HealthSupport

Akash Kumar
Mtech IIIT Delhi
MT23012
akash23012@iiit.ac.in

Harsh Choudhary
Btech IIIT Delhi
2020433
harsh202043@iiit.ac.in

Jafreen Rizvi
Mtech IIIT Delhi
MT23040
jafreen23040@iiit.ac.in

Prakhar Sharma
Mtech IIIT Delhi
MT23060
prakhar23060@iiit.ac.in

Shazra Irshad
Mtech IIIT Delhi
MT23089
shazra23089@iiit.ac.in

Mo Rashid
Mtech IIIT Delhi
MT23047
rashid23047@iiit.ac.in

ABSTRACT

People often get ill and cannot understand whether their symptoms are signs of a disease. The medicines that they take are actually good for them. These are all because of a lack of medicinal knowledge. The search of this information is hectic and time-consuming. There comes our HealthSupport system. This project addresses healthcare accessibility by developing two components. The first utilizes NER to extract symptoms from text, predicts diseases using machine learning, and provides treatment details. The second component extracts medicine names from images and offers insights into their advantages and disadvantages. Techniques include symptom preprocessing, expansion using synonyms, and disease prediction via ML and cosine similarity. The system aims to enhance healthcare delivery by streamlining symptom analysis and medicine recognition.

ACM Reference Format:
Akash Kumar, Harsh Choudhary, Jafreen Rizvi, Prakhar Sharma, Shazra Irshad, and Mo Rashid. 2024. HealthSupport. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 PROBLEM DEFINITION

The project aims to develop a user-friendly system for symptom-based disease prediction and medicine recognition. Users can input unstructured symptoms or select suggested symptoms, receiving a list of probable diseases in return. Upon selecting a disease, the system provides comprehensive information on symptoms, causes, diagnosis, and treatment options. Additionally, the system suggests related symptoms based on user input. It caters to individuals with limited medical knowledge, facilitating early disease detection and diagnosis. Particularly beneficial for those hesitant to seek medical attention for minor symptoms, the system offers insight into the severity of potential diseases. By providing accessible and informative healthcare assistance, the project addresses the need for efficient symptom analysis and medicine recognition, ultimately contributing to improved healthcare accessibility and early intervention.

2 BACKGROUND

Machine Learning applications in healthcare have revolutionized early disease detection and diagnosis, significantly improving patient care. Despite the abundance of health-related information available on the internet, accessing relevant and accurate insights remains challenging due to scattered and overwhelming data. Existing disease prediction systems cater to specific conditions like heart diseases, neurological disorders, and skin ailments, yet a comprehensive system for universal disease prediction based on symptoms is lacking. Such a system is crucial for early diagnosis, enabling medical professionals to initiate timely treatment. However, predicting diseases solely based on user-input symptoms presents challenges, as users often express symptoms in non-technical terms, complicating the prediction process. Thus, the project aims to develop an innovative architecture integrating techniques like query expansion, synonym matching, and symptom suggestion to enhance disease prediction accuracy. By leveraging web-scraped data, the project lays the groundwork for future research in this domain, emphasizing the importance of accurate disease prediction without the need for extensive medical tests.

3 DATASET USED

The project initiated data collection by scraping disease names from the Illinois Department of Public Health website (https://dph.illinois.gov/topics-services/diseases-and-conditions/diseases-a-z-list.html) and retrieved corresponding symptoms from Wikipedia. Subsequently, the dataset underwent cleaning procedures, ensuring consistency and reliability. Each disease entry was then encoded in a binary format to represent the presence or absence of symptoms, facilitating efficient symptom-based disease prediction.

Similarly, datasets for precautions, medications, diets, and workouts were compiled to provide comprehensive information for each disease. Precautions encompassed preventive measures individuals could take to mitigate disease risk, while medication datasets detailed prescribed drugs for treatment. Dietary guidelines and recommended workouts were tailored to address specific disease management needs.

This meticulous structuring of data allows the system to offer holistic insights into various diseases, empowering users with valuable information for disease prevention, management, and

Conference'17, July 2017, Washington, DC, USA
© 2024 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

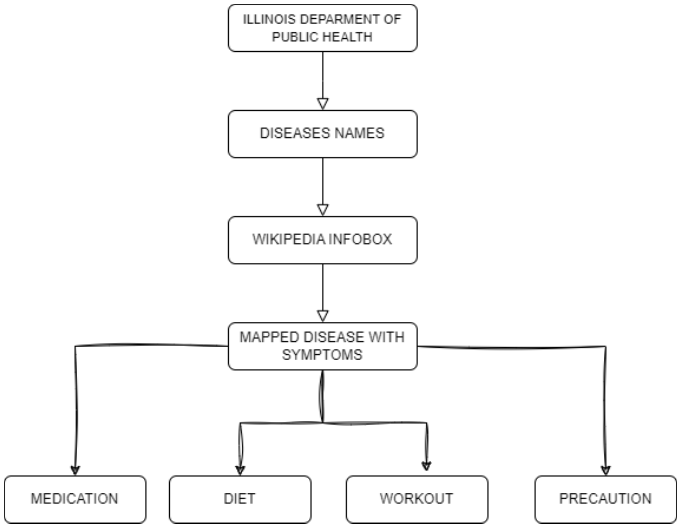

ILLINOIS DEPARMENT OF PUBLIC HEALTH
DISEASES NAMES
WIKIPEDIA INFOBOX
MAPPED DISEASE WITH SYMPTOMS
MEDICATION
DIET
WORKOUT
PRECAUTION

Figure 1: Database creation

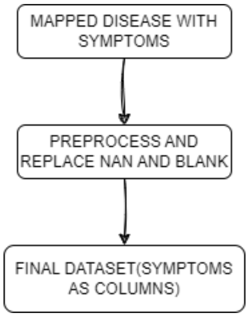

MAPPED DISEASE WITH SYMPTOMS
PREPROCESS AND REPLACE NAN AND BLANK
FINAL DATASET(SYMPTOMS AS COLUMNS)

Figure 2: Medicine Information Retrieval

treatment. By leveraging curated datasets sourced from reputable sources, the project ensures accuracy and reliability in disease-related information, enhancing the effectiveness of symptom-based disease prediction and healthcare delivery.

4 PROPOSED SOLUTION SKETCH

Our proposed solution for the health support system begins with users inputting their symptoms using free text. We employ advanced Natural Language Processing (NLP) techniques, specifically Named Entity Recognition (NER), to extract specific symptoms from the user's input. This process ensures that we capture accurate and relevant symptoms for disease prediction.

4.1 Dataset Cleaning and Symptom-Disease Mapping

The dataset undergoes rigorous cleaning to ensure accuracy and reliability. We meticulously map diseases with their corresponding symptoms, establishing a robust foundation for precise disease prediction.

4.2 Using Machine Learning

The heart of our system lies in the application of the Support Vector Classifier (SVC) algorithm. Trained on the cleaned and mapped dataset, the SVC model learns intricate patterns and relationships between symptoms and diseases. This enables the model to accurately predict diseases based on the input symptoms provided by the user, significantly enhancing diagnostic capabilities.

Support Vector Classifier (SVC) is particularly effective in scenarios where the dataset has clear margin of separation between classes. This algorithm works by finding the hyperplane that best separates the classes while maximizing the margin, which leads to robust generalization performance. Moreover, SVC is less prone to overfitting, especially in high-dimensional spaces, due to its ability to utilize only the support vectors for decision boundary determination.

Furthermore, SVC is versatile and can handle both linearly separable and non-linearly separable datasets efficiently by using different kernel functions such as linear, polynomial, and radial basis function (RBF). This flexibility enables SVC to capture complex relationships within the data and make accurate predictions.

Overall, the decision to continue with SVC was driven by its superior performance, robustness, and adaptability to the dataset characteristics, making it a suitable choice for the task at hand.

4.3 Comprehensive Disease Prediction and Recommendations

Upon successful disease prediction, our system generates comprehensive recommendations tailored to each user. These recommendations encompass various aspects such as disease diagnosis, personalized diet plans, workout routines, medication suggestions, and precautionary measures. This holistic approach ensures that users receive a well-rounded health support system that addresses their unique needs.

4.4 Integration of Gemini Pro AI for Medication Analysis

Furthermore, our system integrates Gemini Pro AI for in-depth medication analysis. This advanced analysis delves into the advantages and disadvantages of various medications, considering both the medication data and the symptoms/diseases predicted by the model. Users receive valuable insights into their prescribed medications, enhancing their understanding of medication choices.

4.5 Optical Character Recognition (OCR) Technology for Medication Insights

To further enhance medication recommendations, we incorporate Optical Character Recognition (OCR) technology. By extracting information from medication labels uploaded by users, our system gains a deeper understanding of medication specifics. This enables us to offer detailed recommendations that include medication details along with their advantages and disadvantages, as well as additional precautions.

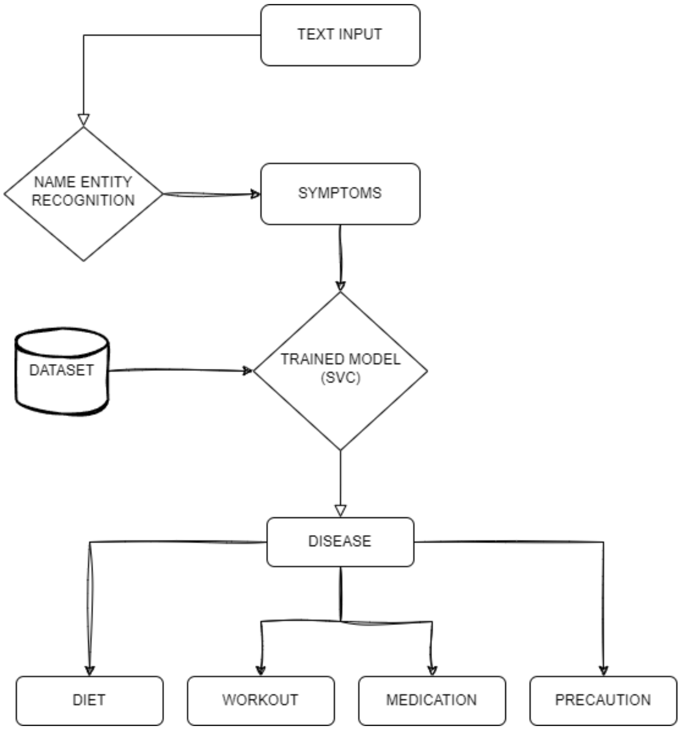


Figure 3: Project Data Flow

4.6 Comprehensive and Personalized Health Management Solution

In conclusion, our health support system encompasses a comprehensive approach, from symptom extraction to disease prediction, personalized recommendations, medication analysis, and OCR-based insights. We provide users with a comprehensive and personalized health management solution that empowers them to make informed decisions about their health and well-being.

4.7 Disease Details

After obtaining the disease name from the SVC model, the system retrieves corresponding data from the precompiled dataset containing information about diseases, including details about diet, precautionary measures, workout recommendations, and medication. This dataset is curated to provide comprehensive insights into each disease, ensuring that users receive relevant and reliable information.

Once the medication names associated with the predicted disease are identified, they are passed to Gemini, a reliable source for pharmaceutical information. Gemini provides detailed information about medications, including their advantages and disadvantages. By leveraging Gemini's extensive database, the system ensures accuracy and completeness in presenting medication-related information to users.

This integration of Gemini enhances the system's functionality by offering users a deeper understanding of the prescribed medications, including their potential benefits and drawbacks. Users can make informed decisions about their treatment plans, considering both the advantages and disadvantages of the prescribed medications.

Overall, this process of fetching data from the disease dataset and leveraging Gemini for medication information enriches the system's capabilities, empowering users with comprehensive insights to manage their health effectively.

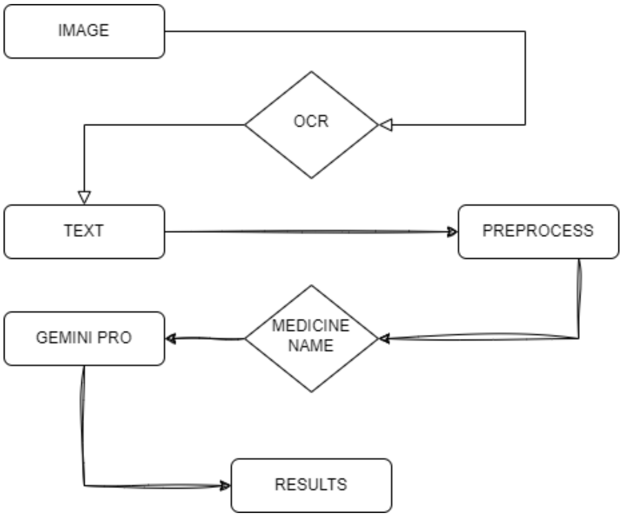


Figure 4: Medicine Information Retrieval

4.8 Medicine Information Retrieval from Image

The developed OCR system utilizes Paddle OCR to process images of medicine labels or strips. Upon receiving an image input, Paddle OCR extracts text data along with the spatial information, including the space occupied by each text element in the image. This information is then sorted in descending order based on the size of the text, with the largest font indicating the medicine name.

This approach ensures accurate identification of the medicine name from the label or strip, as larger font sizes typically signify key information such as the name of the medication. Once the medicine name is extracted, the system formulates a query and sends it to Gemini Pro, a trusted source for pharmaceutical information.

Gemini Pro provides detailed insights into the identified medication, including its advantages and disadvantages. By leveraging Gemini Pro's extensive database, the system ensures that users receive comprehensive and reliable information about the prescribed medication, enabling them to make informed decisions about their treatment.

This integration of Paddle OCR and Gemini Pro enhances the system's functionality by automating the extraction of medicine names from images and providing valuable pharmaceutical information to users. It streamlines the process of accessing medication-related data, contributing to improved healthcare decision-making and patient care.

5 LITERATURE REVIEW

Extensive research efforts have been dedicated to the domain of disease prediction, particularly focusing on symptoms and healthcare data. This extensive exploration spans various methodologies and technologies aimed at enhancing diagnostic capabilities and improving patient outcomes.

This paper proposed by Md Tahmid's [3], a model has been proposed that follows a structured approach. The model is designed to handle user input in a specific format, where each symptom is described in a single sentence, and subsequent symptoms are added in new lines. After receiving all the input symptoms, their model makes the prediction based on the symptoms. The system's core function is using machine learning for disease prediction. It learns patterns and relationships between symptoms and diseases from tagged input data, creating an accurate predictive model.

There is also a research paper by Shamsher Bahadur Patel [2] for Coronary Heart Disease Using Neuro-Fuzz. In their project, they've developed an automated tool to assess the likelihood of coronary heart disease in patients. Their approach employs a layered neuro-fuzzy system that accepts patient information and medical test results. The training data includes key parameters like age, sex, blood sugar, cholesterol, blood pressure, max heart rate, exercise-induced angina, ECG results, and chest pain type.The system generates fuzzy-based rules across multiple parameters organized into two layers. It creates a fuzzy database and query engine for data processing. Ultimately, the system provides graphical representations of the chances of heart disease occurrence based on the input data.

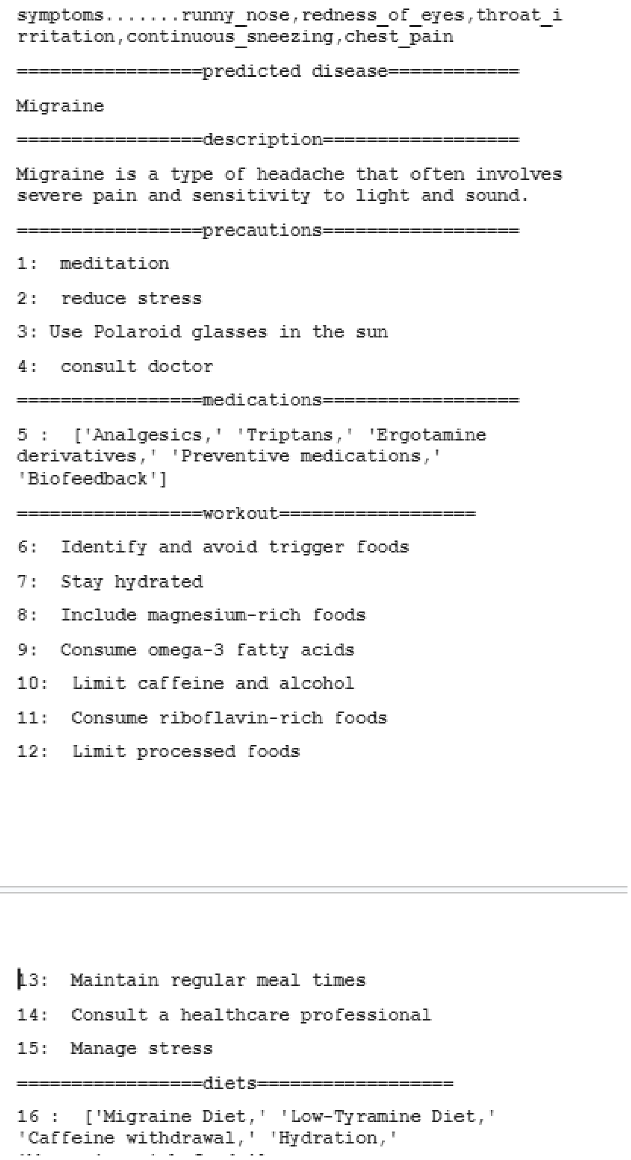
Many online platforms, like WebMD, provide disease prediction services based on user input symptoms. However, these systems face significant challenges when users input multiple symptoms within a single sentence, as seen in examples like "I have high fever, cold, and pain." The complexity of accurately parsing and interpreting natural language inputs increases when symptoms lack clear separation. Furthermore, these systems often struggle to capture the relationships between different symptoms, which is crucial for accurate disease prediction.

Y. Zhang and B. Liu's paper on "Semantic Text Classification of Disease Reporting" (2007) [5] introduces an innovative approach to disease prediction. They trained a model that leverages both word-level features and semantic information to enhance accuracy in predicting infectious diseases, achieving notable results in their study.By incorporating sentence-level semantics into the model, Zhang and Liu were able to capture nuanced relationships between words and extract deeper contextual meaning from text data. This integration of semantic features alongside traditional word-based features significantly improved the model's ability to classify disease-related text accurately.

This is another paper [4] focuses on acquiring and discovering medical knowledge embedded in clinical narrative reports. The authors apply a Natural Language Processing (NLP) system called MedLEE to extract and encode clinical entities from narrative clinical reports obtained from New York-Presbyterian Hospital (NYPH). They specifically focus on disease-symptom associations. The evaluation indicates an overall recall of 90 and a precision of 92 for disease-symptom associations.

Petrov, Das, and McDonald's paper[1] underscore the crucial role of linguistic resources and multilingual tagging systems in advancing computational linguistics. These resources not only aid in tackling challenges associated with free word order languages but also pave the way for improved cross-linguistic analysis and the development of more robust natural language processing solutions.

6 BASELINE



```
symptoms.......runny_nose,redness_of_eyes,throat_i
rritation,continuous_sneezing,chest_pain
=================predicted disease============
Migraine
=================description==================
Migraine is a type of headache that often involves
severe pain and sensitivity to light and sound.
=================precautions==================
1:  meditation
2:  reduce stress
3: Use Polaroid glasses in the sun
4:  consult doctor
=================medications==================
5 :  ['Analgesics,' 'Triptans,' 'Ergotamine
derivatives,' 'Preventive medications,'
'Biofeedback']
=================workout==================
6:  Identify and avoid trigger foods
7:  Stay hydrated
8:  Include magnesium-rich foods
9:  Consume omega-3 fatty acids
10:  Limit caffeine and alcohol
11:  Consume riboflavin-rich foods
12:  Limit processed foods
```

```
13:  Maintain regular meal times
14:  Consult a healthcare professional
15:  Manage stress
=================diets==================
16 :  ['Migraine Diet,' 'Low-Tyramine Diet,'
'Caffeine withdrawal,' 'Hydration,'
```

Figure 5: Result of Baseline

In the baseline, we defined the problem correctly, collected the data required and preprocessed it and did eda to get the insights of the data. We applied multiple machine learning models to determine which works best for our needs. Evaluated and finalized SVC. Then we made to for query process. Certainly. Initially, we are also working with the structured format where we take the symptoms from the user one by one and add them to the list. After preprocessing and correcting each symptom input using a spell checker, we transform them into TF-IDF vectors and we also transform the disease symptoms vector to tf idf vector.

The TF-IDF formula is given by:

$$\text{TF-IDF}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D)$$

where

- $\text{tf}(t, d)$ is the term frequency of term t in document d,
- $\text{idf}(t, D)$ is the inverse document frequency of term t in the document set D.

The term frequency $\text{tf}(t, d)$ is typically calculated using the formula:

$$\text{tf}(t, d) = \frac{\text{Number of times term } t \text{ appears in document } d}{\text{Total number of terms in document } d}$$

The inverse document frequency $\text{idf}(t, D)$ is calculated as:

$$\text{idf}(t, D) = \log\left(\frac{\text{Total number of documents in } D}{\text{Number of documents containing term } t}\right) + 1$$

Now we calculate the cosine similarity between the symptoms vector and the disease symptoms corpus vector and return the top result.

The cosine similarity between two vectors $\mathbf{v}$ and $\mathbf{w}$ is given by:

$$\text{Cosine Similarity}(\mathbf{v}, \mathbf{w}) = \frac{\mathbf{v} \cdot \mathbf{w}}{\|\mathbf{v}\| \|\mathbf{w}\|}$$

where $\mathbf{v} \cdot \mathbf{w}$ represents the dot product of vectors $\mathbf{v}$ and $\mathbf{w}$, and $\|\mathbf{v}\|$ and $\|\mathbf{w}\|$ denote the Euclidean norms of vectors $\mathbf{v}$ and $\mathbf{w}$, respectively.

After predicting the disease from the symptoms provided by the user, we extract detailed descriptions of the disease and list out precautions, including actions to take and avoid, from our stored database.

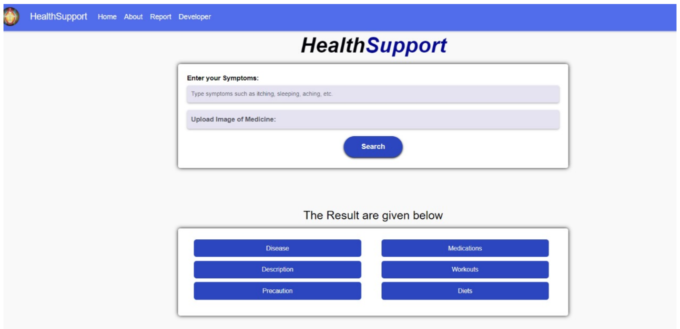
7 NOVELTY

The system features a user-friendly interface allowing users to input symptoms in free-text format, providing flexibility and convenience in describing their health issues. Based on the symptoms provided, the system generates comprehensive health recommendations encompassing medication, diet, and workout plans tailored to the user's specific needs.

Moreover, the platform employs advanced analysis techniques, leveraging Gemini AI to delve deeper into the provided symptoms. Gemini AI offers valuable insights into the advantages and disadvantages of medications and treatments, aiding users in making informed decisions about their healthcare.

Additionally, the system supports OCR-based image analysis, enabling users to upload images of medication labels for evaluation. Through Optical Character Recognition (OCR), the system extracts and analyzes information from the images, facilitating quick and accurate assessment of medication details.

By integrating symptom-based diagnostics, personalized treatment recommendations, and image-based analysis, the platform serves as a comprehensive health resource for users. It offers a holistic approach to healthcare management, empowering users to take control of their health and make well-informed decisions regarding their treatment and lifestyle choices.



HealthSupport

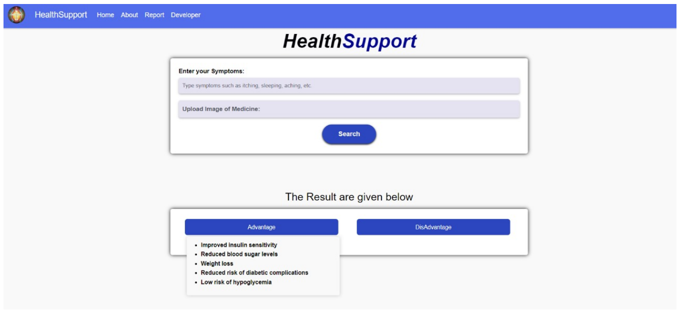
Enter your Symptoms:

Type symptoms such as itching, sleeping, aching, etc.

Upload Image of Medicine:

The Result are given below

Figure 6: Image of Medical Support



HealthSupport

Enter your Symptoms:

Type symptoms such as itching, sleeping, aching, etc.

Upload Image of Medicine:

The Result are given below

- Improved insulin sensitivity
- Reduced blood sugar levels
- Weight loss
- Reduced risk of diabetic complications
- Low risk of hypoglycemia

Figure 7: Image Retrieval

8 RESULTS

Our health support system begins with users inputting symptoms via free-text, utilizing advanced NLP techniques like Named Entity Recognition (NER) for symptom extraction. Rigorous data cleaning ensures accuracy, followed by mapping diseases to corresponding symptoms for precise prediction. The Support Vector Classifier (SVC) algorithm is the core, learning symptom-disease patterns for accurate predictions. Recommendations are then generated, covering diagnosis, personalized diet/workout plans, medications, and precautions. Integration of Gemini Pro AI enables in-depth medication analysis, supplemented by OCR for medication label insights. In summary, our system offers a holistic health management solution, encompassing symptom extraction, disease prediction, personalized recommendations, medication analysis, and OCR-based insights.

REFERENCES

[1] Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A Universal Part-of-Speech Tagset. *arXiv preprint arXiv:1104.2086* (2012). http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf

[2] Dr. D. P. Shukla, Kumar Sen, and Shamsher Bahadur Patel. 2013. A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy. *International Journal Of Engineering And Computer Science* 2 (2013), 2663–2671.

[3] Md. Tahmid, Md. Tahmid, A. Raihan, and N. Rashid. 2016. Automated Disease Prediction System (ADPS): A User Input-based Reliable Architecture for Disease Prediction. *IJCA* 133, 15 (Jan. 2016), 24–29. https://doi.org/10.5120/ijca2016908193

[4] Xue Wang, Arthur Chused, Noémie Elhadad, Carol Friedman, and Marianthi Markatou. 2008. Automated knowledge acquisition from clinical narrative reports. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, 783–787.

[5] Yi Zhang and Bing Liu. 2007. Semantic text classification of disease reporting. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Amsterdam, The Netherlands) *(SIGIR '07)*. Association for Computing Machinery, New York, NY, USA, 747–748. https://doi.org/10.1145/1277741.1277889