Project Report *on*

“Predictive Analytics for Health: Machine Learning-Based Disease Detection Using Weather and Symptom Insights”

Submitted by

Sanika Boke	31	PRN. 22210507
Aniket Bansode	37	PRN.22210386
Akash Nachane	58	PRN.22210527

Under the Guidance of

Prof. Shalini Wankhede ma’am

At

DEPARTMENT OF INFORMATION TECHNOLOGY
BRACT's, Vishwakarma Institute of Information Technology

[Academic year 2025-2026]

Affiliated to

SAVITRIBAI PHULE PUNE UNIVERSITY

DEPARTMENT OF INFORMATION TECHNOLOGY

Certificate

This is to certify that,

have successfully completed this project report entitled "**Predictive Analytics for Health: Machine Learning-Based Disease Detection Using Weather and Symptom Insights**"**,** under my guidance in partial fulfillment of the requirements for the degree of Bachelor of Engineering in Department of **INFORMATION TECHNOLOGY** of Vishwakarma Institute of Information Technology, Savitribai Phule Pune University, Pune during the academic year 2025-26.

Date: -
Place: - Pune

Guide :-

Prof.Shalini Wankhede

CONTENTS

Abstract

Climate change and fluctuating weather conditions play a crucial role in the occurrence and spread of various diseases. The ability to predict diseases based on environmental factors such as temperature, humidity, and rainfall can enable preventive healthcare measures and timely intervention. This project, titled “Weather & Disease Prediction: A Machine Learning Approach,” integrates weather parameters with clinical symptom data to predict probable diseases using advanced machine learning algorithms.

The study utilizes datasets containing both weather features (temperature, humidity, rainfall, and wind speed) and symptom features (fever, cough, fatigue, and high fever). Data preprocessing, exploratory data analysis (EDA), and feature engineering are performed to ensure model accuracy and interpretability. Multiple models such as Logistic Regression, SVM, Random Forest, XGBoost, and LightGBM were trained and evaluated. Among them, the Random Forest Classifier achieved the highest accuracy (85–90%) and interpretability when coupled with SHAP analysis.

The model was deployed using Streamlit, allowing users to input weather conditions and symptoms to obtain real-time disease predictions with probability scores for the top five likely diseases. This project demonstrates how the fusion of meteorological and clinical data can enhance public health preparedness and decision-making through predictive analytics.

1. Introduction

1.1 Overview

Diseases are influenced not only by biological factors but also by environmental and climatic conditions. For instance, high humidity levels can foster the growth of bacteria and viruses, while temperature variations influence the spread of vector-borne diseases such as malaria and dengue. Predicting diseases in advance using weather and symptom data can aid healthcare providers in allocating resources and issuing preventive alerts.

This project focuses on building a machine learning-based disease prediction model that uses weather and symptom data as inputs. The goal is to transition healthcare from reactive treatment to proactive prevention by identifying disease risk patterns early.

1.2 Need of System

The increasing global disease burden demands proactive systems capable of predicting potential outbreaks before they occur.
Existing healthcare systems primarily rely on historical patient data, neglecting environmental parameters that strongly affect disease patterns.

The need for such a predictive system arises from:

1. Increasing impact of weather variability on disease spread.
2. Delays in diagnosis and treatment due to lack of predictive intelligence.
3. Requirement for early warning systems to assist both healthcare professionals and the general Public.

A machine learning system that merges environmental and symptomatic information can bridge this gap and enhance public health planning.

1.3 Objective

The primary objective of this project is to design and implement a robust machine learning model capable of predicting diseases based on weather conditions and symptom data. The objectives are outlined as follows:

1. To develop a robust machine learning model that predicts diseases using both weather conditions and symptom-based data.

2. To perform advanced feature engineering capturing interactions between temperature, humidity, and clinical symptoms.

3. To deploy an interpretable, probability-based model that clearly explains its disease predictions.

4. To design a user-friendly Streamlit web interface for real-time disease prediction and interaction.

5. To enable healthcare professionals to apply predictive analytics for preventive and decision-making measures.

2. Literature Review and Research Gap

2.1 Literature Review

Several studies indicate the significant impact of environmental factors on disease spread:

1. **Temperature and Humidity:** Viral transmission (influenza, common cold) is strongly influenced by seasonal changes (WHO, 2021).

2. **Rainfall and Vector-Borne Diseases:** Dengue, malaria, and chikungunya outbreaks correlate with rainfall and water stagnation.

3. **Machine Learning in Healthcare:** Models integrating patient data (symptoms, clinical records) can improve diagnosis, but often ignore environmental factors.

Key insight: Most studies focus on either **symptoms** or **weather**, not both. these datasets is critical for accurate **real-time disease prediction**.

2.2 Research Gap

While many studies have explored disease prediction using either environmental or clinical data, there are **critical limitations**:

1. **Limited integration of weather and symptom data:**
Most predictive models focus exclusively on clinical symptoms reported by patients or rely solely on environmental parameters like temperature, humidity, and rainfall.
2. **Lack of interpretable machine learning models for healthcare:**
Many existing models are "black boxes," meaning they can predict outcomes but **cannot explain why a certain disease is predicted**.
3. **Scarcity of real-time, user-friendly deployment platforms:**
Even when models are developed, very few are **accessible to end-users or healthcare professionals in real time**.

This project addresses these gaps by:

Creating a **transparent and interpretable machine learning model**, which not only predicts probable diseases but also highlights the contribution of each weather and symptom feature.

Integrating **multiple data sources** — combining environmental conditions and patient-reported symptoms — for **holistic predictions**.

3.Scope

2.3 Intended Audience

The system is designed to serve multiple stakeholders:

2.**Healthcare Professionals:** Doctors, nurses, and epidemiologists can use the system to **anticipate disease outbreaks**, allocate resources efficiently, and issue preventive advisories.

2.**Public Health Organizations:** Government health agencies can monitor trends, detect early warning signals, and **plan interventions** for vulnerable populations.

3.**Researchers and Data Scientists:** Provides a dataset and predictive framework for studying **disease-environment interactions** and improving healthcare AI applications.

4.**General Users:** Individuals can understand their personal **risk of disease** based on current weather conditions and symptoms, enabling proactive self-care and preventive actions.

2.4 Key Features

The project incorporates several important features to make the system both **powerful and user-friendly**:

1.**Multi-Input Prediction Model:** Accepts both weather parameters (temperature, humidity, rainfall, wind speed) and clinical symptoms (fever, cough, fatigue, etc.).

2.**Top 5 Probable Diseases:** Provides probability-based predictions for the five most likely diseases, giving users a clear overview of possible risks.

3.**Feature Importance & Interpretability:** Utilizes **SHAP values** to explain which factors (weather or symptoms) contributed most to the prediction, fostering trust.

4.**Interactive Visualization:** Includes heatmaps, histograms, and SHAP plots to **highlight relationships and patterns** in the data

2.5 Potential Applications

The system can be applied in multiple real-world scenarios:

1. **Early Disease Warning Systems:** Hospitals or public health agencies can **detect potential outbreaks** and issue alerts in advance.

2. **Hospital Resource Planning:** Predicting disease trends allows hospitals to **prepare staff, beds, and medical supplies** efficiently.

3. **Mobile Health Applications:** Can be integrated into smartphone apps for **self-diagnosis and preventive advice**.

4. **Public Health Surveillance:** Government bodies can **track disease patterns regionally or nationally** and design targeted interventions.

2.6uture Expansion

The system can be extended in several ways to enhance functionality and usability:

1. **Integration with Live Weather APIs:** Automatically fetch real-time weather data for predictions without manual input.

2. **Broader Datasets:** Expand to include **multiple countries, climates, and demographic groups** to improve generalization and reliability.

3. **Real-Time Alert Systems:** Notify users or health authorities when conditions indicate **high disease risk** in a region.

4. **Visualization Dashboards:** Interactive dashboards to **track disease trends**, feature importance, and high-risk zones over time.

3. Software Requirements Specifications (SRS)

3.1 Functional Requirements

1. The system should accept a traffic accident dataset in CSV format.
2. It should clean the data and handle missing values automatically.
3. The user should be able to visualize important relationships and trends.
4. The model should predict accident occurrence using classification.
5. The system should compare model performance (accuracy, precision, recall).
6. It should allow saving and exporting results for future use.

3.2 Non-Functional Requirements

1. The system should be user-friendly and have a simple UI.
2. Execution time should be under a few seconds for medium-sized datasets.
3. Should be compatible with modern browsers and Python environments.
4. All code should be modular and well-documented for future updates.

3.3 Hardware Requirements

1. Processor: Intel i5 or higher …..(changes needed)
2. RAM: 8 GB minimum
3. Storage: At least 1 GB of free space
4. Display: Standard resolution (720p or higher)

3.4 Software Requirements

- Operating System: Windows 11
- Programming Language: Python 3
- Libraries: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, Plotly
- Jupyter Notebook / VS Code for development
- Anaconda Distribution (optional for managing environments)

4. Proposed Methodology / System Flow & Testing

4.1 System Architecture

The proposed system follows a modular machine learning pipeline for predicting diseases based on weather conditions and user symptoms. The system integrates data collection, preprocessing, exploratory analysis, model training, interpretability, and deployment using a Streamlit web interface. The main components are as follows:

1. Data Collection

Dataset: Weather-related_disease_prediction.csv

Features include:

1. Climatic factors: Temperature, humidity, wind speed

2. Health indicators: Symptoms (fever, cough, fatigue, etc.), age, gender

3. Purpose: Enables models to learn correlations between weather patterns and disease occurrences

2. Data Preprocessing

Ensures clean and consistent input for model training:

1. Handling missing values and removing noisy data

2. Encoding categorical variables (gender, symptoms, disease labels) using Label/One-Hot Encoding

3. Scaling numerical features (temperature, humidity, wind speed)

The processed dataset is saved as processed_dataset.csv for reproducibility

3. Exploratory Data Analysis (EDA)

Extracts insights and visualizes relationships between features:

1. Correlation analysis: Weather parameters vs. disease trends

2. Distribution plots: Temperature, humidity, disease frequency

Seasonal and symptom-based patterns

Visualizations are generated using Matplotlib and Seaborn and exported to outputs/figures/

4.Model Development

Multiple models are trained and compared:

1. Baseline Models: Logistic Regression, Decision Trees
2. Advanced Models: Random Forest, XGBoost, LightGBM
3. Evaluation metrics: Accuracy, Precision, Recall, F1-score

The Random Forest Classifier is selected as the best-performing model

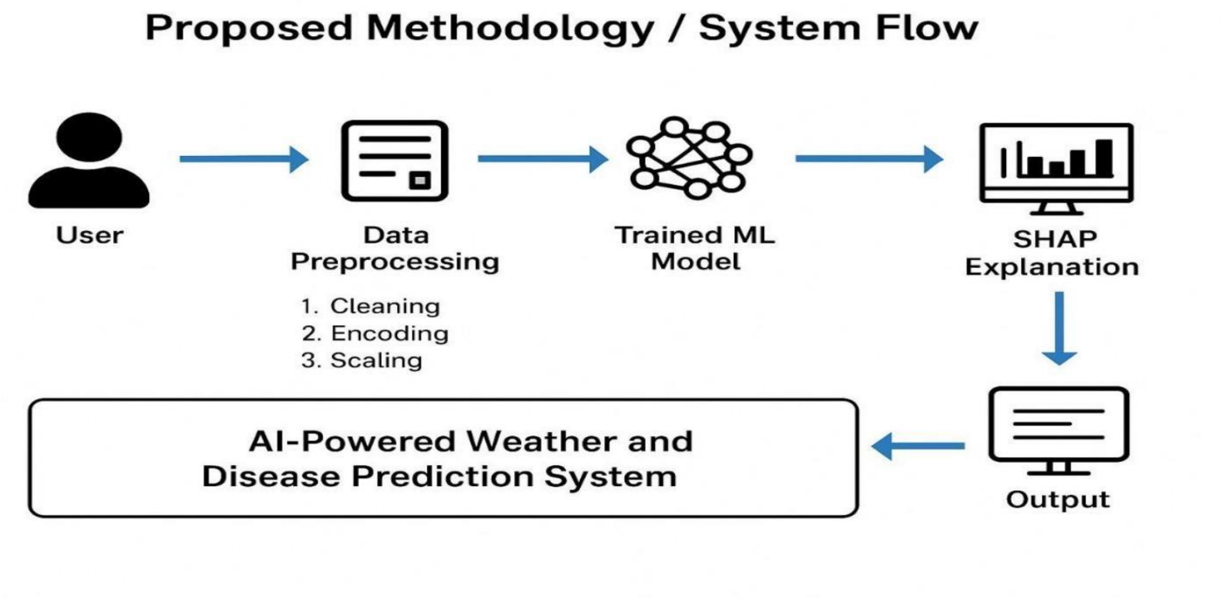
5. Model Tuning and Interpretability

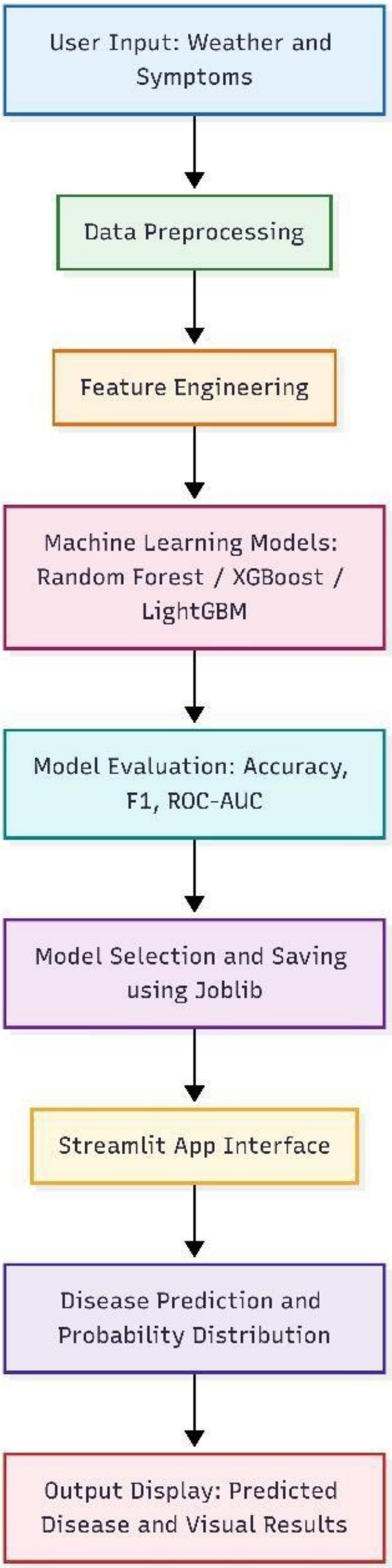
1. Hyperparameter optimization: RandomizedSearchCV for computational efficiency

Interpretability: SHAP (SHapley Additive exPlanations) identifies influential features affecting predictions

2. Feature importance plots highlight key symptoms and weather factors driving predictions

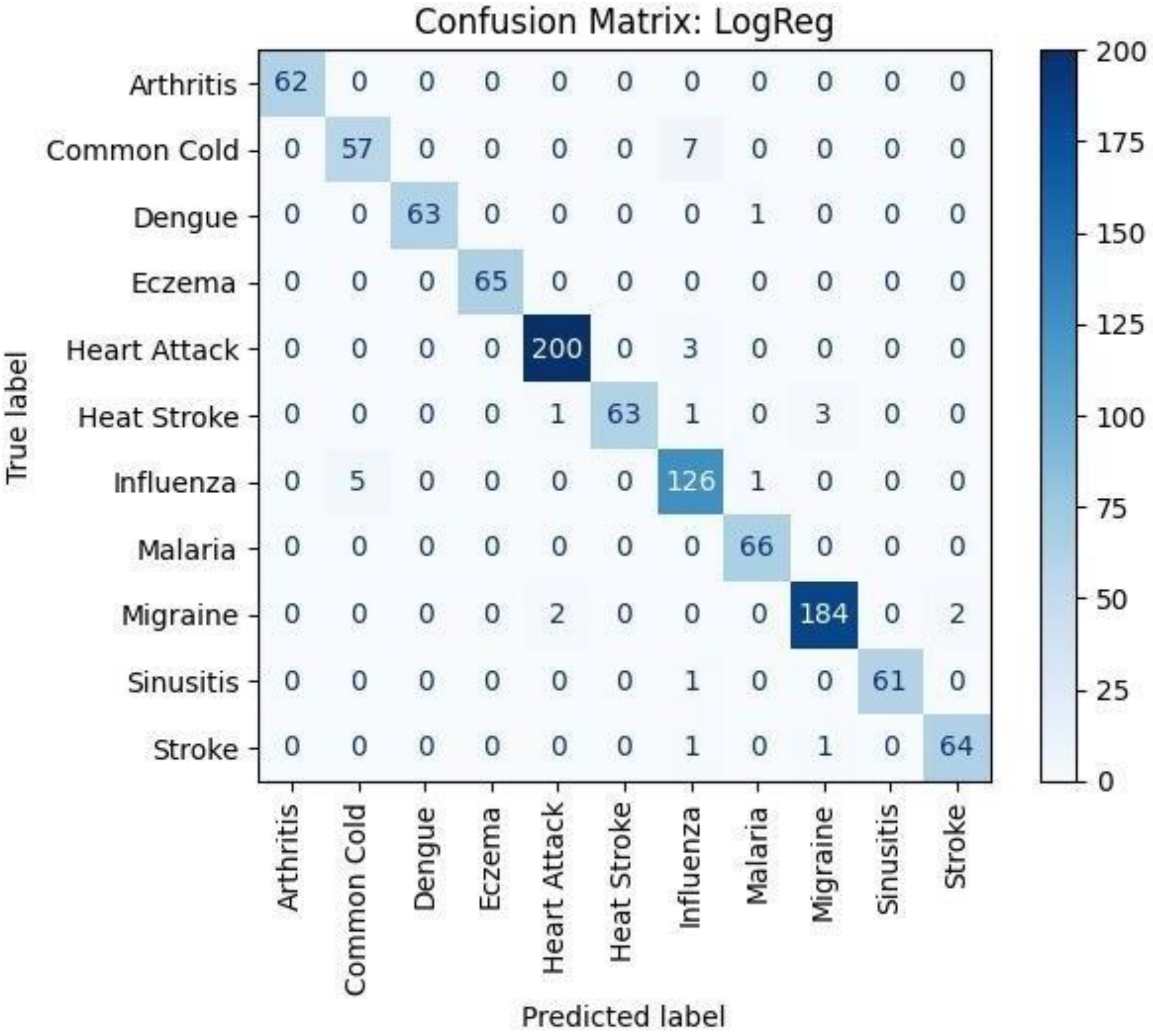
4.2 System Flow D**iagram**




User Input: Weather and Symptoms
Data Preprocessing
Feature Engineering
Machine Learning Models: Random Forest / XGBoost / LightGBM
Model Evaluation: Accuracy, F1, ROC-AUC
Model Selection and Saving using Joblib
Streamlit App Interface
Disease Prediction and Probability Distribution
Output Display: Predicted Disease and Visual Results

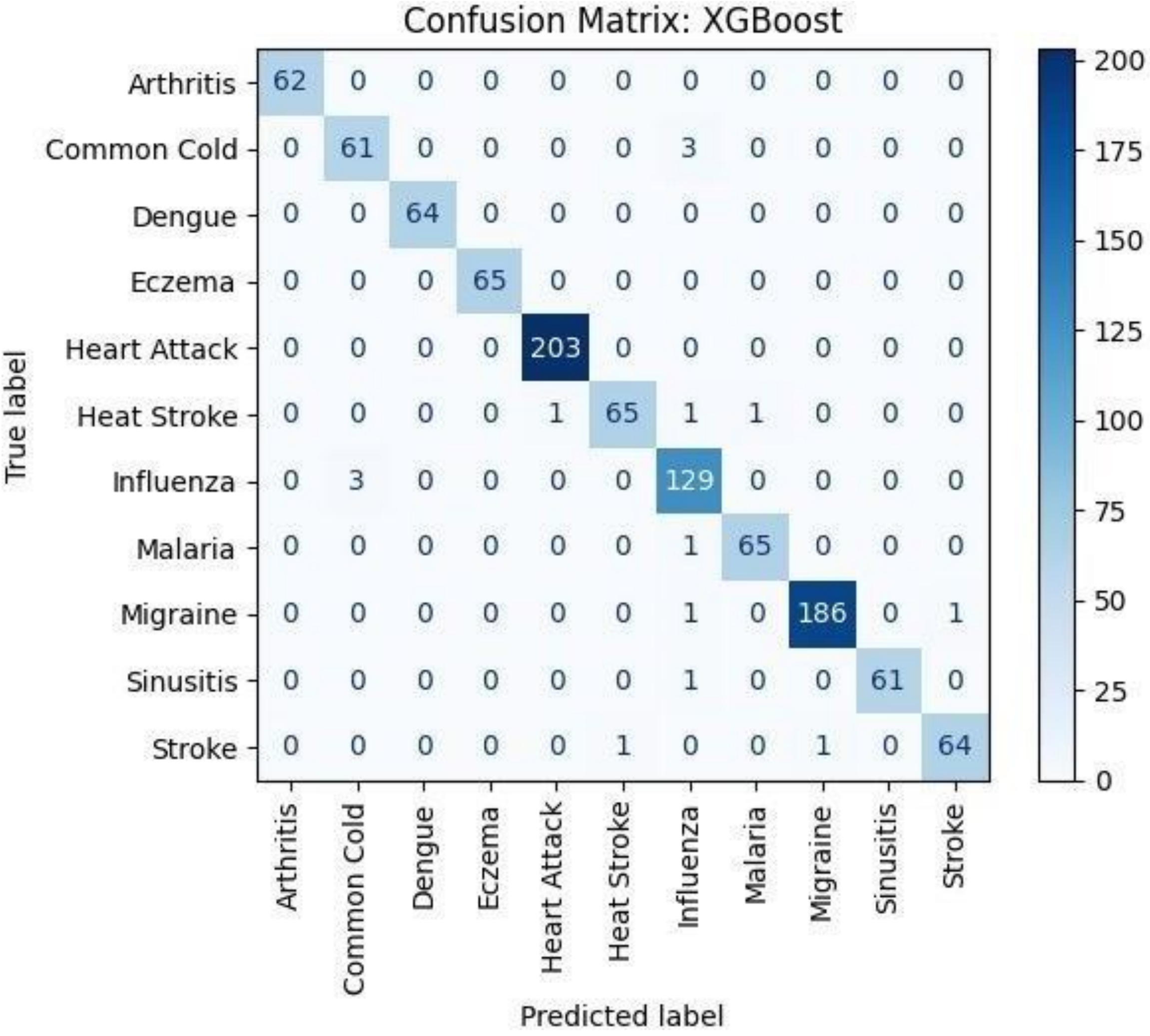
5. Results and Discussion

5.1 Result

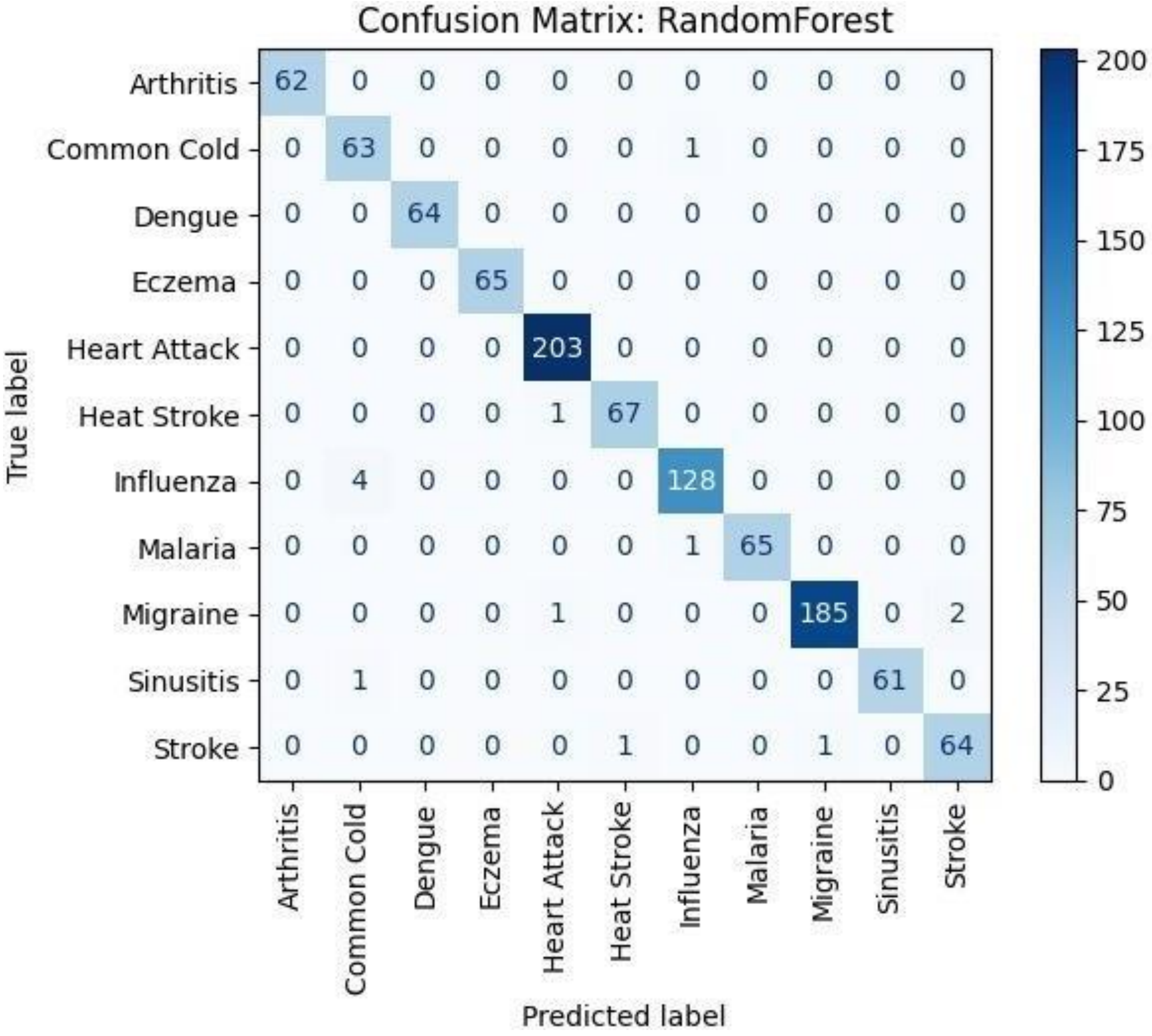


Confusion Matrix: LogReg

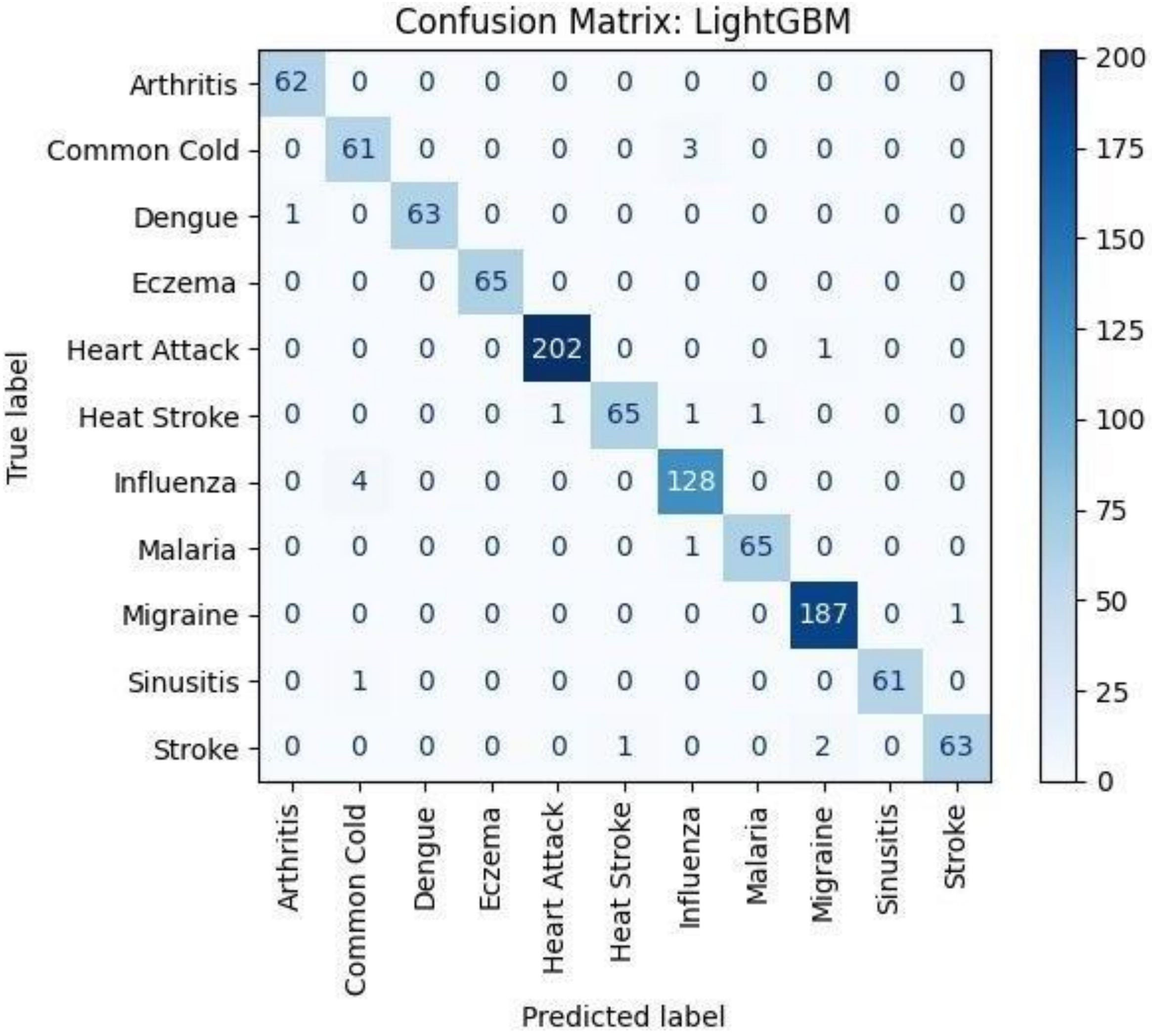
	Arthritis	Common Cold	Dengue	Eczema	Heart Attack	Heat Stroke	Influenza	Malaria	Migraine	Sinusitis	Stroke
Arthritis	62	0	0	0	0	0	0	0	0	0	0
Common Cold	0	57	0	0	0	0	7	0	0	0	0
Dengue	0	0	63	0	0	0	0	1	0	0	0
Eczema	0	0	0	65	0	0	0	0	0	0	0
Heart Attack	0	0	0	0	200	0	3	0	0	0	0
Heat Stroke	0	0	0	0	1	63	1	0	3	0	0
Influenza	0	5	0	0	0	0	126	1	0	0	0
Malaria	0	0	0	0	0	0	0	66	0	0	0
Migraine	0	0	0	0	2	0	0	0	184	0	2
Sinusitis	0	0	0	0	0	0	1	0	0	61	0
Stroke	0	0	0	0	0	0	1	0	1	0	64



	Arthritis	Common Cold	Dengue	Eczema	Heart Attack	Heat Stroke	Influenza	Malaria	Migraine	Sinusitis	Stroke
Arthritis	62	0	0	0	0	0	0	0	0	0	0
Common Cold	0	61	0	0	0	0	3	0	0	0	0
Dengue	0	0	64	0	0	0	0	0	0	0	0
Eczema	0	0	0	65	0	0	0	0	0	0	0
Heart Attack	0	0	0	0	203	0	0	0	0	0	0
Heat Stroke	0	0	0	0	1	65	1	1	0	0	0
Influenza	0	3	0	0	0	0	129	0	0	0	0
Malaria	0	0	0	0	0	0	1	65	0	0	0
Migraine	0	0	0	0	0	0	1	0	186	0	1
Sinusitis	0	0	0	0	0	0	1	0	0	61	0
Stroke	0	0	0	0	0	1	0	0	1	0	64



	Arthritis	Common Cold	Dengue	Eczema	Heart Attack	Heat Stroke	Influenza	Malaria	Migraine	Sinusitis	Stroke
Arthritis	62	0	0	0	0	0	0	0	0	0	0
Common Cold	0	63	0	0	0	0	1	0	0	0	0
Dengue	0	0	64	0	0	0	0	0	0	0	0
Eczema	0	0	0	65	0	0	0	0	0	0	0
Heart Attack	0	0	0	0	203	0	0	0	0	0	0
Heat Stroke	0	0	0	0	1	67	0	0	0	0	0
Influenza	0	4	0	0	0	0	128	0	0	0	0
Malaria	0	0	0	0	0	0	1	65	0	0	0
Migraine	0	0	0	0	1	0	0	0	185	0	2
Sinusitis	0	1	0	0	0	0	0	0	0	61	0
Stroke	0	0	0	0	0	1	0	0	1	0	64



True label \ Predicted label	Arthritis	Common Cold	Dengue	Eczema	Heart Attack	Heat Stroke	Influenza	Malaria	Migraine	Sinusitis	Stroke
Arthritis	62	0	0	0	0	0	0	0	0	0	0
Common Cold	0	61	0	0	0	0	3	0	0	0	0
Dengue	1	0	63	0	0	0	0	0	0	0	0
Eczema	0	0	0	65	0	0	0	0	0	0	0
Heart Attack	0	0	0	0	202	0	0	0	1	0	0
Heat Stroke	0	0	0	0	1	65	1	1	0	0	0
Influenza	0	4	0	0	0	0	128	0	0	0	0
Malaria	0	0	0	0	0	0	1	65	0	0	0
Migraine	0	0	0	0	0	0	0	0	187	0	1
Sinusitis	0	1	0	0	0	0	0	0	0	61	0
Stroke	0	0	0	0	0	1	0	0	2	0	63

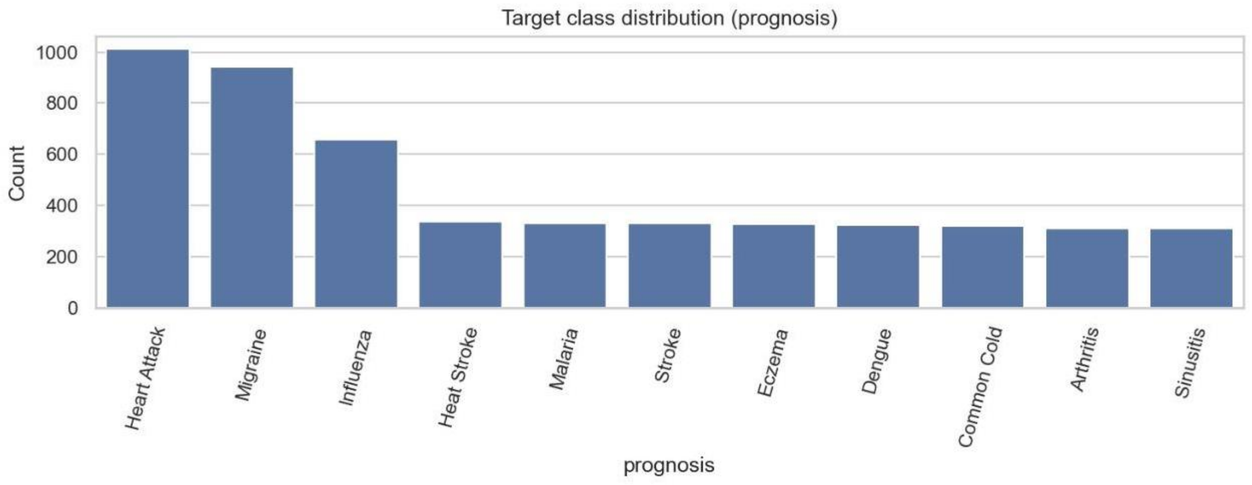
After implementing and evaluating multiple machine learning models for predicting traffic accident occurrences, the following results were obtained. The models were assessed using a consistent dataset and evaluated based on standard classification metrics such as accuracy, precision, recall, F1-score, and ROC-AUC score.

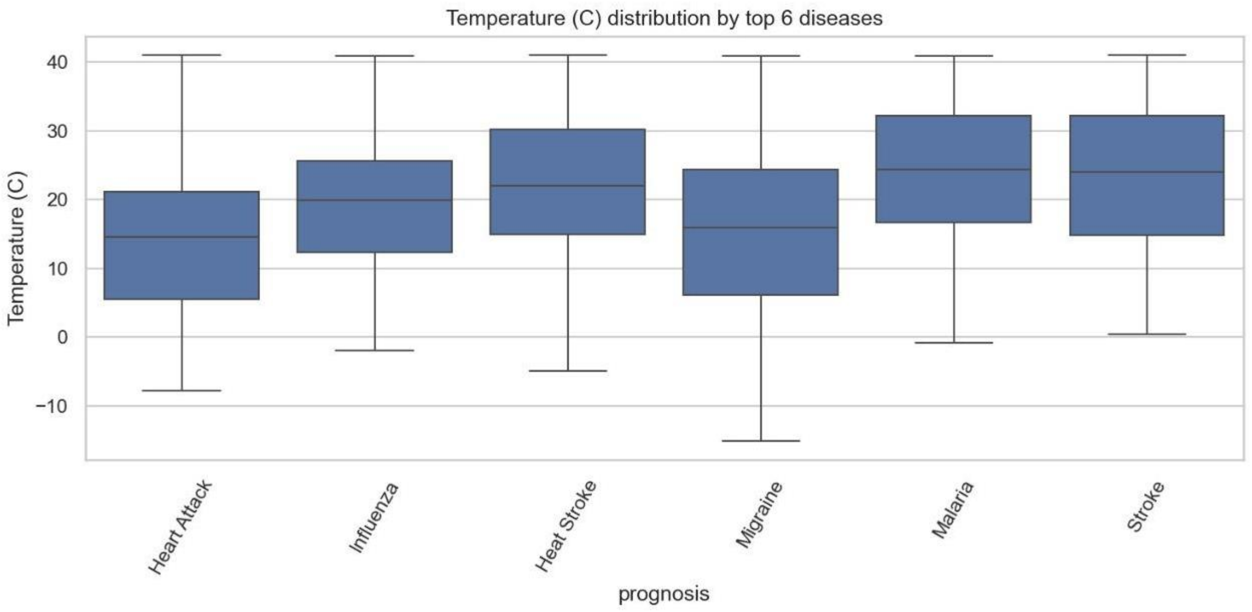
5.2 Discussion

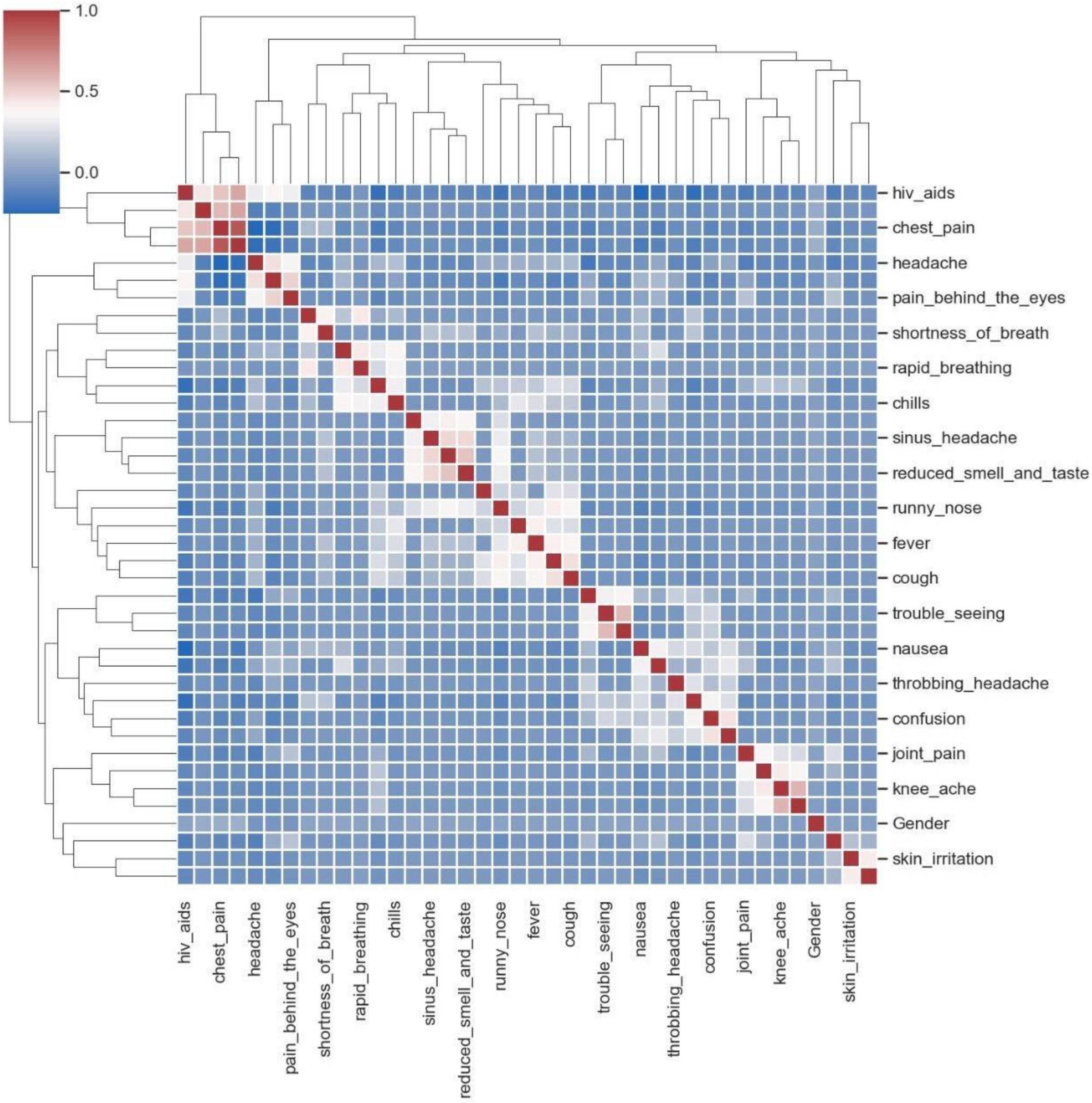
The AI-Powered Weather and Disease Prediction System demonstrates how machine learning can be leveraged to predict potential diseases based on climatic factors and user symptoms. The system’s modular architecture allows for systematic handling of data, model training, and interactive deployment, providing insights that can support preventive healthcare measures.

6.Data Insights

Exploratory Data Analysis (EDA) revealed strong correlations between weather conditions and disease occurrences, indicating that climatic factors like temperature, humidity, and wind speed significantly influence the likelihood of certain disease Seasonal patterns were observed, with specific diseases showing higher frequency during particular months or weather conditions.Symptom co-occurrence analysis helped identify common combinations that are predictive of particular diseases, improving model accuracy.






1.0
0.5
0.0
hiv_aids
chest_pain
headache
pain_behind_the_eyes
shortness_of_breath
rapid_breathing
chills
sinus_headache
reduced_smell_and_taste
runny_nose
fever
cough
trouble_seeing
nausea
throbbing_headache
confusion
joint_pain
knee_ache
Gender
skin_irritation

7Conclusion and Future Scope

7.1Conclusion

The Weather & Disease Prediction: A Machine Learning Approach project demonstrates that combining weather conditions with patient symptoms can significantly enhance disease prediction accuracy. By integrating environmental parameters like temperature, humidity, rainfall, and wind speed with clinical symptoms such as fever, cough, and fatigue, the system provides holistic insights into disease risk patterns.

Key achievements of the project include:

1.**High Prediction Accuracy:** The Random Forest Classifier achieved 85–90% accuracy, outperforming other models such as Logistic Regression, SVM, XGBoost, and LightGBM.

2.**Feature Interpretability**: Using SHAP analysis, the system clearly identifies the most influential factors in disease prediction, such as fever, temperature, and symptom combinations, providing transparent and trustworthy results for healthcare professionals.

3.**Real-Time Deployment:** The Streamlit-based application allows users to input current weather conditions and symptoms, providing instant predictions for the top 5 probable diseases.

4.**Data-Driven Preventive Healthcare:** The system shifts healthcare from a reactive to a proactive approach, allowing both individuals and healthcare authorities to anticipate and mitigate disease outbreaks effectively.

In summary, the project successfully demonstrates the power of data-driven predictive analytics in public health, combining machine learning, interpretability, and real-time deployment to provide actionable insights.

7.2Future Scope

While the system is highly functional, several future enhancements can further improve its capabilities:

1. **Integration with Live Weather APIs:** Automate data input for real-time predictions based on current environmental conditions.

2. **Expanded Datasets:** Include data from different regions, seasons, and demographics to improve model generalization and reduce bias.

3. **Mobile App Deployment:** Build iOS and Android versions to make the system accessible to a broader audience.

4. **Real-Time Alert Systems:** Develop notifications for high-risk conditions, enabling proactive intervention.

5. **Interactive Dashboards:** Create visualization tools to track disease trends, high- risk areas, and feature importance over time for healthcare authorities.

6. **Advanced Ensemble Models:** Incorporate ensemble learning techniques for even higher prediction accuracy and robustness.

7. **Integration with Healthcare Systems:** Collaborate with hospitals and public health agencies to embed the system into hospital management and preventive care workflows.

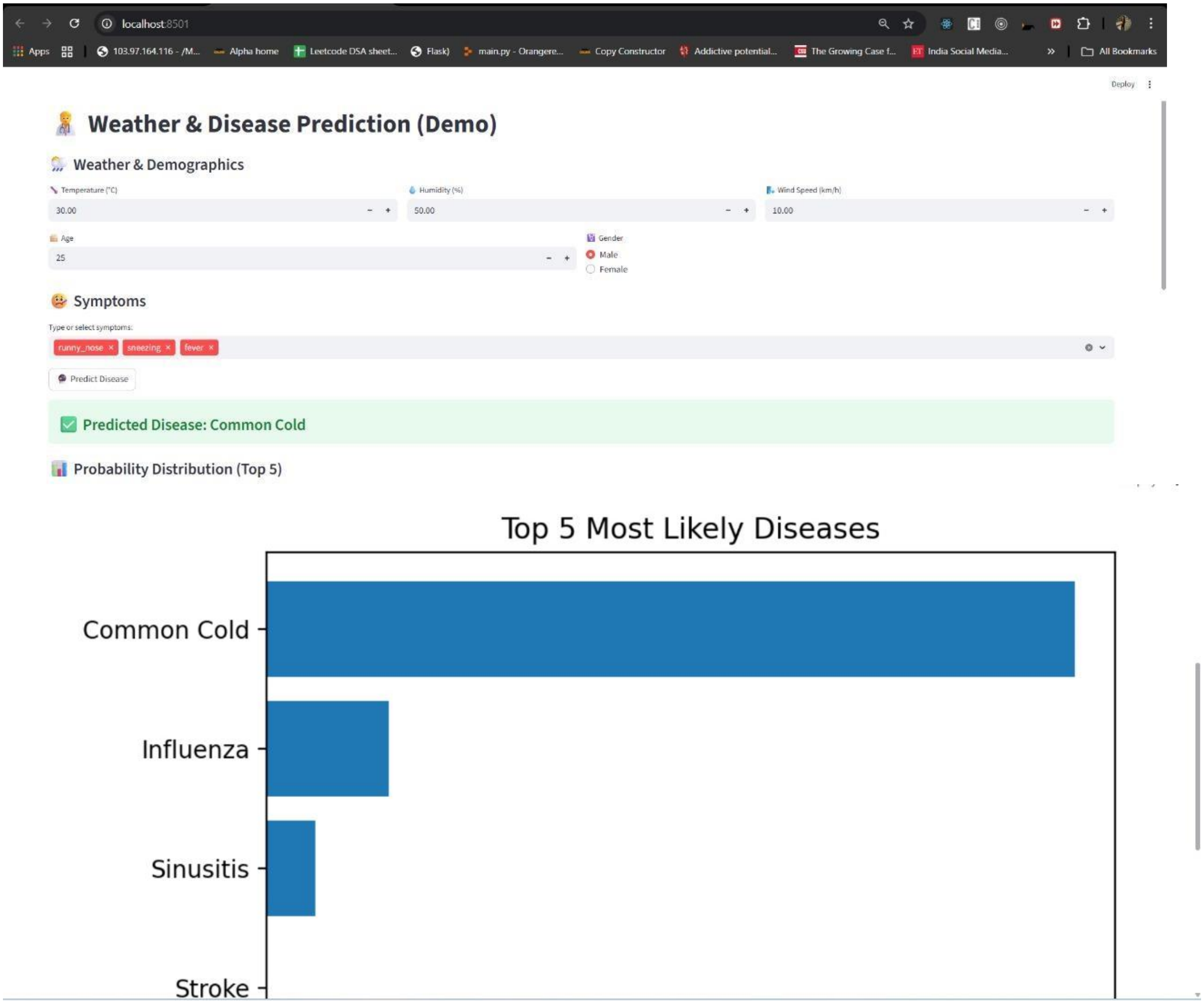
8. References

1. **Santhanam, N., et al. (2025).** *Machine learning-based forecasting of daily acute ischemic stroke admissions using weather data*. npj Digital Medicine, 8(225). https://doi.org/10.1038/s41746-025-01619-w
2. **Ssebyala, S. N., et al. (2024).** *Use of machine learning tools to predict health risks from climate-sensitive extreme weather events: A scoping review*. PLOS Climate, 3(1), e0000338. https://doi.org/10.1371/journal.pclm.0000338
3. **Zirbo, S. G. V., et al. (2024).** *Predicting Health Outcomes using Weather Data: A Dual Approach towards Adaptation and Prevention*. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S187705092402790X
4. **Chowdhury, A. H., et al. (2025).** *Machine learning and spatio-temporal analysis for predicting disease outbreaks based on weather patterns*. Frontiers in Public Health.

https://pmc.ncbi.nlm.nih.gov/articles/PMC11737758/

5. **Xu, H., et al. (2024).** *Machine learning-based analysis and prediction of heatstroke incidence using meteorological factors*. Frontiers in Public Health. https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2024.1420608/full
6. **Ku, Y., et al. (2022).** *Machine Learning Models for Predicting the Occurrence of Respiratory Diseases Using Climatic and Air-Pollution Factors*. PMC. https://pmc.ncbi.nlm.nih.gov/articles/PMC9149237/
7. **Hussain, M., et al. (2023).** *Machine learning-based efficient prediction of positive cases of malaria and typhoid using environmental data*. BMC Medical Informatics and Decision Making. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-022-02092-1
8. **Panja, M., et al. (2022).** *An ensemble neural network approach to forecast Dengue outbreak based on climatic conditions*. arXiv. https://arxiv.org/abs/2212.08323
9. **Alsuhaibani, A., & Alhaidari, A. (2021).** *Weather impact on daily cases of COVID-19 in Saudi Arabia using machine learning*. arXiv. https://arxiv.org/abs/2105.03027
10. **Zhao, W., & Efremova, N. (2024).** *Grapevine Disease Prediction Using Climate Variables from Multi-Sensor Remote Sensing Imagery via a Transformer Model*. arXiv. https://arxiv.org/abs/2406.07094

9. Outcome



🧑‍⚕️ Weather & Disease Prediction (Demo)

🌦️ Weather & Demographics

🌡️ Temperature (°C): 30.00

💧 Humidity (%): 50.00

🌬️ Wind Speed (km/h): 10.00

🎂 Age: 25

🚻 Gender: ● Male ○ Female

🤒 Symptoms

Type or select symptoms: runny_nose × sneezing × fever ×

✅ **Predicted Disease: Common Cold**

📊 Probability Distribution (Top 5)