

Air Pollution Forecasting and Monitoring System Using Machine Learning

Darshan S. Mali¹, Sanjana P. Bangar², Aditi R. Ugale³

^{1,2,3}Department of Computer Science and Design, MET's Institute of Technology (P) B.Tech., Nashik, Maharashtra, India

Affiliated to Dr. Babasaheb Ambedkar Technological University, Lonere

Abstract - Air pollution has become a major environmental and public health concern in urban regions. Traditional monitoring systems are limited by high cost and low spatial coverage. This paper presents a machine learning-based air pollution forecasting and monitoring system that predicts air quality index (AQI) using meteorological and historical pollutant data. A Random Forest regression model is used to estimate pollutant concentrations and compute AQI values. The system integrates real-time data and provides short-term forecasts along with visualization through a web-based dashboard. Experimental results show that the proposed model achieves improved accuracy compared to traditional methods, making it suitable for real-time environmental monitoring and decision-making.

Key Words: Air Quality Index, Random Forest, Air Pollution Forecasting, Machine Learning, Real-Time Monitoring, Flask

1. INTRODUCTION

Air pollution has become a critical environmental and public health issue in urban areas due to the continuous rise in industrial activity, transportation, and energy consumption. Pollutants such as Nitrogen Dioxide (NO₂) and Ozone (O₃) significantly affect air quality and contribute to respiratory and cardiovascular diseases.

Conventional air quality monitoring systems depend mainly on fixed monitoring stations that are expensive to install and maintain. Although these systems provide accurate observations, their limited spatial coverage reduces their effectiveness for real-time monitoring across large urban regions.

Recent studies have applied machine learning techniques for pollutant prediction, but many existing systems lack seamless real-time data integration and practical user-oriented visualization. In addition, short-term forecasting remains challenging when pollutant behavior varies across time and location.

This work proposes a machine learning-based air pollution forecasting and monitoring system that predicts pollutant levels and computes AQI using meteorological and historical air quality data. The system combines

prediction capability with a web-based dashboard for visualization and supports short-term forecasting for improved environmental monitoring. The main contributions of this work are:

- Development of a machine learning-based AQI prediction system
- Integration of real-time meteorological data
- Implementation of a web-based interactive dashboard
- Generation of short-term air quality forecasts

2. LITERATURE REVIEW

The World Health Organization (2021) presented global air quality guidelines on pollutant exposure and identified major health risks associated with poor air quality. However, the report is limited by its policy-oriented nature and does not provide a predictive forecasting framework.

Kumar and Singh (2022) used ground-based air quality monitoring network analysis on urban monitoring infrastructure and achieved a clear assessment of spatial and operational limitations. However, the approach is limited by its dependence on static monitoring systems without predictive capability.

The European Space Agency (2022) used Sentinel-5P TROPOMI atmospheric observations on large-scale pollutant datasets and achieved wide spatial coverage of air pollution indicators. However, the approach is limited by the difficulty of directly mapping satellite observations to ground-level pollutant concentrations.

Chen, Wang, and Li (2019) used Random Forest on meteorological and pollutant datasets and achieved better prediction accuracy than several conventional machine learning methods. However, the approach is limited by restricted focus on model performance without real-time system integration.

Zhang et al. (2022) used deep learning on multi-pollutant temporal datasets and achieved strong performance in capturing complex nonlinear relationships. However, the approach is limited by high

computational cost and greater dependence on large-scale training data.

Goldberg (2020) used satellite-based remote sensing on atmospheric pollutant measurements and achieved effective large-area monitoring of pollutant distribution. However, the approach is limited by the requirement for additional processing to estimate localized surface conditions.

Copernicus Climate Change Service (2023) used ERA5 meteorological reanalysis data and achieved reliable large-scale meteorological support for air quality studies. However, the approach is limited by the need for integration with pollutant prediction models for real-time forecasting.

Central Pollution Control Board (2022) used standardized AQI methodology on Indian air quality data and achieved a structured framework for pollutant-based air quality assessment. However, the approach is limited by its focus on index calculation rather than predictive monitoring.

Existing methods lack real-time integration, interactive visualization, and accurate short-term forecasting, which motivates the proposed system.

3. METHODOLOGY AND SYSTEM DESIGN

3.1 System Architecture

The proposed system follows a web-based architecture for forecasting and monitoring air pollution levels. It accepts meteorological inputs and historical pollutant data, processes them using machine learning models, computes AQI values, and displays the results through an interactive dashboard. The system also supports short-term forecasting and visual representation of pollutant trends for multiple monitoring locations.

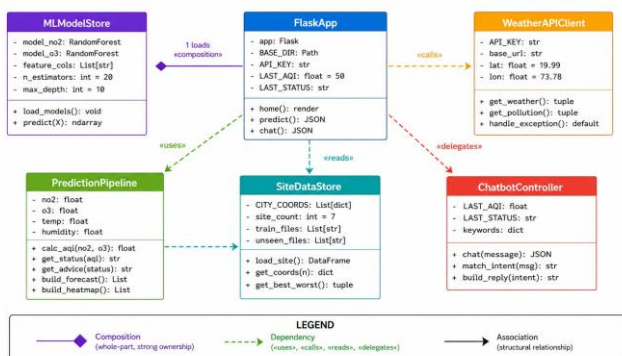


Fig -1: UML Diagram of Proposed System

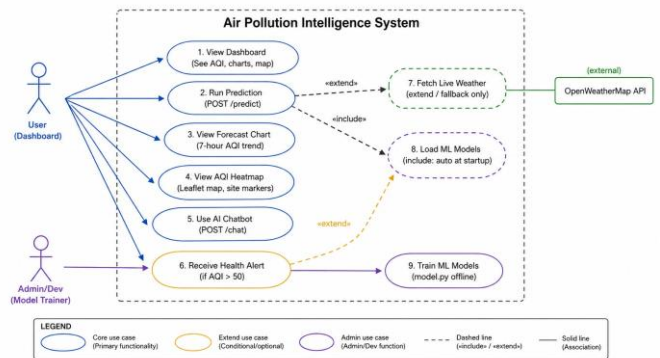


Fig -2: Use Case Diagram

3.2 Dataset and Data Preparation

The dataset consists of historical air quality and meteorological data collected from multiple monitoring locations. The data includes temperature, humidity, wind speed, and pollutant concentrations. The dataset is divided into training (80%) and testing (20%) sets.

Before model training, the collected data is preprocessed to handle missing values, improve consistency, and prepare the input variables for prediction. Meteorological variables and pollutant measurements are organized into a structured format suitable for regression analysis.

3.3 Machine Learning Model

Random Forest is an ensemble learning method that combines multiple decision trees to improve prediction accuracy and reduce overfitting. In this work, separate regression predictions are generated for Nitrogen Dioxide (NO₂) and Ozone (O₃) using the available meteorological and historical pollutant features. The predicted pollutant concentrations are then used for AQI computation and further forecasting.

3.4 AQI Computation

The Air Quality Index is computed using the predicted pollutant values for Nitrogen Dioxide (NO₂) and Ozone (O₃). The AQI calculation used in the system is:

$$AQI = \max(I_{NO_2}, I_{O_3})$$

The resulting AQI value is used to indicate the overall air quality condition and to support health-related interpretation in the monitoring dashboard.

3.5 Forecast Generation

The system generates a 7-hour forecast by using predicted values iteratively for future time steps. This forecasting mechanism helps estimate near-future pollutant behavior and supports early awareness of possible air quality deterioration. The generated forecast values are displayed through the web-based dashboard for easier interpretation.

4. IMPLEMENTATION

4.1 Technology Stack

The system is developed using Python and Flask for backend processing and web deployment. Machine learning functionality is implemented using standard Python libraries, and the dashboard provides visualization support for air quality monitoring results.

4.2 Dashboard and Monitoring

The dashboard presents pollutant concentrations, AQI values, and forecast trends in an interactive format. This improves accessibility of air quality information for users and supports real-time environmental observation across monitoring locations.

5. RESULTS AND ANALYSIS

The developed system was evaluated using standard regression performance metrics for Nitrogen Dioxide (NO₂) and Ozone (O₃) prediction. The obtained results indicate that the model provides reliable predictions for air quality monitoring and short-term forecasting.

The Random Forest model outperforms linear regression by providing lower RMSE and higher R² values. The model performs well across different monitoring locations, although slight errors are observed during peak hours.

Table -1: System Performance Summary (NO₂ and O₃ Models)

Metric	NO ₂ Model	O ₃ Model
RMSE	4.2	5.8
MAE	3.1	4.3
R ²	0.87	0.82

6. APPLICATIONS

The proposed system can support public health monitoring, smart city management, and environmental analysis by providing timely air quality predictions. It can also be used to assist in decision-making for authorities and the public during periods of poor air quality.

7. FUTURE SCOPE

Future improvements may include extension to additional pollutants, enhancement of forecast duration, and integration with a larger network of monitoring stations. The system can also be expanded with advanced machine learning models and mobile-based alert support.

8. CONCLUSIONS

This paper presents a machine learning-based air pollution forecasting and monitoring system for predicting Nitrogen Dioxide (NO₂) and Ozone (O₃) concentrations and computing AQI values. The use of Random Forest with real-time meteorological inputs and dashboard-based visualization makes the system suitable for practical short-term air quality monitoring.

ACKNOWLEDGEMENT

The authors acknowledge the Department of Computer Science and Design, MET's Institute of Technology (P) B.Tech., Nashik, and Dr. Babasaheb Ambedkar Technological University, Lonere, for institutional support. The authors also acknowledge the OpenWeatherMap API, Sentinel-5P TROPOMI, and ERA5 data providers for publicly accessible atmospheric datasets.

REFERENCES

- [1] World Health Organization, "WHO global air quality guidelines," WHO, 2021.
- [2] A. Kumar, R. Singh, "Limitations of ground-based air quality monitoring networks," *Journal of Environmental Science*, 2022.
- [3] European Space Agency, "Sentinel-5P TROPOMI data user guide," ESA Publications, 2022.
- [4] X. Chen, Y. Wang, Z. Li, "Air quality prediction using Random Forest and meteorological data," *IEEE Access*, 2019.
- [5] S. Zhang et al., "Deep learning for air quality prediction: A survey," *IEEE Transactions on Knowledge and Data Engineering*, 2022.
- [6] D. Goldberg, "Satellite remote sensing of atmospheric pollutants," *IEEE Geoscience and Remote Sensing Magazine*, 2020.
- [7] Copernicus Climate Change Service, "ERA5 reanalysis data documentation," 2023.