Medical Expert System For Disease Prediction

Ritesh Kumar Gupta[1], Pankti Salvi[2], Sneha Choudhary[3], Gaurav Singh[4], Tushar Kumar[5]

Department of Information Technology.
Indian Institute of Information Technology, Allahabad.

Contributing authors: iit2021135@iiita.ac.in; iit2021134@iiita.ac.in; iit2021132@iiita.ac.in; iit2021120@iiita.ac.in; iit2021203@iiita.ac.in;

Abstract

The Medical Expert System is a culmination of efforts to predict the probability of occurrence of diseases and recommend suitable medications based on reported symptoms. Designed to counter prevalent challenges in healthcare such as limited access and frequent misdiagnoses, this system emphasizes the necessity for accurate and prompt disease identification.

The inception of this project stemmed from motivations centered around enhancing healthcare accessibility, minimizing misdiagnoses, educating users, optimizing resource allocation, and fostering continuous learning within the medical realm. Notably, the implementation journey involved an iterative process, initially exploring decision tree and random forest methods.

However, due to the high model complexity encountered with these approaches, the system transitioned to employing rule-based methods. This strategic shift allowed for a streamlined, interpretable, and efficient framework, overcoming complexities observed in earlier methods.

Keywords: Medication recommendation, symptoms-based diagnosis, healthcare accessibility, misdiagnosis prevention, iterative development, decision tree, random forest, rule-based methods, model complexity.

1 Introduction

The burgeoning field of medical informatics has witnessed remarkable advancements in recent years, revolutionizing the landscape of healthcare delivery. In the pursuit of enhancing healthcare accessibility, accuracy, and timely disease identification, the development of a Medical Expert System has emerged as a pivotal endeavor. This project represents a concerted effort to harness technological innovations in predicting diseases and recommending appropriate medications based on reported symptoms. Rooted in the recognition of pervasive challenges within healthcare systems, such as limited access, frequent misdiagnoses, and the critical need for prompt disease identification, the Medical Expert System serves as a beacon of hope in overcoming these hurdles.

The genesis of this project was motivated by a multidimensional approach focused on broadening healthcare access, reducing misdiagnoses, educating users about potential ailments, optimizing resource utilization, and fostering a continuous learning paradigm in the medical domain. The system's development journey encompassed a meticulous exploration of various methodologies, initially delving into complex decision tree and random forest methods. However, encountering challenges associated with high model complexity, the project pivoted towards a pragmatic approach, leveraging rule-based methods to streamline the system's framework.

This introduction sets the stage for a comprehensive exploration of the Medical Expert System, chronicling its evolution, methodologies employed, and its crucial role in addressing critical healthcare challenges.

2 Literature Review

The extraction of pertinent information from Electronic Health Records (EHRs), clinical notes, and discharge summaries has been a focal point in healthcare, serving as a pivotal input for physicians in disease diagnosis. Initially, knowledge bases were pivotal in this extraction process. However, contemporary advancements have witnessed a shift towards the utilization of rule-based learning, machine learning (ML), and deep learning (DL) concepts to extract and diagnose diseases automatically. This transition signifies a paradigm shift in leveraging sophisticated methodologies to enhance diagnostic accuracy and efficiency.

Electronic Medical Records (EMRs) and Electronic Health Records (EHRs) offer benefits like instant access, comprehensive patient tracking, and enhanced healthcare quality. Yet, challenges like system downtime, technical expertise limitations, communication gaps, and data security concerns persist. Transitioning to digital records has eased access to patient information but extracting specific data from extensive records remains a cumbersome task, emphasizing the need for automated disease diagnosis systems.

To address these challenges, healthcare systems have adopted Natural Language Processing (NLP)-based computational phenotyping, using rule-based, machine learning, and deep learning methods to extract insights from raw text. These systems aim to streamline data analysis and aid in disease diagnosis by transforming textual records into valuable information, particularly notable in Case-Based Reasoning (CBR) Systems. This paper presents a comprehensive survey of automatic disease diagnosis techniques from electronic records, providing an overview of different systems, methodologies, advantages, limitations, current trends, and future directions.

3 Proposed Methodology

Throughout the development of our medical expert system, we embarked on a comprehensive journey exploring varied methodologies to enhance disease prediction and recommendation accuracy. Initially, our pursuit led us through the terrain of decision tree and random forest models. The decision tree approach offered a structured framework, yet as our dataset expanded, the resulting complexity at times hindered its efficacy. Subsequently, leveraging the random forest ensemble method revealed promising outcomes, reducing overfitting concerns encountered with decision trees(Fig. 4). However, the amplified model complexity persisted, prompting a strategic pivot.

Acknowledging the need for a more streamlined and interpretable model, we transitioned towards a rule-based system. This marked a crucial turning point in our implementation strategy, as rule-based methods facilitated a more intuitive framework. The simplicity and transparency inherent in rule-based systems not only addressed the complexity challenges but also aligned closely with the interpretability crucial in medical decision-making processes.

3.1 Data Analysis

Utilizing the powerful functionalities offered by Matplotlib, Pandas, and Seaborn libraries, our exploration into visual representations delved into a diverse array of graphical depictions.

Among these were horizontal bar graphs meticulously crafted to showcase symptom frequencies across a spectrum of diseases. Heatmaps emerged as valuable tools, vividly illustrating the presence or absence of symptoms within each disease profile. Additionally, intersection matrix(Fig. 2) plots proved instrumental in elucidating overlaps between symptoms and diseases, offering nuanced insights into their intricate relationships.

In our quest for comprehensive visualization, vertical bar (Fig. 3) graphs took center stage, effectively communicating the symptom count associated with each disease. Meanwhile, the intuitive depiction offered by pie charts facilitated a clear understanding of disease or symptom distributions, thoughtfully categorized by their respective target organs.

To unravel complex interconnections, we engineered network graphs meticulously linking target organs, diseases, and symptoms. These network graphs served as visual bridges, enabling a holistic comprehension of the intricate web of relationships among these crucial medical elements.

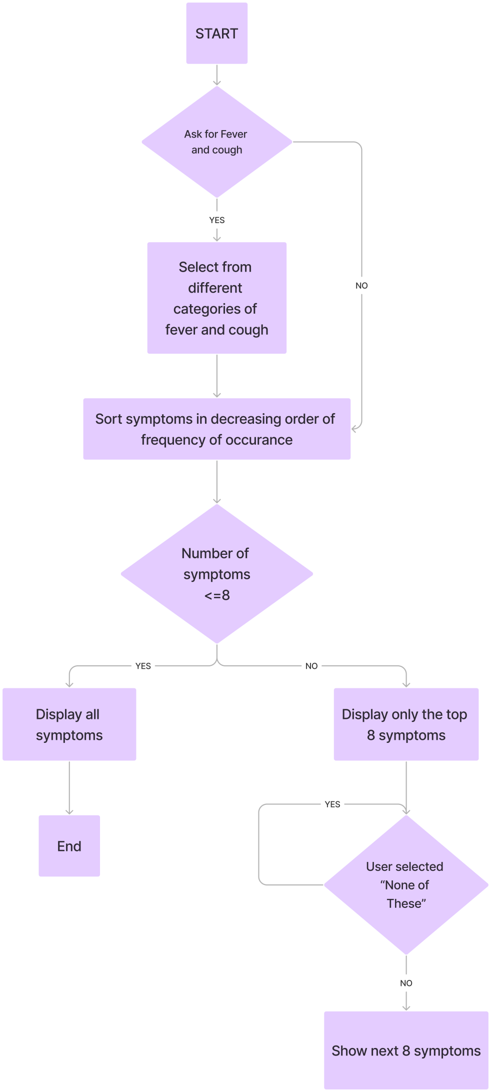

START
Ask for Fever and cough
YES
NO
Select from different categories of fever and cough
Sort symptoms in decreasing order of frequency of occurance
Number of symptoms <=8
YES
NO
Display all symptoms
Display only the top 8 symptoms
End
YES
User selected "None of These"
NO
Show next 8 symptoms

Fig. 1: Flowchart of our system [1]

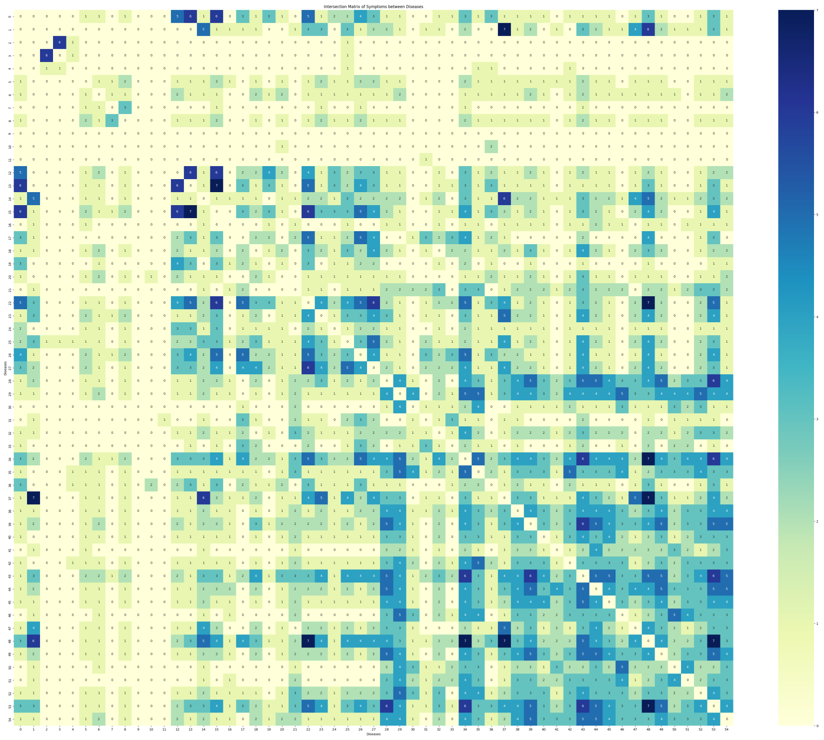


Fig. 2: Intersection Matrix Plot (Symptoms between Diseases)

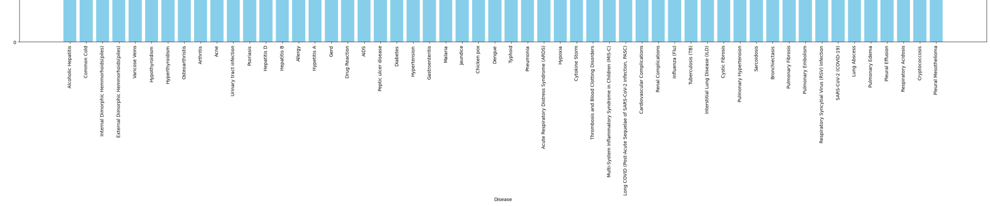


Fig. 3: Vertical Bar Graph Plot (Number of Symptoms vs Diseases)

3.2 Using Decision Tree:

The decision tree, while initially promising, resulted in an excessively deep tree structure, reaching 19 levels. This intricate depth raises several concerns. Firstly, the model's complexity escalates significantly, potentially leading to overfitting, where the model learns too much from the training data and loses generalizability. Secondly, interpretability becomes challenging with such a deep tree, making it harder to comprehend the decision-making process.

In response to these challenges, the exploration of random forest—an ensemble learning technique—has emerged. Random forest excels in mitigating overfitting by aggregating multiple decision trees and averaging their predictions. By introducing randomness during tree construction and feature selection, random forest diversifies the individual trees, leading to a more robust and less prone-to-overfitting model. Moreover, while maintaining high predictive performance, random forest often retains a level of interpretability that is more manageable compared to a deeply intricate decision tree.

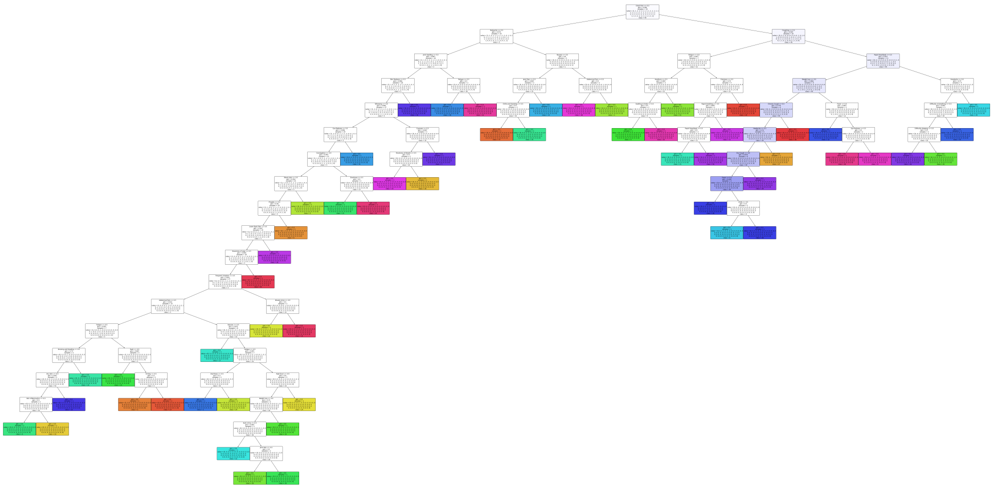


Fig. 4: Decision tree of depth 19 constructed for our dataset

3.3 Using Random Forest:

Indeed, transitioning from a decision tree to a random forest marked progress in mitigating complexity, yet the random forest's inherent intricacy, albeit reduced from the initial decision tree, persisted as a concern. This lingering complexity drove the need for a more refined and interpretable solution, culminating in a strategic turn towards rule-based implementation.

The pivot to a rule-based system signifies a deliberate shift towards prioritizing simplicity and interpretability without sacrificing predictive accuracy. Rule-based approaches, renowned for their structured and transparent nature, present a distinct advantage in medical applications. These systems operate on a framework of explicit rules, fostering a clear understanding of the decision-making process, a critical requirement in healthcare settings.

3.4 Using Rule-Based Implementation:

In our rule-based implementation(Fig. 1), user interaction optimization takes center stage, driven by the objective of minimizing query count while progressively narrowing down potential outcomes[2]. This streamlined approach strategically guides users through a series of questions tailored to efficiently refine the search space. By dynamically structuring questions and continuously adapting search criteria based on user inputs, we ensure each query significantly reduces the pool of potential diseases.

Initially, the system presents the top 8 symptoms most commonly associated with various diseases, offering users a starting point sorted by prevalence. As users make selections aligning with these symptoms, subsequent iterations reveal the top 8 symptoms linked to diseases featuring the chosen symptom. This iterative process significantly reduces the search space, focusing on symptoms closely tied to the user's selections and refining potential disease options progressively.

Furthermore, our diagnostic system calculates probabilities associated with specific diseases based on reported symptoms, aiding in evaluating illness severity. When any calculated probability exceeds or equals 0.5, the system presents the top 5 probabilities to the user. Conversely, if all probabilities remain below this threshold, the system displays all potential probabilities. This comprehensive approach provides users with vital insights into the likelihood and severity of their condition, empowering them to make informed decisions regarding their healthcare journey. Through these streamlined interactions and informative outputs, our system aims to offer users a more intuitive and insightful diagnostic experience.

4 Dataset Generation

The dataset creation process underwent a meticulous journey starting with an expansive compilation of 132 diseases associated with 189 symptoms. These diseases encompassed a wide spectrum, touching various physiological systems such as respiratory, ocular, auditory, hepatic, and systemic conditions. To capture their relationship, a binary system represented symptom presence or absence, offering a straightforward framework for each disease-symptom profile.

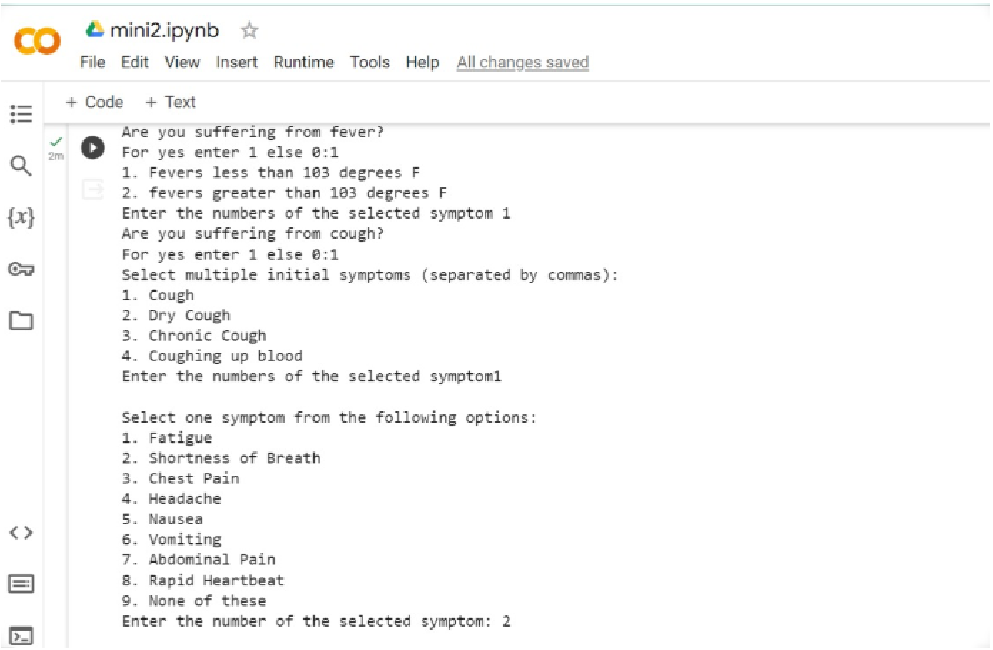
Transitioning from this comprehensive dataset, our focus honed in on a refined subset, emphasizing respiratory ailments alongside post-COVID conditions and sequelae. This subset expanded to include common diseases and systemic conditions, augmenting the dataset's breadth. Notably, an upgrade in the coding system occurred, introducing a nuanced approach. This revised system utilized values of 0, 1, 2, and 3, denoting probability levels reflecting symptom manifestation within each disease profile. This adaptation aimed to offer a more intricate representation by assigning varying probability values to symptoms, enriching the dataset for heightened analytical depth and precision.

The dataset's creation was a rigorous process, commencing with exhaustive information compilation from diverse sources, including robust online searches and reference materials. To ensure data accuracy and reliability, a stringent cross-verification process ensued. This verification phase involved meticulous comparisons[3] and validations against multiple reputable datasets[4], guaranteeing the dataset's fidelity and reinforcing its integrity as a reliable reference in the realm of medical diagnostics and analysis.

5 Results

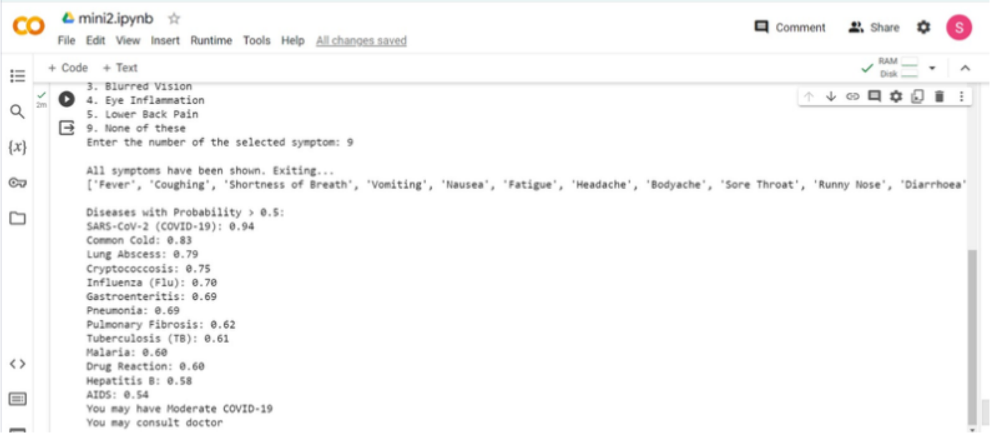
The medical expert system operates through a multifaceted approach to disease prediction and management. Firstly, it employs a dynamic method to predict potential diseases based on user-input symptoms(Fig. 5), calculating probabilities for various conditions. To streamline the process, the system employs a threshold-based strategy. If any disease surpasses a probability threshold of 0.5, the top 5 probable diseases are presented to the user, aiding in focused consideration or further investigation. In cases where predicted probabilities for diseases remain below the threshold, the system generates and displays a comprehensive list encompassing all potential diseases(Fig. 6). Although these conditions might be less probable based on the symptoms entered, this comprehensive list provides a broader view for consideration.

Furthermore, the system specifically addresses COVID-19 cases by categorizing their severity into low, medium, and high levels. This classification allows for a more nuanced approach to managing cases based on their severity levels. For cases classified as low severity, the system provides primary medication recommendations, including general advice, over-the-counter medications, and tailored self-care instructions designed for managing mild cases. This holistic approach combines probabilistic disease prediction with severity-based categorization, offering tailored guidance and recommendations to users based on the presented symptoms and the severity of the identified condition, especially in the context of COVID-19.



```
Are you suffering from fever?
For yes enter 1 else 0:1
1. Fevers less than 103 degrees F
2. fevers greater than 103 degrees F
Enter the numbers of the selected symptom 1
Are you suffering from cough?
For yes enter 1 else 0:1
Select multiple initial symptoms (separated by commas):
1. Cough
2. Dry Cough
3. Chronic Cough
4. Coughing up blood
Enter the numbers of the selected symptom1

Select one symptom from the following options:
1. Fatigue
2. Shortness of Breath
3. Chest Pain
4. Headache
5. Nausea
6. Vomiting
7. Abdominal Pain
8. Rapid Heartbeat
9. None of these
Enter the number of the selected symptom: 2
```

Fig. 5: Taking input of symptoms from user



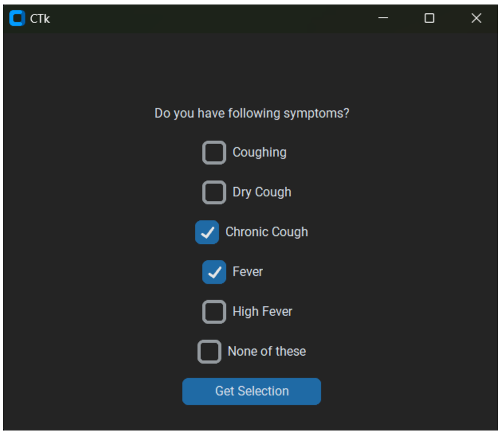
```
3. Blurred Vision
4. Eye Inflammation
5. Lower Back Pain
9. None of these
Enter the number of the selected symptom: 9

All symptoms have been shown. Exiting...
['Fever', 'Coughing', 'Shortness of Breath', 'Vomiting', 'Nausea', 'Fatigue', 'Headache', 'Bodyache', 'Sore Throat', 'Runny Nose', 'Diarrhoea'

Diseases with Probability > 0.5:
SARS-CoV-2 (COVID-19): 0.94
Common Cold: 0.83
Lung Abscess: 0.79
Cryptococcosis: 0.75
Influenza (Flu): 0.70
Gastroenteritis: 0.69
Pneumonia: 0.69
Pulmonary Fibrosis: 0.62
Tuberculosis (TB): 0.61
Malaria: 0.60
Drug Reaction: 0.60
Hepatitis B: 0.58
AIDS: 0.54
You may have Moderate COVID-19
You may consult doctor
```

Fig. 6: Probability of occurrence of a particular disease

5.1 GUI Implementation

The GUI implementation(Fig. 7) of our project utilizing the Tkinter toolkit has been a pivotal aspect of our Medical Expert System's user interface. Tkinter, being a powerful and intuitive library for building graphical user interfaces in Python, has allowed us to craft an interactive and user-friendly platform.

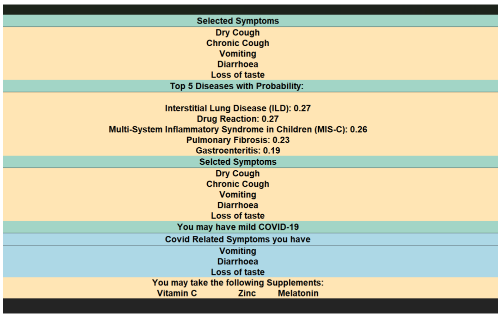
The GUI design encompasses strategically positioned input fields, buttons, and interactive elements, providing a streamlined pathway for users to enter their symptoms and receive real-time feedback(Fig. 8).This GUI implementation enhances the accessibility and usability of our Medical Expert System, ensuring a smooth and intuitive interaction between users and the system's diagnostic functionalities.



Do you have following symptoms?

☐ Coughing
☐ Dry Cough
☑ Chronic Cough
☑ Fever
☐ High Fever
☐ None of these

Fig. 7: Taking input of symptoms from user using GUI



```
Selected Symptoms
Dry Cough
Chronic Cough
Vomiting
Diarrhoea
Loss of taste
Top 5 Diseases with Probability:

Interstitial Lung Disease (ILD): 0.27
Drug Reaction: 0.27
Multi-System Inflammatory Syndrome in Children (MIS-C): 0.26
Pulmonary Fibrosis: 0.23
Gastroenteritis: 0.19
Selcted Symptoms
Dry Cough
Chronic Cough
Vomiting
Diarrhoea
Loss of taste
You may have mild COVID-19
Covid Related Symptoms you have
Vomiting
Diarrhoea
Loss of taste
You may take the following Supplements:
Vitamin C          Zinc      Melatonin
```

Fig. 8: Output produced by the GUI Implementation

6 Conclusion

In essence, our medical expert system operates through a synergy of rule-based algorithms and probability assessments, adeptly analyzing user-input symptoms to swiftly pinpoint potential diseases. Beyond this, the system goes a step further by categorizing the severity of COVID-19 cases and tailoring recommendations accordingly. For instances classified as low severity, the system offers guidance on primary medications, ensuring users have access to initial treatment options.

However, for cases falling within the medium to high severity spectrum, the system strongly emphasizes the urgency of seeking professional healthcare guidance. This dual approach aims to streamline the initial assessment process, empowering users with preliminary insights while underscoring the critical importance of consulting healthcare professionals, especially in situations demanding immediate attention and expert intervention. Ultimately, this technology represents a harmonious blend of automated assessment and human-centered care, striving to provide timely support while prioritizing the necessity of professional guidance in critical healthcare scenarios.

7 Future Scope

To bolster our medical expert system's diagnostic capabilities, we aim to expand the disease database by encompassing a wider spectrum of ailments across various organs and bodily systems. This expansion will enable a more comprehensive assessment of symptoms, ensuring a broader range of potential diseases are considered during diagnosis.

Additionally, to enhance the accuracy of disease prediction, we plan to integrate medical imaging data such as X-rays, MRI scans, and laboratory reports into our machine learning algorithms. This integration will provide a more holistic approach to diagnosis, leveraging visual and quantitative data for a more refined and accurate prediction of diseases.

Moreover, recognizing the importance of tailored healthcare, we aim to focus on specific demographics within our system. For pediatric healthcare, our goal is to develop specialized algorithms adept at predicting diseases specific to children, considering their unique physiology and health challenges.

Addressing women's health concerns, particularly gynecological conditions and reproductive health, is another key aspect of our strategy, aiming to better serve the specific healthcare needs of women. Furthermore, our system will be tailored to detect age-related illnesses, geriatric syndromes, and chronic conditions prevalent in the elderly, catering to the distinctive healthcare requirements of the elderly population. By incorporating these focused demographic approaches, our system aims to provide more precise, personalized, and effective healthcare solutions across diverse age groups and specific health concerns.

8 Acknowledgements

We express our heartfelt gratitude to Vrijendra sir and Trapti ma'am for their unwavering guidance, steadfast support, and continuous motivation throughout our project journey. Their invaluable insights have served as a beacon, guiding us through the intricate implementation phases of our medical expert system.

Furthermore, our sincere appreciation extends to the esteemed medical professionals: **Dr. Shambhavi Kumari**, [M.B.B.S.] , **Dr. Bijal Salvi**[M.B.B.S], **Dr. Praful Salvi**,[M.D. Anaesthesia], **Dr. Daraksha Naaz**,[M.B.B.S.], and **Dr. Rupali Singh**,[M.B.B.S.], each with their specialized expertise denoted by their qualifications. Their profound contributions were instrumental, providing profound feedback and invaluable guidance during the development of our disease-symptoms dataset.

Their meticulous assistance in assigning probabilities to individual symptoms has significantly enriched the quality and precision of our system. Their collective expertise has been a cornerstone in refining our system's accuracy, ensuring a more robust and reliable platform for assessing and predicting diseases based on user-input symptoms.

References

[1] Yadav, A.K., Shukla, R., Singh, T.R.: Machine learning in expert systems for disease diagnostics in human healthcare, pp. 179–200 (2021). https://doi.org/10.1016/B978-0-12-821777-1.00022-7

[2] Latif, J., Xiao, C., Tu, S., Rehman, S.U., Imran, A., Bilal, A.: Implementation and use of disease diagnosis systems for electronic medical records based on machine learning: A complete review (2020) https://doi.org/10.1109/ACCESS.2020.3016782

[3] Health Topics. https://www.msdmanuals.com/en-in/home/health-topics

[4] Nayak, R., Boloor, A.: Exam Preparatory Manual for Undergraduates—Medicine, 2nd edn. (2018). https://doi.org/10.5005/jp/books/18461