# DATA REPORT
# DOC_AI MEDICAL CHATBOT
## CAPSTONE PROJECT


GROUP 7

AUTHORS

1. FELIX MUSAU
2. VICTOR ONGAKI
3. ROSE MATOKE
4. DAISY KERUBO THOMAS
5. VICKER IVY MIMI

# CONTENTS

## EXECUTIVE SUMMARY

The DOC_AI project aims to develop a conversational medical chatbot to enhance healthcare accessibility and efficiency by accurately classifying patient diseases based on reported symptoms.

The chatbot is designed to function as a 24/7 virtual health assistant to provide preliminary health guidance, symptom analysis, and efficient case routing in a hospital setting.

## BUSINESS UNDERSTANDING & OBJECTIVES

### BUSINESS PROBLEM

Hospitals face challenges with overcrowded emergency room, limited consultation time, and high administrative burdens from non-critical inquiries. Patients often struggle to receive timely, reliable preliminary medical advice, leading to potential delays in care.

### MAIN OBJECTIVE

To develop and implement a conversational medical chatbot system that enhances healthcare guidance for patients and efficiently routes cases to the appropriate healthcare professionals by categorizing patient diseases based on their symptoms.

### SPECIFIC OBJECTIVES

- Provide 24/7 automated medical support for quick answers to patient queries as regards to symptoms.
- Reduce the workload of healthcare staff (nurses, doctors, receptionists) by automating routine inquiries.
- Enhance patient experience and engagement by delivering empathetic, accurate, and easy-to-understand responses.
- Efficiently collect patient information and symptoms to reduce waiting times.

### METRIC OF SUCCESS

The success of the disease prediction chatbot project was measured by its ability to deliver safe and responsible preliminary diagnoses, quantified by achieving high Recall (patient safety) and high Precision (responsible advice). The final model chosen is based on the highest F1-score and the most trustworthy probability outputs i.e. lowest Log Loss.

### DATA LIMITATION

The dataset is large and noted as computationally expensive.

## DATA UNDERSTANDING

The data was sourced from Hugging Face; it is focused on diseases and their associated symptoms. It contains 246,945 rows and 378 columns (1 disease column and 377 symptom columns).

### DATA CLEANING AND FEATURE ENGINEERING

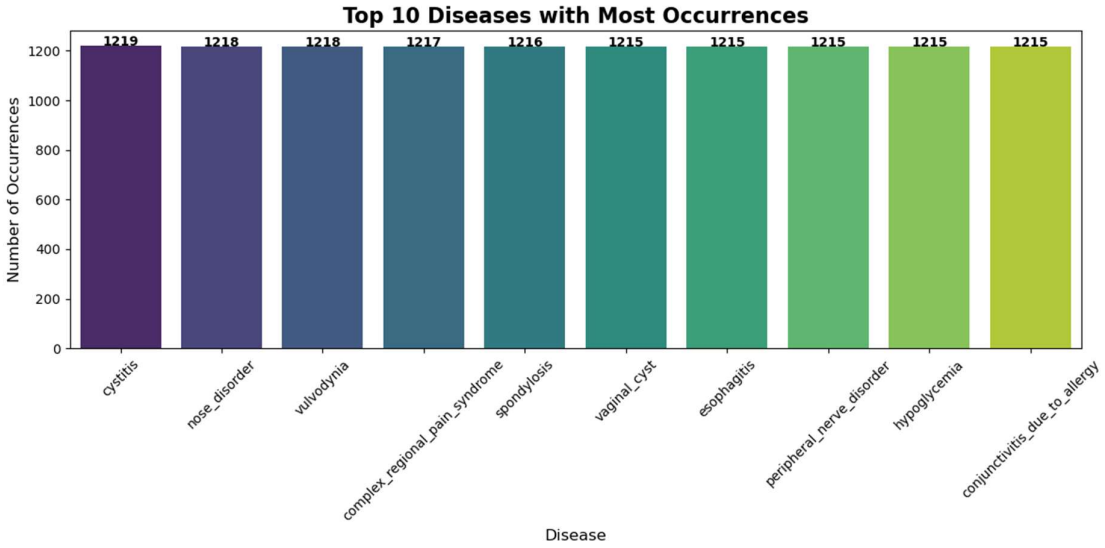
- Missing Values: No missing values were found.

- Standardization: The column names indicating symptoms were cleaned by converting them to lowercase and replacing special characters with underscores for uniformity i.e. 'anxiety and nervousness' became 'anxiety_and_nervousness'.

## EXPLORATORY DATA ANALYSIS (EDA)

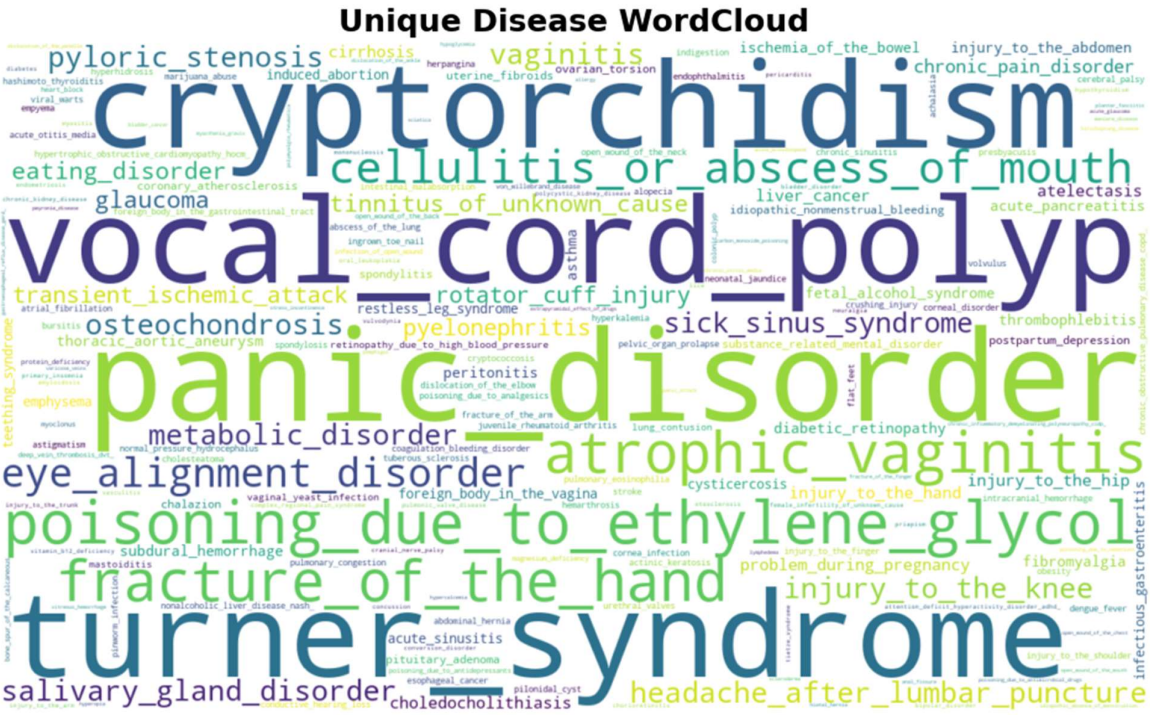
### DISEASE DISTRIBUTION

The initial dataset showed extreme imbalance, with Top 10 diseases having counts around 1,215–1,219 and the least occurring ones with very low counts, as low as 1.




The unique/rare conditions required careful handling like data augmentation and stratified sampling to prevent the model from ignoring them.

As per the word cloud the size of each word is proportional to the number of times a disease appears with the largest words representing the most frequent of the unique diseases.




## SYMPTOM CENTRALITY ANALYSIS

### 1. DEGREE CENTRALITY

These are symptoms that are most connected.

The top 5 symptoms are: Headache, Sharp abdominal pain, Vomiting, Dizziness, Cough. These are the most common symptoms, as they frequently occur alongside the largest number of other symptoms in the patient data.

### 2. BETWEENNESS CENTRALITY (BRIDGE)

Top 5 Symptoms: Headache, Sharp abdominal pain, Back pain, Abnormal appearing skin, Dizziness. These symptoms are crucial for differential diagnosis because they act as "bridges" connecting different, otherwise distinct, clusters of symptoms in the network.

### 3. CLOSENESS CENTRALITY (QUICKLY REACHABLE)

Top 5 Symptoms: Headache, Sharp abdominal pain, Vomiting, Dizziness, Cough. These symptoms are central to the overall network. Being quickly reachable from other symptoms, they are excellent indicators for early screening and initial assessment.

## SYMPTOMS CO OCCURRENCE

A Symptom Co-occurrence Network Analysis was done to identify and leverage distinct symptom clusters, these are groups of symptoms that frequently appear together. Symptoms don't appear randomly; they form predictable medically relevant groups. For instance, there was a tight cluster of headache, sharp abdominal pain, vomiting, dizziness and cough which suggests a shared root cause or a

recognized syndrome. Instead of learning these symptoms individually, our model recognizes this entire group as a single highly predictive feature.

The model was also trained to recognize isolated symptoms like *vaginal bleeding after menopause* as highly specific guiding it immediately toward rare targeted diagnoses. By training our model to recognize these pre-validated clusters, we significantly enhanced their ability to capture the underlying etiology of symptoms, making the DOC_AI chatbot a more accurate and efficient diagnostic assistant.

## DISEASE CLASSIFICATION

We grouped the diseases into different categories as per the table below

| | Category | Description | Diseases |
|---|---|---|---|
| 1 | Infectious Diseases | Caused by bacteria, viruses, fungi, or parasites. | tuberculosis, dengue fever, cryptococcosis, infectious gastroenteritis, hepatitis, pneumonia, abscess of lung, otitis media, conjunctivitis, cellulitis, infection of wound. |
| 2 | Respiratory Diseases | Affects the lungs and breathing process. | asthma, COPD, emphysema, chronic/acute sinusitis, bronchitis, pulmonary embolism, pulmonary congestion, atelectasis, thoracic injury, pulmonary eosinophilia. |
| 3 | Cardiovascular Diseases | Disorders of heart and blood vessels. | hypertension, ischemic heart disease, heart block, atrial fibrillation, pericarditis, coronary atherosclerosis, thoracic aortic aneurysm, deep vein thrombosis (DVT), hemorrhage. |
| 4 | Endocrine & Metabolic Disorders | Hormonal and metabolic imbalances. | diabetes, hypothyroidism, Hashimoto thyroiditis, obesity, hyperkalemia, hypokalemia, vitamin B12 deficiency, metabolic disorder, Cushing's syndrome, Addison's disease. |
| 5 | Neurological Disorders | Brain, nerve, and coordination disorders. | epilepsy, stroke, Parkinson's, multiple sclerosis, myasthenia, neuralgia, migraine, dizziness, myoclonus, restless leg syndrome, cerebral palsy, tic disorder, concussion. |
| 6 | Psychiatric / Mental Health Disorders | Mental, mood, or behavioral health conditions. | depression, anxiety, panic disorder, bipolar disorder, ADHD, psychotic disorder, substance-related mental disorder, insomnia, chronic pain disorder, eating disorder, stuttering. |
| 7 | Musculoskeletal Disorders | Disorders of bones, joints, or muscles. | spondylosis, spondylitis, osteochondrosis, bursitis, gout, arthritis, back pain, osteoporosis, bone spur, flat feet, Tietze syndrome. |
| 8 | Gastrointestinal Diseases | It affects the stomach, intestines, liver, or pancreas. | GERD, gastritis, ulcerative colitis, intestinal malabsorption, pancreatitis, choledocholithiasis, liver disease, cirrhosis, peritonitis, diarrhea, colonic polyp, pyloric stenosis. |
| 9 | Reproductive & Genitourinary Disorders | Disorders of reproductive or urinary organs. | endometriosis, vaginitis, vulvodynia, infertility, ovarian torsion, uterine fibroids, cryptorchidism, urethral valves, cystitis, pyelonephritis, bladder disorder, priapism. |
| 10 | Cancers / Neoplasms | Malignant and benign tumors. | liver cancer, breast cancer, leukemia, lymphoma, pituitary adenoma, melanoma, esophageal cancer, brain cancer. |
| 11 | Dermatological Disorders | Affect the skin, hair, or nails. | eczema, psoriasis, seborrheic dermatitis, seborrheic keratosis, actinic keratosis, alopecia, dermatitis, skin abscess, viral warts, fungal infection, hyperhidrosis. |

| | Category | Description | Diseases |
|---|---|---|---|
| 12 | Hematologic Disorders | Blood or bone marrow diseases. | anemia, von Willebrand disease, coagulation disorders, hemarthrosis, polycythemia and bleeding disorder. |
| 13 | Congenital & Genetic Disorders | Present from birth or inherited. | Down syndrome, Turner syndrome, tuberous sclerosis, spina bifida, cysticercosis, fetal alcohol syndrome, cryptorchidism. |
| 14 | Autoimmune Disorders | Immune system attacks body tissues. | lupus (SLE), rheumatoid arthritis, Hashimoto's thyroiditis, vasculitis, psoriasis. |
| 15 | Chronic Pain & Functional Disorders | Long-term pain and functional syndromes. | fibromyalgia, chronic back pain, irritable bowel syndrome (IBS), chronic fatigue, teething syndrome and pain disorder. |
| 16 | Injuries & Trauma | Physical injuries to tissues or bones. | fractures (arm, rib, hand), dislocations, crushing injuries, open wounds (neck, back, mouth), contusions, concussion, thoracic injury. |
| 17 | Pregnancy & Perinatal Conditions | Conditions affecting pregnancy or childbirth. | preeclampsia, gestational diabetes, ectopic pregnancy, postpartum depression, problem during pregnancy, induced abortion. |
| 18 | Eye Disorders | Eye and vision-related diseases. | glaucoma, cataract, corneal disorders, conjunctivitis, chorioretinitis, endophthalmitis, cornea infection. |
| 19 | Ear, Nose, Throat (ENT) | Infections or disorders related to the ear, nose, or throat. | otitis media, sinusitis, laryngitis, mastoiditis, otosclerosis, presbycusis, tinnitus, nose deformity, and sore in nose. |
| 20 | Other / Miscellaneous | Conditions that do not fit other categories. | poisoning (antidepressants, analgesics, ethylene glycol), allergic reaction, vitamin deficiencies, drug intoxication, withdrawal syndrome, hyperhidrosis. |

Infectious Diseases (25.9%) form the largest category, indicating their wide diversity and global prevalence. These diseases, caused by bacteria, viruses, fungi, and parasites, often evolve rapidly and affect multiple body systems, contributing to their dominance.

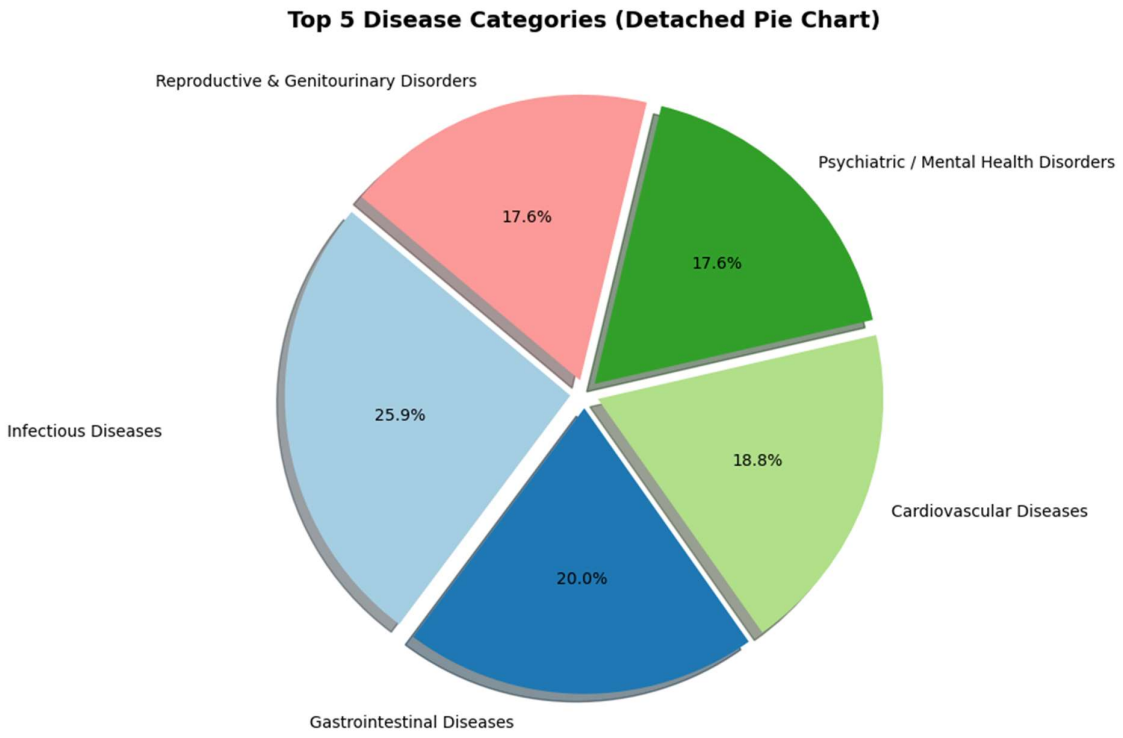
Gastrointestinal Diseases (20.0%) rank second, showing the significant impact of digestive system disorders such as ulcers and liver disease. Their high proportion reflects how lifestyle, diet, and infections contribute to widespread gastrointestinal problems.

Cardiovascular Diseases (18.8%) follow closely, representing the major global burden of heart and blood vessel disorders. This category's prominence aligns with the growing incidence of hypertension, stroke, and heart disease worldwide.

Psychiatric / Mental Health Disorders (17.6%) account for a considerable share, emphasizing the increasing recognition of mental health issues such as depression and anxiety as key components of global health.

Reproductive & Genitourinary Disorders (17.6%) also represent a significant portion, highlighting the diversity of conditions affecting reproductive and urinary health, including infertility and infections.

Overall, the top five categories show that diseases affecting essential body systems — immunity, digestion, circulation, mental health, and reproduction — dominate the global disease landscape.




.

## MODELING AND EVALUATION

Modeling focused on building a robust multi-class classification engine to predict the disease based on the combination of symptoms.

### DATA PREPARATION FOR MODELING

To tackle the data imbalance while managing computational complexity, the following steps were taken:

**Class Filtering**: Diseases with less than 800 samples were removed, reducing the number of classes from the initial count to 114 unique diseases. This resulted in a filtered dataset of 114,312 rows and 377 columns.

**Feature Preparation**: The 377 symptom features were verified as binary indicators and down cast from int64 to **int8** to optimize memory usage for large-scale training.

**Class Balancing**: To further address the remaining imbalance among the 114 classes, SMOTE was applied to the filtered data increasing the dataset size to 138,966 rows.

### DATA SPLITTING

The balanced dataset was split into Training, Validation, and Test sets with stratification to ensure consistent disease distribution across all subsets.

- **Training Set**: This set contains 97,276 rows, representing approximately 70% of the total data. Its primary purpose is for model fitting and learning, where the algorithm learns patterns from the data.
- **Validation Set**: This set contains 20,845 rows, making up approximately 15% of the total data. It is used for hyperparameter tuning and validation during the model development process to assess how well different model configurations perform.
- **Testing Set**: This set also contains 20,845 rows, approximately 15% of the total data. Its purpose is for the final, unbiased performance evaluation of the model after training and validation are complete.

## MODEL PERFORMANCE SUMMARY

Model Performance Comparison

| | Accuracy | Precision | Recall | F1-score | Log Loss |
|---|---|---|---|---|---|
| **Naive Bayes** | 0.8859 | 0.8954 | 0.8857 | 0.8869 | 0.5729 |
| **XGBoost** | 0.8847 | 0.8903 | 0.8843 | 0.8856 | 0.2732 |
| **Baseline Logistic Regression** | 0.8845 | 0.8903 | 0.8845 | 0.8852 | 0.2558 |
| **Word2Vec NN** | 0.8772 | 0.8883 | 0.8758 | 0.8740 | 0.2553 |
| **Random Forest** | 0.8728 | 0.8858 | 0.8728 | 0.8754 | 1.6225 |
| **Neural Network Classifier** | 0.8485 | 0.8883 | 0.8758 | 0.8740 | 0.3450 |

Six models were evaluated across five metrics. While several models achieved high F1-scores of 0.88 a critical divergence was observed in the Log Loss metric, which determines how trustworthy the models stated probability of disease is.

The Word2Vec NN was chosen for deployment because it's excellent performance on Log Loss guarantees that the chatbot's advice is clinically reliable and trustworthy, which is more valuable than the slight increase in F1-score.

## FEATURE IMPORTANCE

The feature-important analysis from the Word2Vec NN model provides insights into the most predictive symptoms.

**Dominant Feature:**

- "Headache" is the most important feature, with the highest average absolute weight (approximately 0.24).

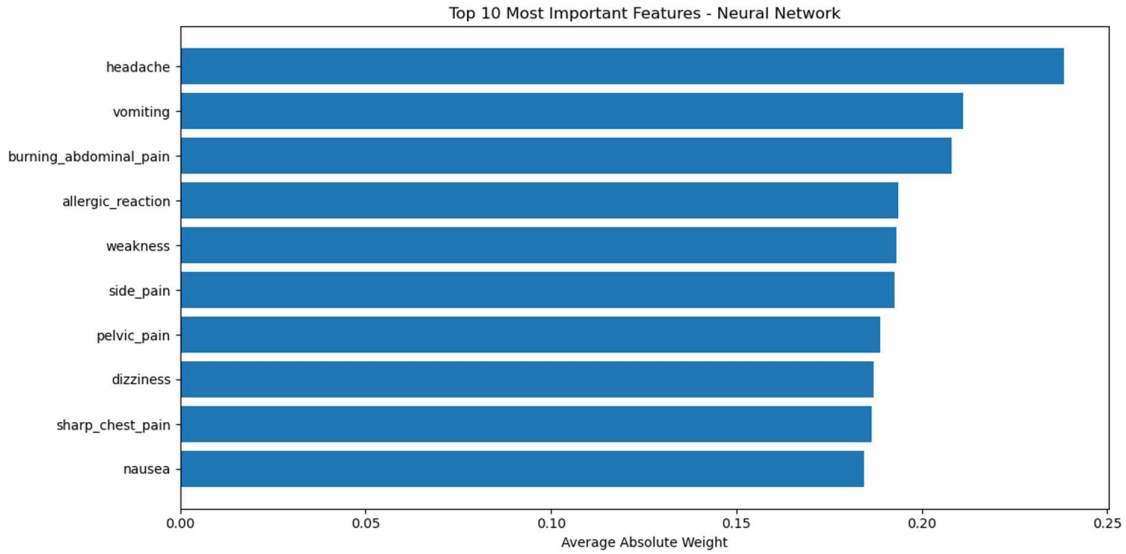
- This suggests that the presence or absence of a headache, and its specific value in the input data, is the single most crucial piece of information the neural network uses to differentiate between the 114 disease classes.

**Highly Influential Symptoms (Top 3)**

- The top three features—headache, vomiting, and burning_abdominal_pain—are significantly more influential than the remaining seven.
- These symptoms likely act as strong discriminators for several common or high-frequency disease classes.

**Cluster of General Symptoms (Bottom 7)**

- Features from "allergic reaction" down to "nausea" have very close average absolute weights, all hovering between approximately 0.17 and 0.21.
- These symptoms which include weakness, side pain, pelvic pain, dizziness, sharp_chest_pain, and nausea are considered important but have a more uniform impact on the model compared to the top three. Their presence contributes substantially, but none stand out dramatically from the others in this group.




## MODEL SELECTION

The Word2Vec NN is the recommended model because it offers the ideal blend of high predictive capability and unquestionable trustworthiness in its output, making it the safest and most reliable choice for a user-facing medical diagnosis application.

## CONCLUSION RECOMMENDATIONS

This project successfully developed and deployed a Word2Vec Neural Network (NN) model to predict potential diseases from user reported symptoms, utilizing over 250,000 symptom-disease records across more than 400 categories. The model proved highly effective, achieving an accuracy of 0.8772, while demonstrating a strong balance of precision and recall. A key strength of Word2Vec NN is it's ability to understand semantic relationships within text symptom data. Analysis highlighted headache, vomiting, and abdominal pain as the most influential diagnostic features across all models tested.

Also, dataset analysis revealed high symptom connectivity, with fever, cough and headache linking many diseases. Finally, the disease landscape reflects diversity, with Cystitis being the most prevalent individual condition, and Infectious diseases leading the overall category prevalence closely followed by Gastrointestinal and Cardiovascular issues.

## FUTURE WORK

1. Integrate with Trusted Medical Databases: Connect the chatbot to verified and reputable medical databases such as the World Health Organization (WHO) and MedlinePlus. This will ensure that the system stays up to date with the latest medical research, treatment guidelines, and disease information, improving the accuracy and reliability of diagnoses.

2. Develop a Mobile or Web-Based Application: Create a mobile or web version of the chatbot to enhance accessibility and user convenience. A user-friendly interface will allow patients and healthcare workers to interact with the system anytime and anywhere, promoting wider adoption and usability.

3. Incorporate Reinforcement Learning: Implement reinforcement learning techniques to enable the model to learn from user interactions and feedback over time. This will allow the chatbot to continuously refine its predictions and improve decision-making accuracy as it gathers more data from real-world use.

4. Design Adaptive Conversational Flows: Introduce a dynamic conversational structure that allows the chatbot to ask targeted follow-up questions when prediction confidence is low. This interactive approach enhances the model's diagnostic precision and builds user trust through personalized, context-aware responses.

5. It is recommended that healthcare systems adopt integrated, preventive, and data-driven approaches that target both infectious and non-communicable diseases, with increased investment in public health awareness, early diagnosis, and mental health support to reduce the global disease burden.

6. Ensure Ethical and Data Privacy Compliance: Adhere strictly to medical data protection laws and ethical AI standards, ensuring patient confidentiality, consent, and transparency. This will build user trust and compliance with healthcare regulations. These improvements would transform the project from a prototype diagnostic model into a scalable, intelligent, and trustworthy medical assistant that evolves with data, adapts to user needs, and supports evidence-based healthcare decision-making.