AI Boot Camp **Project 3**

# Symptom-Based Disease Prediction Model

Team Members:
Anand Bhagwat
Jean Clark
Usha Hariharan
Alexander Iruthaya

# Project Description

- This project aims to create a machine-learning model capable of predicting diseases based on user input symptoms.
- The model will use a reliable dataset, the Symptom-Disease Prediction Dataset (SDPD), to train and evaluate predictive algorithms.
- Output will include a customer-facing app that provides preliminary diagnoses and possible remedies based on the symptoms provided by the user.

# Project Goals

- Develop a disease prediction model, based on symptoms
- Pilot a UI/UX application for users to input symptoms

# Dataset/ Data Extraction

- **Dataset**:  Symptom-Disease Prediction Dataset (SDPD)
- **Source**: Tucker, Jay (2024), "SymbiPredict", Mendeley Data, V1, doi: 10.17632/dv5z3v2xyd.1
- **Description**: A comprehensive and structured dataset linking symptoms to various diseases, rated as "reliable" by medical institutions and professionals, including the CDC.
- **Format**: CSV

# Data Cleaning & Preprocessing

- Different format data was looked at
- Cleaned data to replace NaN values with 0s
- Visualized dataset for diseases, symptoms, occurrence frequencies)
- Normalizing and encoding symptom data.

# Cleaned Dataset: DiseasePredict

| | Disease | Symptom_1 | Symptom_2 | Symptom_3 | Symptom_4 | Symptom_5 | S |
|---|---|---|---|---|---|---|---|
| **4379** | Hypoglycemia | vomiting | fatigue | anxiety | sweating | headache | |
| **393** | Psoriasis | skin rash | skin peeling | silver like dusting | small dents in nails | inflammatory nails | |
| **1164** | Osteoarthritis | joint pain | neck pain | hip joint pain | swelling joints | 0 | |
| **4478** | Bronchial Asthma | fatigue | cough | high fever | breathlessness | family history | |
| **731** | Hyperthyroidism | fatigue | mood swings | weight loss | restlessness | sweating | |

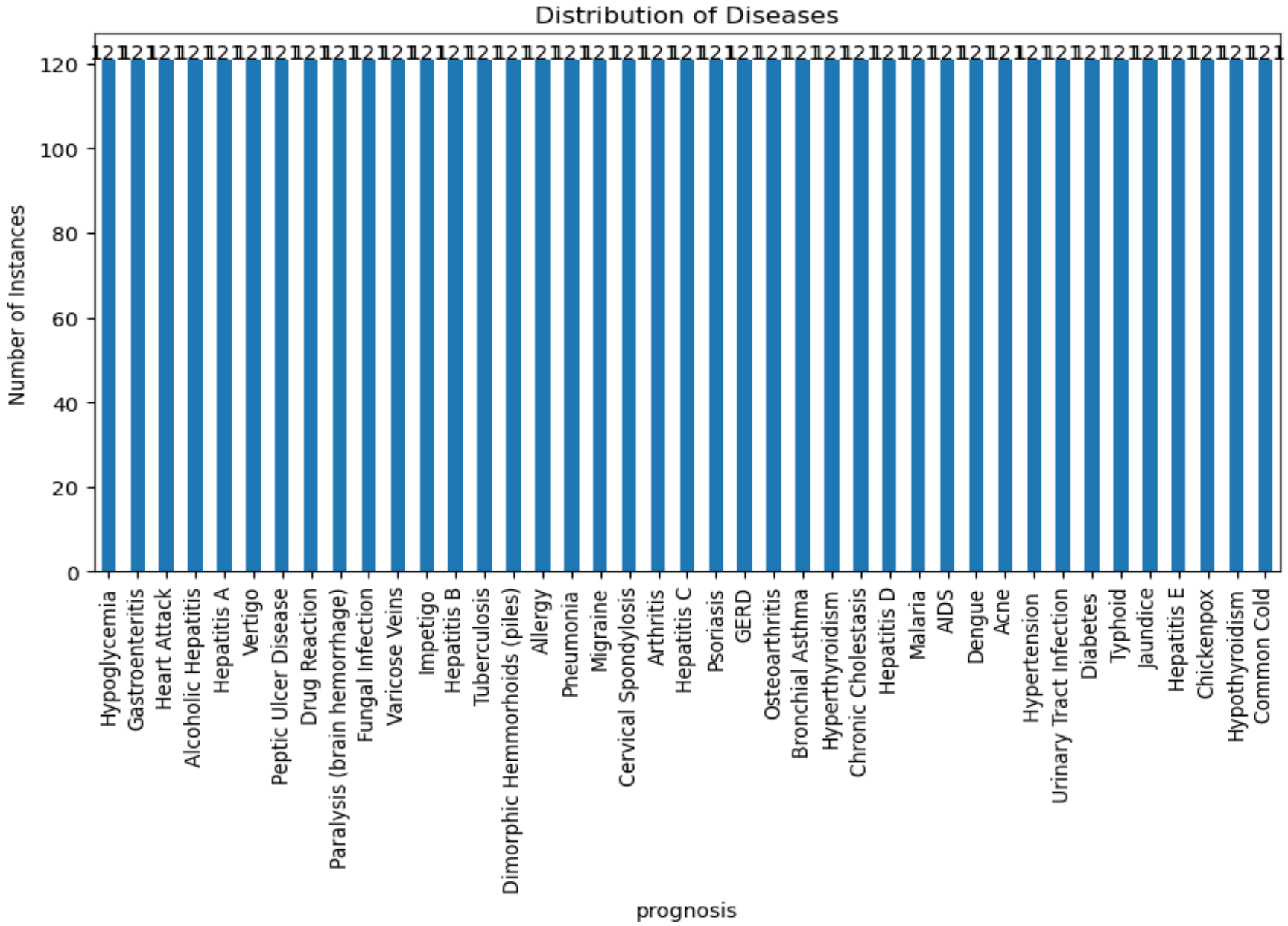
# Cleaned Dataset: SymbiPredict

| silver_like_dusting | small_dents_in_nails | inflammatory_nails | blister | red_sore_around_nose | yellow_crust_ooze | prognosis |
|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | Hypoglycemia |
| 1 | 1 | 1 | 0 | 0 | 0 | Psoriasis |
| 0 | 0 | 0 | 0 | 0 | 0 | Osteoarthritis |
| 0 | 0 | 0 | 0 | 0 | 0 | Bronchial Asthma |
| 0 | 0 | 0 | 0 | 0 | 0 | Hyperthyroidism |

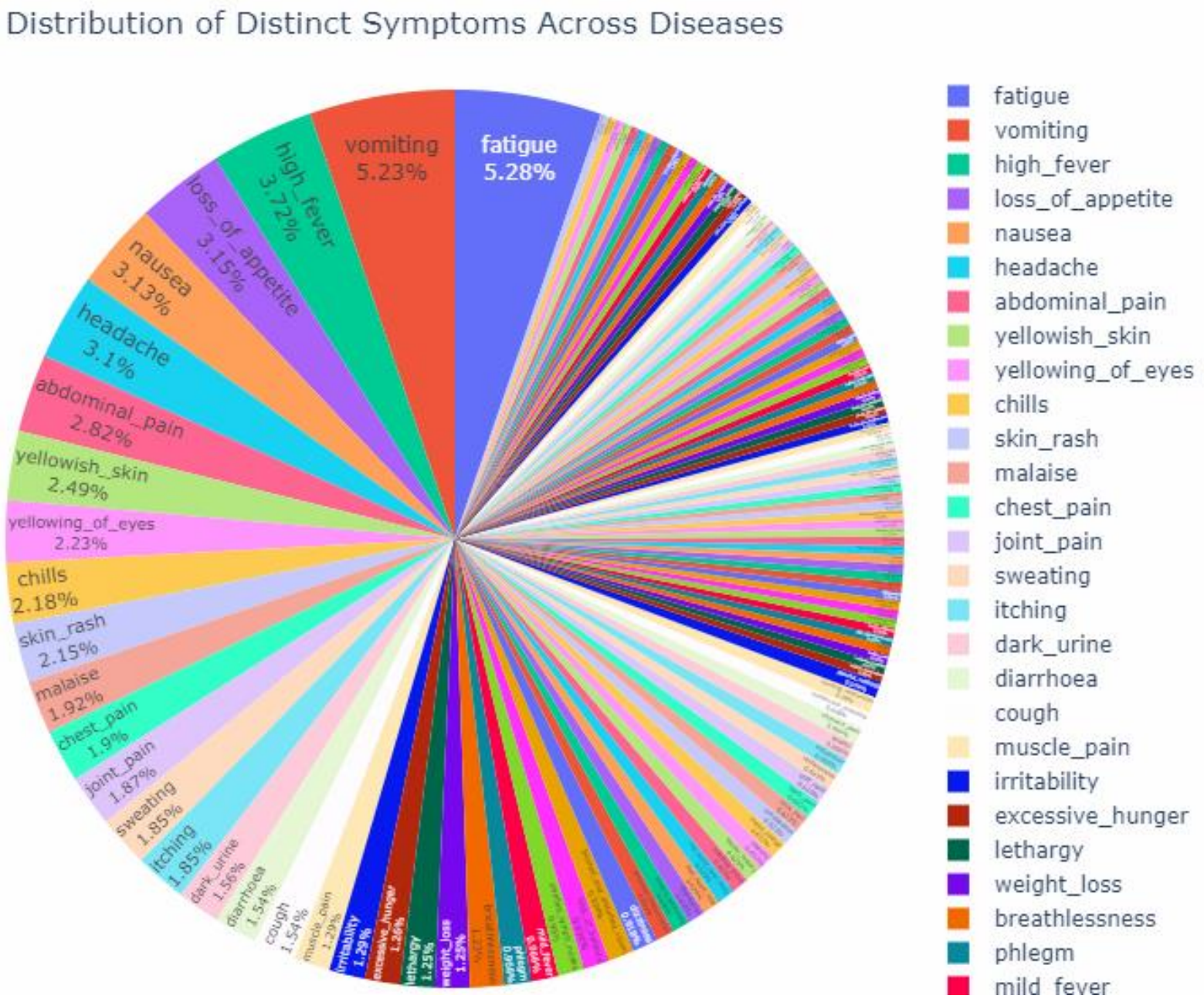
# Exploratory Data Analysis

- Visualized key feature distributions and their relationships
- Examined correlations between variables.
- Identified top attributes such as ` disease’ and ‘symptom out of 41 diseases and 132 symptoms
- Visuals on next slides: (title—model output 1)

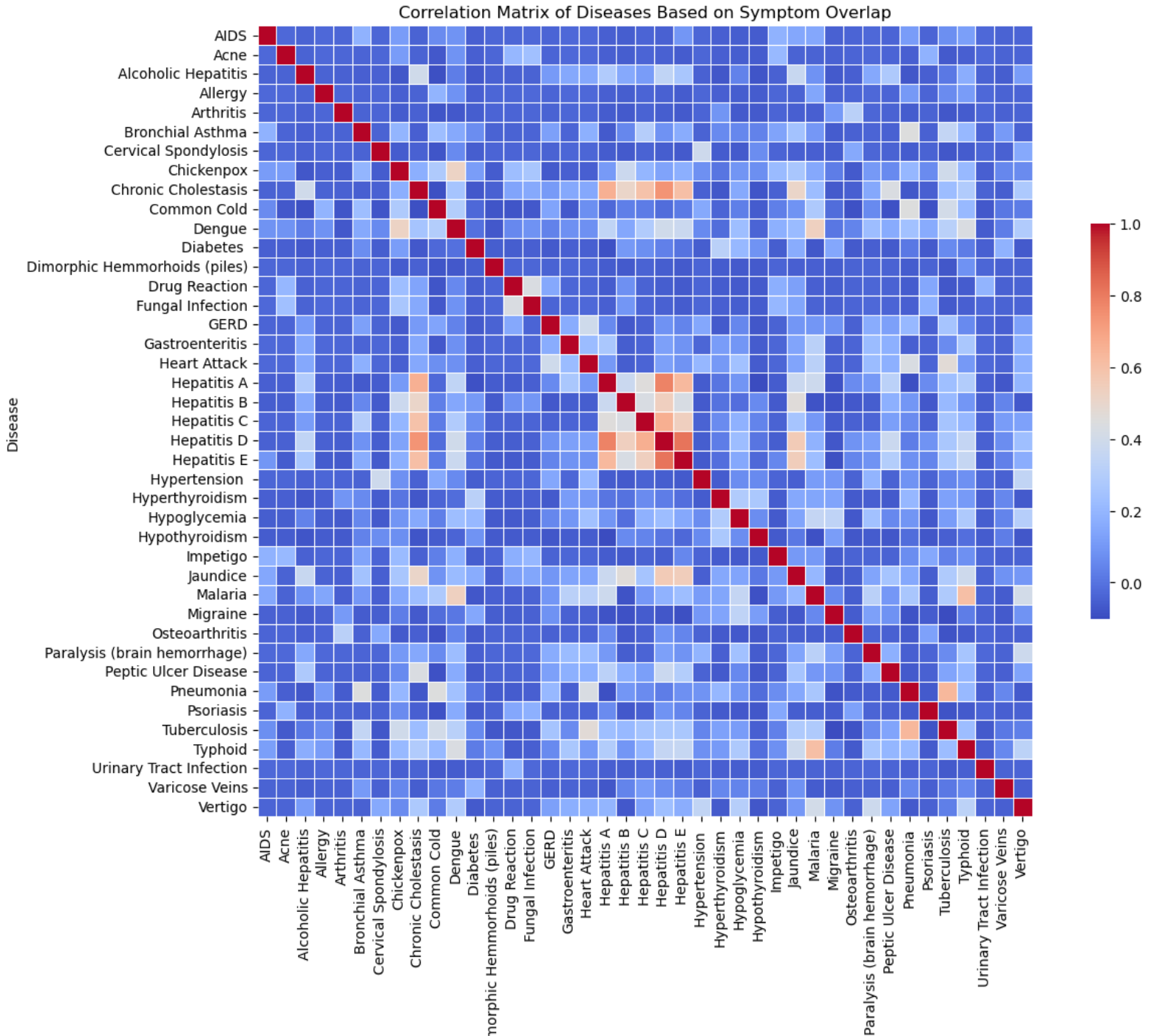
# Model Output



# Model Output



# Model Output



# Project Methodology

## Approach taken to achieve goals

- Split data into training and testing sets.
- Feature Importance
- Splitting the dataset into training, validation, and testing subsets.

# Model Optimization & Evaluation

## Model Optimization and Evaluation (accuracies of all models) 1.0

- Three models—MLP, CNN, RNN, RandomForest with LSTM

Scaled features using ‘Label Encoder’---ordinal variables(copy)

# Model Optimization & Output

| Model | Training | | Testing | |
|---|---|---|---|---|
| | Accuracy | Loss | Accuracy | Loss |
| **MLP** | 1.00 | 6.40E-05 | 1.00 | 5.34E-05 |
| **CNN** | 1.00 | 2.94E-06 | 1.00 | 2.44E-06 |
| **RNN with LSTM** | 0.0282 | 3.71 | 0.0243 | 3.713 |

-1ML

# Result/Conclusion Of Models -- MLP, CNN, RNN with LSTM

- MODEL PERFORMANCE: MLP, CNN, RNN with LSTM
- CNN Model performed best--high accuracy, low loss
- MLP also performed well, with slightly lower test accuracy
- RNN with LSTM did not perform well, perhaps not the right choice for this dataset

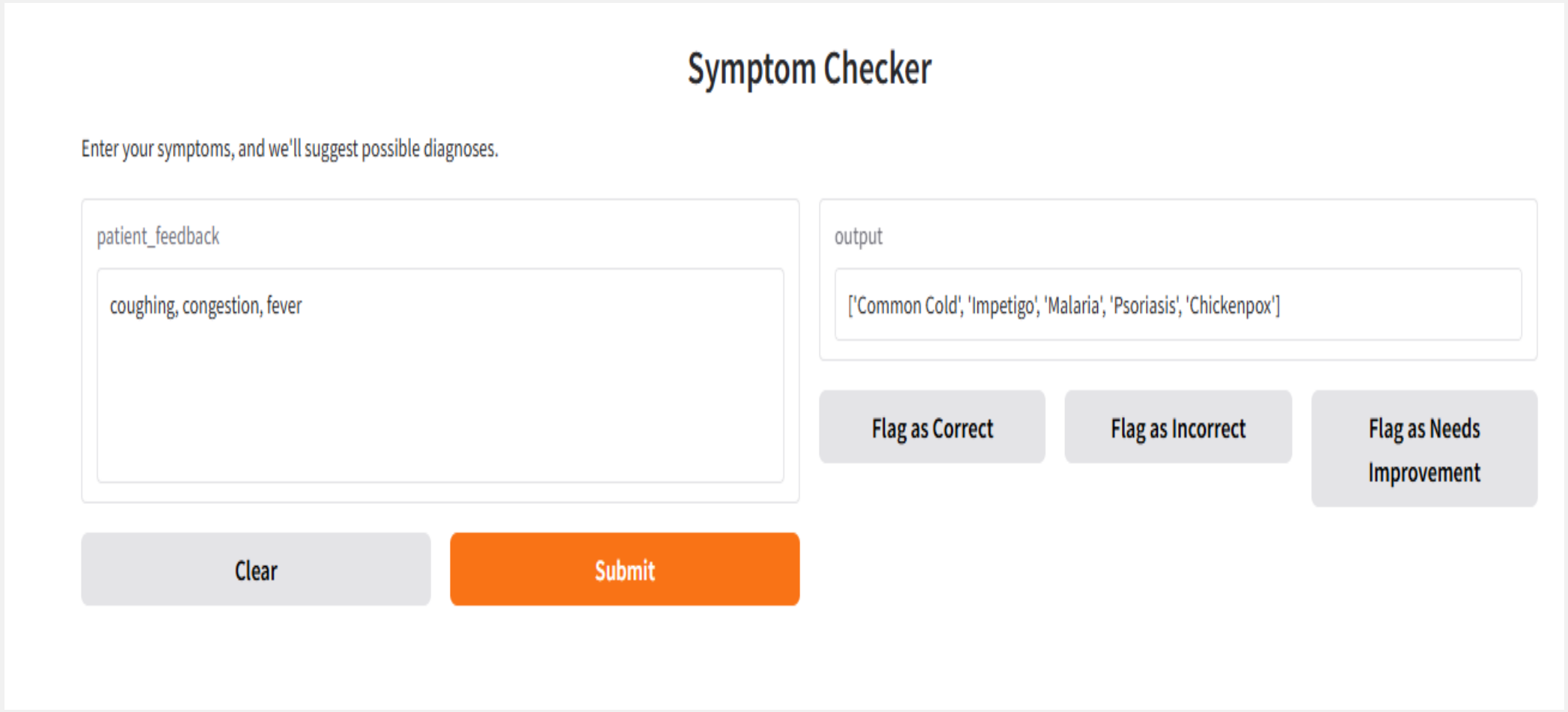
# Disease Prediction Application

Our application allows for patients to input some symptoms into a symptom tracker. This information is then associated with some possible diagnoses.

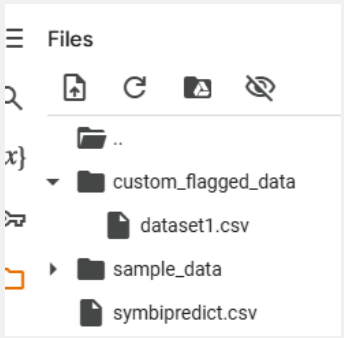
Some new technologies we used (we did not cover in our boot camp) are:

1. **SentencePiece** which is a supplement to our NLTK. This supplement is needed to assist in translating medical terms or more complex words.
2. **%%capture** which is unique to Google Colab. This allows for the !pip installs to run without generating all the responses, which clutter up the application.
3. **sqlite** which is a lightweight database management system. Given that we are dealing with large dataset(s) for our model, sqlite allows our application to store and retreive data using SQL (structured query language.) We are using this for efficiency and speed of use.
4. **Flagging** we added this feature to our gradio interface. It is used to collect information from users about how the application is working. It is part of improving the model over time.

# Disease Prediction Application



# Disease Prediction Application
# Output from our ‘flagging’ tool built into Gradio




dataset1.csv

1 to 10 of 11 entries Filter

| patient_feedback | output | flag | timestamp |
|---|---|---|---|
| skin rash, itching, swelling | ['Chickenpox', 'Jaundice', 'Chronic Cholestasis', 'Hepatitis B', 'Drug Reaction'] | Needs Improvement | 2025-02-11 00:03:18.063648 |
| skin rash, itching, swelling | ['Chickenpox', 'Jaundice', 'Chronic Cholestasis', 'Hepatitis B', 'Drug Reaction'] | Needs Improvement | 2025-02-11 00:06:35.291604 |
| congestion, fever, nausea, vomiting | ['Common Cold', 'Hepatitis D', 'Hepatitis A', 'Dengue', 'Hepatitis E'] | Needs Improvement | 2025-02-11 00:07:06.364648 |
| congestion, fever, nausea, vomiting | ['Common Cold', 'Hepatitis D', 'Hepatitis A', 'Dengue', 'Hepatitis E'] | Correct | 2025-02-11 00:07:45.178327 |
| rash, itching, swelling, | ['Chickenpox', 'Jaundice', 'Chronic Cholestasis', 'Hepatitis B', 'Drug Reaction'] | Correct | 2025-02-11 00:08:09.806250 |
| rash, itching, swelling, sneezing, skin irritation, | ['Chickenpox', 'Jaundice', 'Chronic Cholestasis', 'Hepatitis B', 'Drug Reaction'] | Incorrect | 2025-02-11 00:10:48.157823 |
| coughing, sneezing, headache, congestion | ['Common Cold', 'Dengue', 'Chickenpox', 'Typhoid', 'Malaria'] | Needs Improvement | 2025-02-11 00:14:37.820036 |
| congestion, coughing, fever, headache, | ['Common Cold', 'Dengue', 'Chickenpox', 'Typhoid', 'Malaria'] | Incorrect | 2025-02-11 00:15:06.861320 |
| nausea, vomiting, fever, headache, | ['Dengue', 'Typhoid', 'Malaria', 'Hypoglycemia', 'Vertigo'] | Incorrect | 2025-02-11 00:21:41.612551 |
| skin rash, itching, redness, swelling | ['Chickenpox', 'Jaundice', 'Chronic Cholestasis', 'Hepatitis B', 'Drug Reaction'] | Needs Improvement | 2025-02-11 00:22:16.215378 |

Show 10 per page

1 2

# Challenges Encountered

1. SpaCy and Gradio had version conflicts, so Transformers (vectorizing) via TF-IDF was used for app development.

2. App was running too slow, so sqlite was used—it allows for data storage and retrieval using SQL for speed and efficiency

# Summary

- ➔ A reliable dataset, the Symptom-Disease Prediction Dataset (SDPD) was used to train and evaluate multiple predictive algorithms.
- ➔ CNN outperformed other models, so was selected for integration into the Gradio app pilot.
- ➔ The **Gradio web app** pilot serves as a **testing interface** for users to input symptoms and receive **real-time disease predictions.**

# Future Considerations

## Future Enhancements Post-Pilot

After piloting, improvements can be made based on user feedback and pilot results. Potential enhancements include:

🔷 **Enhancing Model Generalization** – Expanding the training dataset with more symptom-disease pairs.

🔷 **Improving Explainability** – Using SHAP or LIME to provide reasons behind predictions.

🔷 **API Integration** – Connecting with external health databases for supplementary information.

🔷 **Deployment on Cloud & Mobile** – Making the app available via cloud services and mobile applications.

# Future Considerations

## FUTURE FORWARD Health AI CONSIDERATIONS:

1. **Ethical, inclusive AI Models**: Bias-resistant diagnoses, culturally adapted AI, fair, affordable AI-based care such as **virtual health screenings** to underserved, rural populations, preventing healthcare disparities

2. **Environmental/social sustainability metrics**: health equity metrics, community and public health-centered AI models and apps

3. **Blockchain and decentralized health records**: smart contacts, fraud prevention, verification, patient control of own health records

4. **Real-time and Dynamic Health scoring**: AI-driven early warning systems, dynamic risk profiling, automated triage, prioritization in ERs

5. **Emerging / Alternative Health Data Sources**: Wearable devices, personalized medicine, DNA sequencing

6. **Network-based Health-Risk modeling**: community-based predictions, symptom tracking, public health policy

7. **Hybrid models**: Traditional rules + AI-based probabilistic models

# Future Considerations

## FUTURE FORWARD HEALTH AI CHALLENGES:

**Key Challenges to Address:**

•**Data Privacy and Security**: HIPPA, EHRs, more

•**Secure AI Model Training**: Preventing AI-powered cyber threats

•**Regulatory Compliance**: Aligning innovations with HHS, FDA, CDC, more

•**User Trust**: Building confidence in AI-powered healthcare

•**Fairness and Bias Mitigation**:  Diverse and Representative Data

•Tackling these challenges is crucial to **unlock the full potential of AI in healthcare**.

• **AI-powered healthcare** has the potential to enhance patient outcomes, improve accessibility, and revolutionize modern medicine.