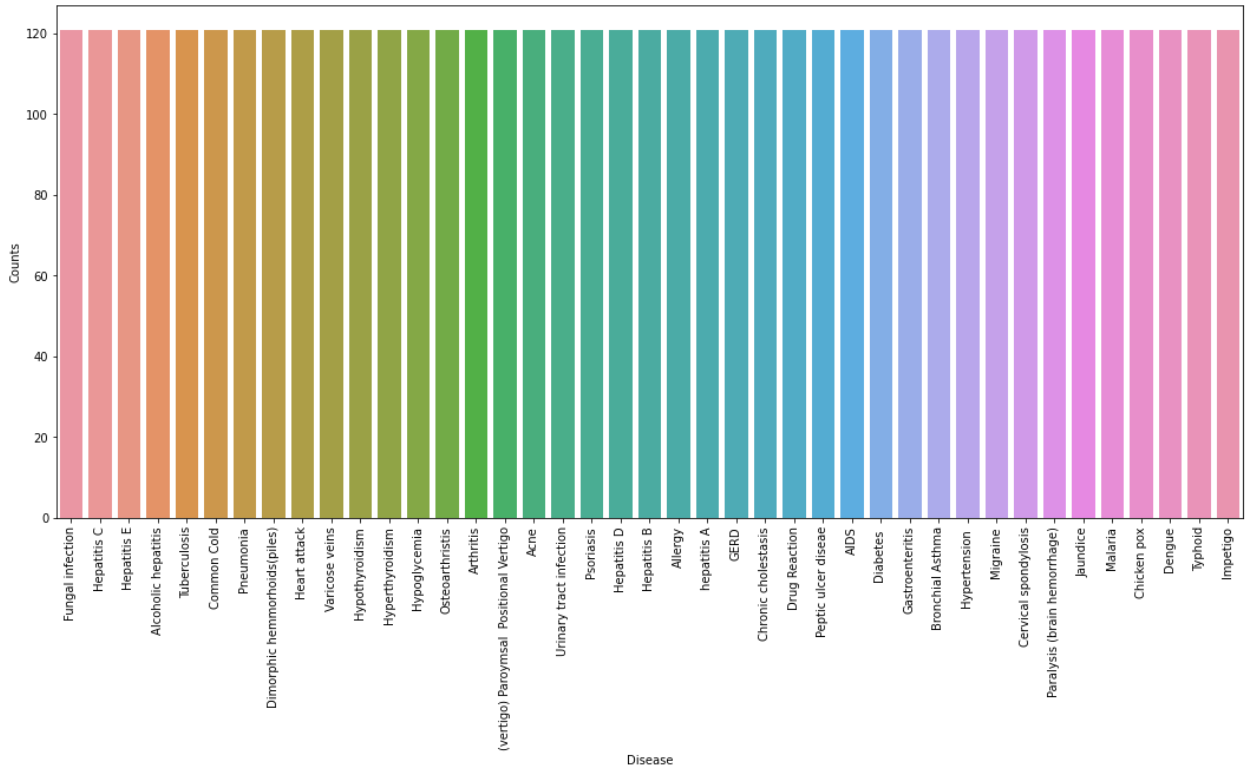
Fig(3): Correlation among all features in Disease prediction dataset

After the feature selection is done, out of the 129 features only 32 features are left.

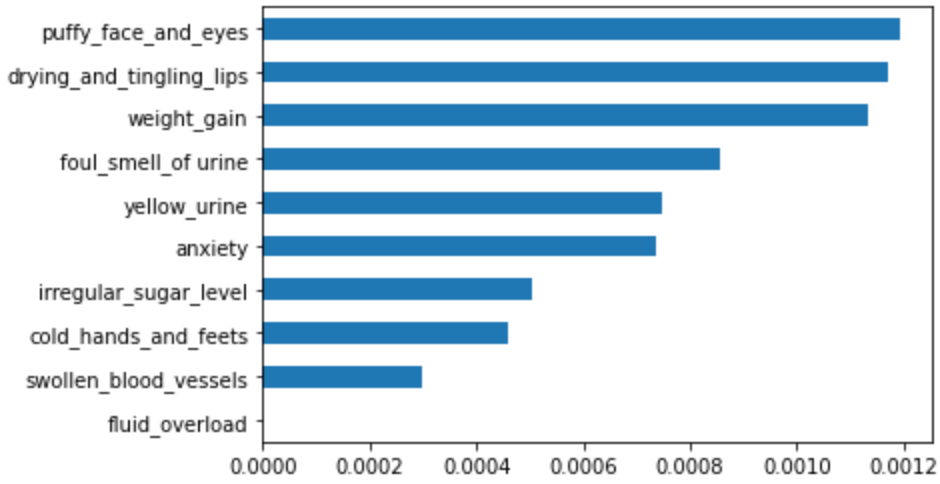
```
Index(['itching', 'skin_rash', 'nodal_skin_eruptions', 'continuous_sneezing',
       'chills', 'joint_pain', 'stomach_pain', 'acidity', 'muscle_wasting',
       'vomiting', 'burning_micturition', 'fatigue', 'weight_gain', 'anxiety',
       'weight_loss', 'cough', 'sunken_eyes', 'headache', 'yellowish_skin',
       'pain_behind_the_eyes', 'constipation', 'diarrhoea',
       'acute_liver_failure', 'fluid_overload', 'swelling_of_stomach',
       'cramps', 'spinning_movements', 'weakness_of_one_body_side',
       'family_history', 'pus_filled_pimples', 'skin_peeling', 'blister'],
      dtype='object') 32
```

Fig(4): The 32 features that are selected for disease prediction

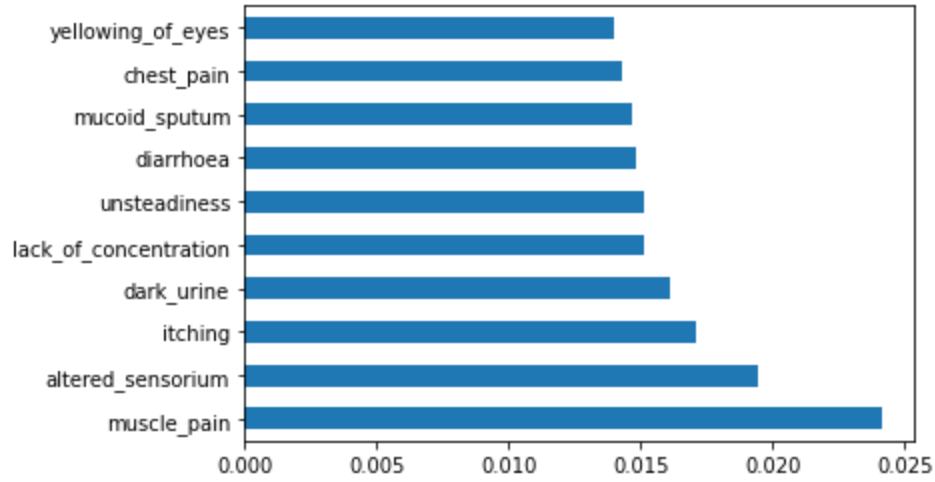
The disease dataset target variable (“prognosis”) is very equally balanced, i.e the count of rows for each target variable is equilateral.



Fig(5): Counts of rows vs Disease names in the target



Fig(6): Features with least importance to the target in disease Prediction



Fig(6): Features with Most importance to the target in disease Prediction

In the preprocessing part of the dataset, we worked on the **Robust scaler** (Use statistics that are resistant to outliers to scale features. The mean is separated, and the data is mounted according to the bivariate range (defaults: IQR). The IQR is the interval linking the first and third quartiles (twenty-fifth and third quantiles) (seventy-fifth quantile).) and **Quantile Transformer** (Quantile data is used to change attributes. The characteristics are transformed into a constant or standard dispensation utilizing this plan).

3) **Machine Learning Classifiers**:

As we get the 32 attributes and one target variable, we divide the data into X and Y, with X.Shape=(4961,32) and Y.Shape=(4961,). X contains 32 features, and Y has only the target variable.

Then to split the dataset into Train and Test, use the Train Test Split method with 80:20 split, such that the Training dataset contains 80% of the data and the testing dataset includes 20% of the data.

Then We apply different machine learning classifiers to the dataset, such as Random Forest, Decision Tree, K-Nearest Neighbor, Logistic Regression, SVM, and CNN [1]. The accuracy of the classifiers was calculated using the confusion matrix. A classifier that the highest accuracy bags can identify as the best classifier.

Logistic regression - A supervised machine learning technique used to address classification issues is logistic regression. True or false, Spam or not Spam, and other discrete values are predicted using classification issues. The sigmoid function is used to model the data in logistic regression.

Random Forest -The RF is created via supervised machine learning. Random forest is a user-friendly and adaptable algorithm. Random forest is a method that, in most cases, produces good results without the use of hyper-parameter tweaking. It creates a forest out of a collection of decision trees.

Decision tree - The DT is a supervised machine learning approach for segmenting data based on specific features. It splits the dataset into smaller and smaller subgroups as it develops the tree. There are two sorts of nodes in the tree: decision nodes and leaf nodes. In a DT, decision nodes indicate outcomes, whereas leaf nodes are where the data is split.

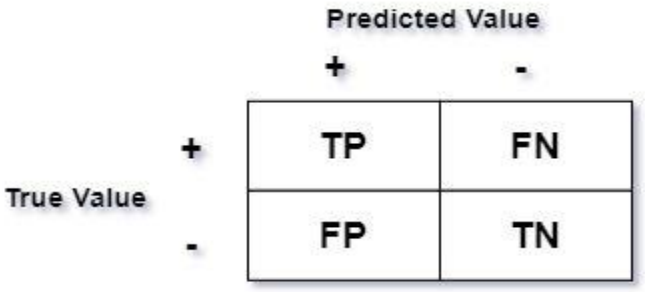
Support Vector Machine(SVM) - SVM, or supervised learning, is the most widely used machine learning method. It's used to solve categorization issues. This technique tries to build a decision boundary that divides n-dimensional space into different classes, allowing new data points to be quickly allocated to the correct category.

K-Nearest Neighbor(KNN) - K-NN is a supervised learning approach that may be used for both regression and classification, however classification is the most prevalent use. The K-NN approach assumes that latest and past data are comparable and allocates latest data to the category closest to the general categories. During the training phase, the K-NN approach saves the data, and when fresh data is received, it is classified into a class that is highly close to the most recent data.

Convolutional Neural network(CNN) - It's a hardware or software system based on the human brain and nervous system operating neurons. CNN is established on artificial neurons.The primary purpose of CNN was to solve problems in a similar manner to a human brain.

Evaluation Parameters - For the assessment, we utilized the confusion matrix, accuracy, area under the receiver operating characteristic (AUC-ROC), sensitivity, specificity, f1-score, Matthews correlation coefficient (MCC) score, and Cohen kappa score. The confusion matrix is a table-like structure with actual and projected values. The below we describe the confusion matrix and evaluation parameters formulas:-

Confusion Matrix -



$$\text{Accuracy} = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

$$\text{Specificity} = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

$$\text{Sensitivity} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

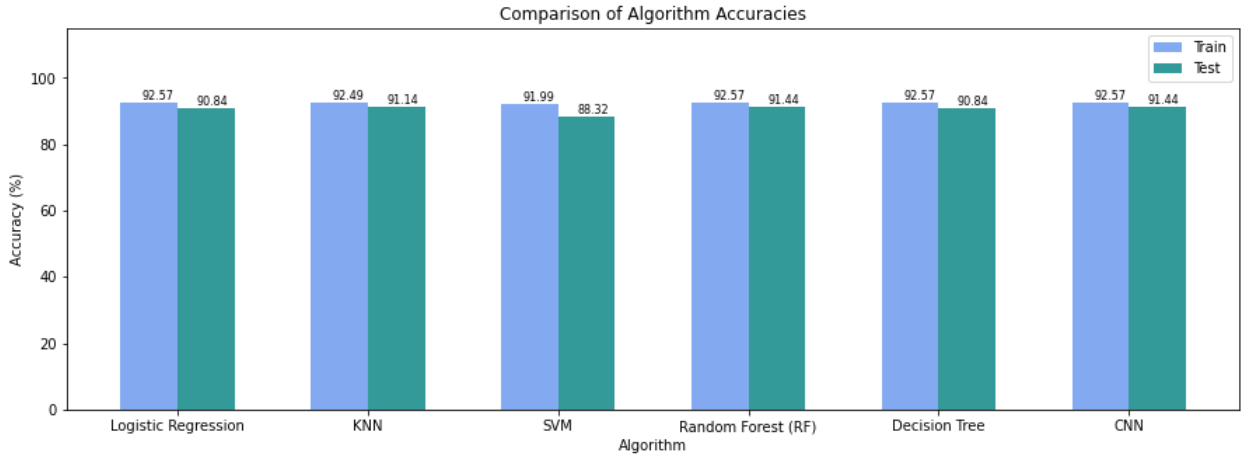
$$\text{F1-Score} = \frac{True\ Positive}{True\ Positive + 1/2(False\ Positive + False\ Negative)}$$

4) **Natural language processing Toolkit**: NLTK is a part of Python language tools for operating with normal speech data. It contains a bag of sentence and word filtering libraries for classification, encoding, markup, parsing, and semantic reasoning, as well as wrappers for the powerful NLP libraries. It is widely accepted due to its easy-to-use interfaces.[19]

RESULTS AND ANALYSIS

For our research, we divided our dataset 80:20 for training and testing. Our work achieves the best accuracy of 91.44% by using the convolutional neural network algorithm. We have used different ML classifiers for our work and the algorithms used are logistic regression, k-NN, SVM, DT, RF, and CNN [1]. By using Logistic Regression, we got an accuracy of 90.84%. When we used the K- Nearest neighbor algorithm, we got an accuracy of 91.14%. The accuracy of the SVM classifier is 88.32%. The accuracy of the use of the Decision tree algorithm is 90.84%. The accuracy of the use of Random Forest is 91.44%.

Classifier	Accuracy	Specificity	Sensitivity	F1-Score	MCC	Kappa
LR	90.84	90.84	90.84	90.84	0.91	0.91
KNN	91.14	91.14	91.14	91.14	0.91	0.91
SVM	88.32	88.32	88.32	88.32	0.88	0.88
RF	91.44	91.44	91.44	91.44	0.91	0.91
DT	90.84	90.84	90.84	90.84	0.91	0.91
CNN	91.44	91.44	91.44	91.44	0.91	0.91



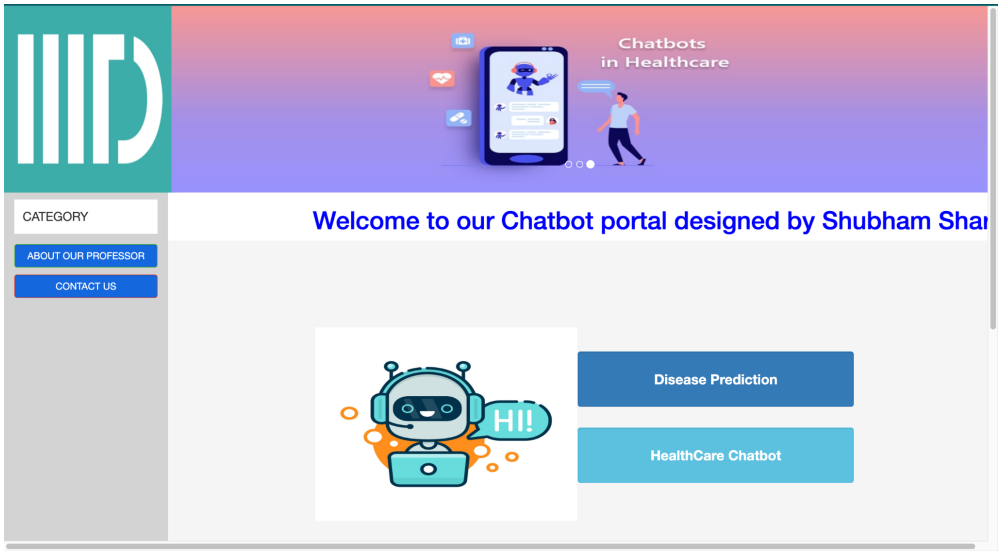
Web Application:

We have used different tool/technologies for building the webpage :

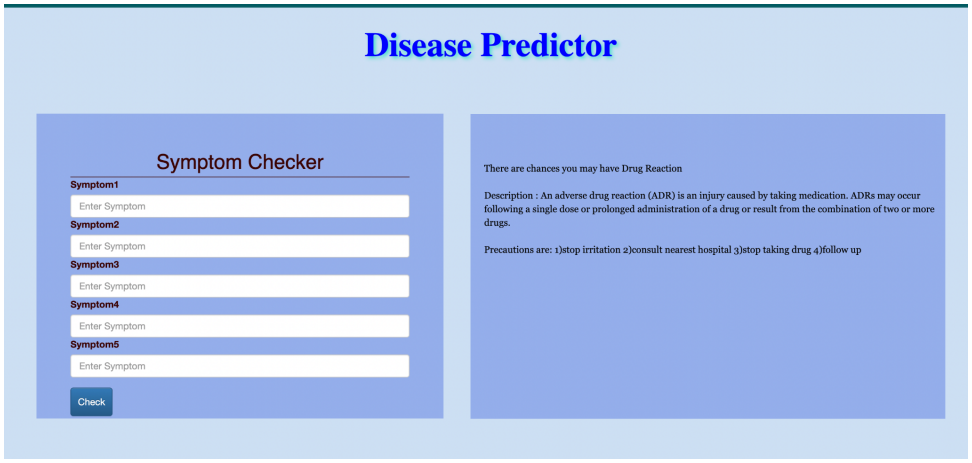
- HTML for displaying our web page on a web browser.
- CSS used for presentation of our webpage and managing styles etc.
- JavaScript used for creating dynamic and interactive web pages.
- Bootstrap for developing responsive websites.

Chatbot WebPage:

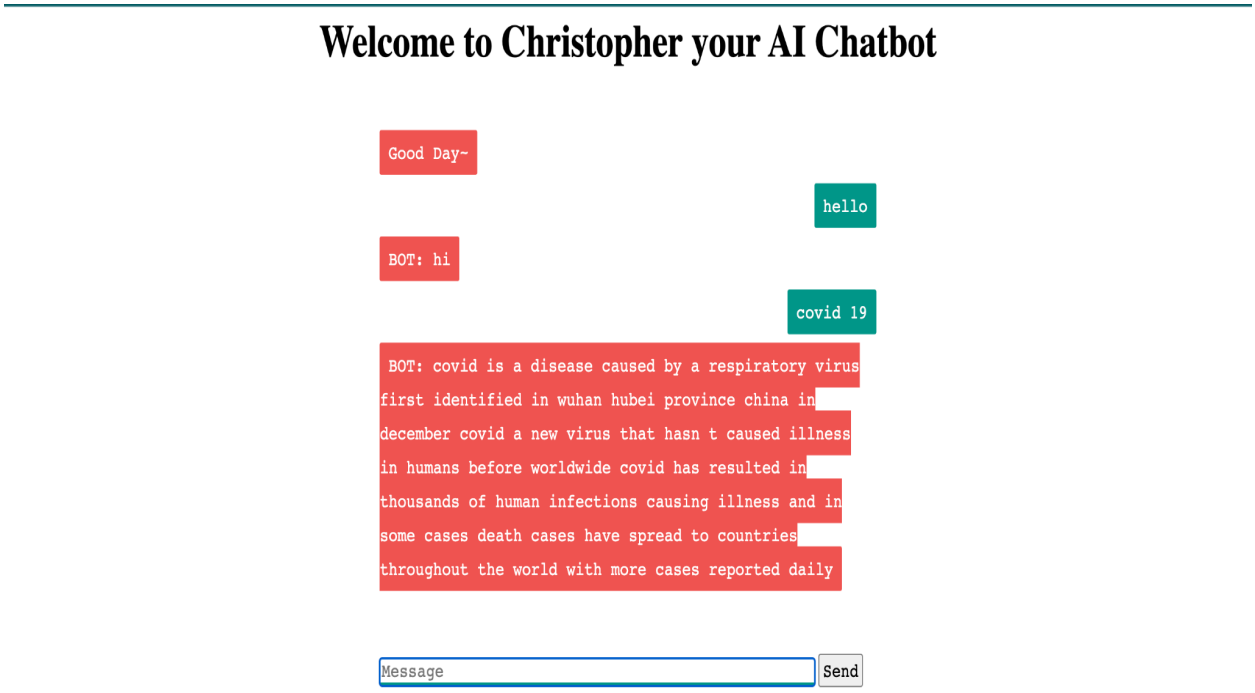
a. Front Page



b) Disease Prediction



c) Healthcare Chatbot



CONCLUSION AND FUTURE SCOPE

We have developed a web interface for both Healthcare Chatbot and Disease Prediction. The web page aims to provide users with details related to different diseases. To access the Healthcare chatbot, the user should input some query related to illness and predict disease, and the user should select the symptoms. This, in turn, helps the users to get instant results without contacting any healthcare professionals. We have used different models such as KNN, RF, etc., and libraries such as NLTK to build our model and run it on the web application as mentioned in the methodology.

There is a scope for making this web application more advance by experimenting with more datasets. Apart from providing results related to the input query, try to implement more useful features such as alerting the user in case of any severe symptoms to contact a particular healthcare professional.

REFERENCES

1) https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

2) "Novel Approach for Medical Assistance Using Trained Chatbot",Divya Madhu, Neeraj Jain C,International Conference on Inventive Communication and Computational Technologies.

3) "Healthcare Chatbot System using Artificial Intelligence",Varun Srivastava, Suraj Kumar Prajapati, Shri Krishna Yadav, Dr. Himani Mittal, Unique Paper ID: 151926, Publication : IJERT EXPLORE, Volume: Volume 8, Issue: Issue 1

4) Mrs. Rashmi Dharwadkar, Dr.Mrs. Neeta A. Deshpande "A Medical ChatBot". International Journal of Computer Trends and Technology (IJCTT) V60(1):41-45 June 2018. ISSN:2231-2803. www.ijcttjournal.org. Published by Seventh Sense Research Group.

5) Kim AJ, Yang J, Jang Y, Baek JS. Acceptance of an Informational Antituberculosis Chatbot Among Korean Adults: Mixed Methods Research. JMIR Mhealth Uhealth. 2021 Nov 9;9(11):e26424. doi: 10.2196/26424. PMID: 34751667; PMCID: PMC8663686.

6) To QG, Green C, Vandelanotte C. Feasibility, Usability, and Effectiveness of a Machine Learning-Based Physical Activity Chatbot: Quasi-Experimental Study. JMIR Mhealth Uhealth. 2021 Nov 26;9(11):e28577. doi: 10.2196/28577. PMID: 34842552; PMCID: PMC8665384.

7) Xu L, Sanders L, Li K, Chow JCL. Chatbot for Health Care and Oncology Applications Using Artificial Intelligence and Machine Learning: Systematic Review. JMIR Cancer. 2021 Nov 29;7(4):e27850. doi: 10.2196/27850. PMID: 34847056; PMCID: PMC8669585.

8) https://github.com/abachaa/MedQuAD

9) Wang H, Gupta S, Singhal A, Muttreja P, Singh S, Sharma P, Piterova A. An Artificial Intelligence Chatbot for Young People's Sexual and Reproductive Health in India (SnehAI): Instrumental Case Study. J Med Internet Res. 2022 Jan 3;24(1):e29969.

10)https://www.kaggle.com/ap1495/american-sexual-health-association

11) https://github.com/merav82/Safebot

12) Chkroun M, Azaria A. A Safe Collaborative Chatbot for Smart Home Assistants. Sensors (Basel). 2021 Oct 6;21(19):6641. doi: 10.3390/s21196641. PMID: 34640960; PMCID: PMC8512550.

13) D. Dahiwade, G. Patle and E. Meshram, "Designing Disease Prediction Model Using Machine Learning Approach," *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, 2019, pp. 1211-1215, doi: 10.1109/ICCMC.2019.8819782.

14)https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset

15)https://www.kaggle.com/narendrageek/mental-health-faq-for-chatbot

16)https://www.kaggle.com/nupurgopali/depression-data-for-chatbot

17)https://www.kaggle.com/xhlulu/covidqa

18)https://github.com/LasseRegin/medical-question-answer-data

19)https://www.nltk.org/