IJE TRANSACTIONS B: Applications Vol. 37 No. 02, (February 2024) 312-322

International Journal of Engineering

Journal Homepage: www.ije.ir

Implementation of Chatbot that Predicts an Illness Dynamically using Machine Learning Techniques

S. Shedthi B.[a], V. Shetty*[b], R. Chadaga[a], R. Bhat[a], B. Preethi[a], P. Kini K.[a]

[a] *Nitte (Deemed to be University), NMAM Institute of Technology (NMAMIT), Department of Computer Science and Engineering, Nitte, India*
[b] *Nitte (Deemed to be University), NMAM Institute of Technology (NMAMIT), Department of Mechanical Engineering, Nitte, India*

PAPER INFO

Paper history:
Received 11 August 2023
Received in revised form 11 September 2023
Accepted 12 September 2023

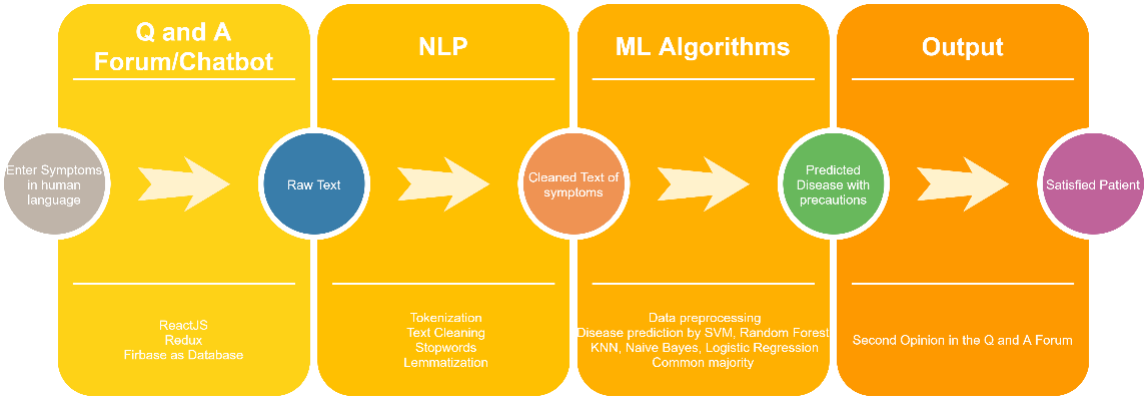
Keywords:
Artificial Intelligence
Machine Learning
Natural Language Processing
Healthcare
Chatbot

ABSTRACT

Timely access to healthcare is crucial in order to maintain a high standard of living. However, obtaining medical consultations can be difficult, especially for those living in remote areas or during a pandemic when face-to-face consultations are not always possible. The ability to accurately diagnose diseases is essential for effective treatment, and recent technological advancements offer a potential solution. Machine learning (ML) and Natural language processing (NLP) enables computer programs to understand human language and extract desired features from responses, allowing for human-like interaction with users. By leveraging these technologies, healthcare professionals can potentially provide more accessible and efficient medical consultations to individuals, regardless of their location. The concept is to establish an online platform where users can ask medical-related queries and receive responses from both medical professionals and fellow users. The platform would feature a Medical Chatbot, which employs advanced ML techniques to analyze user-provided symptoms and provide initial disease diagnosis and related information prior to consulting with a doctor. This disease prediction chatbot interacts dynamically with the users to enter the symptoms of the diseases and based on syntactic and semantic similarity response is given. In this work the threshold of similarity score is kept of 0.7. K-Nearest neighbors, Random forest, Support vector machine, Naive bayes and Logistic regression algorithms are used for prediction of disease based on symptoms which are faced by users. The syntactic similarity, fuzzy string matching and semantic similarity using all-MiniLM-L6-v2 model is used to improve the efficiency of the result.

***doi**: 10.5829/ije.2024.37.02b.08*

Graphical Abstract



*Corresponding Author Email: vidyasagar.shetty@gmail.com (V. Shetty)

Please cite this article as: Shedthi B. S, Shetty V, Chadaga R, Bhat R, Preethi B, Kini K P. Implementation of Chatbot that Predicts an Illness Dynamically using Machine Learning Techniques, International Journal of Engineering, Transactions B: Applications. 2024;37(02):312-22.

1. INTRODUCTION

Machine learning is an area of artificial intelligence (AI) with a concept that a computer program can learn and adapt to new data without human intervention (1). When a machine learning system receives new data, it guesses the outcome using the prediction models it has built using prior data. The amount of data used determines how well the output is anticipated, as a larger data set makes it easier to create a model that predicts the outcome more precisely. Natural Language Processing (NLP) is a subfield of AI which uses ML algorithms to train the computers to understand, interpret human language and derive meaningful insight from it (2). NLP uses computer algorithms with machine learning to process and analyze textual data which helps to analyze the sentiments (3-5). NLP combines techniques from linguistics, computer science, and machine learning to process and analyze natural language data, such as text or speech has numerous applications in areas such as information retrieval, sentiment analysis, Chatbots, and language translation. By enabling computers to process and understand natural language, NLP is bridging the gap between humans and machines and opening up new channels for connection and communication between the people (6). Chatbots are an intelligent system being developed using AI and NLP algorithms (7).

A chatbot can also be known as digital assistants that understand human capabilities and it interpret the user intent, process their requests, and give prompt relevant answers. Chatbots have become increasingly popular as useful tools for companies and institutions. Many companies recognize the potential for using Chatbots in healthcare to support patients. Obtaining consultations with doctors for every health issue can be challenging. The concept is to develop a medical chatbot using Artificial Intelligence, which can diagnose diseases and provide essential information about them before consulting a doctor. This approach aims to decrease healthcare expenses and enhance access to medical knowledge through the use of a medical chatbot (8). Chatbots are computer programs that interact with users using natural language. They store data in a database to identify keywords in sentences and provide appropriate responses to user queries.

In 1950, Alan Turing introduced the concept of Chatbots by posing the question, "Can machines think?" (9). The first known chatbot, ELIZA, was created to act as a psychotherapist and utilized pattern matching and template-based responses to engage in question-based conversations (10). Another well-known chatbot, ALICE, was developed later and employed a pattern-matching technique to retrieve example sentences from output templates and avoid inappropriate responses (11). This concept was then expanded to various fields using machine learning techniques. Among the most popular applications of Chatbots in healthcare are diagnostics, patient support and health promotion (12).

A medical Chatbot is a conversational AI powered solution specifically designed to make healthcare much more interactive and proactive. Objective of the proposed work is to develop an effective question-and-answer forum where users can discuss the queries on the medical field and also develop a medical Chatbot that predicts an illness dynamically based on user symptoms using NLP and machine learning techniques.

Section 2 provides an overview of previous research conducted on disease diagnosis using machine learning and also about Chatbots. In Section 3, we discuss the fundamental principles of Chatbot and the construction of our proposed work. Section 4 is dedicated to the implementation and results of the Question and Answer Forum and Chatbot model. Finally, in section 5, we present a summary of our findings and offer suggestions for future work.

2. RELATED WORK

In this literature survey section, a detailed summary of the different kinds of research that have been conducted that are related to disease diagnosis using machine learning and Chatbots are considered.

Problem-Based Learning (PBL) is a key point for learning activities (13) and technology plays an important role in the learning process (14). Computer science and health care field is correlated and lots of research is going on with combination of these two fields (15-17). Ahsan et al. (18), this paper is a comprehensive review of various algorithms in Machine Learning for popular diseases based on symptoms. There is also a great raise in the number of research papers published from 2012 to 2021 on different diseases. This review also mentions the possible data related problems like data scarcity, noisy data and manipulation of training data making us aware to be careful of the data set selection, as well as disease diagnosis related challenges like misclassification and confusion that algorithms predict. This paper is very insightful as it provided history of many diseases related classifications and where algorithms could be inaccurate. Shaji et al. (19) investigated on heart diseases which are getting extremely common and there are a variety of attributes such as age, sex, blood pressure, and the levels of urea, sodium and potassium will help in predicting, if a person is suffering from a heart disease. The factors like stress level or if any person is a smoker or not plays a major role in determining the disease. These attributes were taken as data and used in classifying algorithms like SVM, RF, KNN and ANN. This paper though very specific to detection of heart diseases and its symptoms, gives a great direction on how the process of data

cleaning is done and compares the accuracy of algorithms and is very helpful for researchers starting out in the medical AI and ML field.

Keniya et al. (20) used various KNN and gave the most accurate prediction with accuracy of 93%. It used attributes like symptoms, age and gender to determine a disease out of 230 possible outcomes. The aim of this work was that sometimes doctors are unavailable and diseases like Ebola are deadly and contagious so we can avoid human contact as much as possible which matches our aim too. This project lacks the User Interface which usually makes access easier to healthcare professionals.

Ferjani et al. (21) proposed supervised ML algorithms for disease diagnosis and aiding medical experts in the early identification of high-risk diseases. The aim of this work was to investigate and analyze already existing algorithms and focuses on the accuracy of commonly used algorithms and find out which algorithm is more precise for which of the diseases. Magoulas et al. (22) insisted that most disease prediction systems are application driven. It can face numerous issues like overfitting and improper scaling. The visualization of the learnt problem will be tough. We also need to deal with large datasets which may have high dimensional input features. Also, there are ethical aspects in the medical field and human intervention is necessary. This work indicates AI and ML in healthcare is no new topic. This was way ahead of the times it was published, almost all statements are true even now. Though there are various python libraries to help us with the visualization. Jain et al. (22) stated that people do not go to the doctor for harmless diseases like cold, sneezing, fever which may be a symptom of a more harmful disease. The person must choose symptoms from groups of it which range from mild, medium, and severe symptoms. It uses DT and deep neural networks to classify the disease as well as compares the two algorithms used. Researchers (21-23) are given an in-depth detail of AI and ML in healthcare, how to deal with data. The algorithms to try on our data set and driving the force that prompted the backend of our project. Pingale et al. (24) used Naïve Bayes, KNN and Logistic Regression to diagnose a disease and the dataset is from real-life hospital data which is in structural format. This work highlights two points that our work also stresses in the sections mentioned below, that if a user enters unrelated or random symptoms or if there are a smaller number of symptoms, there are high chances of obtaining less accurate results.

Bharti et al. (25) proposes a multilingual conversational bot which intends to give free primary healthcare and education. It is implemented on google cloud platform and uses NLP to implement the system. The paper also talks about voice-based user interface as well as a chat-based interface. It's based on DialogFlow and uses only 255 intents which may be a limitation if put into production. Though this work has a great interface including a message based and voice-based AI doctor; this is not very helpful to healthcare professionals who need to deal with more complicated diseases. It focuses on the use where possible patients need advice. A review paper by Tjiptomongsoguno et al. (26) which makes a logical table comparing the Chatbots based on the methodology, algorithms and the datasets used. It also highlights the various pros and cons found in each Chatbot. This work highlights the necessity of medical websites in today's digital world and was the root cause that a question-and-answer forum/website exists in our work. Reshma et al. (27) uses SVM to predict health status and it also uses the google API to convert speech to text and text to voice. The input is sent to a Chatbot, which responds with relevant information and displays it on the stand-alone app. Prayitno et al. (28) proposed a Chatbot for frequently asked questions in a company which uses NLP to successfully diagnose the user's illness with 87 percent accuracy.

Ayanouz et al. (29) performed a detailed survey and examined many publications which are related to Chatbots using the AI concepts needed to build an intelligent conversational agent based on Deep Learning Models. One of the Chatbots is Casper which helps Insomniacs pass the night. Another is One Remission Chatbot which has the goal of helping those involved in the fight against cancer. This paper also mentions that selecting the right engine for NLP is the most important step of creating a Chatbot. Programmers also must decide whether they want structured or unstructured conversations. They also tell the basics of NLP and finally propose a general architecture of a smart Chatbot.

Caldarini et al. (30) mainly classifies Chatbots into rule-based Chatbots and Artificial intelligent Chatbots. Though rule-based models were easy to design and implement, they had limited capabilities. AI Models are based on ML Algorithms, and they allow them to learn from an existing database of human conversations. It also talks about different datasets used like Open Subtitles, Cornell, Daily Dialog. A variety of evaluation metrics used are F1 score, Perplexity and Bilingual evaluation understudy (BLEU). Kumar et al. (31) concentrates on NLP and neural network and summarizes the most efficient implementation techniques that have been carried out in the previous years. It then proposes a methodology to develop a state-of-the-art Chatbot application that can be personalized easily according to customer needs. It is implemented using tools such as Dialog Flow, Tensor Flow, Android Studio and Firebase. Chaudhary et al. (32), in this paper presented the pros and cons along with comparisons of various available medical Chatbots. A multilingual Chatbot called 'HEALTHBOT' was also proposed which will interact with patients in English and Marathi and will note down their symptoms and pathological test reports and will also

prescribe the patients with further medical tests and suggest them basic medications, diets, and lifestyle changes. Also is highly useful for the medical practitioners as personal assistant and not only that it would save valuable time for both the parties and reduce the treatment time.

Vasileiou and Maglogiannis (33) developed a health Chatbot predict and diagnosis and limitation of this work is it is used for only two diseases i.e., Covid-19 and heart problem. In future work we can extend this work with more diseases. Shaikh et al. (34) presented an adolescent oriented intelligent conversational chatting system called "HappySoul", which acts as a virtual friend who can assist to encourage, understand, comfort, and guide stressful adolescents to pour out their bad and negative feelings, thereby releasing the stress. Chatbot will allow a user to simply ask questions in the same way that they would address a human. The technologies used were NLP, RNN and client server architecture with the help of Android GUI. Above works made us realize that using NLP even though more difficult than DialogFlow and TensorFlow (They need us to define intents) gives us more independence and flexibility in the design of the Chatbot. PhaniRaghavaa and Kumarb (35) aimed to create an artificial conversation entity for healthcare treatment using python. Two algorithms, i.e., fuzzy support vector machine algorithms are compared with the Decision tree algorithm. The results indicate that the suggested fuzzy support vector machine algorithm will outperform the current approach in terms of output.

Eslam, et al. (36) presented a smart Chatbot system that can communicate with people, and it gives answers about the COVID-19. The model used the pretrained Google BERT language model. This technique employs the BERT Transformer to categorize text input into various categories based on the meaning of the words. The second stage involves using the BERT model and choosing the query domain for the replies. Their proposed system is trained and tested on Stanford University's SQuADV2.0, a well-known question-answering dataset. Kumar et al. (37) have used the classification algorithms such as Naive Bayes, Random Forest, Logistic Regression, and KNN to predict the disease based on the symptoms inputted by the patient. These algorithms have different accuracies and are chosen based on the specific disease being predicted. Moreover, the system can also determine whether a patient is suffering from a specific disease or not by predicting "True" or "False". Once the disease is predicted, the system can recommend the patient which type of doctor to consult for the specific disease. The entire model has been implemented using Django and connected to the Django server. Tamizharasi et al. (38) proposed to make accurate predictions and increase the model's effectiveness. In their study, they addressed the use of a support vector machine learning method in a Chatbot for healthcare. The conversational style is achieved using natural language processing. In addition to receiving free or low-cost services, this helps people spend less time.

In recent work lots of improvement done related to healthcare digitalization in that Chatbot is a prominent one (39). There is no doubting the extent to which the use of AI, including Chatbots, will continue to grow in public health (40). In an ideal scenario, research should play a crucial role in shaping the progress of digital innovations in public health. This involves assessing the effects of technologies like Chatbots on health interventions, gaining understanding of user experiences, and giving utmost importance to the safety and welfare of users.

From the above literature survey observed that NLP and machine learning technology can help to give solution to most healthcare problems. But in most of the works they have used dialogflow and tensor flow for Chatbot creation and the final test result is derived through one best training algorithm. Further work can be extended on developing a robust experimental design that can effectively demonstrate the efficacy and efficiency of the Chatbot. Additionally, it is necessary to create a user-friendly Question and Answer Forum to facilitate better interaction.

3. PROPOSED SYSTEM

There sometimes arises a situation where a person is feeling ill but cannot access a doctor in a clinic or hospital because of odd timings or a person is in a remote area. Instead of panicking and self-diagnosing randomly it is better to have knowledge of possible diseases a person is suffering from and take precautions for the same.

The aim of this research work is to design a platform where users can ask health related queries is important to have a wider view on the conditions. Search, posting questions, answering questions if relevant are the additional features which can escalate its use. A medical Chatbot can also help in predicting the disease given the symptoms. Our project aims at developing a website where users can have an interaction with fellow users on health-related aspects, and an integrated medical Chatbot created using machine learning and natural language processing techniques. People can interact with the Chatbot just like they do with another human and through a series of queries; Chatbot will identify the symptoms of the user and thereby, predict the disease using a machine learning algorithm. The proposed system hence reduces hospital wait times, consultation time, hospital readmission time in case patients need to be connected to the correct healthcare professionals and helps them understand their conditions and possible treatment option without even visiting a doctor. Front end design of the

proposed system is shown in Figure 1. Redux and ReactJS used for Chatbot front end.

Redux and ReactJS are both popular tools for managing the state of a web application, but they approach the problem in different ways. ReactJS has its own built-in state management system, known as local state. Local state is managed within a component and is not shared with other components. Redux, on the other hand, is a standalone state management library that can be used with React or other frameworks. Redux provides a central store that holds the entire state of the application, which can be accessed and modified by any component that subscribes to the store. To update the state, you dispatch an action to the store, which triggers a series of reducers that modify the state in a predictable and consistent way. React is one of the most popular frontend JavaScript Library. Even though there are other popular frameworks like Angular JS, Vue JS, Ember JS and many more. There were certain characteristics of React and Redux giving them the upper hand and hence they were chosen as our framework. React JS uses a virtual DOM (Document Object Model) that updates only parts of the DOM that were changed whereas Angular JS uses real DOM which will be slower while rendering large amounts of data. Thus React and Redux have better performance than Angular JS. React and Redux also have more flexibility than Vue JS. React JS has better support for mobile development. ReactJS and Redux are scalable, a small application can be gradually converted to a larger application while Ember JS is for already scaled systems.

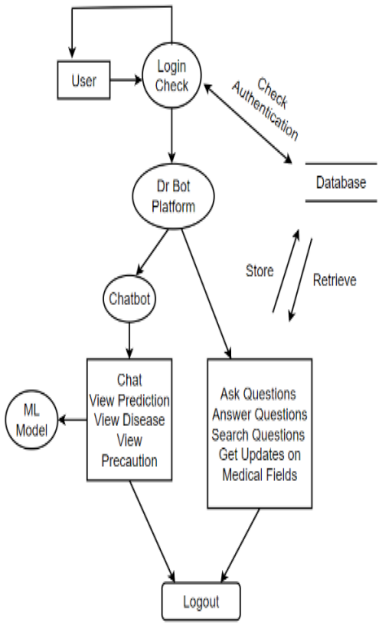


Figure 1. Chatbot front end using Redux and ReactJS

Chatbot Backend done using NLP and steps of that is shown in Figure 2. The purpose of using NLP for the chat bot is that NLP ensures that the Chatbot can understand the vocabulary, sentiments and meaning that users use when naturally conversing. There was a possibility of that the project could be done without the use of NLP by presenting the symptoms as options to select by the users. This would remove the ease of use and make it harder for the users to choose their symptoms as people may not even have certain symptoms. For example, the symptom nausea means having discomfort and feeling the need to vomit. A person who does not know “nausea” may input “discomfort” and “vomiting” and NLP can make semantic relation between them to choose “nausea” as the symptom. The SpaCy Library, which is a free, open source library is used for the Natural Language processing in this work. We have downloaded a trained Pipe line en_core_web_md which returns a language object that contains all components to process the text. The raw text is the original text that we want to process, and Tokenization is the process of breaking down the raw text into smaller units, called tokens. These can be words, phrases, or even characters. Text Cleaning is the process of removing unwanted characters, symbols, numbers, and other noise from the text. This can include removing HTML tags, punctuation marks, and special characters. POS Tagging that is Part-of-speech (POS) tagging is the process of assigning grammatical tags to each token in the text. This helps to identify the syntactic structure of the text. Stopwords are common words that are often removed from the text during preprocessing, as they do not add much meaning to the text. Examples include "the," "and" "of," etc. The SpaCy library also consists of a stopwords list which by default consists of 326 words, for each word in the input text, if it consists of one of these 326 stopwords, it will be removed. Lemmatization is the process of reducing each word to its base or root

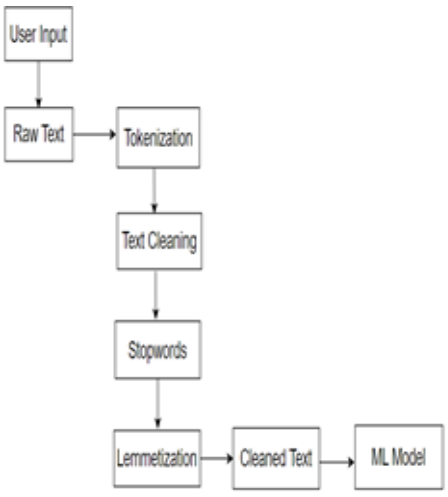


Figure 2. Steps in NLP

form, called a lemma. This helps to reduce the overall vocabulary size and to normalize the text. Cleaned Text is the final output of the NLP pipeline, which is a cleaned and pre-processed version of the original raw text.

Various machine learning algorithms are used in experiments for disease identification based on symptoms. Machine learning algorithms which are used for this experiment and disease prediction process is shown in Figure 3.

Machine learning algorithms which is used for this experiment are listed below:

Support Vector Machine (SVM): It is a powerful supervised learning algorithm and separates data into classes by finding the optimal hyperplane that maximally separates the classes. It is effective for handling high-dimensional data and also handles outliers effectively. In this work "rbf" kernel is used.

Random forest (RB): It is a popular ensemble learning algorithm in machine learning. It combines multiple decision trees to create a robust and accurate model. Each tree in the forest is trained on a random subset of the data and features, reducing overfitting and improving generalization. This is good for handling complex datasets, and providing reliable predictions in various domains.

K-nearest neighbors (KNN): It operates by assigning a new data point to the majority class of its K nearest neighbors in the feature space. It can handle both classification and regression tasks, but it is particularly popular for classification. KNN is easy to implement and understand but the result depends on how parameter K is set (41, 42). Here K=3 used. Algorithm can be computationally expensive for large datasets since it requires calculating distances between data points.

Naïve Bayes (NB): It is a classification algorithm based on Bayes' theorem and the assumption of independence among input features. It is widely used in text classification and spam filtering. It is particularly suitable for large datasets and works well even with limited training data. Used in many real-world applications due to its simplicity and effectiveness. Gaussian Naive Bayes is used in this experiment.

Logistic regression (LR): It is a statistical model used for classification and it predicts the probability of an instance belonging to a specific class Logistic regression is widely used in various domains, including finance, healthcare, and social sciences. In this work class weight balanced parameter is used.

The user inputs their symptoms or medical condition into the system. Here it is in the form of text. In Data preprocessing the input data is pre-processed and cleaned to remove inconsistencies, and irrelevant information. The pre-processed data is the set of symptoms along with their correlated symptoms of threshold more than 0.8 are then fed into machine learning models for classification. Including correlated features along with the given set of symptoms improves prediction accuracy, reduces overfitting, and enhances feature importance. The ML models generate an output that provides a diagnosis or recommendation for the user. The disease which is predicted the maximum that is by the most algorithms is taken into consideration. This may include the likelihood of a particular disease or other relevant information. The system provides the output to the user, and the process is complete.

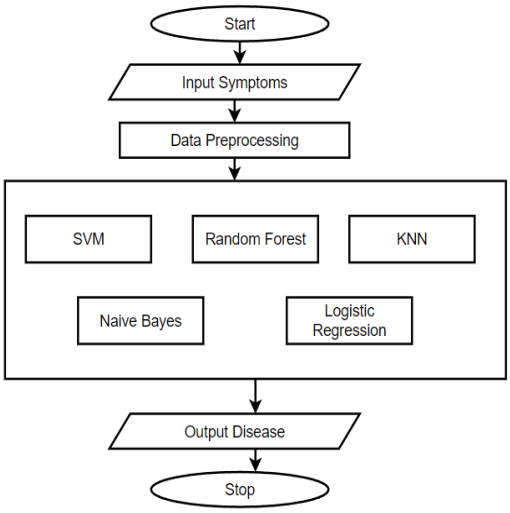


Figure 3. Prediction of diseases using machine learning models

4. RESULTS AND DISCUSSIONS

Natural Language Toolkit (NLTK) is used for implementation, and it has text processing libraries for sentence detection, tokenization, lemmatization, stemming and parsing. Based on the symptoms, most accurate algorithms will be used to classify and using a simple majority the Chatbot must output the diagnosis. At least 4 symptoms must be needed by the Chatbot for the diagnosis to be accurate.

4. 1. Question and Answer Forum The user initially has to login to the website using email. Login can be done through google authentication or by email id and password entered manually. Users can register using email and password for the first time. User is directed to the main page that is the app component in redux. Different components present are:

Post component: the questions asked by all the users are each wrapped in a post component. Buttons are provided with each post to open a modal to answer the query and another which takes the answer component.

Answer component: displays all the answers or discussions on the particular query.
User component: displays all the questions asked by the user logged in.

Search Component: Displays questions with the keyword provided by the user.

Navbar Component: Contains links to user, search, Chatbot, logout options.

Sidebar Component: Contains various specialization in medical fields on clicking on each will take to the news component which displays latest news on the topic. The news is extracted using the news api.

All the data are stored in the firebase database. Adding, extracting of data is done dynamically from the firestore.

Login Page:

Login page is a component rendered by React which is accompanied with its own stylings. There are two options to login and register which is shown in Figure 4, i.e.,

Sign in with email and password.

Sign in with Google Authentication.

This is an answer forum where users can post questions, and other users or doctors can provide relevant answers. The purpose of this forum is to facilitate the exchange of information and knowledge on a variety of topics, including health, wellness, medicine, and other related subjects. Figure 5 shows Q&A forum.

Once a user clicks on add a question, the user can post a query in the forum which is shown Figure 6. Also, the user can attach any image or link along with the question while submitting it.

Figure 7 shows all the questions posted by the user and the recently posted questions appears on the top.

4. 2. Chatbot Results The Dr.Bot a disease prediction Chatbot interacts dynamically with the user. The user needs to provide the symptoms they face one after the other and then they get the information about the disease they would probably be facing. The conversation starts by the Chatbot instructing the users the response they need to provide for the prediction to be generated.

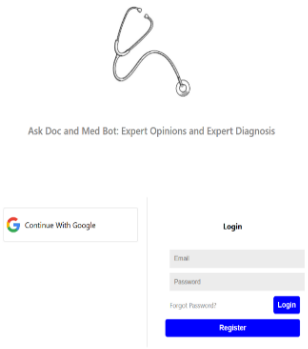


Figure 4. Login and Register page

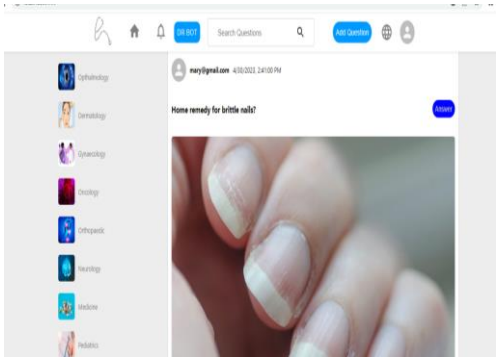


Figure 5. Question and Answer Forum

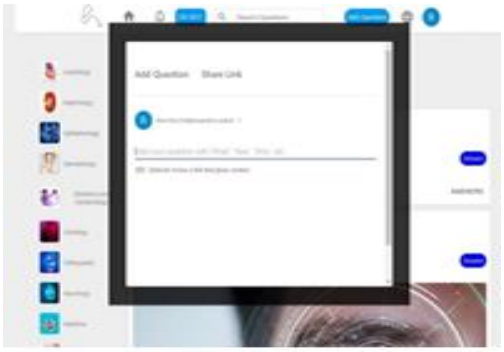


Figure 6. Adding a Question

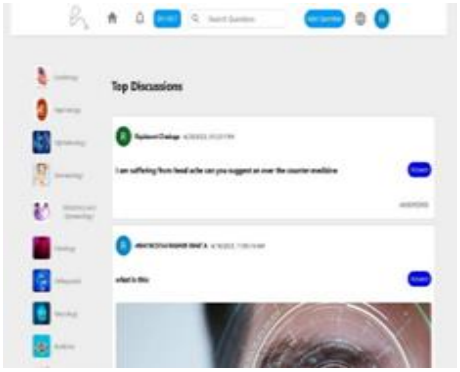


Figure 7. Posted Questions

The reply from the user cannot be used directly by the Chatbot form machine learning prediction. That is the input must be converted into other form accepted by the ML model. For each input from the user the response is Preprocessed: removal of stop words, pronouns are removed e.g., rash in skin is processed to rash skin.

Finding syntactic similarity: The user might sometimes enter a full sentence with the input symptom in it. For e.g.: I am suffering from muscle pain. The Jaccard set is used to find the intersection between the response and the set of all symptoms in the dataset. If more than one symptom has an intersection with the input provided the list of possible symptoms is provided to the

user to select one from the list. If no intersection is found the next step continues.

Jaccard Similarity Between two sets i.e.

X and Y is J (X, Y) = (X∩Y) / (XUY)

If the similarity index is 1 for any symptom in the dataset (i.e., both A and B are the same) that symptom is considered for prediction. If no similarity index of 1 is present with any of the symptoms, all other symptoms which have a similarity score other than 0 are listed and the user is asked to choose the symptom which they are suffering from; e.g., if the user enters pain, all the symptoms in dataset with pain in it is matched and these symptoms have partial similarity with the one user entered (pain in this case)

If all the symptoms have 0 Jaccard similarity with the one user entered, fuzzy string matching is done of the pre-processed user response with all symptoms in dataset. Fuzzy string matching is the technique of finding strings that match with a given string partially and not exactly. A fuzzy similarity score is assigned to the comparison with every symptom, and whichever has the highest similarity score and is above threshold similarity is considered. This is to consider the condition when the user has entered a wrong spelling for a symptom, but the fuzzy string gives a good similarity if not too many spelling errors are there. Example if a user has entered 'chils' instead of 'chills' a good similarity score is given by the algorithm since they are only one letter apart.

If any of the symptoms do not fuzzy match with the user input, then semantic similarity is found between the input and every symptom.

Finding semantic similarity: If the input response does not match with any symptom in corpus, semantic search is done for the response. The semantic search is done using the all-MiniLM-L6-v2 model of sentence transformer from hugging face. Sentence transformers are models which are pre-trained using a large dataset to find semantic similarity or gives a similarity scores based on their meaning and context. The model assigns similarity score for the two sets provided. In our work we consider a threshold of 0.7 the symptom with similarity passing the threshold is taken into consideration.

If the user enters 'ache in head' instead of 'headache', the sentence transformer assigns a good similarity score to the above two words since they mean the same thing.

If the response does not meet any of the above criteria, an error message is sent to the user to input again.

A minimum of three symptoms are required for accurate prediction.

The symptoms are then passed to the ML model which evaluates and predicts the most probable disease. Along with the disease, its description and precautions are also provided to users if required. As we know the disease predicted by ML model has the possibility of being a misdiagnosis, the Chatbot after predicting its output always states to please consult a doctor for best opinion. It indicates that the chatbot is just making its prediction for the possible set of symptoms and is always better to reach out to a doctor if the predicted output is something serious. It reiterates the fact that it is just an AI doctor.

All the algorithms used in the project give accuracy above 90% for the test data. For the input symptoms entered by the user the output is accurate when the symptoms are related. In case of unrelated symptoms, the model may not give the best result. All algorithms are applied to test data and out which algorithm gives better results that is considered for final display.

Data Pre-processing Machine Learning Algorithms:

The data is from an open-source site Kaggle[1]. This data has 133 columns and 4920 rows. Each column is a symptom except the last one which represents the prognosis (Class label). Each cell consists of 0 or 1 indicating if the symptom was present or not. If a cell consists of 0 it indicates that the symptom was not seen in that instance. In data preprocessing null values are removed, duplicate rows are removed and few symptoms which won't give much weightage to overall disease prediction are also removed. We then divide the data set into training data set and testing data set where 75% of the data is used for training the model and 25% is used as the testing data. The Correlation matrix in Figure 8 which consists of 132 rows and 132 columns is hard to understand because of the sheer volume of the symptoms. Selection of highly relevant features is very important to enhance effectiveness of algorithm (43). The code filters out the symptoms which have a correlation more than 0.5 to understand what symptoms are related to each other. For example, the main symptoms of the disease AIDS are if a person was engaging in extramarital contact, had muscle wasting and there occurred patches in the throat.

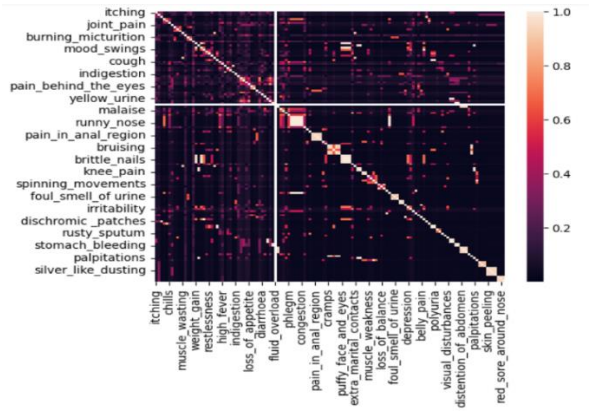


Figure 8. Correlation Matrix

[1] https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset?select=dataset.csv

In Figure 9 we see that extra marital contact is related to muscle wasting and patches in throat with the absolute value 0.89. In Figure 10 we see that phlegm is a common symptom with two diseases, with Common Cold and as well as Malaria with different Correlation values. The more the symptoms are correlated to each other, more the probability that the disease predicted is accurate.

Machine learning algorithms like SVM, RF, KNN, NB and LR algorithm results are evaluated using statistical measures, such as accuracy, precision, recall, and F1-score, to quantify the performance of the chatbot. From the above results observed that the random forest algorithm performs better than all other algorithms for this data (Table 1).

Figure 11 shows results of prediction i.e., common cold using machine learning. To predict diseases, the Chatbot may ask the user a series of questions about their symptoms, medical history, lifestyle habits, and other relevant information. Here we have taken two examples where the Chatbot has predicted diseases that are common cold (Figure 12).

```
extra_marital_contacts | muscle_wasting | 0.89
extra_marital_contacts | patches_in_throat | 0.89
```

Figure 9. Example 1 for Correlation Matrix

```
phlegm | chills | 0.59
phlegm | cough | 0.73
phlegm | breathlessness | 0.52
phlegm | swelled_lymph_nodes | 0.64
phlegm | malaise | 0.66
```

Figure 10. Example 2 for Correlation Matrix

TABLE 1. Evaluation of machine learning algorithm

	SVM	RF	KNN	NB	LR
Accuracy (%)	90.79	97.37	90.79	96.05	94.74
Precision	0.91	0.98	0.89	0.98	0.97
Recall	0.91	0.97	0.91	0.96	0.95
F1-score	0.90	0.97	0.89	0.97	0.94

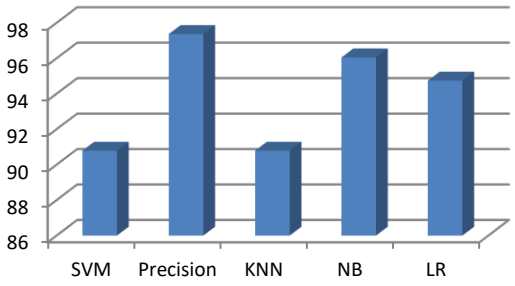


Figure 11. Representation of accuracy of ML model

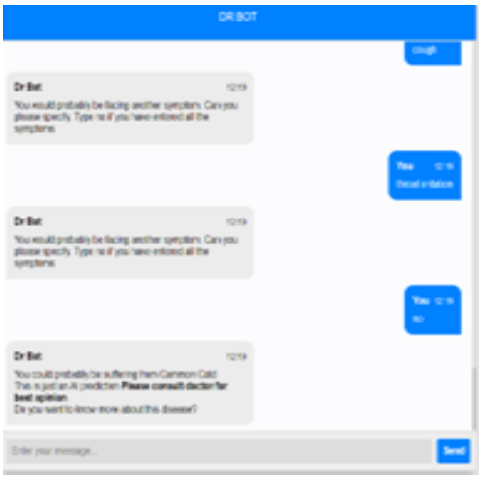


Figure 12. Chatbot predicting Common Cold

5. CONCLUSION AND FUTURE SCOPE

In this paper, the disease prediction Chatbot engages with users in real-time to input disease symptoms, and provides a response based on both syntactic and semantic similarity. In syntactic search if input response does not match with any symptom in the corpus, semantic search is done for response. Semantic search seeks to improve search accuracy by understanding the content of the search query and for semantic search all-miniLM -L6v2 sentence transformer model is used. A similarity score threshold of 0.7 is used in this study and if criteria not met means the model asks the user to enter more symptoms. Minimum three symptoms are required to get for better prediction. The symptoms are then passed to the ML model which evaluates and predicts the most probable disease. The Chatbot takes the input as symptoms and predicts the disease for it. Also gives the description of the disease if required. Support vector machine, random forest, K-nearest neighbors, naive bayes and logistic regression algorithms are used for disease prediction, where random forest outperforms with 97.37% accuracy. All algorithms are applied on new unknown test data and out which disease which is predicted by the majority of the algorithms that are taken into consideration and displayed for the user. To enhance the efficiency of the result fuzzy string-matching technique is used. The backend of the Chatbot has been designed where data preprocess of the input from the user is pre-processed to extract only the required symptoms. The frontend development for the login and registration pages, as well as the design of the Chatbot, were completed for the web platform that allows users to post questions and answers. Integration of the Chatbot with the frontend and the backend of the question-and-answer platform are done. In this platform where prediction can be further discussed, a patient can be directed to the actual doctor and get a second opinion there itself.

In future additional features for Question-and-Answer Forum like upvote, downvote etc., can be added to the developed model to make it more interactive. The presence of more upvotes on an answer will make it known to users that the clarity of the answer is good, implying that it is accurate and concise. It will also mean that the answer was useful and helpful to the users. The presence of downvotes would imply that the answers were difficult to understand and have been out of topic to the question asked. The upvotes and downvotes can later be used to filter out the answers there by helping the users access the most relevant answer to their question if it has already been asked before. More diseases and symptoms can be added to the existing dataset to make it more accurate. If addition of data results in more symptoms (features) without new examples (rows), there may be chances of overfitting the training data. But overall, the more assortment of examples we have, our model becomes more general and thus decreases the generalization error. While adding more data there are chances of duplication of the data and inaccurate information creeping into our data set.

7. REFERENCES

1. Michalski RS, Carbonell JG, Mitchell TM. Machine learning: An artificial intelligence approach: Springer Science & Business Media; 2013.
2. Liddy ED. Natural language processing. 2001.
3. Ahangari M, Sebti A. A Hybrid Approach to Sentiment Analysis of Iranian Stock Market User's Opinions. International Journal of Engineering, Transactions C: Aspects,. 2023;36(3):573-84. 10.5829/IJE.2023.36.03C.18
4. Harimi A, Esmaileyan Z. A database for automatic Persian speech emotion recognition: collection, processing and evaluation. International Journal of Engineering, Transactions A: Basics,. 2014;27(1):79-90. 10.5829/idosi.ije.2014.27.01a.11
5. Sadjadi S, Mashayekhi H, Hassanpour H. A two-level semi-supervised clustering technique for news articles. International Journal of Engineering, Transactions C: Aspects,. 2021;34(12):2648-57. 10.5829/IJE.2021.34.12C.10
6. George AS, George AH, Baskar T, Martin AG. Human Insight AI: An Innovative Technology Bridging The Gap Between Humans And Machines For a Safe, Sustainable Future. Partners Universal International Research Journal. 2023;2(1):1-15. https://doi.org/10.5281/zenodo.7723117
7. Lalwani T, Bhalotia S, Pal A, Rathod V, Bisen S. Implementation of a Chatbot System using AI and NLP. International Journal of Innovative Research in Computer Science & Technology (IJIRCST) Volume-6, Issue-3. 2018. http://dx.doi.org/10.2139/ssrn.3531782
8. Athota L, Shukla VK, Pandey N, Rana A, editors. Chatbot for healthcare system using artificial intelligence. 2020 8th International conference on reliability, infocom technologies and optimization (trends and future directions)(ICRITO); 2020: IEEE. 10.1109/ICRITO48877.2020.9197833
9. Turing AM. Mind. Mind. 1950;59(236):433-60.
10. Weizenbaum J. ELIZA—a computer program for the study of natural language communication between man and machine. Communications of the ACM. 1966;9(1):36-45. 10.1145/365153.365168
11. AbuShawar B, Atwell E. ALICE chatbot: Trials and outputs. Computación y Sistemas. 2015;19(4):625-32. 10.13053/cys-19-4-2326
12. Xu L, Sanders L, Li K, Chow JC. Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. JMIR cancer. 2021;7(4):e27850. 10.2196/27850
13. Saengrith W, Viriyavejakul C, Pimdee P. Problem-Based Blended Training via Chatbot to Enhance the Problem-Solving Skill in the Workplace. Emerging Science Journal. 2022;6:1-12. 10.28991/ESJ-2022-SIED-01
14. Mendonça YV, Naranjo PGV, Pinto DC. The Role of Technology in the Learning Process. Emerging Science Journal. 2022;6(Special Issue):280-95. 10.28991/ESJ-2022-SIED-020
15. Roshani S, Coccia M, Mosleh M. Sensor technology for opening new pathways in diagnosis and therapeutics of breast, lung, colorectal and prostate cancer. medRxiv. 2022:2022.02. 18.22271186. 10.28991/HIJ-2022-03-03-010
16. Hamidi H, Daraee A. Analysis of pre-processing and post-processing methods and using data mining to diagnose heart diseases. International Journal of Engineering, Transactions A: Basics 2016;29(7):921-30. 10.5829/idosi.ije.2016.29.07a.06
17. Ye EZ, Ye EH, Bouthillier M, Ye RZ. DeepImageTranslator V2: analysis of multimodal medical images using semantic segmentation maps generated through deep learning. bioRxiv. 2021:2021.10. 12.464160. 10.28991/HIJ-2022-03-03-07
18. Manjurul Ahsan M, Siddique Z. Machine learning based disease diagnosis: A comprehensive review. arXiv e-prints. 2021:arXiv: 2112.15538.
19. Shaji SP, editor Predictionand diagnosis of heart disease patients using data mining technique. 2019 international conference on communication and signal processing (ICCSP); 2019: IEEE. 10.1109/ICCSP.2019.8697977
20. Keniya R, Khakharia A, Shah V, Gada V, Manjalkar R, Thaker T, et al. Disease prediction from various symptoms using machine learning. Available at SSRN 3661426. 2020. http://dx.doi.org/10.2139/ssrn.3661426
21. Ferjani MF. Disease Prediction Using Machine Learning. Bournemouth, England: Bournemouth University. 2020. 10.13140/RG.2.2.18279.47521
22. Magoulas GD, Prentza A. Machine learning in medical applications. Advanced course on artificial intelligence: Springer; 1999. p. 300-7.
23. Jain R, Chotani A, Anuradha G. Disease diagnosis using machine learning: A comparative study. Data Analytics in Biomedical Engineering and Healthcare: Elsevier; 2021. p. 145-61.
24. Pingale K, Surwase S, Kulkarni V, Sarage S, Karve A. Disease prediction using machine learning. International Research Journal of Engineering and Technology (IRJET). 2019;6(12):831-3.
25. Bharti U, Bajaj D, Batra H, Lalit S, Lalit S, Gangwani A, editors. Medbot: Conversational artificial intelligence powered chatbot for delivering tele-health after covid-19. 2020 5th international conference on communication and electronics systems (ICCES); 2020: IEEE. 10.1109/ICCES48766.2020.9137944
26. Tjiptomongsoguno ARW, Chen A, Sanyoto HM, Irwansyah E, Kanigoro B. Medical chatbot techniques: a review. Software Engineering Perspectives in Intelligent Systems: Proceedings of 4th Computational Methods in Systems and Software 2020, Vol 1 4. 2020:346-56. 10.1007/978-3-030-63322-6_28
27. Reshma R, editor An Improved Chatbot for Medical Assistance using Machine Learning. 2022 International Conference on Inventive Computation Technologies (ICICT); 2022: IEEE. 10.1109/ICICT54344.2022.9850470

28. Prayitno PI, Leksono RPP, Chai F, Aldy R, Budiharto W, editors. Health Chatbot Using Natural Language Processing for Disease Prediction and Treatment. 2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI); 2021: IEEE. 10.1109/ICCSAI53272.2021.9609784
29. Ayanouz S, Abdelhakim BA, Benhmed M, editors. A smart chatbot architecture based NLP and machine learning for health care assistance. Proceedings of the 3rd international conference on networking, information systems & security; 2020.
30. Caldarini G, Jaf S, McGarry K. A literature survey of recent advances in chatbots. Information. 2022;13(1):41. https://doi.org/10.3390/info13010041
31. Kumar R, Ali MM. A review on chatbot design and implementation techniques. Int J Eng Technol. 2020;7(11).
32. Chaudhary J, Joshi V, Khare A, Gawali R, Manna A. A Comparative Study of Medical Chatbots. International Research Journal of Engineering and Technology (IRJET). 2021;8(02).
33. Vasileiou MV, Maglogiannis IG. The Health ChatBots in Telemedicine: Intelligent Dialog System for Remote Support. Journal of Healthcare Engineering. 2022;2022. https://doi.org/10.1155/2022/4876512
34. Shaikh A, More D, Puttoo R, Shrivastav S, Shinde S. A survey paper on chatbots. International Research Journal of Engineering and Technology (IRJET). 2019;6(04):2395-0072.
35. PhaniRaghavaa B, Kumarb SA. An Improved Chatbot for Predicting Disease and Medicines Using Natural Language Processing with Fuzzy Logic. Advances in Parallel Computing Algorithms, Tools and Paradigms. 2022;41:258. 10.3233/APC220035
36. Amer E, Hazem A, Farouk O, Louca A, Mohamed Y, Ashraf M, editors. A proposed chatbot framework for COVID-19. 2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC); 2021: IEEE. 10.1109/MIUCC52538.2021.9447652
37. Kumar A, Sharma GK, Prakash U. Disease prediction and doctor recommendation system using machine learning approaches. International Journal for Research in Applied Science & Engineering Technology (IJRASET). 2021;9:34-44. https://doi.org/10.2214/ijraset.20221.36234
38. Tamizharasi B, Livingston LJ, Rajkumar S, editors. Building a medical chatbot using support vector machine learning algorithm. Journal of Physics: Conference Series; 2020: IOP Publishing. 10.1088/1742-6596/1716/1/012059
39. Giansanti D. The Chatbots Are Invading Us: A Map Point on the Evolution, Applications, Opportunities, and Emerging Problems in the Health Domain. Life. 2023;13(5):1130. https://doi.org/10.3390/life13051130
40. Wilson L, Marasoiu M. The development and use of chatbots in public health: scoping review. JMIR human factors. 2022;9(4):e35882. 10.2196/35882
41. Fooladi S, Farsi H, Mohamadzadeh S. Segmenting the lesion area of brain tumor using convolutional neural networks and fuzzy k-means clustering. International Journal of Engineering, Transaction B: Applications. 2023;36(8):1556-68. 10.5829/IJE.2023.36.08B.15
42. Pourbahrami S, Balafar MA, Khanli LM, Kakarash ZA. A survey of neighborhood construction algorithms for clustering and classifying data points. Computer Science Review. 2020;38:100315. https://doi.org/10.1016/j.cosrev.2020.100315
43. Kakarash ZA, Mardukhia F, Moradi P. Multi-label feature selection using density-based graph clustering and ant colony optimization. Journal of Computational Design and Engineering. 2023;10(1):122-38. https://doi.org/10.1093/jcde/qwac120

Persian Abstract

چکیده

دسترسی به موقع به مراقبت های بهداشتی برای حفظ استاندارد بالای زندگی بسیار مهم است. با این حال، دریافت مشاوره پزشکی می تواند دشوار باشد، به ویژه برای کسانی که در مناطق دور افتاده زندگی می کنند یا در طول یک بیماری همه گیر که مشاوره حضوری همیشه امکان پذیر نیست. توانایی تشخیص دقیق بیماری ها برای درمان موثر ضروری است و پیشرفت های تکنولوژیکی اخیر یک راه حل بالقوه ارائه می دهد. یادگیری ماشینی (ML)و پردازش زبان طبیعی (NLP)برنامه‌های رایانه‌ای را قادر می‌سازد تا زبان انسان را بفهمند و ویژگی‌های مورد نظر را از پاسخ‌ها استخراج کنند و امکان تعامل انسان‌مانند با کاربران را فراهم کنند. با استفاده از این فناوری‌ها، متخصصان مراقبت‌های بهداشتی به طور بالقوه می‌توانند مشاوره‌های پزشکی در دسترس و کارآمدتری را برای افراد، صرف‌نظر از موقعیت مکانی آن‌ها، ارائه دهند. مفهوم ایجاد یک پلت فرم آنلاین است که در آن کاربران می توانند سوالات مربوط به پزشکی را بپرسند و پاسخ هایی را هم از متخصصان پزشکی و هم از کاربران دیگر دریافت کنند. این پلتفرم دارای یک Chatbot پزشکی است که از تکنیک های پیشرفته ML برای تجزیه و تحلیل علائم ارائه شده توسط کاربر و ارائه تشخیص اولیه بیماری و اطلاعات مرتبط قبل از مشورت با پزشک استفاده می کند. این چت بات پیش بینی بیماری به صورت پویا با کاربران وارد تعامل می شود تا علائم بیماری ها را وارد کند و بر اساس شباهت نحوی و معنایی پاسخ داده می شود. در این کار، امتیاز آستانه تشابه ۰.۷ حفظ شده است. -Kنزدیکترین همسایه، جنگل تصادفی، ماشین بردار پشتیبان، سیسنم ساده و الگوریتم های رگرسیون لجستیک برای پیش بینی بیماری بر اساس علائمی که کاربران با آن مواجه هستند استفاده می شود. شباهت نحوی، تطبیق رشته فازی و شباهت معنایی با استفاده از مدل -all-MiniLM L6-v2 برای بهبود کارایی نتیجه استفاده می شود.