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 and deliver relevant information, as well as diagnose the disease using disease prediction and provide basic information about the disease before contacting a doctor. Through the use of a medical chatbot, this will assist to minimize healthcare expenses and enhance access to medical information. Chatbots are computer programmes that communicate with users using natural language. The data is stored in the database so that the chatbot can recognise the sentence keywords, make a query choice, and respond to the inquiry. The n-gram, TFIDF, and cosine similarity are used to rank and calculate sentence similarity. From the supplied input sentence, a score will be calculated for each sentence, and additional comparable sentences will be found for the query. The chatbot has two different parts one of them is a conversing agent that converses with the user and another one is disease prediction that takes input from the user and predicts what is the disease and what precautions can be taken. For the disease prediction part we have used multiple machine learning classifiers such as logistic Regression, K-Nearest Neighbors, SVM, Random Forest, Decision Tree and Convolutional Neural Network [1] on the dataset and we got highest accuracy of 91.44 using CNN classifier on test dataset.

INTRODUCTION

Chatbots are natural language processing structures that function as a digital conversational agent, simulating human interactions as part of AI devices.
While this generation is still in its early stages of development, healthcare chatbots are likely to be a growth driver. Improve access to healthcare, improve doctor–patient and clinic–patient communication, or help manage the increased demand for fitness solutions such as remote testing, medication adherence tracking, and teleconsultations. These are some of the main functionalities 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 key 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, which can advise physicians and even patients on what to do in such instances, eventually saving the lives of hundreds of people [3]. The scheme's major goal is to bridge the communication gap between users and healthcare practitioners by providing prompt responses to questions posed by users. People nowadays are more likely to be addicted to the internet, yet they are unconcerned about their own health. They avoid going to the hospital for minor issues that might turn into severe diseases in the future. Rather than looking through a list of possibly relevant documents on the web, creating question-and-answer forums is becoming an easy approach to answer such queries [4].

Making a user interface to provide input and get responses is used to construct our chatbot application. It is a system that interacts with the user by keeping track of the current state of the interaction and recalling previous commands in order to provide functionality. Artificial intelligence algorithms can be used to create medical chatbots that examine user inquiries, recognise them, and respond to similar queries. Not only this but if a user wants to know about the specific disease that he has using the symptoms that he has, this can also be done using the disease prediction part of our application.

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 knowledge base was obtained from the information provided by the Korea Disease Control and Prevention Agency. [5]

The study aims to investigate the feasibility, usability, and effectiveness of a machine learning–based physical activity chatbot. Participants wore a Fitbit Flex 1 (Fitbit LLC) and connected to the chatbot via the Messenger app. Fitbit activity data were synced from the Fitbit platform to the chatbot Platform. Strategy is Dialog Flow (Google Inc), an advanced Google machine learning platform for creating conversational artificial intelligence applications. [6]

The paper basically focuses on the recent trends in the chabot technology, it gives us an overview of the brief history of chatbots in medical science and 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]

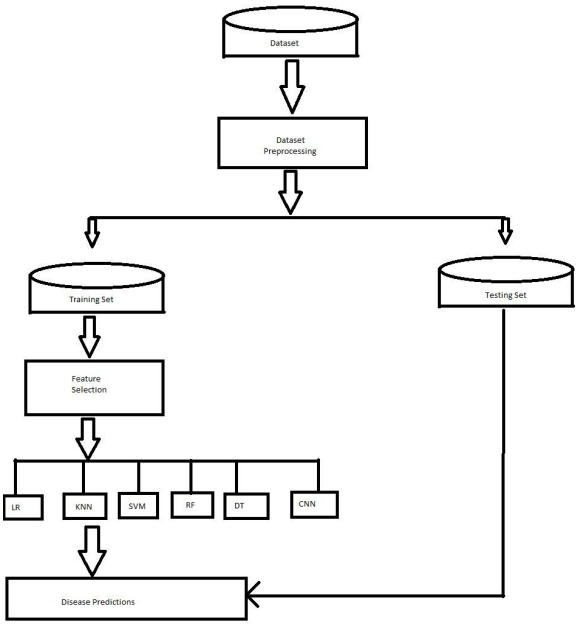
The research focuses on making a artificial intelligence chatbot for identifying social and behavioral changes regarding sexual and reproductive health in India. SnehAI chatbot aims to provide a safe space for Indian youth to have conversations about SRH, dispel sex-related myths and taboos, offer accurate information about safe sex and contraceptive choices, and address mental health concerns. They are using Natural language processing, Azure LUIS app, commonly used in AI chatbots, receives user input in the form of conversational and natural language texts, treats them as user utterances, processes the information by extracting keywords and predicting user intentions, and then uses the JSON response to make decisions about how to fulfill the user's requests. [9,10]

The paper presents Safebot, a chatbot that permits users to show it new reactions in natural language. Safebot depends on its own trusted users to distinguish malicious users and use the information acquired to decrease the action of malicious users later on. The analyses show the viability of Safebot. They plan to situate SafeBot as an augmentation of the intelligent home aide and permit genuine cooperative learning by multiple users. Since Safebot education completely depends on the natural language, it can be placed at the core of Toy, such as a talking robot (or talking parrot). [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

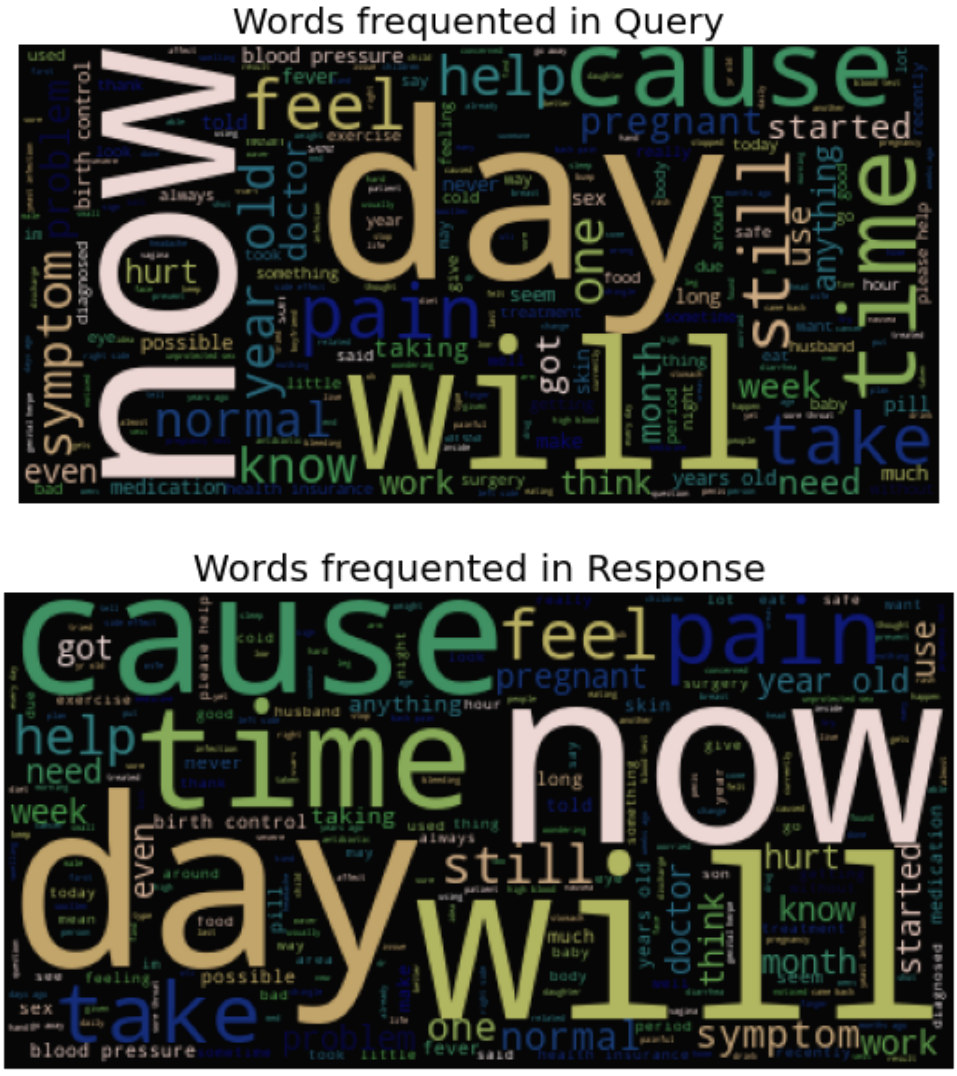
We have done two different types of work, one of them is healthcare chatbot in which we have merged many different datasets [15,16,17,18]. The generated dataset will have question and answer pairs. After that we put the dataset to work using the chatbot.
In the second part of our work we used the Disease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



Fig(1): Disease Prediction Methodology

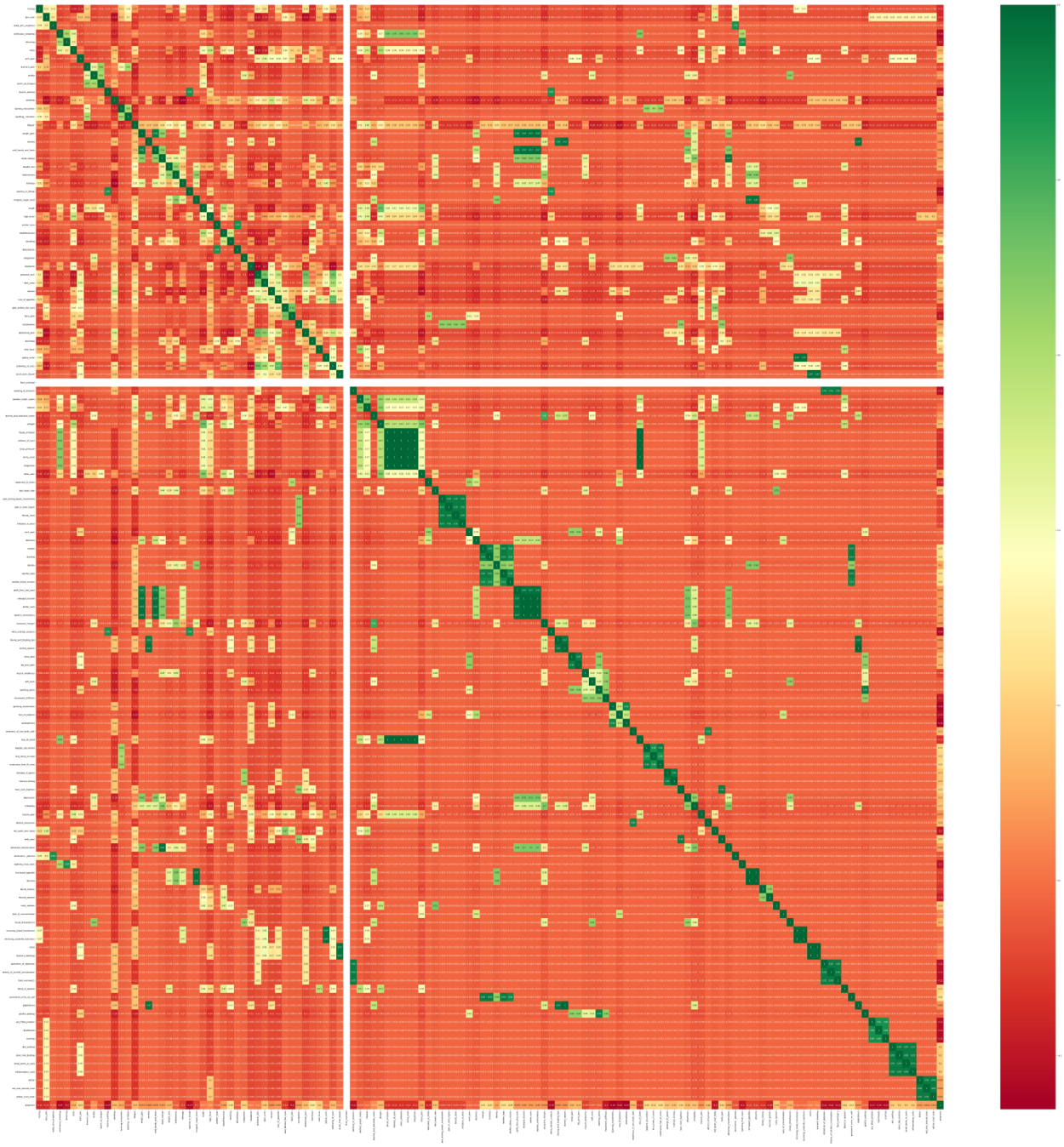
1) **Dataset Description**: The dataset used for chatbot is a merged dataset that contains 11278 question answer pairs [15,16,17,18] regarding covid, mental health, depression and general doctor-patient quotes. The dataset used for disease prediction is disease symptom dataset [14]. The dataset contains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



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.



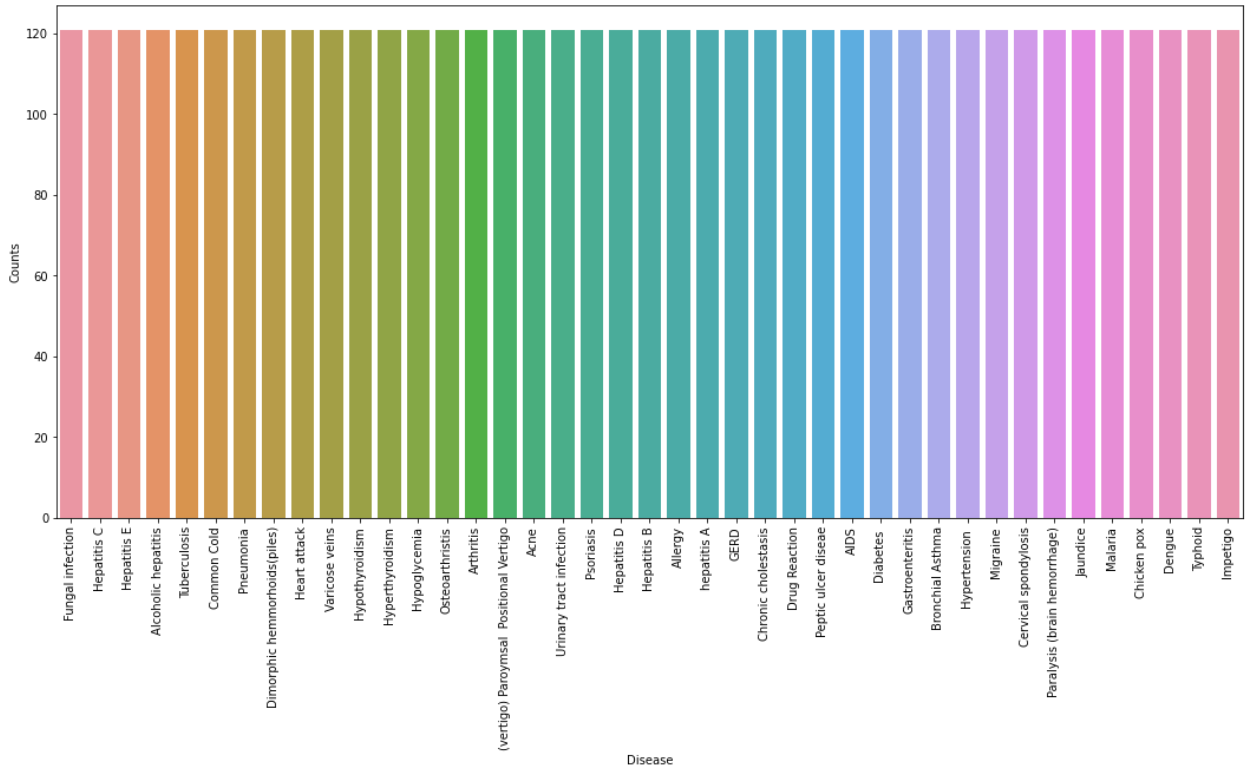
Fig(3): Correlation among all features in Disease prediction dataset

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

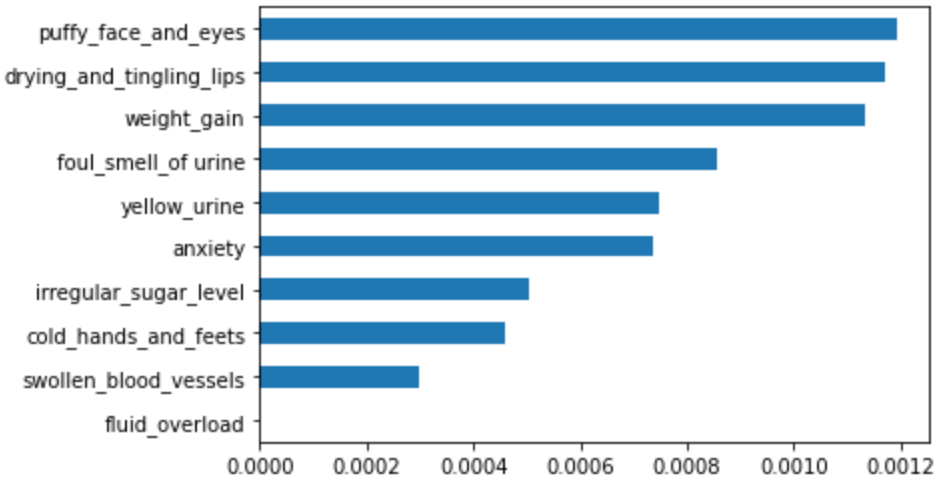
```
Index(['itching', 'skin_rash', 'nodal_skin_eruptions', 'continuous_sneezing',
       'chills', 'joint_pain', 'stomach_pain', 'acidity', 'muscle_wasting',
       'vomiting', 'burning_micturition', 'fatigue', 'weight_gain', 'anxiety',
       'weight_loss', 'cough', 'sunken_eyes', 'headache', 'yellowish_skin',
       '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

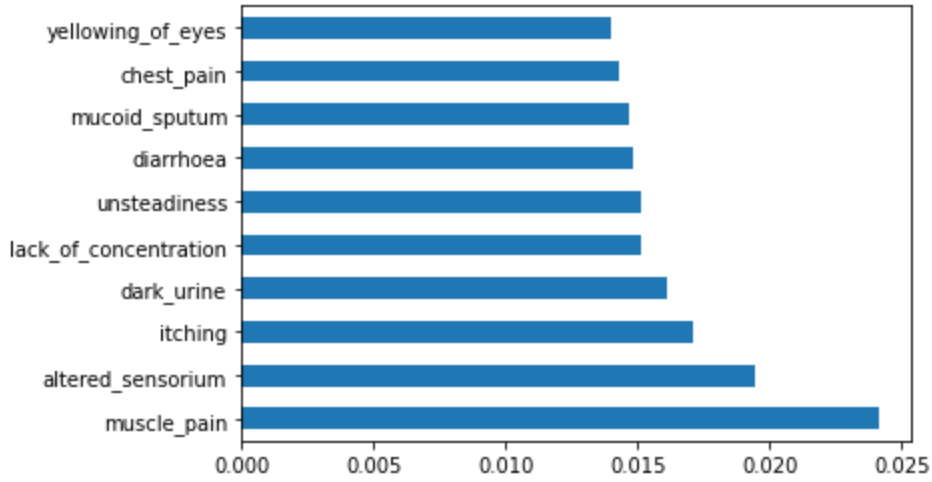
The disease dataset target variable (“prognosis”) is very equally balanced, i.e the count of rows for each target variable is equilateral.



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 have also worked on the **Robust scaler** (Use statistics that are resistant to outliers to scale features. The median is removed, and the data is scaled according to the quantile range (defaults to IQR: Interquartile Range). The interquartile range (IQR) is the distance between the first and third quartiles (25th and 3rd quantiles) (75th quantile).) and **Quantile Transformer** (Quantile information is used to transform features. The characteristics are transformed into a uniform or normal distribution using this procedure.).

3) **Machine Learning Classifiers**:

As we get the 32 features and 1 target variable, we divide the dataset in two parts X and Y with X.Shape=(4961,32) and Y.Shape=(4961,). Such that X contains all the 32 features and Y contains only the target variable.

Then we use the Train_Test_Split function to split the dataset into Train and Test with 80:20, such that the Training dataset contains 80% of the data and the testing dataset contains 20% of the data.
Then We apply different machine learning classifiers on 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 can divide n-dimensional space

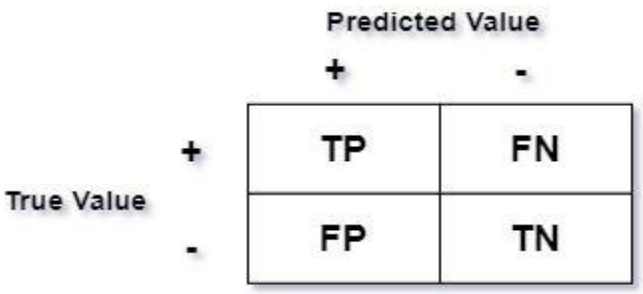
into distinct classes so that fresh data points may be rapidly assigned to the appropriate category. We forecast the incidence of stroke across a pre-defined time period in our stroke prediction issue, making it a binary classification problem that fits within the SVM framework.

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 technique implies that new and old data are similar, and it assigns new data to the category that is 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 true and projected values. In the below we describe confusion matrix and evaluation parameters formulas :-

Confusion Matrix -



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

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

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

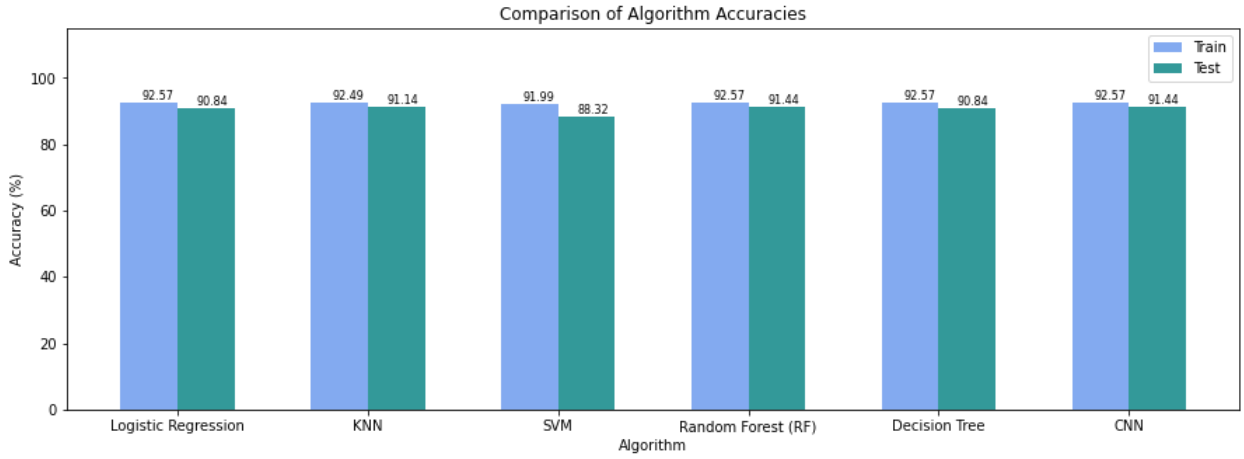
$$\text{F1-Score} = \frac{True\ Positive}{True\ Positive + 1/2(False\ Positive + False\ Negative)}$$

4) **Natural language processing Toolkit**: NLTK is a popular Python programming language for working with human language data. It includes a set of text processing libraries for categorization, tokenization, stemming, tagging, parsing, and semantic reasoning, as well as wrappers for industrial-strength NLP libraries and easy-to-use interfaces to over 50 corpora and lexical resources like WordNet.[19]

RESULTS AND ANALYSIS

For the sake of our research, we divided our dataset 80:20 for training and testing. We can see that 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 the use of Logistic Regression, we got an accuracy of 90.84%. When we use the K- Nearest neighbor algorithm, then we got an accuracy of 91.14%. The accuracy of the SVM classifier is 88.32%. The accuracy by the use of the Decision tree algorithm is 90.84%. The accuracy by 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:

CONCLUSION AND FUTURE SCOPE

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. doi: 10.2196/29969. PMID: 34982034; PMCID: PMC8764609.

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/