Early Disease Detection through Machine Learning

Kundan Sai Chowdary Sannapaneni*, Deepak Kandikattu *, Sai Preetham Vinnakota *, Osita Onyejekwe*
*Department of Data Science, University of Colorado Boulder, CO, USA

Abstract—In the contemporary world of constant rush, health is the least bothering topic for most people however, negligence in this respect can have some unfathomable consequences. In order to detect various diseases at the first stage; however big or small, it is very important to know about them. Diagnose as we all know initially depends on the symptoms of any disease however, symptoms of various diseases are much similar. We have done our research in the same domain, i.e. Early disease detection using an algorithmic approach by analyzing symptoms and comparing them with those of diseases to find patterns of specific disease and detect it earlier than normal. Our project is based on developing a model of Machine Learning, that uses random forest, multinomial naive Bayes, decision trees, and logistic regression to detect diseases from the given set of symptoms, we have used the dataset of disease symptoms which we have trained on a model and then predicted results based on the dataset.

Index Terms—Logistic Regression, Prediction, Smart Healthcare

1 Introduction

The healthcare sector is currently experiencing a period of significant change and growth. This is characterized by a noticeable increase in medical facilities and a surge in the amount of data collected from treatments for a wide range of diseases [1], [2]. The data, which is both extensive and varied, holds the key to potentially groundbreaking improvements in how we care for patients [2]. As healthcare facilities expand, they are equipped with advanced technology and systems capable of collecting and storing more patient information than ever before [3]. This includes data from clinical trials, patient medical histories, treatment outcomes, and even real-time health monitoring. The richness and diversity of this data provide an unparalleled opportunity for healthcare professionals and researchers to gain deeper insights into disease patterns, treatment effectiveness, and patient outcomes. Analyzing this data effectively is a critical step towards unlocking its potential. Through careful examination and interpretation, healthcare professionals can identify trends and correlations that were previously unnoticed. For instance, they might discover that certain treatments are more effective for specific patient demographics, or identify early warning signs of a disease that could lead to earlier intervention [4]. The impact of this data-driven approach in healthcare is far-reaching. It can lead to more personalized patient care, where treatments are tailored to the individual needs and characteristics of each patient. It also holds the promise of improving the overall quality of healthcare services, making them more efficient and effective. For patients, this means better health outcomes, quicker recovery times, and potentially lower healthcare costs. Furthermore, this abundance of data supports continuous learning and improvement in the healthcare sector. As more data is collected and analyzed, it leads to a cycle of ongoing enhancement in patient care strategies and treatment methodologies. Healthcare professionals are better equipped to make informed decisions, backed by concrete data, leading to better patient care and overall health management [5].

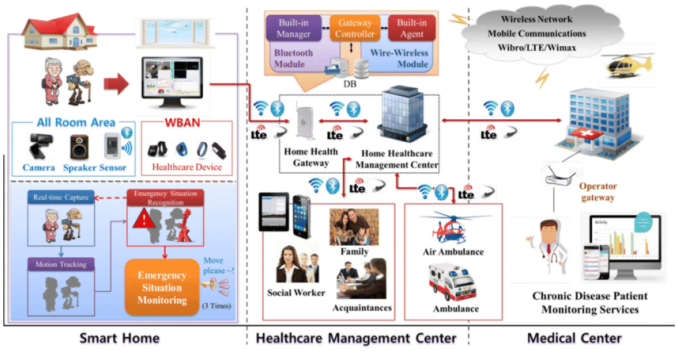

Fig. 1. Conceptual Overview of Smart Healthcare

Our initiative is focused on enhancing healthcare services by integrating the insights of applied sciences with the advancements in computer science. We employ data mining and machine learning techniques to manage and interpret the large, complex datasets that healthcare institutions gather [6]. These techniques are vital in sifting through vast amounts of data to unearth valuable information that might otherwise remain hidden. The use of machine learning is particularly transformative in our project [7]. By applying these algorithms to healthcare data, we aim to simplify and improve the accuracy of disease prediction. This advancement is not just about identifying diseases; it's about breaking down the barriers that currently make disease prediction challenging. Machine learning's ability to analyze patterns and make predictions can significantly aid in early disease detection. This early detection is crucial as it allows for timely intervention, potentially reducing the severity of diseases and improving patient outcomes. Our project specifically targets the development of a predic-

tive model for two prevalent diseases: Diabetes and Heart Disease [8]. These conditions were chosen due to their widespread impact and the critical need for early detection and intervention. The model we're developing relies on a large dataset, which undergoes a rigorous process of data cleansing and processing. This process is essential to ensure the accuracy and reliability of our model. By cleaning and organizing the data, we remove any errors or irrelevant information, which in turn enhances the performance of our machine learning algorithms. In essence, our project is guided by the principle that "prevention is better than cure." By focusing on early detection through predictive analysis, we can provide patients with timely and effective treatment options. This approach not only helps in managing the diseases more effectively but also plays a crucial role in improving the overall quality of life for patients. Our goal is to leverage the power of technology and data to create a healthcare system that is more responsive, accurate, and patient-focused.

1.1 Challenges in Tabular Data Inference

One significant challenge in high-dimensional classification is known as the **Curse of Dimensionality.** This term refers to the problems that arise when the number of features in the dataset is very high. As features increase, the data becomes spread out over an increasingly large space. Imagine having to find specific points in a vast room instead of a small box; it's much harder because there's more space to cover. This spread leads to sparsity, where data points are far apart from each other, making it difficult to find patterns or relationships between them. Another problem related to having so many features is the risk of overfitting. Overfitting is like memorizing answers to a test rather than understanding the subject. If a model is overfitted, it performs very well on the data it was trained on but fails to do well on new, unseen data. It learns the noise and random details in the training data instead of the actual patterns. Handling high-dimensional data also requires a lot of computational power and memory. This means that both building and using these models can be resource-intensive and time-consuming. Furthermore, in high-dimensional scenarios, not every feature is equally important. Some might not contribute much to the classification task and can even make things more confusing by introducing irrelevant information. It's like having too many clues in a mystery novel, where some of them lead nowhere. To manage these challenges, machine learning experts use various techniques:

1) **Regularization:** This is like adding a penalty for complexity. It discourages the model from becoming too complicated and focuses on the main patterns.
2) **Feature Selection:** This involves choosing only the most relevant features for the task. It's like picking the right tools from a toolbox; unnecessary ones are left out.
3) **Dimensionality Reduction:** Techniques like Principal Component Analysis (PCA) help reduce the number of features while keeping the essential information. It's like summarizing a long book into key points without losing the main story.
4) **Cross-validation:** This is a method to ensure the model works well on different sets of data, not just the one it was trained on.

2 Methodology

2.1 Data preprocessing

Our dataset, crucial for the study of diseases and their symptoms, has been assembled from a variety of reliable sources, each contributing unique and valuable information. Initially, we started with a dataset from **Columbia University**'s website, which included basic data on diseases and their associated symptoms. However, we realized that this dataset was somewhat limited, covering only around 1300 diseases and 22 symptoms. To expand our dataset, we turned to additional sources. We explored two different websites to enrich our dataset. The first was **Notemyhealthcare.com**, but it mostly contained information on diseases that were already present in our initial dataset. The more significant contribution came from the **National Health Portal of India**, overseen by the **Lloyd's Register Foundation** at the Centre for Health Informatics. This source was particularly valuable as it is known for its reliability and comprehensive coverage of diseases prevalent in India. From here, we could gather extensive disease data, which we then supplemented with a predefined list of symptoms to create a more robust dataset.

To collect detailed symptom information, we employed a tool called **Google-search-script**. This allowed us to scrape symptom data efficiently. A notable contribution came from the diseases page on Wikipedia, which lists over 100 diseases along with their symptoms. Our script was designed to navigate these Wikipedia pages, extract symptom information, and link them to the respective diseases. Once we had gathered all this data, our next step was to organize it effectively. We structured the dataset into two columns: "Diseases" and "Symptoms". We then undertook the meticulous task of combining diseases with their corresponding symptoms. This process included creating all possible combinations of symptoms for each disease. However, just combining this data was not enough. We needed to refine it further to ensure its usefulness. We focused on separating combinations of symptoms more clearly and removing any duplicate entries. To enhance the accuracy of our dataset, we introduced a method for comparing the similarity of symptoms between different diseases. We used the **Jaccard similarity coefficient**, setting a threshold of 0.75. This meant that for two diseases to be considered similar in our dataset, their symptoms had to have at least a 75% match. Additionally, we utilized synonyms and the wordnet tool from Princeton University to identify and include various similar words. This helped in recognizing and linking diseases that might have been listed under different names or described with slightly different symptoms.

2.2 User Symptom Preprocessing

In our group project, we have developed a methodical approach to process and analyze symptom data provided by users. This process ensures that the symptom data is accurately interpreted and utilized for disease prediction.

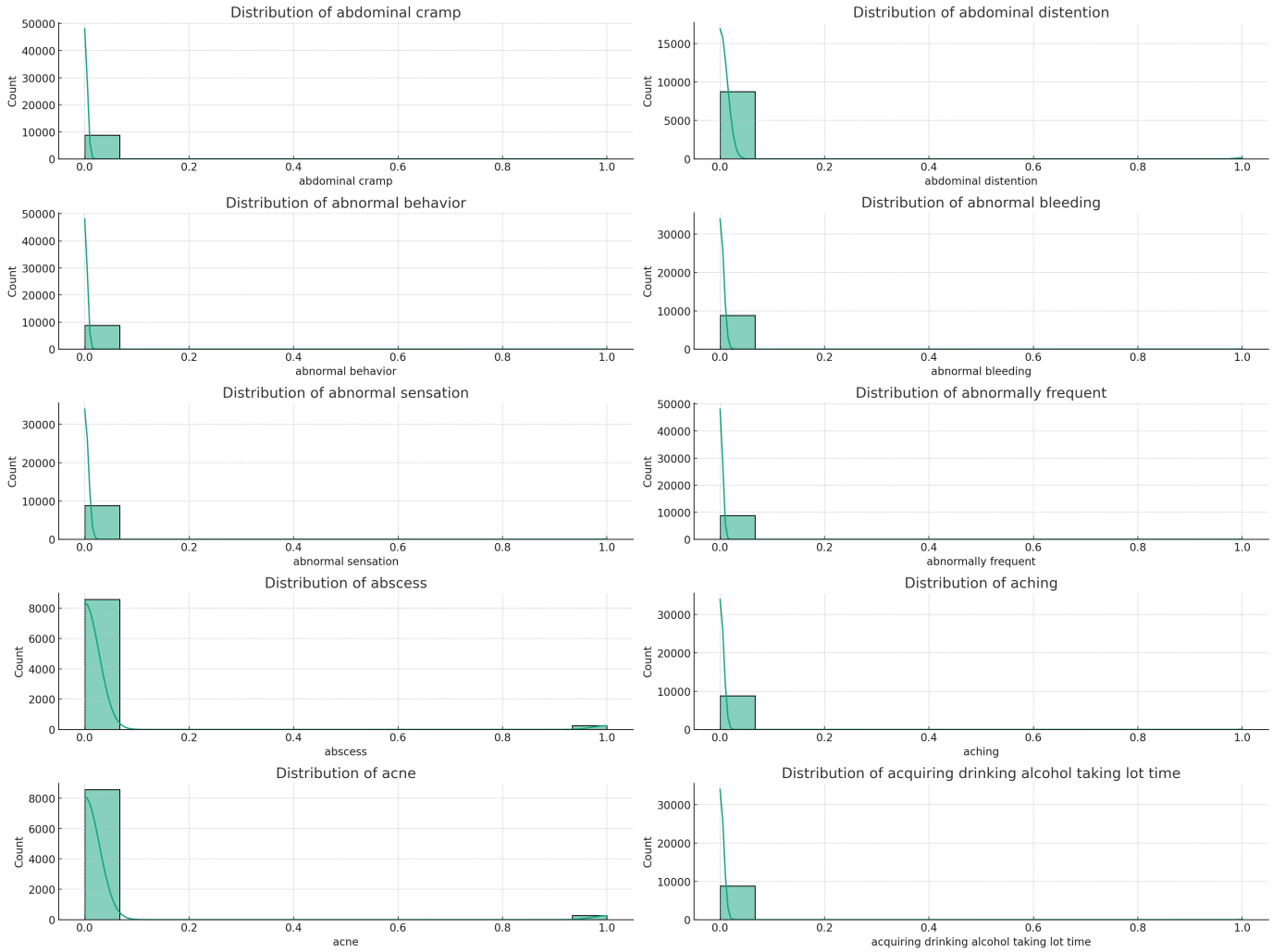

Fig. 2. Distribution of data in key columns

Here's a step-by-step breakdown of how we handle this data:

1) **Symptom Separation:** When users submit their symptoms, they typically list them in a sentence, separated by commas. Our first step is to split this sentence into individual symptoms using the commas as dividing points.
2) **Converting to Lowercase:** To maintain uniformity, we convert all the symptoms into lowercase. This avoids any discrepancies that might arise from different users using different cases (like uppercase, lowercase, etc.).
3) **Removing Stop Words:** Certain common words, known as stop words, often do not add significant meaning to the symptoms (such as 'and', 'or', etc.). We remove these words to focus on the more meaningful parts of the symptoms.
4) **Tokenizing Symptoms:** We then tokenize the symptoms. This process involves removing any punctuation and breaking down the text into smaller units, or tokens, making it easier to analyze.
5) **Lemmatization of Tokens:** Lemmatization is a process where we convert words to their base or root form. For instance, "running" would be lemmatized to "run." This helps in making the symptom data more consistent and easier to compare.

After processing the symptoms in this manner, we then expand the input. Here's how:

- For every symptom entered by a user, we add a list of synonyms to create a broader query. This is done to ensure that we don't miss any similar symptoms in our dataset.
- We tokenize the symptoms in our dataset and compare them with the user's expanded query. If a symptom from our dataset has a similarity score of more than 0.75 (indicating a 75% similarity) with the user's query, we consider it a match.

Next, we present the users with these matched symptoms from our dataset and invite them to select the ones that best describe their condition. Based on their selections, we also suggest additional symptoms that are commonly associated with the ones they have chosen. Finally, for disease prediction:

- We use the final list of symptoms selected by the user to create specific data vectors. These vectors are

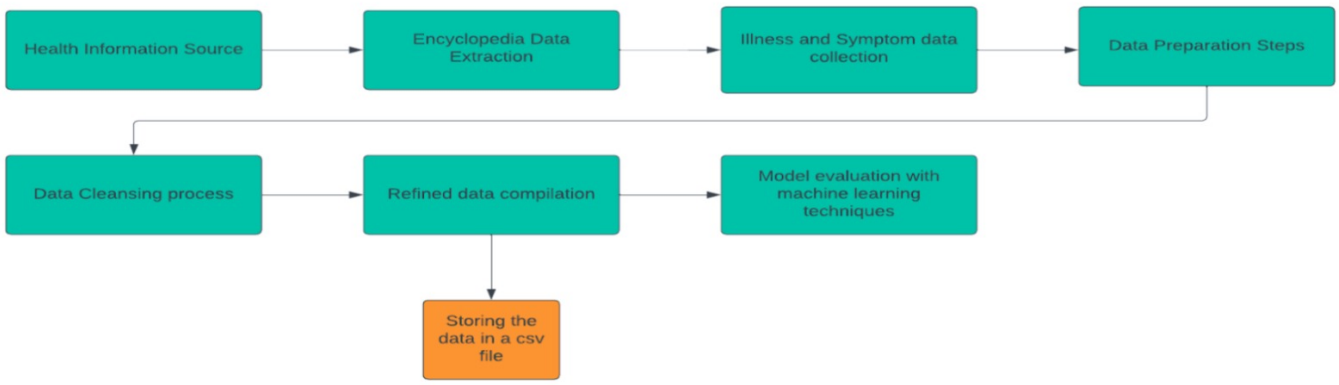

Fig. 3. Data Acquisition Flowchart

structured representations of the symptoms, formatted in a way that our disease prediction model can interpret.

- These vectors are then inputted into our disease prediction model. The model analyzes these vectors and outputs a list of possible diseases that correlate with the symptoms, ranked by the likelihood or probability of each disease being the correct diagnosis.

In essence, our approach is designed to be thorough, user-friendly, and efficient, ensuring that we can provide users with accurate disease predictions based on their reported symptoms.

2.3 Training the model

In our project, we have a specific way of training our model to predict diseases based on symptoms reported by users. Here's a more detailed explanation of our process:

1) **Creating Binary Vectors for Symptoms:** When users provide their symptoms, we represent each symptom as a binary vector. This means we use '1' to indicate that a symptom is present and '0' to show it's not. For example, if someone has a fever and cough, but no headache, and we're considering these three symptoms, their vector might look like [1, 1, 0] [9].
2) **Training Machine Learning Models:** We use these binary vectors to train various machine learning models. We've found that Logistic Regression, a statistical method for predicting binary outcomes, is often the most effective for our needs. This model looks at the symptoms (represented by the binary vectors) and learns to predict the likelihood of various diseases based on these symptoms [10].
3) **Using Cosine Similarity and TF-IDF Scores:** To enhance our disease prediction, we also employ techniques like cosine similarity and TF-IDF (Term Frequency-Inverse Document Frequency) scores. TF-IDF is a statistical measure used to evaluate the importance of a word (in our case, a symptom) in a document (here, a list of diseases). It helps us understand which symptoms are more significant for certain diseases.
4) **Calculating Scores:** We calculate TF-IDF scores for both the symptoms in our dataset and those in the users' queries. This helps us identify which diseases are most relevant based on the symptoms.
5) **Using Cosine Similarity for Disease Prediction:** Cosine similarity is used to compare the similarity between two vectors. In our case, it helps us compare the disease vectors and the user's symptom vectors. We create matrices (tables) with diseases and symptoms based on TF-IDF scores and then calculate the cosine similarity. Diseases with the highest similarity scores to the user's symptoms are considered the most likely diagnoses.
6) **Providing Detailed Disease Information:** If users want to know more about any of the diseases we've identified, we have a system to provide up-to-date information. We use a script to scrape Wikipedia in real time, ensuring that the information we provide is current [11]. This includes details about the disease's symptoms, causes, complications, risk factors, diagnostic methods, and treatment options.

2.4 Evaluation

We used various machine learning models, including Logistic Regression, Decision Trees, Random Forests, and Multinomial Naive Bayes. These models are tools that help us sort through patient symptom data to find patterns that might indicate specific health conditions [12]. A key part of our approach was using a technique called 5-fold cross-validation to check how well our models were working. Cross-validation is a method used to test the accuracy of a model. In 5-fold cross-validation, the data is divided into five parts. Each part, in turn, is used as a test set while the model is trained on the remaining four parts. This process is repeated five times, with each part being used as a test set once. This method is really useful because it helps reduce bias in the model – it makes sure that every bit of data is used for both training and testing the model [13].

Accuracy: Accuracy is a measure of how often the model is correct. It is the ratio of the number of correct predictions to the total number of input samples. It is calculated as follows:

$$Accuracy = \frac{NumberofCorrectPredictions}{TotalNumberofPredictions} \quad (1)$$

Recall: Recall, also known as sensitivity or the true positive rate, is the measure of our model correctly identifying True Positives. Thus, it is the ratio of the number of True Positives to the number of actual total positives (True Positives + False Negatives). The formula for recall is:

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \quad (2)$$

Cross Validation: Cross-validation is a statistical method used to estimate the skill of machine learning models. It is commonly used in applied machine learning to compare and select a model for a given predictive modeling problem because it is less biased than other methods.

3 Results and Experimental Analysis

In the results and experimental analysis section, we focus on two parts: performance metrics as described above and secondly the predictive modeling of all the diseases present in the dataset. For the first part, Our research revealed some interesting results. The Multinomial Naive Bayes model sees a modest drop off from its overall accuracy of 86.07% to an 84.50% accuracy. This shows it's fairly accurate overall but it is the least accurate of the models. The Random forest model does see a more significant drop of 5.39% from 92.31% overall to 86.92% shows there's potentially overfitting in the random forest model however it still performs quite well. The Logistic Regression appears to be the most consistent and generalizable out of the models with an 89.19% cross validation accuracy showing only a slight drop off from the overall accuracy indicating that the model may predict well on new data. The Decision tree is the most heavily overfitted model with a cross validation accuracy of 83.62% from an overall accuracy of 91.97% [14]. It seems that the decision tree is very good at predicting the initial data set it's shown however it may not necessarily predict well on new unseen data. Overall, the logistic regression is the model that offers the best balance of complexity and generalization and so may be the most accurate with new datasets out of the models tested.

TABLE 1
Performance metrics for machine learning models

Model	Accuracy	Recall	Cross Validation Accuracy
Multinomial Naive Bayes (MNB)	85.29%	85.29%	84.50%
Random Forest (RF)	92.31%	92.31%	86.92%
Logistic Regression (LR)	91.74%	91.74%	89.19%
Decision Tree (DT)	91.97%	91.97%	83.62%

Enhancements to the system now allow for greater adaptability when users input a list of symptoms by introducing a feature that proposes associated co-occurring symptoms. Previously, detailed information on potential illnesses and recommended treatments was not included. By integrating this additional data, the model has evolved to provide a more comprehensive medical advisory platform. The results derived from the predictive modeling suggest that the application of our model in symptom-disease inference is robust and can offer valuable insights. The model successfully identified a list of potential conditions based on the user's reported symptoms, with Influenza being a highly probable diagnosis given the presence of fever, coughing, and headache. The ability of the model to associate these symptoms with respiratory infections underscores its potential utility in aiding preliminary self-diagnosis. The additional information provided by the model about Influenza, such as its symptoms, transmission, prevention, and treatment options, can serve as an informative resource for users. It enhances the model's practicality by not only predicting possible illnesses but also educating the user about them. It is, however, crucial to emphasize that the predictions made by the model should be interpreted with caution. While the model can suggest probable diagnoses, it does not replace the expertise of a healthcare professional. Users are encouraged to seek medical advice for a definitive diagnosis and treatment plan, especially if symptoms are severe or persistent.

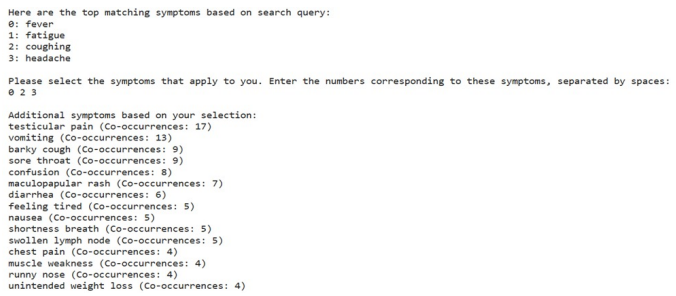

Fig. 4. Output 1

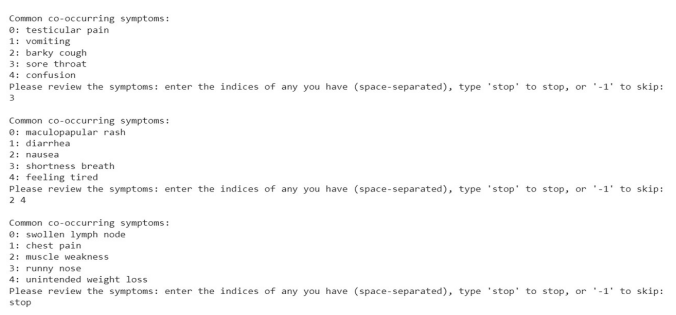

Fig. 5. Output 2

4 Future Work and Summary

Our future endeavors for this project are set to significantly broaden its scope and enhance its precision. We are poised to fortify the dataset by incorporating medical history, which is instrumental in providing context to current symptoms. The integration of demographic details such as age and gender is anticipated to sharpen the accuracy of our diagnostic outcomes. Moreover, we are equipped to delve into the realm of predictive analytics by examining the interplay between pre-existing medical conditions and the emergence of new ailments [15]. This analysis is pivotal in lending depth to our predictions and, as we refine our model with each iteration, our confidence in its reliability swells. We are also exploring

```
Top 5 diseases predicted based on symptoms
0 Disease name: Rubella, Probability: 66.89%
1 Disease name: Influenza, Probability: 66.89%
2 Disease name: Hepatitis E, Probability: 44.6%
3 Disease name: Hepatitis D, Probability: 44.6%
4 Disease name: Fibromyalgia, Probability: 44.6%

More details about the disease? Enter index of disease or '-1' to discontinue and close the system:
1

Influenza
Other names - flu, the flu, grippe (French for flu)
Specialty - Infectious disease
Symptoms - Fever , runny nose , sore throat , muscle pain , headache , coughing , fatigue
Usual onset - 1–4 days after exposure
Duration - 2–8 days
Causes - Influenza viruses
Prevention - Hand washing , flu vaccines
Medication - Antiviral drugs such as oseltamivir
Frequency - 3–5 million severe cases per year
Deaths - >290,000–650,000 deaths per year
```

Fig. 6. Output 3

the potential to venture into dermatological diagnostics by employing image analysis. This could revolutionize the way skin diseases are identified, offering an additional layer of validation through visual examination. Such a feature would not only facilitate the diagnostic process but also enrich the doctor's understanding of the patient's immediate and historical conditions. In addressing pragmatic needs, we envision the system to dispense preliminary diagnosis tips based on recurrent patient inputs. Harnessing data processing capabilities, the system could suggest over-the-counter remedies like ibuprofen for common complaints such as headaches, albeit with a clear disclaimer about the limitations inferred from the dataset. To encapsulate, this project represents a seminal step in the fusion of medical science and machine learning, enabling disease prediction with burgeoning accuracy. Each facet of the project is designed with the potential for refinement, ensuring that the diagnoses provided to patients are continually optimized.

REFERENCES

[1] S. Tian, W. Yang, J. M. Le Grange, P. Wang, W. Huang, and Z. Ye, "Smart healthcare: making medical care more intelligent," *Global Health Journal*, vol. 3, no. 3, pp. 62–65, 2019.

[2] H. Yin, A. O. Akmandor, A. Mosenia, N. K. Jha, *et al.*, "Smart healthcare," *Foundations and Trends® in Electronic Design Automation*, vol. 12, no. 4, pp. 401–466, 2018.

[3] M. M. Baig and H. Gholamhosseini, "Smart health monitoring systems: an overview of design and modeling," *Journal of medical systems*, vol. 37, pp. 1–14, 2013.

[4] J. Soni, U. Ansari, D. Sharma, S. Soni, *et al.*, "Predictive data mining for medical diagnosis: An overview of heart disease prediction," *International Journal of Computer Applications*, vol. 17, no. 8, pp. 43–48, 2011.

[5] T. Bhanuteja, K. V. N. Kumar, K. S. Poornachand, C. Ashish, and P. Anudeep, "Symptoms based multiple disease prediction model using machine learning approach," *International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN*, pp. 2278–3075.

[6] A. S. Kwekha-Rashid, H. N. Abduljabbar, and B. Alhayani, "Coronavirus disease (covid-19) cases analysis using machine-learning applications," *Applied Nanoscience*, vol. 13, no. 3, pp. 2013–2025, 2023.

[7] K. Sujatha, K. K. Kishore, B. S. Rao, and R. Rajasekaran, "Diabetes disease prediction based on symptoms using machine learning algorithms," *Annals of the Romanian Society for Cell Biology*, vol. 25, no. 6, pp. 3805–3817, 2021.

[8] G. Choudhary and S. N. Singh, "Prediction of heart disease using machine learning algorithms," in *2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)*, pp. 197–202, IEEE, 2020.

[9] R. Keniya, A. Khakharia, V. Shah, V. Gada, R. Manjalkar, T. Thaker, M. Warang, and N. Mehendale, "Disease prediction from various symptoms using machine learning," *Available at SSRN 3661426*, 2020.

[10] D. J. Park, M. W. Park, H. Lee, Y.-J. Kim, Y. Kim, and Y. H. Park, "Development of machine learning model for diagnostic disease prediction based on laboratory tests," *Scientific reports*, vol. 11, no. 1, p. 7567, 2021.

[11] S. Grampurohit and C. Sagarnal, "Disease prediction using machine learning algorithms," in *2020 International Conference for Emerging Technology (INCET)*, pp. 1–7, IEEE, 2020.

[12] D. Dahiwade, G. Patle, and E. Meshram, "Designing disease prediction model using machine learning approach," in *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 1211–1215, IEEE, 2019.

[13] Y. Deepthi, K. P. Kalyan, M. Vyas, K. Radhika, D. K. Babu, and N. Krishna Rao, "Disease prediction based on symptoms using machine learning," in *Energy Systems, Drives and Automations: Proceedings of ESDA 2019*, pp. 561–569, Springer, 2020.

[14] K. Arumugam, M. Naved, P. P. Shinde, O. Leiva-Chauca, A. Huaman-Osorio, and T. Gonzales-Yanac, "Multiple disease prediction using machine learning algorithms," *Materials Today: Proceedings*, vol. 80, pp. 3682–3685, 2023.

[15] K. N. R. Challa, V. S. Pagolu, G. Panda, and B. Majhi, "An improved approach for prediction of parkinson's disease using machine learning techniques," in *2016 international conference on signal processing, communication, power and embedded system (SCOPES)*, pp. 1446–1451, IEEE, 2016.

YouTube Link Click Here