ISSN No. 0976-5697

VOLUME 17 SPECIAL ISSUE 1, JULY 2026
International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

AI-Based Air Quality Prediction and Pollution Source Analysis System for Delhi NCR

Vinay Lodha
Department of Computer Science & Engineering
Geetanjali Institute of Technical Studies (GITS)
Udaipur, India
E-Mail: vinaylodhar21@gmail.com

Gourav Menaria
Department of Computer Science & Engineering
Geetanjali Institute of Technical Studies (GITS)
Udaipur, India
E-Mail: gouravmenaria667@gmail.com

Ronak Choudhary
Department of Computer Science & Engineering
Geetanjali Institute of Technical Studies (GITS)
Udaipur, India
E-Mail: ronakchoudhary504@gmail.com

Dr. Vijendra Kumar Maurya
Department of Computer Science & Engineering
Geetanjali Institute of Technical Studies (GITS)
Udaipur, India
E-Mail: ritesh.jain@gits.ac.in

Abstract: Delhi NCR is severely affected by pollution from Vehicles, industries, construction dust, and seasonal crop burning. Although data is available, it is dispersed across various systems and does not provide a clear picture of the primary cause of pollution or its potential effects. Decisions are often made only after the situation gets worse.

This paper describes an AI-based system for identifying pollution sources, predicting AQI levels, and supporting decision-making in Delhi NCR. The system gathers data from sensors, satellite sources, weather conditions and traffic pattern in a single platform. It utilizes Machine Learning models to identify pollution hotspots, major sources, and predict future AQI levels. The results are presented using a real-time dashboard that helps authorities to monitor the condition and take action accordingly. The system is aligned with initiatives by the Ministry of Environment and related pollution control bodies in India.

This platform is built using Python for Machine Learning, React.js for frontend of dashboard and cloud services for data management and storage. It supports role-based access, where users can view and use the system according to their responsibilities where administrators monitor data, analysts study trends, and policymakers make decisions. The system supports environment sustainability goals and help in proper pollution management.

Keywords*:* artificial Intelligence, Air Pollution, AQI Prediction, Delhi NCR, Machine Learning, Policy Dashboard, Environmental Monitoring

I. INTRODUCTION

Air pollution in Delhi NCR has become a major concern which will affect the daily lives and health of millions of people. Every year during winters there is a significant drop in air quality which make it hard to breath and cause major health issues This issue is caused due to combination of factors such as Vehicles, industries, construction dust, and seasonal crop burning. It is challenging to understand and manage this problem, as pollution sources keep on changing and interacting over time, making it difficult to control the overall situation.

There are many systems that provide information about air quality, however, the information contained within each of these systems tend to be different and frequently do not tie together. The first system, Air quality monitoring stations, provide information on the quality of the air around us; the second system, Satellite imagery, provides a large-scale view of air quality; the third system, Weather systems, provide a weather forecast which will ultimately cause pollutant(s) to migrate from one location to another. There is typically very little attempt to combine these 3 sources of information, therefore local government officials do not know the exact source of where the pollution is coming from nor how it will change in the near future, which can create challenges for taking action in a timely and effective manner.

Utilizing artificial intelligence and data-driven techniques to solve issues related to the environment has been an increasing trend in recent times. The capabilities of machine-learning techniques allow for the evaluation and interpretation of large amounts of data to facilitate the identification of hidden patterns and therefore, to generate accurate future predictions.

Unfortunately, most of the currently available methods used to make environmental assessments focus only on prediction and do not yet include source identification, prediction and decision support within one integrated process. The purpose of this paper is to present an AI-Derived Pollution Source Identification, Prediction and Policy Administration Dashboard System that has been specifically developed for use within the National Capital Region of Delhi. This system combines multiple data sources, applies machine-learning models to evaluate pollution trends, and provides a 24-hour complete dashboard for use as a tool for monitoring, reporting and making decisions. The intended use of this Dashboard System is to provide public authorities with increased knowledge and experience in order to take action at the earliest possible time so that specific actions will be taken prior to the occurrence of more severe environmental conditions.

II. LITERATURE REVIEW

A. Air Pollution Challenges in Delhi NCR

Delhi NCR is one of the most polluted cities in the world. The AQI often reaches dangerous and unhealthy levels. Some of the major sources of pollution in Delhi NCR include transport emissions, industrial activity, construction dust, and the burning of crop residues by farmers in neighbouring states. However, not only are there high levels of pollution in the area, but it is also very difficult to determine the exact contribution from each of the different sources. Weather and seasonal changes also introduce complexity into understanding air quality, resulting in dynamic and unpredictable patterns of pollution throughout the region. Most of the air pollution studies conducted to date have focused on measuring the concentration of pollutants, with little attention paid to identifying the various sources of pollution. Therefore, the lack of a source-based understanding on air pollution makes it difficult for policymakers to make effective interventions and reach timely decisions.

B. Air Quality Predictions Using Machine Learning

Machine learning has become increasingly popular in the area of predicting air quality. In the past, time-series forecasting using a traditional model such as ARIMA has been implemented. Unfortunately, these types of time-series forecasting models do not account for the more complex, non-linear patterns typically associated with environmental data. However, more sophisticated techniques that employ machine learning (e.g., Random Forests, Support Vector Machines (SVMs) and Neural Networks (NNs)) have demonstrated significantly greater accuracy in predicting AQI levels.

C. Pollution Source Identification Techniques

Identifying sources of pollution is an essential factor when managing air quality effectively. Clustering methods such as K-means have been employed to identify areas with the highest concentrations of pollutants (hot spots), by grouping together areas with similar levels of contamination. Feature importance methods within models like Random Forest can help understand what drives the most significant amounts of pollution in a specific area. In addition, satellite monitoring has been performed for many years to study large-scale patterns of pollution (e.g. crop burning) and to identify sources of pollution - such as industrial facilities or urban areas - that are contributing to the overall levels of pollutants. However, integrating satellite and ground-based sensor data has been difficult because of the difference in data resolution and data frequency.

D. Smart Dashboards for Environmental Monitoring

Furthermore, with the emergence of smart city projects, there has been an increased focus on creating dashboards for visualizing environmental data. Dashboards provide real-time information (i.e. what is happening now), graphics, and alerts to offer users better information about pollution trends. However, the majority of dashboards existing today are descriptive; that is, they show pollutants at the time of display but do not forecast future pollutant levels nor how changes in policy might affect future pollutant levels. There will be an increasing demand for intelligent dashboards that not only combine data visualization and presentation techniques but also enable predictive analytics and enable users to make informed decision.

Aadhaar, India's biometric-based national identity system, has been explored as a mechanism for patient deduplication in healthcare settings. Its high penetration among India's population, including migrant workers, makes it a practical unique identifier for preventing duplicate registrations across facilities [5]. Using Aadhaar as a primary key ensures that a single longitudinal record exists per patient regardless of which hospital first registers them.

E. *Research gap*

While air quality monitoring and prediction systems have made great advances towards creating a predictable model and evaluating those predictions, there remains a large gap in integrated solutions. The majority of the existing technologies available today provide either predictive capabilities or visualization of air quality data with little or no emphasis on both functions or pollution source identification.

There is also a significant lack of solutions that utilize machine learning models that incorporate multi-source data and the opportunity for decision-makers to view this information in real-time, to support their policy development decisions. Thus, there is an opportunity to develop an integrated solution that will enable policy makers to identify pollution sources, predict air quality index levels, and facilitate actionable outcomes through a user-friendly interface.

III. PROBLEM STATEMENT

A. *Challenges in Air Pollution Management*

Numerous contributors produce air pollutant levels in the Delhi NCR region. These include; vehicle emissions, industrial emissions, construction dust emissions, and crop burning emissions. However, identifying the specific source(s) responsible for creating levels of pollutants at any one point in time is a large obstacle. Additionally, weather can influence the level of pollutants, therefore complicating the management of these pollutants. Due to the above circumstances, it is difficult to effectively manage the air quality within the Delhi NCR region. Furthermore, there continues to be a need to accurately document all of the contributors to air quality issues in the region.

B. *Gaps in Existing Systems*

Most current systems only show AQI levels and basic pollution data. They do not clearly explain the source of pollution or predict future trends accurately. Data is available but scattered across different platforms and not connected. This makes it difficult for authorities to take timely action. As a result, decisions are often delayed or ineffective.

C. *Need for the Proposed System*

There is a need for a system that combines data from multiple sources into one platform. It should be able to identify pollution hotspots and predict AQI levels in advance. The system should also provide a simple dashboard for monitoring and decision-making. This will help authorities take timely and better actions. An integrated AI-based approach can improve pollution management significantly.

IV. SYSTEM ARCHITECTURE

The proposed system follows a multi-layered architecture designed to efficiently monitor, analyze, and predict air pollution in the Delhi-NCR region using freely available and scalable technologies.

1. Data Collection Layer
The system collects data from multiple publicly available and low-cost sources, including satellite datasets, government air quality monitoring stations, weather APIs, and optional low-cost IoT sensors. This ensures comprehensive coverage of environmental conditions without significant financial investment.

2. Data Ingestion and Processing Layer
The collected data is passed through a lightweight data processing pipeline where it is cleaned, filtered, and transformed into a structured format. Open-source tools and scripts are used to handle missing values and remove inconsistencies, ensuring data reliability.

3. Data Storage Layer
Processed data is stored using open-source databases such as PostgreSQL or MongoDB. These databases provide scalable storage solutions suitable for handling both historical and real-time data efficiently without additional licensing costs.

4. AI/ML Processing Layer
The system utilizes open-source machine learning frameworks to analyze pollution data. Time-series models are applied to identify patterns, detect anomalies, and understand pollution trends. This layer performs key tasks such as pollution source identification and AQI prediction.

5. Source Contribution Analysis Module
The system estimates the contribution of various pollution sources, including vehicular emissions, industrial activities, construction dust, and crop burning. This analysis helps in understanding the root causes of pollution.

6. Prediction and Forecasting Layer
Using trained models and weather parameters, the system predicts future AQI levels for the next 24 to 72 hours. It also analyzes how pollutants may spread across regions based on atmospheric conditions.

7. Backend and API Layer
A lightweight backend system is used to manage communication between different components. RESTful APIs are implemented using efficient and open-source frameworks to ensure smooth data flow between the server, machine learning models, and user interfaces.

8. Policy Decision and Simulation Layer
This layer enables decision-makers to analyze pollution data and simulate different scenarios. It helps evaluate the impact of potential actions, such as traffic restrictions or industrial controls, before implementation.

9. User Interface Layer
The system provides intuitive and user-friendly interfaces:

- A citizen interface for real-time AQI updates, health alerts, and safety suggestions
- A policy dashboard for monitoring trends, identifying pollution sources, and analyzing data visually

10. Alert and Notification System
The system generates alerts when pollution levels exceed safe limits. Notifications are delivered through cost-effective messaging services or application-based alerts, ensuring timely communication with users.

11. Continuous Feedback Mechanism
User feedback and new incoming data are continuously integrated into the system. This feedback loop improves model accuracy and system performance over time.

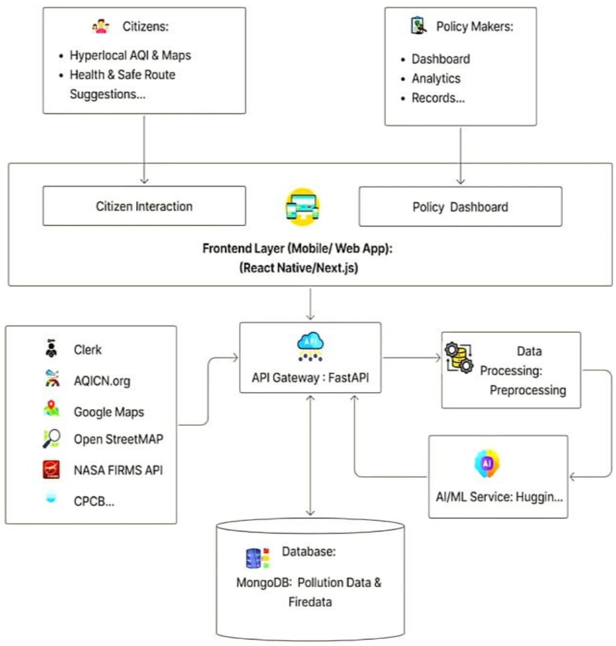


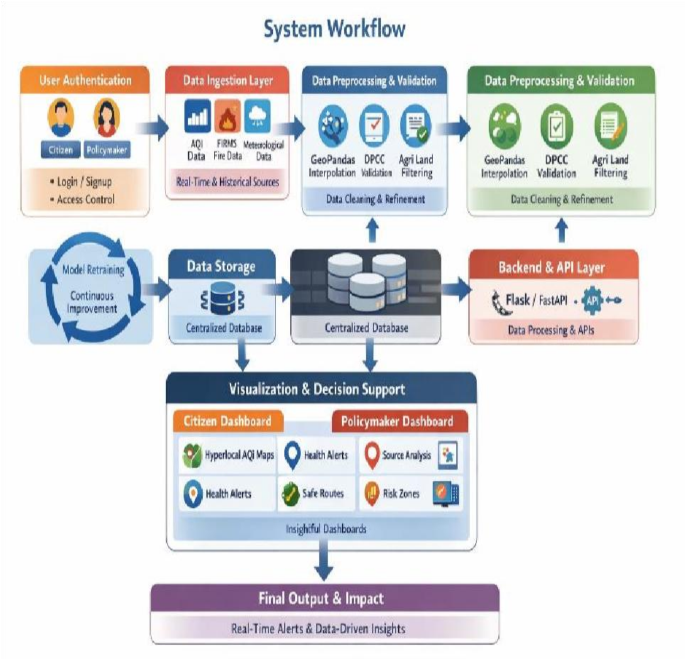
Fig. 3.1 — System Architecture Diagram

V. SYSTEM WORKFLOW

The workflow for the system gathers, processes, and converts environmental data into actionable insights for both people and government decision-makers. The workflow begins with user authentication, which consists of logging into or signing up on the system. The authentication of users is validated, which then gives valid users access to their individual dashboards and leads invalid users to a re-enter screen. This authentication process assures that all users will have secure access to the system and that they will also have restricted access to only those functions related to their role.

Environmental data is collected from numerous sources (i.e., AQI APIs, FIRMS fire data, and various weather data). The environmental data is then pre-processed (i.e., cleaned, structured, and validated) before creating hyperlocal AQIs (1 km) via interpolation of AQIs by using GeoPandas and validating those AQIs with DPCC to ensure accuracy. The identification of pollution sources (i.e., stubble burning) occurs during the processing phase of the data through agricultural land filtering.

After data has been pre-processed, they are processed by using AI to perform AQI prediction, AQI trend analysis, AQI source identification, and generation of heat maps for displayed AQIs. The results of the above processes are stored in a central database and provided to users via an API layer on the back end of their dashboards. Based on the results, citizens will receive AQI maps and advice on their health, as well as providing those responsible for making policy decisions, such as providing effective policy recommendations. The system provides a way to make decisions regarding the quality of the environment based on data, leading to increased public awareness about the environment.



VI. SYSTEM MODULES

A. Dashboard

The system's primary dashboard, which offers a real-time summary of the air quality in the Delhi-NCR area, is represented by this screen.

The dashboard uses a speedometer-style indicator at the top to display the current AQI value (208). This makes it easier for users to comprehend the pollution level, which is classified as Poor. A brief health message is also shown to let people know about potential consequences and safety measures. The system ensures that the data is correct and current by gathering real-time data from several monitoring stations.

The region's overall pollution level is shown by the average AQI.

Levels of PM2.5: displays the amount of PM2.5, or fine particles, which are bad for human health.

Maximum AQI: shows the highest recorded level of pollution.

The lowest AQI : displays the region's least polluted reading.

Fig. 4.1 — Forecast Module (Air Quality Prediction)

AQI Trend Chart: The changes in air quality over the past seven days are depicted in this graph. It aids in determining whether pollution is rising or falling.

Chart of Station Status: shows the various AQI categories (Good, Moderate, Poor, etc.) that different monitoring stations fall under.

B. *Forecast*

The forecast module is designed to predict air quality in advance, specifically for the Delhi-NCR area. It utilizes AI models to estimate the Air Quality Index (AQI) for the upcoming days, which I think is quite helpful for anticipating what to expect. The main display features a 7-day forecast, providing the predicted AQI for each day along with its corresponding category, such as "moderate" or "poor." This format makes it easy to identify whether air quality is improving or deteriorating over the week.

Additionally, there is a weekly trend graph that tracks changes in air quality, displaying the minimum, average, and maximum AQI values. This visual representation helps illustrate how pollution levels may fluctuate, although it does not delve into extensive detail right away.

Fig. 4.2 — Forecast Module (Air Quality Prediction)

C. *Source Analysis*

The source of the pollution is explained on this screen.The system examines various sources of pollution in this module, including: Automobiles, Construction, and Domestic Activities Burning biomass. It is simple to determine which source contributes the most when a pie chart displays each source's percentage contribution. Every source is compared side by side in a bar graph.

Fig. 4.3 — Source Analysis Module (Pollution Source Identification)

Additionally, there is a thorough section that provides precise percentage values (vehicles, for instance, contribute about 38%). This module assists in determining the primary causes of pollution, which is crucial for making the appropriate decisions.

E. *Policy Hub*

This screen displays the project's policy recommendation system.
The system recommends pollution-reduction measures based on pollution data and forecasts. It displays: The quantity of policies in effect , Anticipated decrease in pollution ,Suggested actions

Among the recommended policies are: Odd-Even Vehicle Rule , Prohibition of construction activities ,Control of industrial emissions.
Every policy is labelled with its anticipated impact and priority level (high/medium). Additionally, there are charts that display the current state and efficacy of the policy. This module assists authorities and the government in making timely and informed decisions to reduce pollution.

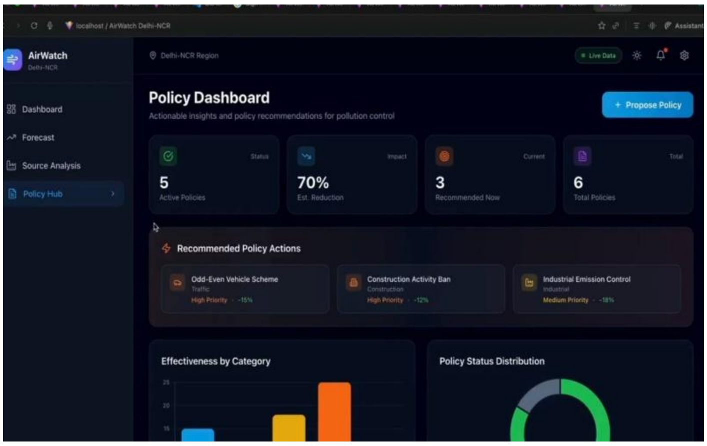


Fig. 4.4 — Policy Hub Module (AI-Based Recommendations)

VII. Implementation And Technology Stack

The implementation of the proposed AI-driven pollution monitoring system is based on a combination of modern, scalable, and cost-effective technologies. The selected tools and frameworks ensure efficient data processing, real-time analysis, and user-friendly visualization.

Frontend Development :
The user interface of the system is developed using **HTML, CSS, and React**. HTML and CSS are used to structure and style the application, while React enables the creation of dynamic and interactive components. This combination allows the system to deliver a responsive and intuitive user experience for both citizens and policymakers.

Backend Development :
The backend is implemented using **FastAPI**, a lightweight and high-performance web framework. FastAPI is used to build RESTful APIs that handle data communication between the frontend, database, and machine learning models. It ensures fast response times and efficient handling of real-time data requests.

AI and Machine Learning :
The core analytical functionality of the system is developed using **Python**, which is widely used for data analysis and machine learning tasks. Various libraries are utilized for data preprocessing, model training, and prediction. The system applies machine learning techniques to identify pollution sources, detect patterns, and forecast AQI levels.

External Data Services:
To ensure accurate and real-time data availability, the system integrates multiple external APIs and data services:

- **NASA FIRMS** is used to detect fire incidents such as crop burning, which significantly contribute to air pollution.
- **OpenAQ** provides real-time air quality data from multiple monitoring stations.
- **OpenWeatherMap** supplies meteorological data such as temperature, humidity, and wind speed, which influence pollution dispersion.
- **Google Maps API** is used to provide location-based services and route suggestions to users.

Cloud Infrastructure :
The system is deployed using the **AWS Free Tier**, which provides a cost-effective cloud environment for hosting the backend services and handling real-time data processing. This ensures scalability and availability without incurring high operational costs.

Database Management:
For data storage, the system uses **MongoDB**, a NoSQL database that efficiently handles large volumes of unstructured and semi-structured data. It is particularly suitable for storing real-time pollution data and supporting flexible queries.

VIII. Results

To evaluate the AI-driven pollution measurement system, both real-time data collected from air quality monitoring stations and historical data downloaded from 3rd party satellite datasets and weather API's have been used to demonstrate how various sources of data can be integrated to give users valuable insight into pollution trends for the Delhi-NCR area.

The machine learning models used to predict AQI levels were also able to provide accurate short-term forecasts of 24 to 72 hours in advance so long as the meteorological data parameters were included within the prediction model. The inclusion of meteorological data improved the AQI prediction model's ability to identify fast-pacing variations in pollution as a result of meteorological changes (e.g., wind speed or temperature).

The source identification module of the proposal was able to illustrate how much each of the major sources of pollution (e.g., vehicle emissions, crop burning, industrial emissions, construction dust) contributed to the overall pollution of the air over various periods of time.

The dashboard interface allowed users to view in real-time or near real-time their area AQI levels, pollution trends and breakdown of what sources contributed to the current level of pollution. Users had the ability to view location-based data, receive alerts when pollution levels exceeded acceptable limits and get some limited indication of expected outcome of various types of government-sponsored policy interventions through the use of simulation tools.

IX. Conclusion And Future Scope

Conclusion : To conclude, the issue of air quality in Delhi-NCR is both complicated and evolving, with multiple factors contributing to its effects. This research established an AI-based system to ingest data coming from various locations; utilize machine learning methods; and form a single point of reference for monitoring, forecasting, and assisting in making appropriate policy decisions for the air quality of the area.

This proposed solution addresses several deficiencies in prior solutions; forecasts air quality, determines where pollution is coming from and provides assistance to support policy-making decisions. Using cost-effective, expandable technology (versus current solutions) allows this proposed solution to have practical applications in "real-life".

The findings indicate that combining AI-models with environmental data can greatly improve the ability of analysts to better understand trends related to air pollution and to respond proactively. Additionally, the dashboard-style interface affords greater ease of access and usability by a variety of users.

Future Scope: The groundwork of this system is solid, but many enhancements are possible going forward: Higher Quality Predictions: Continued improvements could be made by utilizing cutting-edge deep learning models and larger sets of training data to improve prediction quality. Support for Real-Time data from the IoT: Additional sensors can be deployed at lower costs; therefore, much more specific, localized real-time pollution monitoring can take place. Development of a Mobile App: Development of an app dedicated to users' interaction with this system will improve access to the system and increase user engagement. More Advanced Simulation Models Evaluating Policy: Improved policy evaluation through advanced simulation models for long-term effects of different policies can be developed. Integration with Government: The integration of this system with the various government systems could provide quicker decision-making for implementing actions. Expansion to Other Cities: This system can be expanded to monitor pollution in other urban areas beyond Delhi/NCR

X. References

SAFAR: reference for pollution app ui.

Daily scale air quality index forecasting using bidirectional recurrent neural networks: Uses BILSTM emphasizing pollutants like PM2.5/NO2.

Satellite Observations for Air Quality (NASA Terra/ MODIS): Explains MODIS for identifying pollution sources like fires and aerosols

MODIS Data Products for Aerosol and Pollution Monitoring:(NASA MODIS Web): (NASA MODIS Web): Official guide to MODIS AOD.

Air pollution prediction with machine learning case study of Indian cities:ML for hyperlocal routing.

AirNow.gov Interactive Map and Dashboard.

Real-time policy tool for urban AQI management and intervention monitoring.