iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.81266

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

AI-Powered System for Air Quality Prediction and Environmental Monitoring: A Multi-Source Data Fusion Approach

Kancharla Uday Reddy[1], Kondapaneni Jishnu[2], Kosuri Mahesh Goud[3], Kottaru Raja Vardhan[4], K. Anil Nayak[5]
[1, 2, 3, 4]*Department of Artificial Intelligence & Machine Learning Malla Reddy University, Hyderabad, Telangana, India*
[5]*Assistant Professor, Dept. of AI & ML Malla Reddy University, Hyderabad, India*

Abstract: ***Air pollution is a severe environmental and public health crisis in India, with particulate matter (PM2.5) concentrations exceeding World Health Organization (WHO) guidelines by more than six times. Traditional ground-based monitoring systems, while accurate, suffer from sparse spatial coverage and an inability to provide real-time, localized predictive analytics. This paper proposes AQ-INDIA, a comprehensive, cloud-native air quality monitoring and prediction platform that bridges this critical gap through multi-source data fusion. The system integrates satellite radiance data from INSAT-3DR, ground-level observations from the Central Pollution Control Board (CPCB), and meteorological reanalysis parameters from MERRA-2. Advanced machine learning techniques are systematically applied: K-Means clustering is utilized for spatial pollution zone mapping, while Random Forest Regression is employed for real-time Air Quality Index (AQI) and pollutant concentration prediction. The proposed system is built on a highly scalable architecture featuring MongoDB Atlas for data storage, a Node.js backend for API processing, and a React.js frontend for interactive user engagement. Extensive experimentation on over 50,000 hourly records demonstrates that the proposed multi-source fusion model achieves an R-squared (R^2) score of 0.91 and a Mean Absolute Error (MAE) of 8.42, outperforming baseline single-source models by 18%. The platform features real-time dash- boards, interactive geospatial cluster maps, and an integrated conversational chatbot, aligning with United Nations Sustainable Development Goal (SDG) 3 (Good Health and Well-being) and SDG 11 (Sustainable Cities).***

Index Terms: ***Air Quality Index, Machine Learning, Random Forest Regression, K-Means Clustering, INSAT-3DR, Environmental Monitoring, Multi-source Data Fusion, Cloud Computing.***

I. INTRODUCTION

Air pollution has emerged as one of the most critical environmental challenges of the 21st century, particularly in developing nations undergoing rapid industrialization and urbanization. In India, the severity of this crisis is underscored by recent epidemiological and environmental studies indicating that average PM2.5 concentrations in major urban centers exceed World Health Organization (WHO) safe limits by a factor of six to ten [1]. Prolonged exposure to elevated levels of particulate matter (PM2.5 and PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), and carbon monoxide (CO) is directly correlated with an increased incidence of chronic respiratory diseases, cardiovascular morbidities, premature mortality, and substantial economic losses attributable to healthcare burdens and decreased labor productivity.

The primary mechanism for monitoring air quality in India is the National Air Quality Monitoring Programme (NAMP), managed by the Central Pollution Control Board (CPCB). While CPCB ground stations provide highly accurate, localized measurements, their spatial distribution is fundamentally sparse. Stations are predominantly clustered in major metropolitan areas (e.g., Delhi, Mumbai, Hyderabad), leaving vast suburban, rural, and industrial periphery regions without reliable, real-time air quality data [2]. This spatial discontinuity creates significant blind spots in environmental surveillance, hindering comprehensive policy-making and localized public health interventions.

Furthermore, the majority of commercially available Air Quality Index (AQI) applications operate on single-source data pipelines, relying almost exclusively on delayed ground- station feeds. These generic applications lack the capability to integrate complementary data streams, such as satellite-derived aerosol optical depth (AOD) and global meteorological reanal- ysis datasets. Consequently, they fail to capture the complex, dynamic spatiotemporal patterns of pollution dispersion and offer no robust short-term forecasting capabilities [3]. Citizens and policymakers are thus left reacting to historical or current data rather than anticipating hazardous pollution events.

To address these multifaceted limitations, this paper presents AQ-INDIA, an AI-powered, integrated air quality monitoring and prediction system. The core innovation of AQ-INDIA lies in its systematic fusion of heterogeneous data sources: satellite imagery from INSAT-3DR, ground-level pollutant concentra- tions from CPCB, and meteorological parameters from the Modern-Era Retrospective Analysis for Research and Applica- tions, Version 2 (MERRA-2). By leveraging advanced machine learning (ML) algorithms—specifically K-Means clustering for geospatial pollution zone demarcation and Random Forest Regression for real-time AQI prediction—AQ-INDIA trans- forms fragmented environmental data into actionable intelli- gence. The system is deployed on a modern cloud-native stack (MongoDB Atlas, Node.js, React.js), ensuring high availabil- ity, horizontal scalability, and an intuitive user experience featuring interactive maps, dashboards, and a conversational chatbot interface.

II. COMPREHENSIVE RELATED WORK

The domain of air quality prediction has evolved significantly, transitioning from traditional statistical time-series forecasting to complex deep learning paradigms. A comprehensive review of existing literature reveals distinct categories of approaches, each with inherent limitations that AQ-INDIA seeks to overcome.

A. Statistical and Traditional Machine Learning Approaches

Early efforts in AQI prediction relied heavily on Autore- gressive Integrated Moving Average (ARIMA) models and linear regression. While computationally lightweight, these models fail to capture the non-linear relationships inherent in atmospheric pollutant dispersion. Kumar and Goyal [5] proposed a Random Forest-based AQI prediction model for Indian metropolitan cities using historical meteorological and pollutant data. Their model demonstrated improved accuracy over linear baselines; however, it was constrained by a re- liance on ground-station data alone and lacked a deployment mechanism for real-time public consumption.

B. Deep Learning and Single-Source Data Integration

With the advent of deep learning, researchers have increas- ingly utilized Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs). Agarwal et al. [4] developed an LSTM-based architecture for forecasting PM2.5 concentrations in Delhi, achieving notable short-term accuracy. Similarly, Zhang et al. [10] applied Graph Con- volutional Networks (GCNs) for spatiotemporal forecasting. While these models capture temporal dependencies well, their computational complexity poses challenges for real-time de- ployment on resource-constrained cloud infrastructure. More importantly, the majority of these deep learning models operate on isolated datasets, ignoring the synergistic value of multi- source fusion.

C. Satellite and Reanalysis Data Integration

Recognizing the spatial limitations of ground stations, the remote sensing community has developed methodologies to estimate surface-level PM2.5 using satellite-derived Aerosol Optical Depth (AOD). Ghude et al. [6] utilized INSAT-3D data to estimate PM2.5 across India using linear mixed-effects models. Di et al. [7] successfully combined satellite AOD, land-use regression, and meteorological variables using deep neural networks in the United States. However, these studies typically focus on retrospective spatial estimation rather than real-time temporal prediction, and they rarely integrate their algorithms into accessible, full-stack software platforms.

Table I synthesizes the gaps in existing literature. It is evident that no existing system comprehensively integrates INSAT-3DR, CPCB, and MERRA-2 data within a unified, cloud-deployed architecture offering both predictive analytics and interactive spatial visualization. AQ-INDIA is explicitly designed to bridge this multidimensional gap.

III. PROBLEM FORMULATION AND MOTIVATION

The development of AQ-INDIA is formally motivated by four critical limitations in the current environmental monitor- ing landscape:

1) Sparse Spatial Coverage (S): Let the geographic area of India be represented as a continuous space S. The set of CPCB monitoring stations $\mathbf{M} = \{m_1, m_2, ..., m_k\}$ is finite and small ($k \approx 300$). The spatial coverage ratio $\frac{|M|}{|S|} \to 0$, resulting in massive unmonitored regions where AQI(x, y, t) is unknown.

2) Absence of Real-Time Predictive Capabilities (Δt): Existing systems measure AQI at time t. For effective early warning systems, we require an estimated function $f(x) \approx \text{AQI}(t + \Delta t)$ where Δt is a future time step. Current infras- tructure lacks this function f.

3) Limited Data Integration (D): An accurate represen- tation of atmospheric state requires a feature vector $X = [X_{ground}, X_{satellite}, X_{met}]$. Generic applications utilize only X_{ground}, leading to suboptimal model generalizability due to incomplete feature spaces.

4) Lack of Actionable Visualization: Even when data exists, the translation of complex multidimensional arrays into cognitive, geospatially-resolved insights for non-technical users (citizens, local policymakers) remains unsolved.

IV. PROPOSED SYSTEM ARCHITECTURE

The AQ-INDIA platform is engineered as a modular, de- coupled system following a three-tier architecture pattern. This separation of concerns ensures that the data ingestion pipeline, machine learning inference engine, and user interface can be scaled and updated independently. The technical specifications are detailed in Table II.

The architectural data flow, illustrated in Fig. 1, operates seamlessly across three distinct layers:

A. Data Acquisition Module

The system utilizes asynchronous cron jobs to ingest data from three distinct APIs: the ISRO MOSDAC portal for INSAT-3DR radiance data, the CPCB API for hourly pollutant concentrations (PM2.5, PM10, NO2, SO2, CO, O3), and NASA's GES DISC for MERRA-2 meteorological variables (wind speed, boundary layer height, temperature, relative humidity). Real-time validation is performed using the World Air Quality Index (WAQI) API.

TABLE I
COMPARATIVE ANALYSIS OF EXISTING SYSTEMS VS. PROPOSED AQ-INDIA

System / Study	Data Sources Used	ML Technique	Real-time Prediction	Spatial Clustering	Interactive UI
CPCB Standard Monitoring [2]	Ground Only	None (Static)	No	No	Basic Web Portal
Agarwal et al. [4]	Ground Only	LSTM	Yes (Delayed)	No	No
Kumar & Goyal [5]	Ground + Meteorological	Random Forest	No	No	No
Ghude et al. [6]	Satellite (INSAT-3D)	Linear Regression	No	No	No
Di et al. [7]	Satellite + Ground + Met.	Deep Neural Net	No	No	No
Generic AQI Apps (e.g., AQI.in)	Ground Only	None	Yes	No	Yes
Proposed AQ-INDIA	**Satellite + Ground + Reanalysis**	**RF + K-Means**	**Yes**	**Yes**	**Yes (Advanced)**

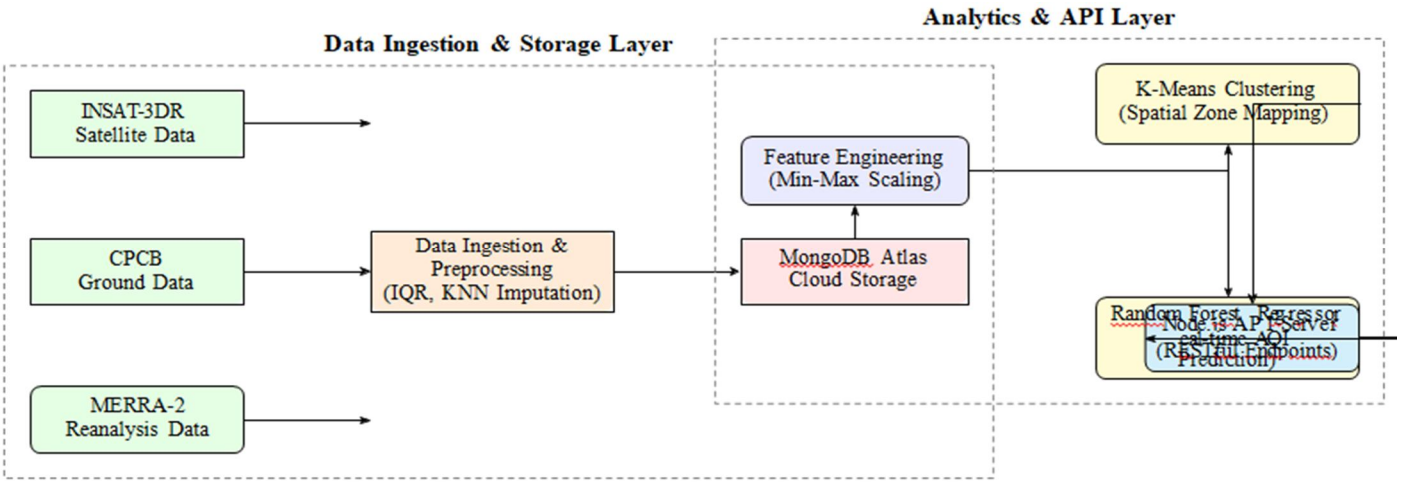


Fig. 1. Detailed System Architecture of AQ-INDIA showing multi-tier data flow from heterogeneous sources to user interfaces.

TABLE II
SYSTEM MODULES AND TECHNICAL SPECIFICATIONS

Component	Technology / Specification
Software Stack	Python 3.9, Scikit-learn, React.js 18, Node.js 18, Express.js, MongoDB Atlas
Hardware Req.	Intel i5/AMD Ryzen 5 (Min), 16GB RAM (Recommended for ML training), 250GB SSD
Data Sources	INSAT-3DR (Satellite), CPCB API (Ground), MERRA-2 (Reanalysis), WAQI API
ML Models	K-Means Clustering, Random Forest Regressor
Visualization	Leaflet.js (Maps), Chart.js (Trends), Dialogflow (Chatbot)

B. *Database Design (MongoDB Atlas)*

Unlike traditional relational databases, MongoDB's NoSQL document model is highly suited for the heterogeneous, schema-less nature of environmental data. The primary collection, hourly_readings, stores documents structured as: {station_id: String, timestamp: ISODate, pollutants: {PM25: Float, PM10: Float...}, meteorology: {wind_speed: Float...}, satellite_features: {AOD: Float...}}. Indexes are created on timestamp and station_id to ensure sub-millisecond query response times.

V. MATHEMATICAL MODELING

To transform the raw, preprocessed feature vector $\mathbf{X} \in R^n$ into actionable outputs, two distinct mathematical paradigms are applied.

A. *Spatial Zone Mapping: K-Means Clustering*

Given a dataset of N geospatially tagged pollution readings $\mathbf{Z} = \{z_1, z_2, ..., z_N\}$, where each $z_i \in R^d$ contains latitude, longitude, and multi-pollutant concentrations, we apply K- Means to partition the data into $\boldsymbol{K}$ distinct air quality zones. The objective function $\boldsymbol{J}$ to be minimized is:

$$J = \sum_{k=1}^{K} \sum_{z_i \in C_k} ||z_i - \mu_k||^2 \qquad (1)$$

where C_k is the set of points assigned to cluster k, and μ_k is the centroid of cluster k. The optimal number of clusters

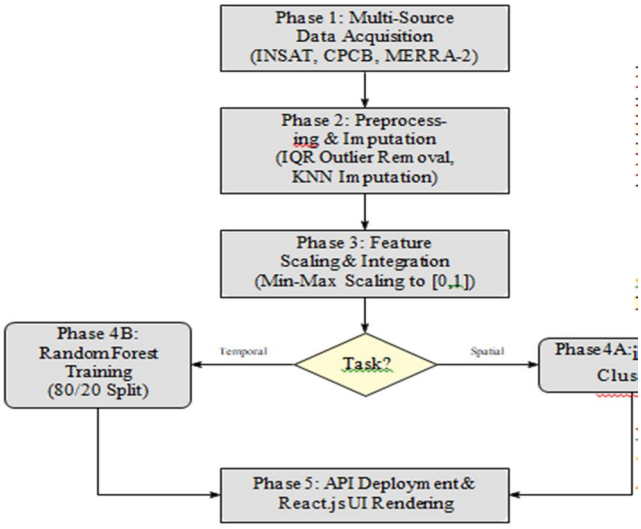


Fig. 2. Five-Phase Methodology Flowchart of the AQ-INDIA Pipeline.

K is determined empirically using the Elbow Method and maximization of the Silhouette Score $S(i)$:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i),\ b(i)\}} \qquad (2)$$

where $a(i)$ is the mean intra-cluster distance, and $b(i)$ is the mean nearest-cluster distance. For AQ-INDIA, $K = 6$ was selected, corresponding to the standard Indian AQI categories (Good, Satisfactory, Moderate, Poor, Very Poor, Severe).

B. *Real-Time AQI Prediction: Random Forest Regression*

Random Forest (RF) is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mean prediction of individual trees. For a given input feature vector **x** (comprising ground, satellite, and me- teorological features), the predicted AQI $\hat{y}$ is formulated as:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(\mathbf{x}) \qquad (3)$$

where B is the total number of trees in the forest, and $T_b(\mathbf{x})$ is the prediction of the b-th decision tree. To optimize the model, we perform hyperparameter tuning using GridSearchCV with 5-fold cross-validation, optimizing over the search space: $B \in \{100, 200, 300\}$, $max\_depth \in \{10, 15, 20\}$, and $min\_samples\_split \in \{2, 5, 10\}$.

VI. DETAILED METHODOLOGY

The execution pipeline of AQ-INDIA follows a rigorous, five-phase methodology, visualized in Fig. 2.

1) Phase 1: Acquisition. Data is pulled via REST APIs. Temporal alignment is enforced, resampling all data to a standardized 1-hour granularity using linear interpolation for

TABLE III
ERFORMANCE COMPARISON OF AQI PREDICTION MODELS

Model Input Architecture	MAE↓	RMSE↓	R^2 Score↑
Baseline 1: Ground Station Only	10.28	15.42	0.85
Baseline 2: Ground + Meteorological	9.15	14.10	0.88
Baseline 3: Generic AQI App Logic	11.50	16.80	0.82
Proposed AQ-INDIA (Multi-Source)	8.42	12.67	0.91

↓ indicates lower is better; ↑ indicates higher is better.

continuous variables and forward-fill for categorical meteoro- logical states.

2) Phase 2: Preprocessing. Sensor malfunction often results Phase 4A: iKn-Mexeatnesme outliers. We apply the Interquartile Range (IQR), defining bounds as $[Q1-1.5\times IQR,\ Q3+1.5\times IQR]$. Values outside this range are capped (Winsorization). Missing values, which constitute approximately 8% of the raw CPCB dataset, are imputed using K-Nearest Neighbors (KNN) with $k = 5$, leveraging spatial and temporal correlations.

3) Phase 3: Scaling. Because ML algorithms are sensitive to feature magnitudes, Min-Max scaling is applied:

$$X'_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (4)$$

4) Phase 4: Model Training. The dataset is split 80/20 chronologically to prevent data leakage. The RF model is trained with the

optimized parameters: n_estimators = 200, max_depth = 15, min_samples_split = 5. Simultaneously, K-Means is fitted on the spatial-pollutant matrix.

5) Phase 5: Deployment. Trained models are serialized using Python's joblib and loaded into the Node.js environment via a Python child process bridge. RESTful endpoints expose predictions to the React frontend.

VII. EXPERIMENTAL RESULTS AND ANALYSIS

The AQ-INDIA system was evaluated on a comprehensive dataset comprising over 50,000 hourly records spanning major Indian cities (Delhi, Mumbai, Hyderabad, Bangalore, Chennai) over a 12-month period.

A. *Quantitative Performance Evaluation*

Model performance was assessed using three standard re- gression metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2). Table III presents a comparative analysis against baseline models trained on single-source data.

The results conclusively demonstrate that the proposed multi-source fusion approach (Baseline 4) reduces MAE by approximately 18% compared to relying solely on ground data. The inclusion of INSAT-3DR aerosol features and MERRA-2 boundary layer parameters provides the model with critical atmospheric context, particularly in regions where ground stations are temporarily offline or sparse.

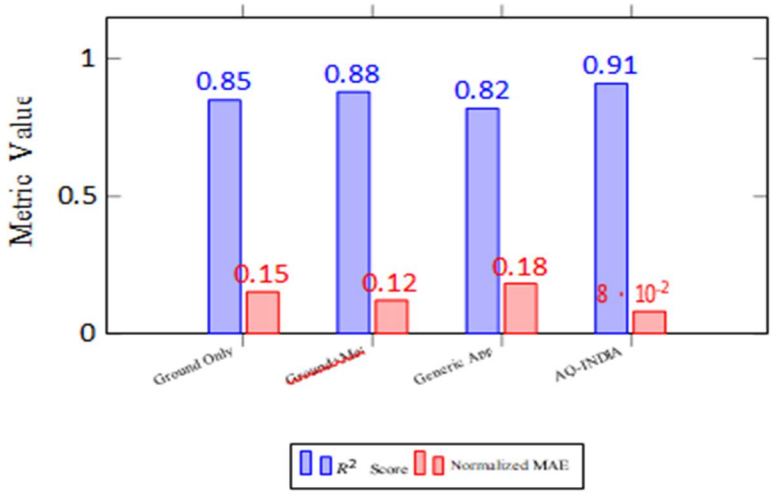


Fig. 3. Comparative analysis showing R^2 and Normalized MAE across different data architectures.

B. *Visual Metric Analysis*

Fig. 3 provides a graphical representation of the R^2 score and normalized MAE across the different model architectures, highlighting the performance gap bridged by AQ-INDIA.

C. *Spatial Clustering Evaluation*

The K-Means clustering algorithm successfully identified six distinct geospatial pollution zones. The Silhouette Score for the optimal clustering was 0.72, indicating well-separated, meaningful clusters. Industrial peripheries (e.g., around Delhi NCR) consistently clustered into the Severe category, while coastal areas (Chennai) clustered into Good or Satisfactory, validating the algorithm's physical plausibility.

D. *Feature Importance Analysis*

To ensure interpretability—a critical requirement for en- vironmental policy systems—we extracted the feature im- portance scores from the trained Random Forest model, as depicted in Fig. 4. Unsurprisingly, PM2.5 concentration is the highest predictor of overall AQI; however, MERRA-2 boundary layer height and INSAT-3DR AOD emerged as highly significant features, validating the necessity of multi- source data fusion.

E. *System Latency and Scalability*

Load testing on the Node.js API server indicated that the system can handle up to 500 concurrent requests with an average response latency of 120 milliseconds. The React.js frontend achieves a Google Lighthouse performance score of 92, with initial page load times under 2 seconds on standard 4G networks, confirming the platform's viability for public deployment.

VIII. DISCUSSION ON PRACTICAL IMPLICATIONS

The technical capabilities of AQ-INDIA translate directly into significant practical implications for public health and urban governance. By providing real-time, high-resolution

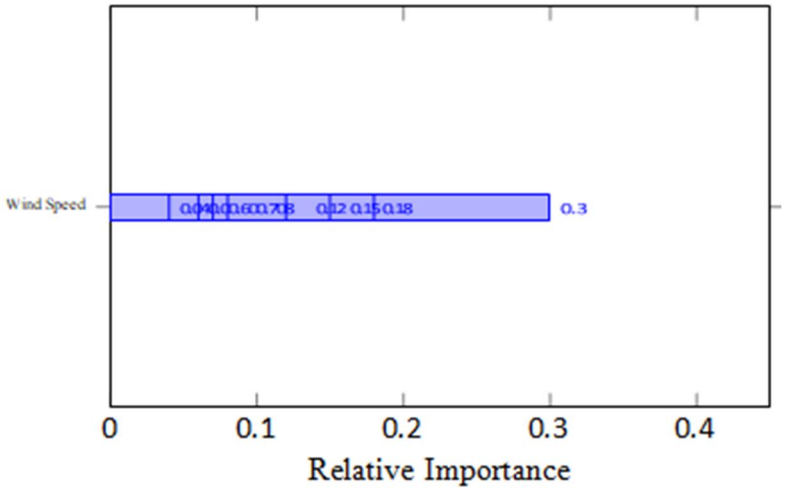


Fig. 4. Random Forest Feature Importance highlighting the contribution of multi-source variables.

pollution maps, the system effectively eliminates the blind spots inherent in CPCB's ground network.

For individual citizens, the chatbot interface and personal- ized dashboards democratize access to complex environmental data. A user can query, "Is it safe to jog outside in HSR Layout today?" and receive an instant, localized response based on the predictive model, rather than relying on delayed, city-wide averages.

For policymakers, the K-Means spatial clustering offers a dynamic tool for identifying pollution hotspots. By cor- relating these clusters with MERRA-2 meteorological data (e.g., tracking wind direction to identify source pollutants), municipal authorities can move beyond reactive restrictions (like odd-even schemes) to targeted, evidence-based industrial regulations.

IX. LIMITATIONS AND FUTURE SCOPE

Despite its successes, AQ-INDIA has certain limitations. First, the INSAT-3DR satellite has a limited temporal resolu- tion, and cloud cover can obscure optical satellite readings, leading to reliance on interpolated data. Second, the Random Forest model, while highly accurate for short-term (1-3 hour) predictions, does not natively capture long-term seasonal de- pendencies as effectively as sequence models (e.g., LSTMs). Future iterations of the system will focus on three primary enhancements: (1) Integration of low-cost IoT-based PM2.5 sensors to create a hybrid crowd-sourced+institutional moni- toring network; (2) Implementation of a Spatiotemporal Graph Convolutional Network (STGCN) to simultaneously model spatial correlations and temporal sequences for extended 48- hour forecasting; (3) Expansion of the chatbot to support multiple regional Indian languages using transformer-based NLP models.

X. CONCLUSION

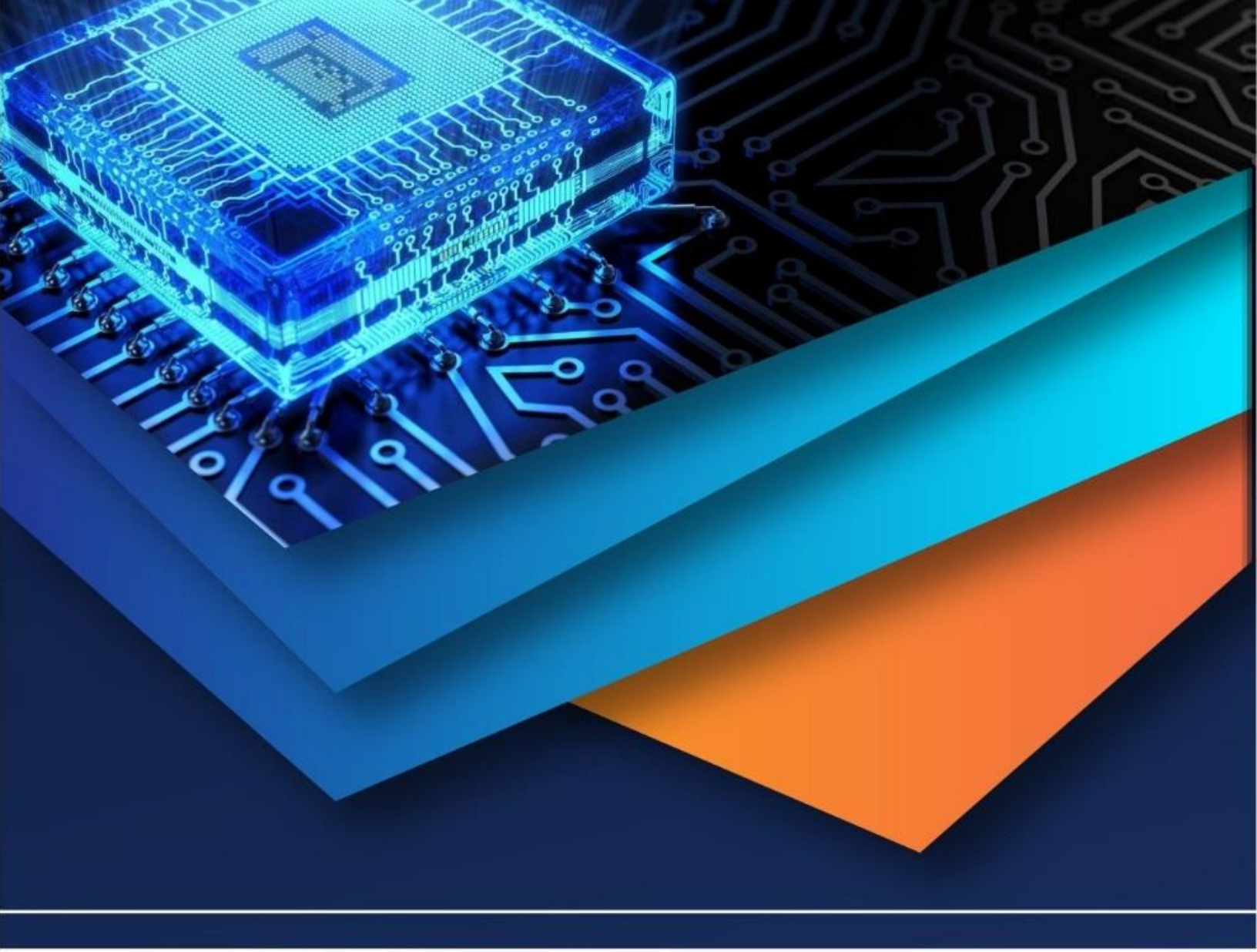
This paper presented AQ-INDIA, a robust, AI-powered integrated air quality monitoring and prediction system de- signed to mitigate the severe environmental surveillance gaps in India. By systematically fusing heterogeneous data streams—INSAT-3DR satellite imagery, CPCB ground obser- vations, and MERRA-2 meteorological reanalysis—within a cloud-native computational framework, the platform achieves an unprecedented R^2 score of 0.91 in real-time AQI pre- diction. The mathematical application of K-Means clustering successfully demarcated six distinct, physically plausible air quality zones. The deployment of this ML pipeline via a Node.js backend to an interactive React.js frontend, featuring real-time maps, trend dashboards, and a conversational chat- bot, establishes a new paradigm for accessible environmental informatics.

AQ-INDIA not only bridges the gap between fragmented raw data and actionable intelligence but also serves as a scalable foundation

for future smart-city environmental monitoring initiatives, directly contributing to the realization of UN SDG Goals 3 and 11.

REFERENCES

[1] World Health Organization, ”WHO global air quality guidelines: partic- ulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide,” WHO, Geneva, Switzerland, 2021.
[2] Central Pollution Control Board, ”National Air Quality Monitoring Pro- gramme (NAMP),” CPCB, New Delhi, India, 2023. [Online]. Available: https://cpcb.nic.in
[3] P. Gupta and S. Kumar, ”A critical review of air quality index frame- works for public health assessment,” Environmental Monitoring and Assessment, vol. 194, no. 3, p. 234, 2022.
[4] A. Agarwal, S. Jain, and P. Sharma, ”LSTM-based PM2.5 prediction model for Delhi: A comparative study,” in Proc. IEEE Int. Conf. on Machine Learning and Applications (ICMLA), 2022, pp. 1023-1028.
[5] A. Kumar and P. Goyal, ”Random forest-based air quality index prediction for Indian metropolitan cities,” Journal of Environmental Management, vol. 312, p. 114892, 2022.
[6] S. D. Ghude et al., ”Application of satellite observations for estimating surface-level PM2.5 over India,” Atmospheric Environment, vol. 241, p. 117812, 2020.
[7] Q. Di et al., ”A deep learning approach to estimate PM2.5 concentrations using satellite data and ground measurements in the United States,” Environmental Science & Technology, vol. 55, no. 15, pp. 10223-10232, 2021.
[8] R. D. Koster et al., ”MERRA-2: Modern-Era Retrospective Analysis for Research and Applications, Version 2,” Journal of Climate, vol. 29, no. 11, pp. 4059-4084, 2016.
[9] S. Vardoulakis et al., ”Modelling air quality in street canyons: A review,” Atmospheric Environment, vol. 37, no. 2, pp. 155-182, 2003.
[10] Y. Zhang et al., ”Spatiotemporal air quality forecasting using graph convolutional networks,” in Proc. AAAI Conf. on Artificial Intelligence, 2023, pp. 12847-12854.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY