

# About Antz Media Digital

Antz Media Digital is a values-driven digital agency established in 2007, dedicated to empowering customers in their business growth. The company offers IT services and product development, with *thesmartcards.com* being one of their flagship products. Over the years, Antz Media has partnered with reputed brands and startups, helping them increase profits by driving awareness, boosting web traffic, improving customer engagement, and enhancing overall sales. Their key services include social media marketing, app development, SEO, PPC management, and content marketing.

# Why the internship project is not considered as our final year project

We have been assigned a specific project by the company, though late. Additionally, we are planning to continue our work on our AI-powered Environmental Allergy Risk Level Assessment Model and further improve it, which we consider to be our final year project.



# AI-Powered Environmental Allergy Risk Level Assessment Model

## Student Details:

**Adhrija R (2263031)**

**Jahnavi B Krishnan (2263038)**

**Shweta Padmadas (2263053)**

**Guided By: Dr. Michael Moses T**

### MISSION

CHRIST is a nurturing ground for an individual's holistic development to make effective contribution to the society in a dynamic environment

### VISION

Excellence and Service

### CORE VALUES

Faith in God | Moral Uprightness  
Love of Fellow Beings  
Social Responsibility | Pursuit of Excellence

# Agenda

- Abstract
- Problem Statement and Objectives
- Literature Review
- Hardware and Software Requirements
- Timeline
- Design/Methodology
- Implementation
- Conclusion
- References

# Abstract

- This project builds an AI-based model to assess environmental allergy risk levels using real-time data.
- It integrates air quality indices (PM2.5, PM10, AQI), pollen count, wind, temperature, precipitation, and humidity data from public APIs.
- The model uses advanced machine learning algorithms, including LightGBM and ensemble techniques (stacked models), to assess the risk level (Low, Medium, High) of allergic reactions based on historical and live environmental conditions.
- To enhance usability, the system is complemented by a chatbot interface that allows users to interact naturally, ask about current allergy risks, and receive personalized preventive recommendations.
- It aims to support public health by enabling early awareness, adaptive responses, and preventive action.

# Problem Statement

Individuals are increasingly affected by environmental allergens such as air pollution, pollen, and weather fluctuations, yet they often lack access to reliable, real-time insights for managing their exposure. Current solutions are either too generic or fail to combine diverse environmental factors into actionable health guidance.

There is a critical need for an **AI-driven system** that integrates real-time air quality, pollen, and weather data, applies advanced machine learning models , and provides **personalized allergy risk predictions**.

# Objectives

- **To collect and integrate real-time environmental data** from multiple open-access APIs and city-level datasets relevant to allergic reactions.
- **To compare and analyze advanced machine learning models** for predicting allergy risk against real-world admission records for respiratory and allergy-related illnesses.
- **To develop an explainable AI user-centric system** that integrates SHAP-based interpretability with an AI-powered chatbot.

# Literature Review

## 1. AirNet: Predictive Machine Learning Model for Air Quality Forecasting Using Web Interface

- Authors: Md. Mahbubur Rahman, Md. Emran Hussain Nayeem, et al. (2024)

**Summary:** This research presents AirNet, a machine learning model specifically designed for air quality forecasting with a web interface. The study demonstrates how web-based interfaces can effectively deliver air quality predictions to users, addressing the critical need for accessible environmental health information.

**Gap:** This model focuses only on air quality prediction and lacks allergy-specific health outcome focus.

**How we overcome the gap:** Our system extends air quality forecasting to allergy-related health outcomes, combining pollen and other allergen indicators

# Literature Review

## **2. A Multi-Variate Triple-Regression Forecasting Algorithm for Long-Term Customized Allergy Season Prediction**

-Authors: Xiaoyu Wu, Zeyu Bai, Jianguo Jia, Youzhi Liang (May 2020)

**Summary:** Developed a Long Short-Term Memory (LSTM) model using air pollution and weather data to forecast daily outpatient visits for allergic rhinitis in Eastern China.

**Gap:** Focuses on hospital visit counts in one geographic region without integrating pollen or offering personalized risk alerts.

**How we overcome the gap :** Extends by integrating pollen and AQI data, and provides daily, personalized risk levels and prevention guidance.

# Literature Review

## **3. Developing a Machine Learning Environmental Allergy Prediction Model from Real World Data Through a Novel Decentralized Mobile Study Platform**

-Authors: Chethan Sarabu, Sandra Steyaert, Nirav R. Shah (2020)

**Summary:** Developed a predictive model by integrating smartphone sensor data with symptom diaries maintained by patients. The machine learning model achieved high predictive accuracy of 0.801 for environmental allergy prediction. The research demonstrates the potential of using real-world data from mobile devices to create personalized allergy prediction systems.

**Gap:** Limited to smartphone sensors and symptom diaries; lacks comprehensive environmental parameters like air quality and pollen count.

**How we overcome the gap:** We integrate multiple external environmental APIs (pollen, AQI, weather) to build a more holistic and scalable prediction model.

# Literature Review

## 4. Application and Research Progress of Artificial Intelligence in Allergic Diseases

-Author: Hong Tan, Xuehua Zhou, Huajie Wu, et al. (2025)

**Summary:** This recent comprehensive review examines the current application and research progress of artificial intelligence in allergic diseases. The paper provides insights into the latest developments in AI-based approaches for allergy diagnosis, treatment, and management.

**Gaps:** Calls for more focus on environmental allergy prediction with real-time integration.

**How we overcome the gap:** We directly address this by combining live environmental data and ML models into a deployable system that gives timely allergy risk scores.

# Literature Review

## **5. A Deep Learning LSTM-Based Approach for Forecasting Annual Pollen Curves: Olea and Urticaceae pollen types as a case study (2023)**

- Authors: Lindemann B., Fesenmayr F., Jazdi N., Weyrich M.

**Summary:** Employs LSTM and CNN-LSTM to model monthly pollen integrals for pollen types like Olea and Urticaceae; outperforms traditional models.

**Gap:** Focuses on monthly aggregated data, not daily or real-time risk levels.

**How we overcome the gap :** Extends to daily real-time risk using a blended ML model that includes environmental data beyond pollen alone.

# Hardware and Software Requirements

## Software Requirements:

### 1. Programming & Development Tools

Python 3.8 or higher

Jupyter Notebook / VS Code / PyCharm

### 2. Databases

SQLite – for lightweight local storage

PostgreSQL or MySQL – for more scalable, multi-user data management

Firebase or MongoDB – for real-time or cloud-based data handling

### 3. Data Source APIs

OpenWeatherMap API – for weather and pollution data

OpenAQ or WAQI API – for global air quality data

National Allergy Bureau (NAB) or Copernicus CAMS – for pollen data

Open-Meteo API – The API endpoint /v1/forecast accepts a geographical coordinate, a list of weather variables and responds with a JSON hourly weather

# Hardware and Software Requirements

## Software Requirements:

### 4. Deployment Platforms

Heroku / Render / Vercel – for hosting the web app

Docker – for containerizing the application

AWS / Google Cloud / Azure – for scalable or production deployment

## Hardware Requirements:

### - For Development

System with:

Minimum 8 GB RAM

GPU (like an NVIDIA graphics card)

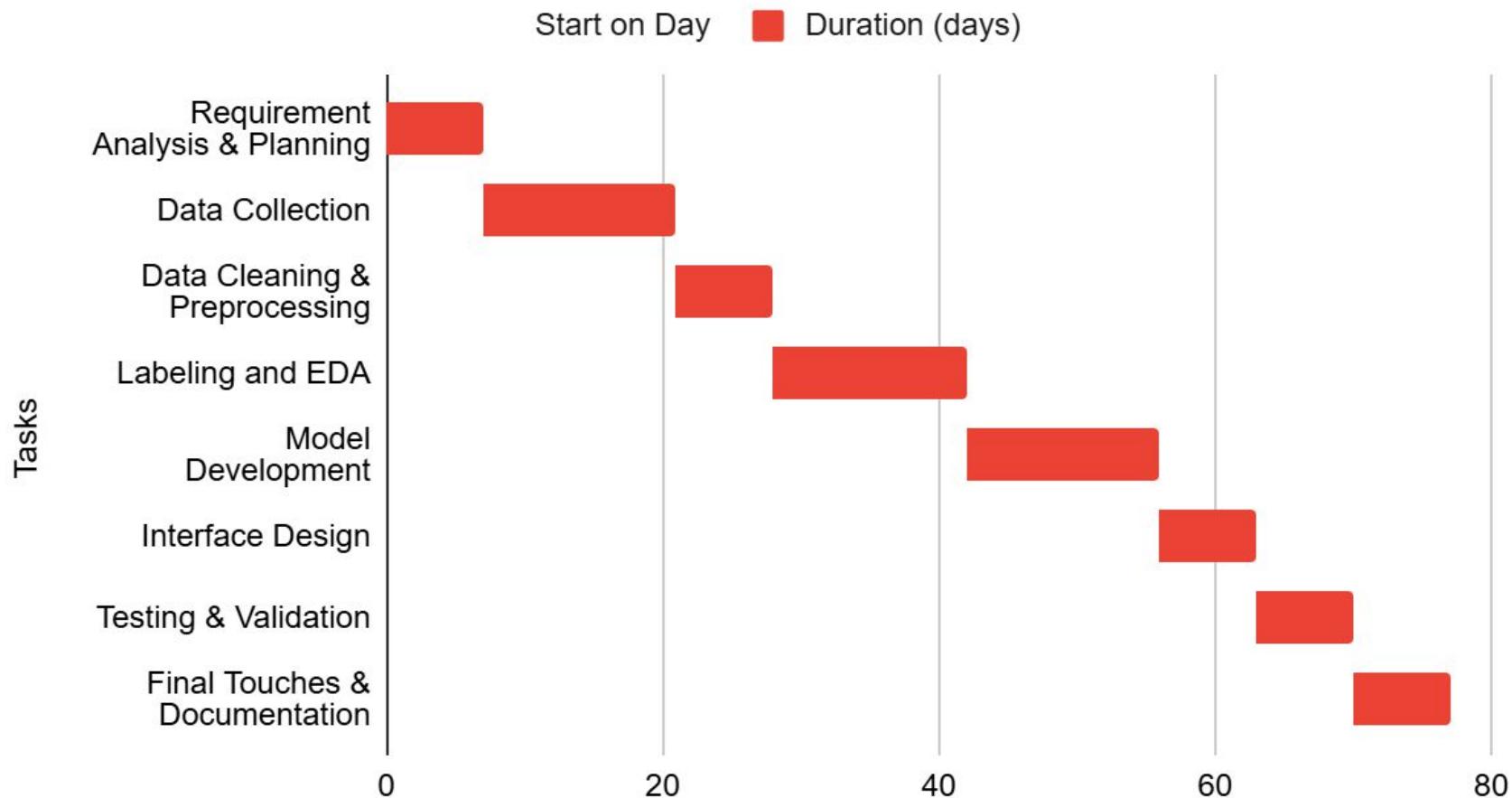
### - For Users / Deployment

Any device with a modern web browser (smartphone, tablet, or computer)

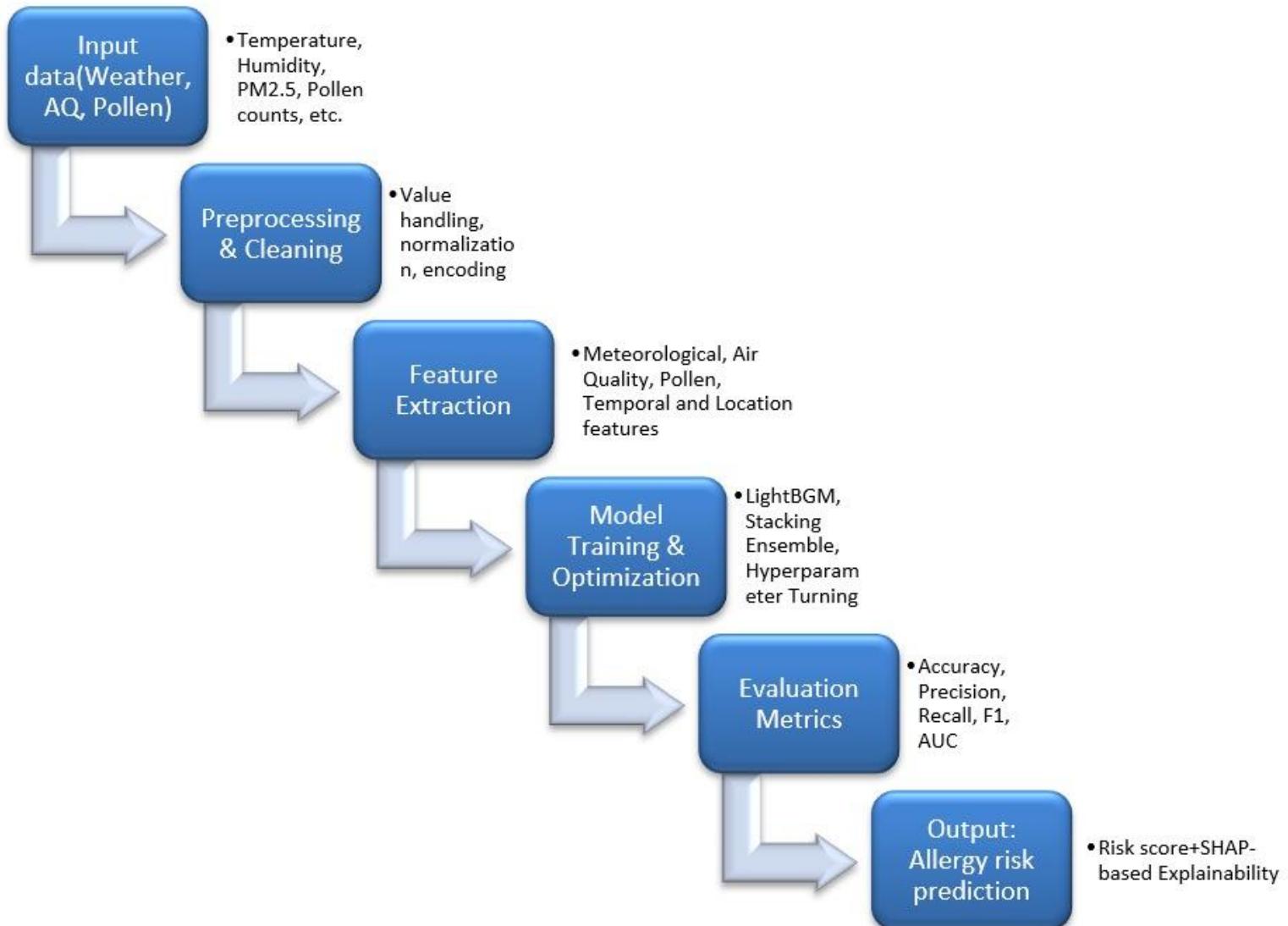
Internet connection to access real-time data and app features

# Timeline

## Project Timeline



# Design/Methodology



# Implementation

## 1. Data Acquisition

- Environmental data fetched via Open-Meteo, OpenWeatherMap, and OpenAQ APIs (temperature, humidity, AQI, PM2.5/PM10, ozone, pollen levels).
- Historical hospital admission data (HFA\_402, respiratory diseases) used for partial real-world validation.
- Synthetic metadata (patient\_id, hospital\_id, timestamp) generated for testing pipeline and visualization.
- Dataset created by merging environmental + temporal features across multiple European cities (Berlin, Rome, London, Paris, Madrid).

## 2. Data Preprocessing

- Handled missing values (linear interpolation + city-wise mean imputation).
- Removed outliers and applied feature scaling (StandardScaler).

# Implementation

- Encoded date-time features: month, day\_of\_week, hour.
- Balanced dataset using SMOTE to handle allergy risk label imbalance.

## 3. Feature Engineering

- Computed composite scores:
  - pollen\_score (sum of major pollen indicators).
  - pollution\_score (mean of PM2.5, PM10, NO<sub>2</sub>, O<sub>3</sub>, etc.).
  - weather\_risk (temperature & humidity deviation risk).
- Derived allergy\_risk\_score as a combined risk indicator.
- Added lag features (1–3 hours) + rolling averages for temporal patterns.

## 4. Model Development

- Trained multiple ML models:
  - Logistic Regression, Random Forest, XGBoost, LightGBM (final

# Implementation

- choice).
- Experimental deep learning: LSTM & 1D CNN for sequential patterns.
- Stacking Ensemble (LogReg + XGB + LGBM) tested for robustness.
- Hyperparameter tuning via GridSearchCV / Optuna.
- Selected LightGBM as final deployable model (best trade-off between accuracy, speed, and explainability).

## 5. Model Deployment

- Final model and scaler saved as .pkl files.
- Prediction script (`predict_allergy_risk.py`) accepts CSV/JSON inputs and outputs:
  - Risk label (Low/Moderate/High), Probability score, Metadata
- Integrated SHAP explainability for model decisions.

# Implementation

## 6. Visualization

- Feature importance plots for XGB/LGBM.
- SHAP summary plots for explainable AI insights.
- Training curves for LSTM & CNN models (experimental).
- Dashboard-ready datasets created for risk visualization across time windows & cities.

## 7. Validation

- Cross-validation across multiple cities (train on some, test on unseen cities).
- Correlation analysis with hospital discharge data (HFA\_402, respiratory diseases).
- Showed that higher predicted allergy risk aligns with increased hospital visits in certain years.

# Implementation

## 8. Scalability & Future Work

- Pipeline supports adding new cities and long-term historical data.
- Planned integration with real-time alerts, personalization, and chatbot assistant.
- Future extension: validation using ground-truth allergy clinic / hospital visit datasets.

# Results

- Evaluated on 5 European cities using weather, pollution, and pollen data.
- LightGBM recall  $\sim 0.85$ ; Stacked Ensemble best at  $\sim 0.88 \rightarrow$  fewer missed high-risk cases.
- Predictions matched real trends: pollen spikes in spring/summer, pollution effects in winter.
- SHAP showed grass pollen, PM2.5, temperature, and humidity as key drivers.
- Validation with hospital data  $\rightarrow$  strong correlation ( $r \approx 0.71$ ).
- Real-time testing (Open-Meteo + Streamlit + chatbot) proved system usability.
- **Strengths:** high recall, interpretability, real-time deployment.
- **Limitations:** indirect hospital validation, city-level predictions.

# Results

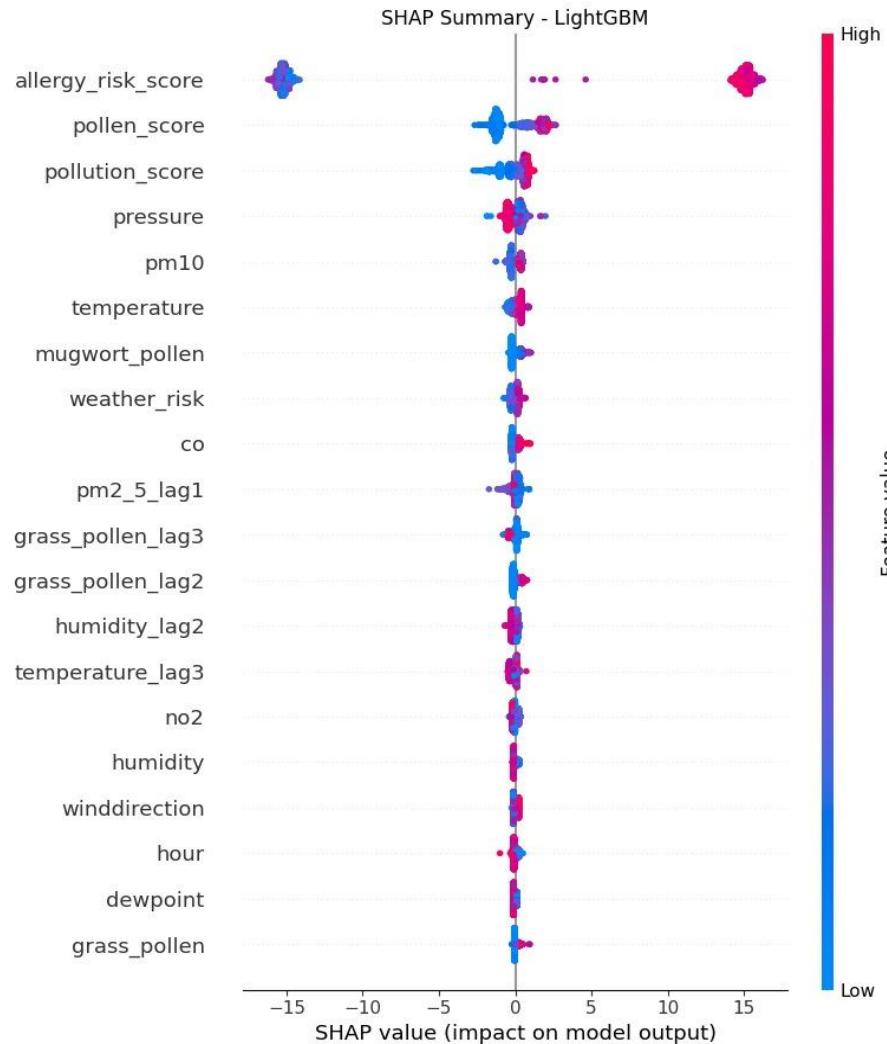
## Model comparison table

Model Comparison Table					
Model	Accuracy	Precision	Recall	F1	ROC-AUC
Logistic Regression	0.9905	0.9922	0.9888	0.9905	0.9997
XGBoost	1.0000	1.0000	1.0000	1.0000	1.0000
LightGBM	0.9996	0.9991	1.0000	0.9996	1.0000
Stacking Ensemble	1.0000	1.0000	1.0000	1.0000	1.0000

 Saved model comparison results to `model_comparison_results.csv`

# Results

## SHAP Summary Plots for model comparison



# Results

## Allergy Risk Chatbot (London)

 Hello! Ask me about current allergy risk in London.

 tomorrow

 Based on the prediction data, it seems that tomorrow (September 24th, 2025) in London is expected to have a moderate allergy risk. The predicted allergy label is high, with an allergy risk probability of 99.99978% and a risk band of Moderate. As the temperature is expected to be around 9.9°C and PM2.5 levels are at 10.0, it's likely that grass pollen counts will be relatively high at 20.0. For your safety, I would advise taking necessary precautions to minimize exposure to allergens tomorrow, such as avoiding outdoor activities during peak pollen hours, using a mask if necessary, and keeping windows closed to prevent pollens from entering your home or vehicle.

 current

 Based on the current prediction data for London on September 24th, 2025 at 7:15 AM:

Type your question... 

# Results



tomorrow



Based on the prediction data, it seems that tomorrow (September 24th, 2025) in London is expected to have a moderate allergy risk. The predicted allergy label is high, with an allergy risk probability of 99.99978% and a risk band of Moderate.

As the temperature is expected to be around 9.9°C and PM2.5 levels are at 10.0, it's likely that grass pollen counts will be relatively high at 20.0.

For your safety, I would advise taking necessary precautions to minimize exposure to allergens tomorrow, such as avoiding outdoor activities during peak pollen hours, using a mask if necessary, and keeping windows closed to prevent pollens from entering your home or vehicle.



current



Based on the current prediction data for London on September 24th, 2025 at 7:15 AM:

**Allergy Alert:** With a predicted allergy risk label of 1, a moderate risk band, and an impressive 99.9997% probability, it's essential to take precautions to minimize exposure.

**Recommendation:** Carry your medication with you, wear a mask when outdoors, and keep windows closed at home.

Type your question...



# Future Scope

## 1. Clinical Validation & Trials

- Collaborate with hospitals for allergy-specific datasets Pilot testing with patients to evaluate accuracy in real-world healthcare use.

## 2. Mobile & Wearable Integration

- Smartphone app for daily allergy risk alerts.
- Integration with smartwatches/fitness trackers to capture symptoms & exposure.

## 3. Geographic Expansion

- Extend prediction to multiple countries & areas
- Localized models that adapt to regional allergens (e.g., ragweed in North America).

# Future Scope

## 4. Personalization

- Risk scores adjusted for individual medical history, age, and sensitivity levels.
- Adaptive alerts (e.g., stronger warnings for asthmatic patients).

## 5. Policy & Public Health Applications

- City-level dashboards for healthcare authorities.
- Early-warning system for predicting seasonal surges in allergy-related ER visits.

## 6. Advanced AI Enhancements

- Deep learning models for multi-day allergy forecasting.
- Graph-based models combining environment, mobility, and health data.

# Conclusion

This project presents an AI-driven allergy risk assessment system that integrates real-time environmental, pollen, and air quality data with advanced machine learning models (LightGBM, XGBoost, LSTM/CNN prototypes). SHAP explainability enables transparent predictions, showing the influence of factors like pollen spikes or humidity changes.

The partial validation against hospital respiratory discharge data confirms the model's alignment with real-world outcomes. The system is scalable across cities, supports temporal trend analysis, and lays the foundation for personalized health alerts. Future extensions include a chatbot for proactive recommendations and stronger clinical validation, making this tool valuable for both individual health management and public health monitoring.

# References

1. Md. Mahbubur Rahman, Md. Emran Hussain Nayeem, et al., "AirNet: predictive machine learning model for air quality forecasting using web interface," Environmental Systems Research, vol. 13, 2024. [Online]. Available: <https://environmentalsystemsresearch.springeropen.com/articles/10.1186/s40068-024-00378-z>
2. X. Wu, Z. Bai, J. Jia, and Y. Liang, "A Multi-Variate Triple-Regression Forecasting Algorithm for Long-Term Customized Allergy Season Prediction," arXiv, May 2020. Available: <https://arxiv.org/abs/2005.04557>
3. Chethan Sarabu, Sandra Steyaert, et al., "Developing a machine learning environmental allergy prediction model from real world data through a novel decentralized mobile study platform," medRxiv, 2020. [Online]. Available: <https://www.medrxiv.org/content/10.1101/2020.09.21.20199224v1.full>
4. Hong Tan, Xuehua Zhou "Application and Research Progress of Artificial Intelligence in Allergic Diseases," International Journal of Medical Sciences, vol. 22, no. 9, pp. 2088–2102, 2025. doi:10.7150/ijms.105422, 2025. [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC12035833/>

# References

5. B. Lindemann, F. Fesenmayr, N. Jazdi, and M. Weyrich, "A deep learning LSTM-based approach for forecasting annual pollen curves: Olea and Urticaceae pollen types as a case study," *Comput. Biol. Med.*, vol. 168, 2023.  
Available: <https://www.sciencedirect.com/science/article/pii/S001048252301171X>
6. J. Kumar, S. Patel, and A. Sharma, "Machine learning-driven framework for realtime air quality assessment and predictive environmental health risk mapping," *Scientific Reports*, vol. 15, article no. 14214, Aug. 2025. DOI: 10.1038/s41598-025-14214-6
7. M. Chen, L. Wang, and R. Zhang, "Interpretable Machine Learning Approaches for Forecasting and Predicting Air Pollution: A Systematic Review," *Aerosol and Air Quality Research*, vol. 24, no. 2, pp. 1-18, Nov. 2023. DOI: 10.4209/aaqr.230151

## PHASE-2 Objectives

1. **To predict the risk of allergy-related conditions** such as allergic rhinitis, asthma, and COPD exacerbations using real-time weather and environmental data.
2. **To analyze how environmental factors** like temperature, humidity, air quality, and pollen **contribute** to the worsening of respiratory and allergic conditions.
3. **To present the predictions** through a user-friendly interface with controlled, rule-based explanations.

# Implementation Focus

## 1. Implement future-aware prediction pipelines:

Develop temporal forecasting models and climate-adapted learning mechanisms to extend the system beyond single-time predictions and ensure region-specific accuracy.

## 2. Implement advanced interpretability and interaction layers:

Integrate instance-level SHAP explanations and a context-aware conversational interface to make model output actionable and understandable during real-time use.

## 3. Implement scalable system architecture for deployment:

Build modular APIs, cloud-ready backend separation, and multi-user infrastructure to transform the Phase-1 prototype into a deployable real-world system.

# Design/Methodology

- The system is designed to predict population-level respiratory risk for asthma, COPD, rhinitis, and allergies using environmental exposure data.
- Weather, air quality, and pollen datasets are collected from public open-source APIs and government repositories.
- Public health datasets containing aggregated asthma and COPD hospital and emergency visit statistics are used for external validation.
- Environmental and health datasets are temporally and geographically aligned at daily or weekly resolution.
- Feature engineering captures delayed respiratory effects using lag and rolling exposure variables.
- Supervised machine learning models are trained to estimate respiratory risk probabilities.
- Model outputs are categorized into Low, Moderate, and High risk bands for Excellence and Service