Disease Diagnosis Assistant

Amit Pareek, Mayur N Sastry

November 13, 2024

1 Introduction

The Disease Diagnosis Assistant is a simple tool designed to support doctors in diagnosing diseases and recommending treatment steps. By interacting with doctors to gather symptom information, the assistant helps narrow down potential diagnoses and suggests relevant next steps for treatment. The aim is to aid early diagnosis by leveraging a machine learning model for symptom analysis and reliable resources for treatment guidance.

2 Literature Review

3 Description of Data

4 Process Workflow

1. Symptom Collection and Model Training:

The assistant is trained on a dataset containing symptoms and related diseases using a Random Forest classifier to predict disease probabilities based on symptom patterns.

The Random Forest classifier is an ensemble machine learning model that builds multiple decision trees during training. Each tree makes predictions

based on random subsets of the data and features, and the final prediction is determined by aggregating the outputs of all trees. This ensemble approach enhances the model's accuracy and reduces the risk of overfitting, making Random Forest a popular choice for classification tasks.

The predict probabilities method of Random Forest provides the probability of each class for a given input. Instead of a direct class prediction, predict probabilities method outputs a probability distribution across all possible classes, showing how confident the model is for each class label. For instance, if predicting between two diseases, predict probalities method might output values like [0.7, 0.3], indicating a 70 percent likelihood of the first disease and 30 percent for the second. This method is valuable in cases where uncertainty or degree of confidence in the predictions is important, such as in medical diagnosis.

2. Symptom Input and Correlation:

The doctor begins by entering the main symptom. The assistant then suggests related symptoms that may correlate with the main symptom, which the doctor can confirm or rule out. There are two ways to get related symptoms from the main symptom. One is to use cooccurence, the other is to use correlation.

The cooccurence between symptoms is calculated as follows:

Let $A_{ij} = 1$ if patient i has symptom j, and 0 if patient i does not have symptom j. The cooccurence between symptom j and symptom k is calculated as:

$$\text{Cooc}(j,k) = \frac{|\{i : A_{ij} = A_{ik}\}|}{8835}$$

This measures the proportion of patients who either have both symptoms or neither, capturing the relationship between j and k. Note that 8835 is the number of patient records in the dataset.

This cooccurence formula can also be extended to multiple symptoms.
Suppose we have symptoms $j_1, \ldots, j_n$ and want to find the correlation of $\{j_1, \ldots, j_n\}$ with another symptom k.
For each patient i, define $N(i) = |\{m : A_{ij_m} = A_{ik}\}|$, which represents the number of symptoms in $\{j_1, \ldots, j_n\}$ matching symptom k. The final formula is given by:

$$\text{Cooc}(j_1, \ldots, j_n, k) = \frac{\sum N(i)}{8835n} = \frac{\sum |\{m : A_{ij_m} = A_{ik}\}|}{8835n}$$

Using this formula, we can determine, for example, the symptoms cooccuring with the combination of fever and cough.

3. Probability Update:

Each symptom added updates the disease probabilities using the Random Forest model's predict probabilities method, progressively refining the potential diagnoses.

In simple terms, the random forest classifier in this system first takes initial symptoms, such as fever and headache, and outputs prediction probabilities for possible diseases. For example, it might indicate a 75% probability of COVID-19, with the remaining 25% probability distributed among other diseases.

Next, we enhance the prediction by identifying the symptom most closely

associated with the existing ones—in this case, perhaps cough. This is done using the previous correlaion analysis. Including this additional symptom in the classifier allows it to update the probabilities. Now, the prediction for COVID-19 might increase to 90%.

Ultimately, once all relevant symptom information has been collected, the disease with the highest prediction probability is determined as the likely diagnosis.

4. Symptom Mapping:

To account for variations in symptom naming, the assistant uses semantic similarity to match doctor-inputted symptoms to those in the dataset, ensuring accurate data interpretation.

For example, considering the "coughing with blood" symptom in the dataset, we would like "blood while coughing" to have a higher semantic similarity score than merely "coughing".

To do this, we use embeddings generated by a Sentence Transformer model to capture the semantic meaning of each phrase, and cosine similarity then quantifies how closely these meanings align in vector space. This approach allows for a more nuanced similarity measure than simple keyword matching, as it captures context and relationships between words.

Results:
Semantic similarity(coughing with blood, blood while coughing) = 0.97196984
Semantic similarity(coughing with blood, coughing) = 0.7482288

5. Treatment Recommendation:

After determining the most probable disease, the assistant uses Retrieval Augmented Generation (RAG) to offer reliable treatment recommendations. This method combines retrieval and generation to give the doctor quick access to accurate, up-to-date medical advice. Here, RAG helps by first searching trusted sources like NHS Inform for treatment options related to the identified condition. The assistant retrieves this relevant information and then uses its language generation capabilities to present the details in a clear, context-appropriate manner, summarizing or rephrasing where needed.

Results (For Hepatitis B):
As a doctor, you can follow these steps for the patient's Hepatitis B treatment: Prescribing medications to help with itchiness, nausea, or vomiting, if needed. Providing advice on painkillers such as paracetamol or ibuprofen for any aches and pains. Recommending a cool, airy environment, loose clothing, and smaller, lighter meals to help reduce feeling sick and vomiting. Advising on the importance of getting plenty of rest. Providing guidance on when to seek medical advice, such as if symptoms get worse or haven't started to improve.

5 Future Improvements

1. Incorporating Additional Health Data:
Including broader health metrics, like age, blood sugar levels, blood pressure, heart rate, and other general health parameters, can provide a more comprehensive basis for predictions. These metrics are crucial because they often correlate strongly with the presence and severity of various diseases. For example, chronic illnesses like diabetes, cardiovascular disease, and hypertension are often linked to these parameters. By integrating this data, the model could improve the accuracy of disease probability predictions, allowing it to identify high-risk cases more precisely.

2. Integrating Explainable AI:
Adding Explainable AI capabilities can make the model's decision-making process more transparent and interpretable for medical professionals. In disease prediction, explainable AI techniques such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) could highlight which symptoms and health parameters contributed most to the prognosis. This transparency can help healthcare providers trust and validate the tool's recommendations, as it provides insights into why certain diseases are being flagged.

6 Conclusion

In this project, a disease diagnosis assistant was developed to help streamline the initial prognosis process for doctors. By using a machine learning model, specifically a Random Forest classifier, the system predicts the probable disease based on symptoms provided by the user. Through the use of semantic similarity models, the assistant can understand various symptom descriptions, even if they differ slightly from those in the dataset. Additionally, treatment steps are generated through retrieval-augmented generation (RAG), accessing reliable resources such as NHS Inform to provide accurate and comprehensive treatment information.

This project has demonstrated the potential for AI-driven tools to support medical professionals in diagnosis and treatment planning, with future improvements likely to increase its effectiveness and trustworthiness in real-world applications.

7 References

1. K. Gaurav, A. Kumar, P. Singh, A. Kumari, M. Kasar, T. Suryawanshi, "Human Disease Prediction using Machine Learning Techniques and Real-life Parameters."
2. NHS Inform, "A to Z Illnesses and Conditions," available at: `https://www.nhsinform.scot/illnesses-and-conditions/a-to-z/`, for treatment information on various common diseases.