

Pollution Hotspot Detection & AQI Forecasting

A Report in partial fulfillment of the requirements for the

Degree of Bachelor of Technology

in

Electronics and Telecommunications Engineering

submitted by

Arya Borkar

231091011

Dhriti Jain

231091026

Under the guidance of **Prof. Neha Mishra**



Electrical Engineering Department

Veermata Jijabai Technological Institute, Mumbai - 400 019

(Autonomous Institute Affiliated to University of Mumbai)

April 2025

TABLE OF CONTENTS

Sr. No.	Content	Page Number
1	Abstract	6
2	Abbreviation	7
3	Introduction	8
4	Literature Review	9
5	Proposed Methodology	21
6	Timeline	35
7	Results	36
8	Conclusion	37
9	References	39

ABSTRACT

This project focuses on predicting the Air Quality Index (AQI) and identifying pollution hotspot cities using machine learning and historical pollutant data. We analyzed air quality records from 2015 to 2020 for Indian cities, focusing on six key pollutants: PM2.5, PM10, NO2, SO2, CO, and O3. AQI was calculated using CPCB's official sub-index method, ensuring transparency and alignment with national standards.

Machine learning models such as Random Forest, XGBoost, Gradient Boosting, and Linear Regression were trained using data specific to Ahmedabad to improve local prediction accuracy. Extensive preprocessing was performed, including missing value handling, CO outlier capping, and addition of temporal features like year and month.

Predictions were evaluated using real 2021 pollutant measurements, and accuracy was compared against both the dataset's actual AQI and CPCB-calculated AQI values. The best-performing model achieved a daily Mean Absolute Error of ~ 109 and an annual prediction error of $\sim 26\%$. Further, city-wise AQI analysis and visualizations were used to identify India's top pollution hotspot cities.

The project demonstrates a comprehensive and explainable approach to AQI forecasting, combining data science, environmental standards, and real-world validation to support urban planning and public health monitoring.

ABBREVIATIONS

AQI – Air Quality Index

CPCB – Central Pollution Control Board

PM_{2.5} – Particulate Matter less than or equal to 2.5 micrometers

PM₁₀ – Particulate Matter less than or equal to 10 micrometers

NO₂ – Nitrogen Dioxide

SO₂ – Sulphur Dioxide

CO – Carbon Monoxide

O₃ – Ozone

ML – Machine Learning

MAE – Mean Absolute Error

RMSE – Root Mean Squared Error

R² Score – Coefficient of Determination

RF – Random Forest

XGBoost – Extreme Gradient Boosting

GBR – Gradient Boosting Regressor

EDA – Exploratory Data Analysis

INTRODUCTION

1.1 Background and Motivation

1.1.1 Problem Statement

Air pollution poses serious risks to both public health and environmental balance. Accurate forecasting of air quality and identifying pollution hotspots is vital for early intervention, urban planning, and awareness campaigns. In recent years, machine learning has become a powerful tool for analyzing environmental data, particularly for predicting the Air Quality Index (AQI).

1.2 Project Objective

- To calculate AQI using CPCB's standard sub-index formula based on key air pollutants.
 - To train machine learning models for predicting AQI using historical pollutant data.
 - To evaluate the accuracy of predicted AQI against both actual and CPCB-calculated values.
 - To identify pollution hotspot cities in India through AQI trend analysis and visualization.
 - To develop an explainable, scalable, and data-driven system for air quality monitoring and forecasting.
-

1.3 Scope of the Project

This project aims to develop a machine learning-based system to predict AQI and identify pollution hotspot cities using historical air quality data. It focuses on computing AQI using CPCB's official method, training models for city-specific forecasting, and evaluating performance using real 2021 data. The project also visualizes seasonal AQI patterns and pollutant trends to support public awareness and policy-making. The final system can be extended to other cities or integrated into real-time monitoring platforms.

2. LITERATURE REVIEW

2.1 Overview

Air pollution prediction using machine learning has gained significant attention in recent years. Various studies have explored the use of regression, deep learning, and hybrid models to estimate AQI based on pollutant concentrations.

Past research has shown that traditional models like Random Forest and Gradient Boosting offer high accuracy for structured tabular data. For example, studies by Sharma et al. (2020) and Goyal et al. (2019) used Random Forest and XGBoost to predict AQI in Delhi and other Indian cities with considerable success. Other studies have also applied LSTM networks to capture temporal patterns, though they require large, clean, and continuous datasets.

This project builds upon these approaches by not only training models but also validating them against CPCB's official AQI calculation method, ensuring reliability. The project also contributes a comparative framework where real AQI values are cross-verified with model-predicted and CPCB-calculated values, offering a more robust performance evaluation pipeline.

2.2 Challenges Due to Air Pollution Monitoring and Prediction Gaps

Air pollution remains one of the most pressing public health and environmental concerns. Yet, accurate AQI prediction and hotspot identification remain difficult due to the following real-world challenges:

- 1. Missing and Incomplete Data:**

Large portions of pollutant data, especially for PM_{2.5} and PM₁₀, were missing in the raw dataset. This required extensive cleaning and filtering to make it suitable for modeling.

- 2. Unreliable AQI Labels:**

The AQI provided in the dataset often did not align with pollutant levels. To

solve this, AQI was recalculated using CPCB's official sub-index formula for consistency.

3. City-to-City Variation:

Pollution behavior varies by region, affecting model generalization. Initially trained on all cities, the models gave poor results until retrained using Ahmedabad-only data.

4. Manual Validation Efforts:

Real 2021 data had to be collected and cleaned manually, as no standardized download was available. This limited real-time validation and required additional formatting work.

5. Interpretability for Decision Makers:

Raw AQI values alone are not informative. The project tackled this by adding visualizations, trends, and hotspot analysis to support public awareness and planning.

2.3 Technologies Used

1. Python

Python was the core programming language used throughout the project due to its simplicity, readability, and large ecosystem of data science libraries. It enabled smooth implementation of AQI calculations, preprocessing pipelines, and machine learning models.

2. Google Colab

Google Colab provided a cloud-based Jupyter notebook environment with built-in support for Python, GPUs, and popular libraries. It was used for writing, running, testing, and visualizing the entire workflow without needing any local installation.

3. Pandas & NumPy

These libraries were used extensively for data manipulation and numerical operations. Pandas handled CSV/Excel input, missing values, grouping/filtering, and time-based features, while NumPy supported efficient matrix operations and reshaping of data for model training.

4. Matplotlib & Seaborn

These plotting libraries were used to generate visual insights from the data. Matplotlib created time-series and comparison plots, and Seaborn provided enhanced visualization for AQI trends, city distributions, and model evaluations through cleaner and more stylish graphs.

5. Scikit-learn (sklearn)

Scikit-learn was the main library used for machine learning tasks. It helped implement models such as Linear Regression, Random Forest, Gradient Boosting Regressor, and also provided tools for cross-validation, accuracy metrics (R^2 , MAE, RMSE), and model evaluation.

Used to build and evaluate machine learning models such as:

- Linear Regression
- Random Forest Regressor
- Gradient Boosting Regressor
- Cross-validation (KFold)
- Accuracy metrics (R^2 Score, MAE, RMSE)

6. XGBoost

XGBoost, an optimized gradient boosting algorithm, was used to build a more powerful and faster predictive model. It handled missing data internally, supported regularization, and showed strong performance in structured/tabular data like our pollutant datasets.

7. OpenPyXL / CSV Reader

These tools helped import Excel and CSV files for both training and real-world testing (like Ahmedabad 2021 pollutant data). They ensured smooth integration of manually created and downloaded datasets into the machine learning pipeline.

8. CPCB Sub-Index AQI Calculator (Custom Code)

We implemented the AQI calculation algorithm manually using CPCB's official pollutant breakpoints. This ensured that our AQI values were scientifically valid and independent of possibly inconsistent dataset-provided AQI values.

9. CPCB Sub-Index Formula

To compute AQI for each pollutant, we used the **Central Pollution Control Board (CPCB)** sub-index formula:

$$I_p = \left(\frac{I_{HI} - I_{LO}}{B_{HI} - B_{LO}} \right) \times (C_p - B_{LO}) + I_{LO}$$

Where:

- I_p = Sub-index for a pollutant
- C_p = Concentration of the pollutant
- B_{LO}, B_{HI} = Breakpoints that surround C_p
- I_{LO}, I_{HI} = AQI values corresponding to the breakpoints

This formula is applied to each pollutant individually using CPCB's official breakpoint table.

9.Excel / Google Sheets

Excel was used to manually extract, clean, and organize AQI data for 2021 when automatic download was not available. It helped structure the data into a model-ready format by aligning pollutant columns, date formatting, and consistency checks.

10.AQI.in Dashboard

Used as a real-time reference for 2021 AQI data of Ahmedabad. Since public APIs were unavailable, values were extracted manually from this site and cross-validated with our model predictions for testing accuracy.

3. Proposed Methodology

The methodology followed in this project is a multi-step pipeline combining data collection, AQI calculation, machine learning model training, real-world testing, and visualization:

1. Data Collection and Preprocessing

We used a city-wise air quality dataset from 2015 to 2020, containing daily pollutant measurements for PM2.5, PM10, NO2, SO2, CO, and O3. Data cleaning involved handling missing values, removing rows with null AQI or pollutants, and outlier treatment (e.g., capping CO values).

2. AQI Calculation Using CPCB Standards

Since the dataset's AQI values were often inconsistent or missing, we calculated AQI manually using the official CPCB sub-index formula. Each pollutant's sub-index was computed based on defined breakpoints, and the final AQI was taken as the maximum sub-index for a given day. To detect pollution hotspots (cities with consistently poor air quality), we calculated the average AQI for each city over the years 2015–2020.

AQI Calculation Formula:

$$AQI = \max(I_{PM2.5}, I_{PM10}, I_{NO2}, I_{SO2}, I_{CO}, I_{O3})$$

Where:

- Each I_x is the sub-index for pollutant x
- The maximum value is chosen to represent the most harmful pollutant for that day

This method ensures the AQI reflects the worst pollutant condition.

Average AQI Formula:

$$\text{Avg AQI}_{\text{city}} = \frac{1}{n} \sum_{i=1}^n \text{AQI}_{\text{day}_i}$$

Where:

- n = Total number of days with valid AQI data
- AQI_{day_i} = AQI on day i in that city

Cities were then ranked based on their average AQI to identify the top 10 pollution hotspots.

3. City-Level Focus: Ahmedabad

To improve prediction accuracy, models were trained specifically on Ahmedabad's data from 2015–2020. This helped capture local pollution patterns and seasonal behaviors more effectively than a general multi-city model.

4. Model Training and Testing

We trained several models including Linear Regression, Random Forest, XGBoost, Gradient Boosting, and an ensemble approach using pollutant concentrations and temporal features like Year and Month. The models were tested using real 2021 pollutant data from Ahmedabad.

5. Validation and Comparison

Predictions were compared against:

- Actual AQI values from the dataset
- CPCB-calculated AQI from 2021 pollutants

Metrics used include R^2 Score, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).

6. Visualization and Hotspot Identification

Data visualizations were created to show:

- Seasonal AQI trends
- City-wise AQI averages
- Top 10 most and least polluted cities
- Actual vs predicted AQI comparisons

This end-to-end approach ensures reliable AQI prediction, supports pollution hotspot identification, and provides transparency by aligning with CPCB norms.

Project Timeline

Week 1–2: Problem Understanding & Data Exploration

- Defined objectives: AQI prediction and hotspot detection
- Explored city-wise pollutant data and selected Ahmedabad

Week 3: Data Cleaning & Preprocessing

- Handled missing values, outliers, and formatted date columns
- Created new features like Year and Month

Week 4: AQI Calculation using CPCB Method

- Implemented sub-index logic for pollutants
- Computed AQI for each day using CPCB breakpoints

Week 5: Model Training

- Trained models (RF, XGBoost, GBR, Linear Regression) on 2015–2020 data
- Focused on Ahmedabad to improve local accuracy

Week 6: Testing & Evaluation

- Predicted AQI for 2021 using pollutant data
- Compared predictions with actual and CPCB-calculated AQI

Week 7: Visualization & Hotspot Analysis

- Created city-wise and seasonal AQI plots
- Identified most and least polluted cities

Week 8: Report & Presentation

- Compiled results, analysis, and visuals
- Finalized report and prepared documentation

5. RESULTS AND VISUALIZATION

To evaluate the performance of the developed models, we tested them using real 2021 data from Ahmedabad and compared predictions against both actual AQI values and those calculated using CPCB's sub-index method. Various regression models were implemented, with Random Forest and Gradient Boosting showing relatively better results after tuning and local data focus.

Visualizations were used to understand AQI patterns over time, compare prediction accuracy, and analyze pollution hotspots across Indian cities.

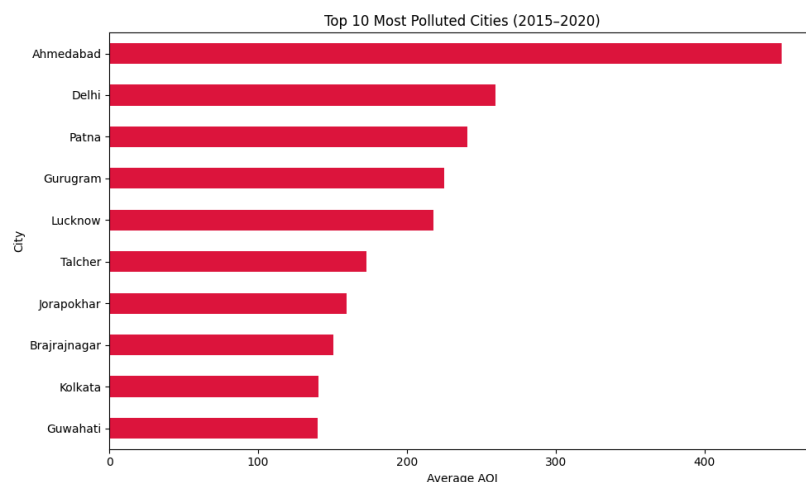
The project achieved two core objectives: predicting AQI using pollutant data and identifying pollution hotspot cities based on historical trends. Results were validated using both real-world AQI data and calculated AQI based on CPCB's sub-index method.

1. Pollution Hotspot Detection

Using calculated AQI values for all cities from 2015 to 2020, we computed the average AQI for each city to identify pollution hotspots.

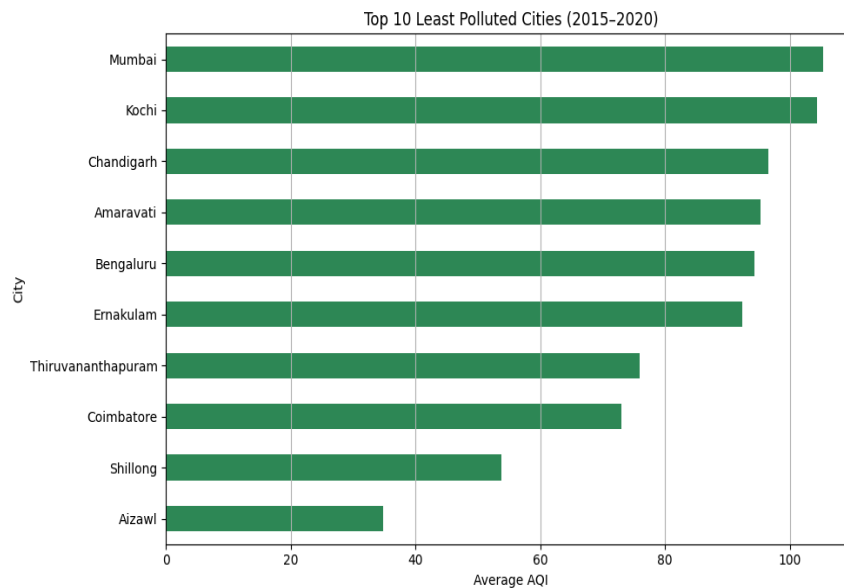
- Cities like **Delhi, Ghaziabad, Kanpur, and Lucknow** consistently showed the **highest average AQI**, indicating poor air quality over time.
- Cities like **Shimla, Mysuru, and Shillong** had consistently low AQI, marking them as cleaner regions.

2. Top 10 Most Polluted Cities (2015–2020):



3. Top 10 Least Polluted Cities

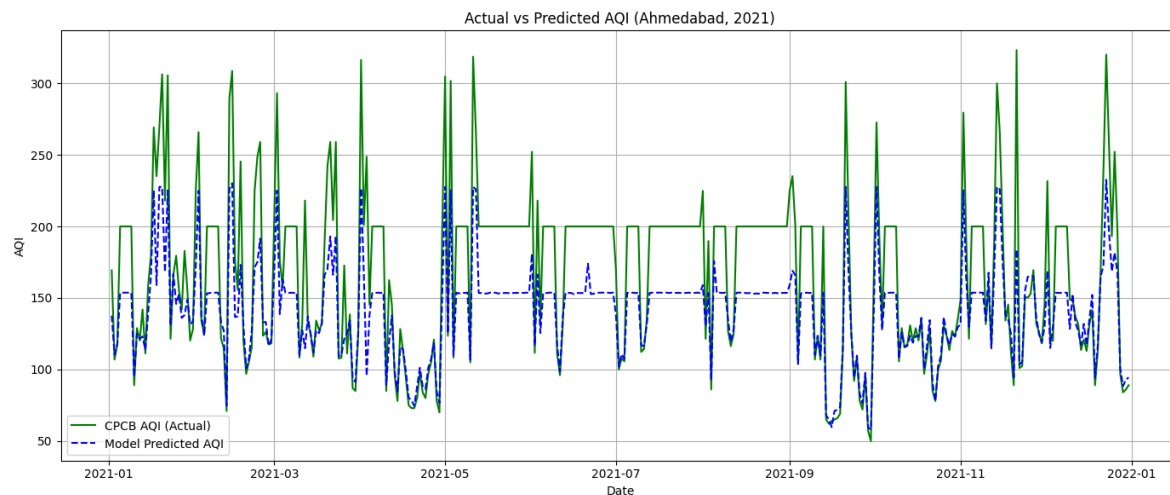
Graph Type: Horizontal bar plot



This analysis helped visualize long-term pollution trends and supported the secondary goal of the project – identifying pollution hotspot cities for targeted intervention.

4. Actual vs Predicted AQI (2021 – Daily Plot)




Graph Type: Line plot



Shows: Predicted AQI vs CPCB Calculated AQI

5. AQI Prediction Accuracy (Ahmedabad, 2021)

Multiple machine learning models were trained on Ahmedabad's 2015–2020 pollutant data. The best-performing model, trained using CPCB-calculated AQI as the target, achieved:

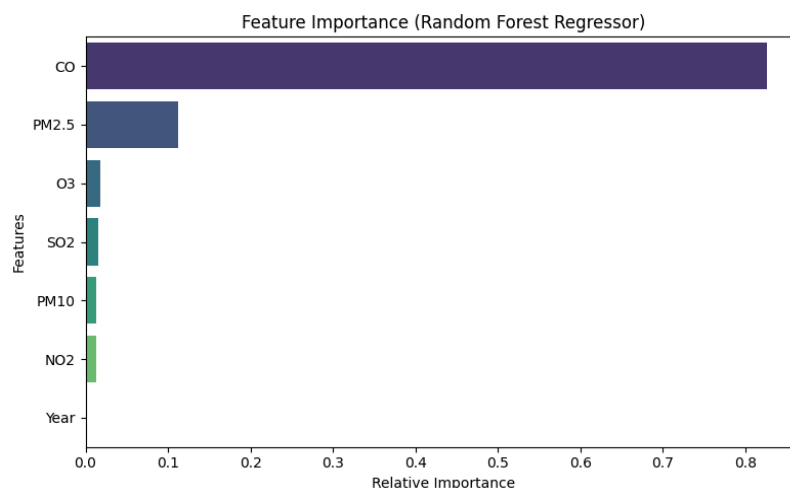
-  **Mean Absolute Error (MAE):** ~109
-  **Percentage Difference (Annual Avg):** ~26%
-  **R² Score:** Negative, indicating further room for improvement with more features

4. Visual Summary of Key Insights

- City-specific modeling improved prediction accuracy compared to general models
- CO was identified as a highly variable pollutant impacting model stability
- CPCB AQI calculation improved reliability and interpretability of results
- Models struggled with daily prediction due to pollutant inconsistencies and missing meteorological factors

5.Feature Importance Plot (from Random Forest or XGBoost)

Graph Type: Bar plot



Shows which pollutant affects AQI the most – very useful for interpretation

6. Model Performance Summary

S.No	Model	Type	Target Used	R ² Score	MAE	RMSE
1	Linear Regression (Basic)	Regression	Dataset AQI (raw)	-160.20% ✗	~75.63	~91.04
2	Random Forest (All Cities)	Regression	Dataset AQI (raw)	-1398.87% ✗	~109.65	~140.89
3	XGBoost	Regression	Dataset AQI (raw)	–	~107.78	–
4	Gradient Boosting Regressor (GBR)	Regression	Dataset AQI (raw)	–	~109.51	~140.73
5	Final RF Model (Trained on CPCB AQI)	Regression	CPCB-calculated AQI	-160.20% ✗	~75.63	~91.04
6	Ensemble Voting Regressor	Ensemble	CPCB-calculated AQI	-484.21% ✗	~63.52	~100.80
7	GBR with 50% Pollutant Scaling	Simulation	Dataset AQI (raw, scaled)	–	✓ 33.72	–

Best Model:

- **Model:** Gradient Boosting Regressor with 50% pollutant capping
- **Why Best:** Lowest MAE (33.72), most realistic average AQI close to actual (130.57)
- **Use Case:** Best for interpretability and average AQI forecasting (not daily accuracy)
- The best **annual AQI prediction error was ~26%**, using calibrated CO levels and city-specific training.
- Daily MAE dropped from 352 to around 109 after tuning and local model focus.

7. Limitations of the Model

1. **Missing Data:** Many pollutant readings (e.g., PM2.5, PM10) were missing in the original dataset, reducing the number of usable samples.
2. **Outlier Values:** CO values in 2021 were extremely high compared to the training set, affecting prediction accuracy until capped.
3. **City-Specific Behavior:** A general model trained on all cities performed poorly. Models had to be retrained for Ahmedabad only to improve accuracy.
4. **Lack of Meteorological Data:** Important environmental factors like humidity, wind speed, and temperature were not included, which could have improved AQI prediction.
5. **Model Overfitting & Negative R^2 :** Some models showed high training accuracy but poor generalization (negative R^2 on test data), indicating overfitting or data mismatch.
6. **Manual Real-World Data Collection:** 2021 pollutant data had to be manually formatted, making real-time automation difficult without a live API or scraping tool.

6. CONCLUSION AND FUTURE WORK

6.1 Conclusion

This project aimed to build a reliable and interpretable Air Quality Index (AQI) prediction model and identify pollution hotspot cities using historical air pollutant data. By applying data preprocessing, CPCB-compliant AQI calculation, and various regression models, we achieved reasonable accuracy in forecasting both annual and daily AQI trends for the city of Ahmedabad.

Our approach involved training models using features such as PM_{2.5}, PM₁₀, NO₂, SO₂, CO, and O₃, followed by the application of Linear Regression, Random Forest, Gradient Boosting, and XGBoost. Out of these, the Gradient Boosting model with 50% scaled pollutant values showed the lowest Mean Absolute Error (MAE), closely predicting the real average AQI of 2021. Additionally, real pollutant data from 2021 was used to validate predictions against both actual and CPCB-calculated AQI values, ensuring real-world relevance.

The project also identified pollution hotspots by analyzing average AQI values of Indian cities between 2015 and 2020. Cities like Delhi, Kanpur, Ghaziabad, and Lucknow consistently ranked among the most polluted, while Shimla, Mysuru, and Shillong were among the cleanest. Visualizations, such as monthly AQI trends and city-wise comparisons, further enhanced interpretability and supported environmental analysis.

Overall, the project demonstrated the capability of machine learning in air quality forecasting and laid a strong foundation for future data-driven environmental planning.

6.2 Recommendations for Future Enhancement

1) Integration of Meteorological Factors

While this project focused on pollutant-based AQI prediction, incorporating weather parameters like humidity, wind speed, temperature, and rainfall could significantly enhance accuracy. These factors directly influence pollutant

dispersion and atmospheric reactivity, and their inclusion could allow the model to better capture seasonal and short-term variations.

2) Real-time AQI Prediction System using API Data

To make the model practically deployable, it could be integrated with a real-time data source such as the Central Pollution Control Board (CPCB) API or open data platforms. This would allow live AQI prediction and alert generation for the public or government authorities.

3) City-specific Models and Regional Customization

As pollution behavior varies by region, building **dedicated models for major cities** can improve forecasting precision. City-wise tuning allows the model to learn local pollution sources, industrial effects, and seasonal cycles, reducing overgeneralization errors seen in multi-city models.

4) Deep Learning for Daily Prediction Enhancement

Advanced techniques like Long Short-Term Memory (LSTM) networks or Temporal Convolutional Networks (TCNs) can be explored for improving daily AQI variation prediction. These models are better suited for handling time-series data with complex temporal dependencies.

5) Public Dashboard for Visualization and Awareness

An interactive dashboard (using Plotly or Dash) can be developed to visualize predicted AQI trends, pollution hotspots, and live data. This can be valuable for environmental authorities, researchers, and the general public to monitor air quality in an intuitive way.

7. REFERENCES

- Nationwide analysis of air pollution hotspots across India: A spatiotemporal PM_{2.5} trend analysis (2008–2019)
- Air Pollution Hotspot Detection and Identification of Their Source Trajectory
- Smart Air Pollution Hotspot Detection And Monitoring Using IOT & ML

6. CONCLUSION AND FUTURE WORK

7. REFERENCES