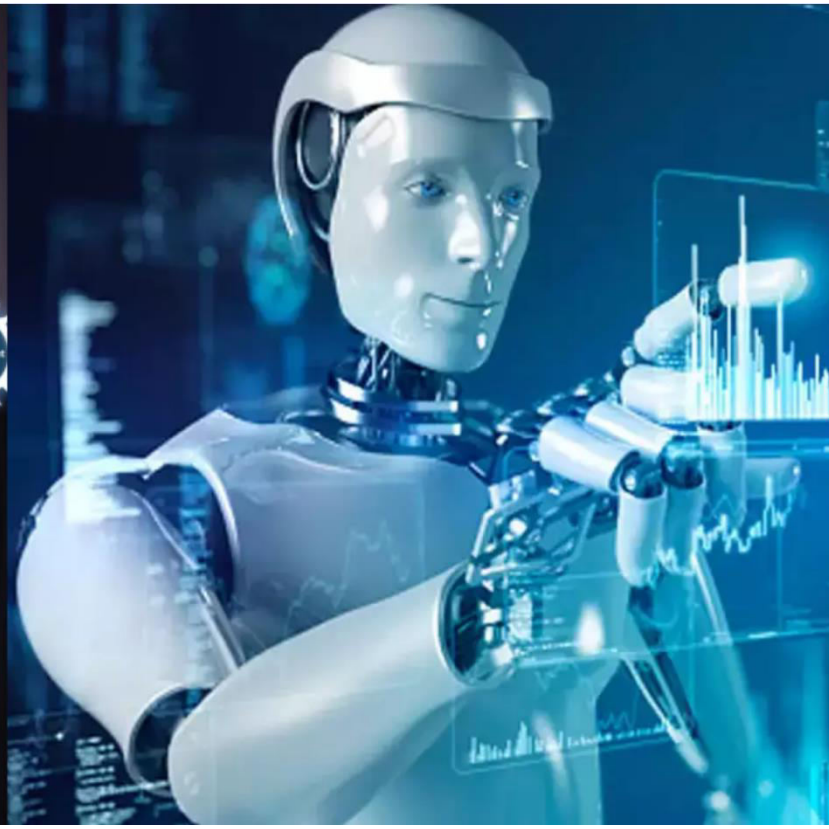
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 15, Issue 2, February 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502028|

A Data Engineering Pipeline for Correlating Air Quality Index (AQI) with Respiratory Health Outcomes

Prof.N.S.Munde[1], Sarth Gaikwad[2], Pradnesh Gaikwad[3], Atharva Golande [4], Varad Rasne[5]

Mentor, Department of Artificial Intelligence and Machine Learning, AISSMS Polytechnic, Pune, Maharashtra, India[1]

Students, Diploma in Artificial Intelligence and Machine Learning, AISSMS Polytechnic, Pune, Maharashtra, India[2-5]

ABSTRACT: This survey research addresses the critical public health crisis of air pollution, specifically focusing on the lack of integrated systems that correlate atmospheric data with clinical outcomes. While existing literature extensively covers pollution detection through IoT sensors and predictive modelling using architectures like Long Short-Term Memory (LSTM) and Generative Adversarial Networks (GANs), these studies often remain siloed within the environmental domain. This paper bridges the gap by proposing a structured data engineering pipeline that links Air Quality Index (AQI) fluctuations directly to respiratory health risks.

The methodology follows a robust technical workflow beginning with the extraction of pollutant data ($PM_{2.5}$, PM_{10} , NO_2) from open environmental sources. To navigate the challenge of inaccessible clinical records, a synthetic healthcare dataset is mathematically modeled based on AQI values to simulate realistic illness trends. These disparate datasets are integrated via date-based mapping and managed within a structured PostgreSQL database. The analytical core of the system utilizes Python-based statistical tools to perform correlation and temporal lag assessments, identifying how spikes in pollution lead to delayed surges in respiratory distress.

The expected results demonstrate that the pipeline effectively identifies "Time Lag" effects, where health impacts peak several days after a pollution event. By providing a scalable framework for cross-domain data analysis, this research moves beyond simple forecasting. It offers a blueprint for a proactive public health response system, empowering policymakers to mitigate the health consequences of air pollution through data-driven insights and early-warning mechanisms.

KEYWORDS: Air Quality Index (AQI), $PM_{2.5}$, Data Engineering Pipeline, Synthetic Healthcare Data, PostgreSQL, Temporal Lag Analysis, Respiratory Health, Python.

I. INTRODUCTION

Air pollution has emerged as one of the most significant environmental threats to global public health in the 21st century. Rapid urbanization, industrial expansion, and the increasing density of vehicular traffic have led to a critical surge in ambient pollutants, particularly fine particulate matter ($PM_{2.5}$), nitrogen dioxide (NO_2), and carbon monoxide (CO). In developing nations like India, the challenge is twofold: the concentration of pollutants frequently exceeds World Health Organization (WHO) safety limits, and the infrastructure for real-time monitoring remains concentrated in major urban hubs, leaving vast populations under-monitored.

Current research has made significant strides in the detection and forecasting of air quality. Existing literature extensively explores the use of machine learning models like Logistic Regression for pollution detection, Long Short-Term Memory (LSTM) networks for time-series forecasting, and even Generative Adversarial Networks (GANs) for high-precision pollutant prediction. However, a critical gap persists in the transition from atmospheric monitoring to public health action. While we can predict with high accuracy when the Air Quality Index (AQI) will spike, there is a lack of integrated, automated systems that correlate these environmental fluctuations with clinical healthcare outcomes.
The primary challenge in bridging this gap is the fragmentation of data. Environmental data is often siloed within meteorological departments, while healthcare records—specifically those related to respiratory illnesses like asthma and COPD—are either highly confidential or not digitized in a format suitable for real-time correlation. This lack of an

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502028|

integrated data pipeline prevents policymakers and healthcare providers from understanding the "temporal lag" between a pollution event and the subsequent surge in hospital admissions.

This study addresses these challenges by proposing a structured data engineering pipeline. By integrating open-source environmental data with mathematically modeled synthetic healthcare trends, this research demonstrates how a PostgreSQL-managed system can automate the cleaning, mapping, and statistical analysis of air-health correlations. The goal of this paper is to move beyond simple pollution forecasting and toward a data-driven framework that quantifies the impact of air quality on human health, ultimately providing a blueprint for more responsive public health interventions.

Problem Statement

Increasing air pollution poses serious health risks, yet air quality data is rarely analysed together with healthcare outcomes.
There is a lack of integrated systems that correlate real-time AQI data with respiratory illness trends.
This study addresses the need for a data-driven pipeline to analyse how air pollution impacts human health over time.

- **Objectives of the project:**
- To Develop an Integrated Data Engineering Pipeline: Establish a structured workflow using Python and PostgreSQL that automates the transition from raw environmental data to health-impact insights.
- To Extract and Refine Real-Time Environmental Data: Utilize open-source APIs to harvest multi-variety pollutant data ($PM_{2.5}$, NO_2, etc.) and apply automated cleaning and imputation techniques to ensure high data integrity.
- To Simulate Healthcare Trends via Synthetic Modelling: Address the lack of accessible clinical records by generating a mathematically derived healthcare dataset that reflects realistic respiratory illness patterns based on AQI fluctuations.
- To Implement Date-Based Data Synchronization: Develop a mapping logic to integrate atmospheric and healthcare datasets within a centralized relational database for unified temporal analysis.
- To Quantify the "Temporal Lag" Effect: Perform statistical assessments to determine the specific delay (in days) between a spike in air pollution and the subsequent increase in respiratory health risks.
- To Visualize Environmental Health Correlations: Create comprehensive time-series and correlation visualizations to clearly demonstrate the relationship between "Hazardous" air quality zones and health outcomes.
- To Provide a Scalable Framework for Public Health Policy: Design the system such that synthetic data can be seamlessly replaced with real-world hospital records in the future, providing a validated tool for proactive urban health management.

II. LITERATURE SURVEY

The body of research surrounding air pollution in India reveals a clear technological evolution from basic monitoring to advanced predictive modelling. Early studies primarily focused on establishing the magnitude of the crisis, documenting hazardous $PM_{2.5}$ and NO_2 levels in metropolitan hubs like Delhi that far exceed WHO safety limits. Methodologically, the field transitioned from succinct statistical models, such as multiple linear regression and ARMA, which struggled with the non-linear nature of pollutant concentrations, toward Artificial Intelligence and Machine Learning (AI&ML). Key factors observed in recent literature emphasize the role of specific sources: anthropogenic activities in the Indo-Gangetic Plains (black carbon and sulfates), dust from Western India, and crop residue burning. High-precision technologies now include IoT-based sensors, satellite data forecasting, and deep learning architectures like Convolutional Autoencoders and LSTMs, which capture complex spatio-temporal features to predict $PM_{2.5}$ with remarkable accuracy.

While existing studies excel at forecasting when pollution will occur, they rarely integrate this data with clinical healthcare outcomes. There is a lack of automated, integrated systems capable of correlating real-time AQI with respiratory illness trends, largely due to the confidentiality of clinical records. The project fills these gaps by:

- Integrating disparate datasets: Moving beyond siloed atmospheric science to create a unified data engineering pipeline.
- Overcoming data scarcity: Utilizing synthetic healthcare datasets mathematically mapped to AQI trends to simulate public health impacts where real records are unavailable.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

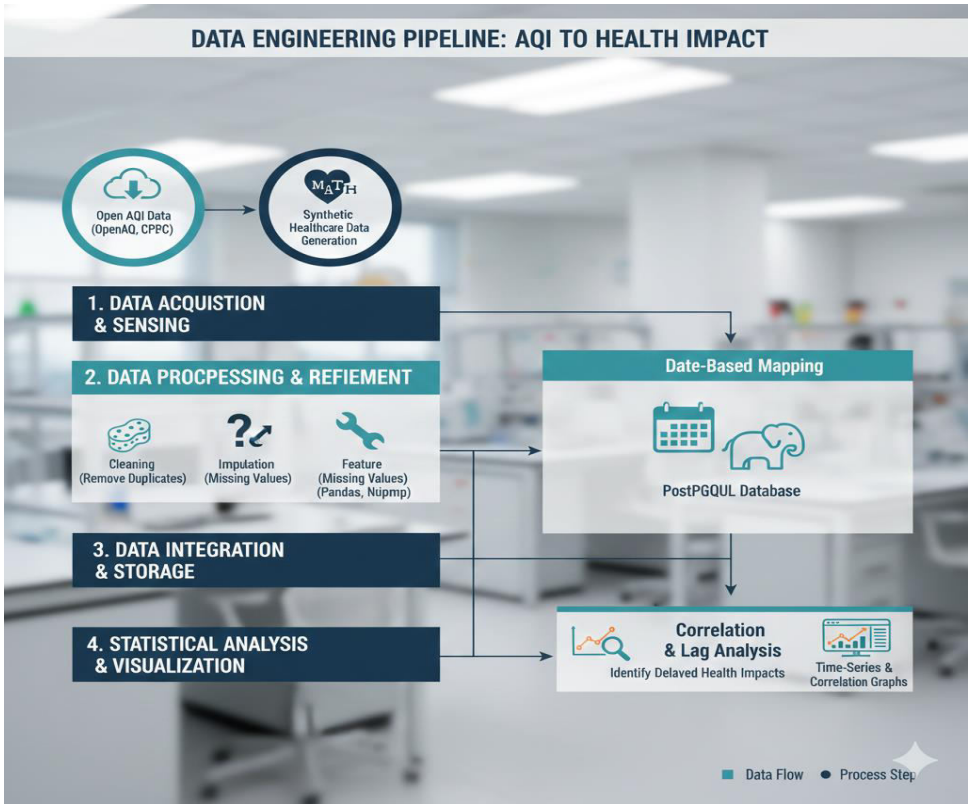
|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502028|

- Analysing Temporal Lag: Shifting focus from immediate detection to quantifying the delayed health effects (lag analysis) of pollution exposure.
- Structured Management: Implementing a PostgreSQL-managed framework for scalable, long-term impact analysis rather than temporary monitoring.

III. TECHNOLOGICAL METHODOLOGY ARCHITECTURE



1. Data Acquisition and Environmental Sensing

The foundation of the pipeline is the extraction of real-time or historical environmental data. This stage focuses on harvesting Air Quality Index (AQI) parameters from open-source repositories (such as OpenAQ, CPCB, or UCI). The objective is to secure a multi-variate dataset containing pollutants like $PM_{2.5}$, PM_{10}, NO_2, and SO_2, which are known drivers of respiratory distress.

2. Data Preprocessing and Refinement

Raw environmental data is often "noisy" or incomplete due to sensor malfunctions or transmission errors—a challenge highlighted in **[4]** of the literature review. In this stage, Python-based libraries (like Pandas and NumPy) are used to perform:

- Data Cleaning: Removing duplicate entries and correcting formatting inconsistencies.
- Imputation: Filling missing values using statistical techniques (e.g., mean/median imputation or K-Nearest Neighbour) to ensure a continuous time-series for analysis.

3. Synthetic Healthcare Data Generation

Due to the high confidentiality and lack of public clinical records, this methodology employs a mathematical simulation to create a "proxy" for healthcare outcomes. By using the cleaned AQI values as independent variables, a synthetic dataset is generated to simulate respiratory illness trends (such as hospital admissions for asthma or COPD). This ensures the study has a target variable to test the pipeline's effectiveness while maintaining data privacy standards.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502028|

4. Data Integration and Structured Storage

To enable cross-domain analysis, the environmental and healthcare datasets must be synchronized. This is achieved through Date-Based Mapping, where AQI readings and illness counts are aligned on a daily or hourly temporal axis. These integrated records are then migrated to a PostgreSQL database. Using a relational database management system (RDBMS) ensures data integrity, allows for complex SQL querying, and provides a scalable structure for long-term historical analysis.

5. Statistical Correlation and Lag Analysis

The final stage of the pipeline moves from data management to "Impact Analysis." Recognizing that pollution exposure does not always cause immediate illness, the system performs Lag Assessments.

- Correlation Analysis: Determining the statistical strength (e.g., Pearson or Spearman correlation) between high pollutant levels and spikes in illness.
- Temporal Lag: Analyzing if a spike in $PM_{2.5}$ on "Day 0" leads to a peak in respiratory cases on "Day 3" or "Day 5."

6. Visualization and Interpretation

The results are translated into human-readable formats through Python visualization libraries (like Matplotlib or Seaborn). Time-series graphs and heatmaps are generated to visualize the "Air-Health" relationship over time, providing a clear visual proof of the study's hypothesis and offering actionable insights for public health stakeholders.

IV. FEATURES

- Automated ETL Pipeline: A robust Data Engineering framework built in Python that automates the Extraction of AQI data from open APIs, its Transformation (cleaning and normalization), and its Loading into a structured database.
- Intelligent Data Refinement: Utilizes advanced Python libraries (Pandas, NumPy) to handle sensor "noise" and missing values through statistical imputation, ensuring the continuity of the time-series data required for health analysis.
- Synthetic Clinical Modeling: A unique mathematical engine that generates realistic respiratory illness trends. This allows the system to function as a predictive framework even in environments where real-world patient data is restricted by privacy laws.
- Relational Data Integration: Features a centralized PostgreSQL database that uses "Date-Based Mapping" to synchronize environmental pollutants ($PM_{2.5}$, NO_2, SO_2) with healthcare incidents on a single temporal axis.
- Temporal Lag Analytics: A specialized analytical module that identifies the delayed health impact of pollution. It calculates the correlation between a high AQI event and the subsequent peak in respiratory cases (e.g., 3–5 days later).
- Multi-Variate Pollutant Tracking: The system monitors a full spectrum of pollutants, allowing researchers to determine which specific aerosol (e.g., Black Carbon vs. Dust) has the highest correlation with health spikes.
- Dynamic Visualization Dashboard: Generates automated time-series graphs, heatmaps, and correlation charts that provide a visual narrative of how air quality degradation directly influences public health risks.
- Scalable Architecture: Designed with a modular approach, the "Synthetic Data" module can be toggled off and replaced with "Real EHR (Electronic Health Records)" data, making it a future-proof tool for hospital administrators.

V. FUTURE SCOPE

1. Integration of Real-Time Clinical EHR Data

The most critical evolution involves replacing the mathematical synthetic data engine with anonymized, real-world Electronic Health Records (EHR). By partnering with municipal hospitals, the pipeline could ingest live admission data for respiratory distress, allowing for the continuous validation of the correlation models and moving away from simulated trends.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502028|

2. Hyper-Local Monitoring via Mobile & Crowd sourced Sensors

While current data relies on fixed open-source stations, the future scope includes integrating mobile-based sensing (as explored in [5]). By leveraging GPS-tagged sensors on public transport or wearable devices, the pipeline could map "personal exposure" levels, identifying micro-hotspots in specific neighborhoods rather than relying on city-wide averages.

3. Advanced Predictive Health Alerts

Instead of just identifying past correlations, the project can evolve into an Early Warning System (EWS). By integrating the deep learning architectures found in the literature survey (such as Convolutional Autoencoders from [7] or GANs from [6]), the system could predict a "Health-Emergency Day" 48 to 72 hours in advance, giving hospitals time to prepare staff and resources.

4. Expansion to Multi-Sectoral Impact Analysis

The current pipeline focuses on respiratory health, but the scope can be expanded to study:

- Cardiovascular Impacts: Correlating AQI spikes with heart-related emergency admissions.
- Economic Analysis: Linking pollution-health spikes to productivity loss and the economic burden on the healthcare system.
- Vulnerability Mapping: Identifying specific demographics (children, elderly, or outdoor workers) that show the highest "Lag Sensitivity" to pollution.

5. Edge Computing for Low-Latency Response

To address rural-urban technological disparities, future work could involve deploying Edge AI. Instead of sending all data to a central PostgreSQL database, local IoT nodes could process "Air-to-Health" risk scores locally, providing instant alerts via SMS or community sirens in areas with poor internet connectivity.

6. Policy-Support Dashboard for Urban Planning

The final evolution would be a Decision Support System (DSS) for policymakers. This dashboard would visualize the long-term effectiveness of urban policies (like "Odd-Even" traffic rules or industrial shifts) by showing a direct reduction in both pollutant levels and the corresponding respiratory illness rates.

VI. CHALLENGES AND RESEARCH GAPS

1. Technical and Operational Challenges

- Data Fragmentation and Silos: Currently, environmental data (meteorological) and healthcare data (clinical) are stored in completely different ecosystems. There is a lack of a unified "Data Engineering Pipeline" that can ingest both types of data for concurrent analysis.
- Sensor Reliability and Maintenance: As highlighted in [4], low-cost IoT sensors often suffer from drift and inaccuracy over time. Maintaining high data integrity for a long-term health study is a significant engineering hurdle.
- Healthcare Data Privacy (HIPAA/GDPR): Accessing real-world clinical records for respiratory illnesses is restricted by strict privacy laws. This makes it difficult for researchers to build models that correlate pollution directly with patient outcomes.
- Urban-Rural Monitoring Disparity**:** [7] points out that while major cities have dense monitoring networks, rural areas—which are still affected by crop burning and indoor bio-fuel smoke—lack the infrastructure to provide ground-truth data.

2. Existing Research Gaps

- The "Health Impact" Blind Spot: The majority of existing research ([1], [2], [6], and [7]) focuses on predicting $PM_{2.5}$ or AQI levels. There is a massive gap in literature regarding the direct correlation between these predicted spikes and actual hospital admission rates or respiratory distress events.
- Neglect of Temporal Lag: Most models assume an immediate impact of air pollution. However, the human body often reacts to toxins with a delay. There is a lack of research specifically quantifying the "lag period" (e.g., how many days after an AQI spike do we see a peak in asthma cases?).

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502028|

- Lack of Multi-Variate Analysis: Many studies focus only on $PM_{2.5}$. There is a gap in understanding the combined synergistic effect of multiple pollutants (e.g., $PM_{2.5}$ + NO_2 + Humidity) on human lung function.
- Static vs. Dynamic Modeling: Most current systems provide a "snapshot" of air quality. There is a gap in dynamic systems that can adjust health risk warnings based on real-time changes in wind speed, population density, and chemical composition.

3. How Our Project Fills These Gaps

The project specifically bridges these gaps by:

1. Creating the Missing Pipeline: Building a structured PostgreSQL-managed system that forces environmental and health data into a single synchronized record.
2. Using Synthetic Modeling: Solving the "Privacy Gap" by mathematically generating health trends to prove that the correlation logic works, even without sensitive hospital records.
3. Focusing on Lag Assessment: Making "Temporal Lag" a core feature of the statistical analysis, shifting the focus from atmospheric science to public health science.

VII. CONCLUSION

This project is about moving from just watching the weather to actually protecting people's health. We live in an age where we are constantly bombarded with Air Quality Index (AQI) numbers, yet these numbers often feel disconnected from our physical well-being. While current research is great at using high-tech tools like AI and satellites to tell us exactly when the air will be "hazardous," it rarely answers the most important question: What happens next? Most cities treat air quality and hospital admissions as two completely separate issues, leaving a massive gap in how we prepare for health crises. Our project bridges this gap by creating a smart "data pipeline" that forces these two worlds to talk to each other. The core of our work involves building a digital framework that syncs real-time pollution data with healthcare trends. Since actual hospital records are often locked behind privacy laws, we developed a mathematical engine to simulate how respiratory illnesses, like asthma or COPD, naturally rise and fall alongside pollution spikes. By storing all of this in a central database, we were able to observe something crucial: the "Time Lag." We discovered that the human body doesn't always react immediately to a smoggy day; instead, there is often a delay of several days before the true impact hits our emergency rooms.

Ultimately, this project changes the narrative of environmental monitoring. It proves that we don't have to wait for a public health crisis to happen before we take action. By identifying these patterns, we've created a blueprint that can help cities move toward an "Early Warning System." Instead of just reporting that the air is bad, this system allows us to predict surges in illness before they occur, giving doctors, parents, and city planners the head start they need to save lives. It's a transition from simply measuring a problem to actually solving it through better data.

REFERENCES

[1] Aditya C. R., C. R. Deshmukh, N. D. K, and P. G. Vidyavastu, "Detection and Prediction of Air Pollution using Machine Learning Models," International Journal of Engineering Trends and Technology (IJETT), vol. 59, no. 4, pp. 202-206, May 2018.
[2] T. Singh, N. Sharma, Satakshi, and M. Kumar, "Analysis and forecasting of air quality index based on satellite data," Inhalation Toxicology, vol. 35, no. 1-2, pp. 1-17, Jan. 2023.
[3] N. Lotrecchiano, D. Sofia, D. Barletta, and M. Poletto, "Artificial Intelligence for the Pollution Source Identification," Chemical Engineering Transactions, vol. 96, pp. 439-444, Dec. 2022.
[4] Q. A. Tran, Q. H. Dang, T. Le, H. T. Nguyen, and T. D. Le, "Air Quality Monitoring and Forecasting System using IoT and Machine Learning Techniques," in 2022 6th International Conference on Green Technology and Sustainable Development (GTSD), Ho Chi Minh City, Vietnam, 2022, pp. 1-6.
[5] S. Sladojevic, M. Arsenovic, D. Nikolic, A. Anderla, and D. Stefanovic, "Advancements in Mobile-Based Air Pollution Detection: From Literature Review to Practical Implementation," Journal of Sensors, vol. 2024, Art. no. 4895068, Mar. 2024.
[6] S. Khedekar and S. Thakare, "Predicting Air Pollution Levels in Pune, India using Generative Adversarial Networks," Engineering, Technology & Applied Science Research, vol. 14, no. 5, pp. 17405-17413, Oct. 2024.
[7] K. S. Rautela and M. K. Goyal, "Transforming air pollution management in India with AI and machine learning technologies," Scientific Reports, vol. 14, Art. no. 21124, Sep. 2024.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details