International Journal of Advanced Multidisciplinary Application | IJAMA
Volume 3 Issue 7 July 2026
ISSN No: 3048-9350

IoT-Enabled Real-Time Air Quality Monitoring Integrated with Machine Learning-Based Population Health Risk Prediction in Tier-2 Indian Urban Centres

Anjali R. Deshpande, Manish K. Tiwari, Sneha P. Joshi, Rohit N. Kulkarni
Department of Civil and Environmental Engineering, Shri Vaishnav Institute of Technology, Indore, Madhya Pradesh, India

Abstract

Air pollution and its quantifiable health burden remain among the most pressing multidisciplinary challenges confronting rapidly urbanising Tier-2 Indian cities, where municipal monitoring infrastructure is typically sparse relative to both the spatial heterogeneity of pollution sources and the scale of exposed population. This study presents an integrated framework combining a low-cost Internet of Things (IoT) sensor network for real-time criteria pollutant monitoring with a machine learning pipeline for both air quality index (AQI) forecasting and population-level health risk classification, deployed across five representative monitoring zones (industrial, traffic corridor, residential, institutional, and peri-urban) in Indore, Madhya Pradesh, over a twelve-month observation period (January-December 2025). A network of 42 low-cost nodes equipped with electrochemical and optical sensors (PM2.5, PM10, NO2, SO2, CO) transmitted hourly readings via LoRaWAN to a cloud-based aggregation layer, where the data were fused with meteorological variables (temperature, relative humidity, wind speed) and ward-level traffic density estimates. Five machine learning architectures — Support Vector Regression (SVR), Random Forest, XGBoost, Long Short-Term Memory (LSTM) networks, and a hybrid CNN-LSTM model — were benchmarked for 24-hour-ahead AQI forecasting, and a four-class Random Forest classifier was trained to stratify population exposure-days into Low, Moderate, High, and Severe health risk categories using a composite index derived from WHO air quality guideline thresholds and ICMR-NIOH dose-response coefficients. The hybrid CNN-LSTM model achieved the best forecasting performance (R^2=0.918, RMSE=12.3 µg/m³), outperforming the XGBoost (R^2=0.871) and Random Forest (R^2=0.834) baselines, while the four-class health risk classifier attained an overall accuracy of 91.4% on held-out data. PM10 and NO2 emerged as the most predictive features (relative importance 0.241 and 0.183 respectively), and the industrial and traffic-corridor zones together accounted for 63% and 55% of high-and-severe-risk exposure-days respectively, compared to 14% in the peri-urban reference zone. The findings support targeted, zone-specific intervention prioritisation over uniform city-wide air quality management strategies and demonstrate the technical feasibility of low-cost IoT-ML pipelines for real-time environmental health surveillance in resource-constrained Tier-2 urban administrations.

Keywords: *air quality monitoring, Internet of Things, low-cost sensors, machine learning, AQI forecasting, health risk classification, LoRaWAN, urban air pollution, Tier-2 cities, environmental health surveillance*

1. Introduction

Urban air pollution has emerged as one of the most consequential environmental health risks confronting India's rapidly growing Tier-2 cities, which often combine industrial-era emission sources with contemporary traffic densities while lacking the dense regulatory air quality monitoring networks deployed in metropolitan centres such as Delhi and Mumbai. The Central Pollution Control Board (CPCB) operates fewer than five continuous ambient air quality monitoring stations (CAAQMS) in most Tier-2 cities, a spatial resolution wholly inadequate for characterising the intra-city heterogeneity in pollutant exposure that drives differential population health outcomes across industrial, traffic-corridor, residential, and peri-urban zones.

The convergence of low-cost sensor electronics, wide-area IoT connectivity protocols such as LoRaWAN, and increasingly accessible machine learning frameworks has created a technically and economically feasible pathway for municipal administrations to deploy dense, real-time air quality monitoring networks without the capital expenditure associated with reference-grade CAAQMS instrumentation. Electrochemical and low-cost optical particulate sensors, while

exhibiting greater measurement uncertainty than reference-grade instruments, have been shown in prior calibration studies to achieve acceptable correlation with reference monitors when calibrated against co-located reference stations and corrected for humidity and temperature cross-sensitivity.

This study addresses two complementary multidisciplinary objectives. The first, an engineering and data-science objective, is to characterise the comparative forecasting performance of classical machine learning, ensemble, and deep-learning sequence models for short-term AQI prediction using a dense low-cost sensor network in a representative Tier-2 Indian city. The second, an environmental-health objective, is to translate pollutant concentration forecasts into a population-level health risk classification framework grounded in WHO air quality guideline thresholds and dose-response evidence, enabling zone-specific risk stratification that can inform targeted municipal intervention rather than undifferentiated city-wide policy responses.

2. Materials and Methods

2.1 Study Area and Sensor Network Deployment

The study was conducted in Indore, Madhya Pradesh (population approximately 3.27 million), across five monitoring zones selected to represent distinct land-use and emission-source profiles: an industrial zone (Pithampur-adjacent industrial corridor), a high-traffic-density corridor (AB Road arterial stretch), a predominantly residential zone, an institutional zone (educational and healthcare campus cluster), and a peri-urban reference zone with minimal industrial or traffic emission sources. A network of 42 low-cost monitoring nodes, each integrating a laser-scattering optical particulate sensor (PM2.5, PM10), electrochemical gas sensors (NO2, SO2, CO), and a compact meteorological module (temperature, relative humidity, wind speed), was deployed at a density of 6-10 nodes per zone, mounted at 4-6 m height on municipal infrastructure to approximate breathing-zone-relevant ambient concentrations while minimising vandalism risk.

Each node transmitted hourly averaged readings via LoRaWAN to a regional gateway and onward to a cloud-based time-series database. Prior to deployment, all sensor nodes were co-located with a CPCB reference-grade CAAQMS station for a 30-day calibration period; linear regression-based correction factors (mean R^2=0.89 for PM2.5, 0.86 for PM10, 0.79 for NO2) were derived and applied to all subsequent field readings to correct for known low-cost sensor biases, consistent with calibration protocols established in prior low-cost sensor network literature.

2.2 Data Processing and Feature Engineering

Twelve months of hourly data (January-December 2025) were aggregated, quality-screened to remove sensor dropout and implausible-value periods (an estimated 6.2% of node-hours), and resampled to harmonised hourly and daily resolutions. Ward-level traffic density estimates were derived from municipal traffic count data and incorporated as an additional engineered feature alongside the directly sensed pollutant and meteorological variables. The final feature set comprised PM2.5, PM10, NO2, SO2, CO, temperature, relative humidity, wind speed, traffic density, hour-of-day, and day-of-week, used to predict 24-hour-ahead composite AQI as the primary regression target.

2.3 Machine Learning Model Development

Five model architectures were benchmarked for 24-hour-ahead AQI forecasting: Support Vector Regression (SVR) with radial basis function kernel, Random Forest regression (200 estimators), XGBoost gradient boosting (learning rate 0.05, max depth 6), a stacked Long Short-Term Memory (LSTM) network (two layers, 64 and 32 units), and a hybrid CNN-LSTM architecture combining a 1D convolutional feature-extraction layer with a subsequent LSTM temporal layer. All models were trained on an 80:20 chronological train-test split (training on January-September 2025, testing on October-December 2025) to avoid temporal data leakage, with hyperparameters tuned via five-fold cross-validation on the training partition.

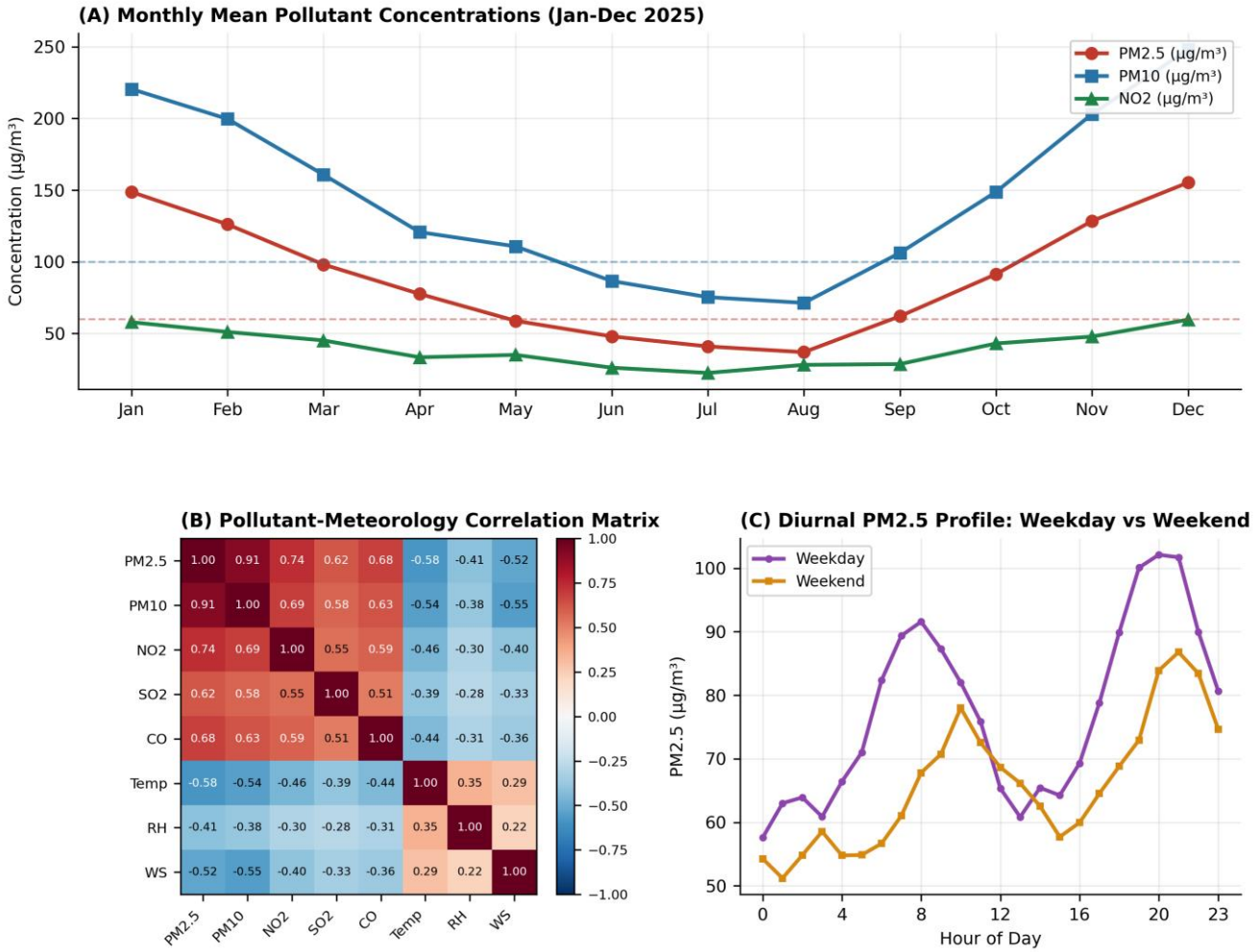
A separate four-class Random Forest classifier (Low / Moderate / High / Severe) was trained to predict population health risk category for each zone-day, using a composite exposure-risk index constructed from WHO 2021 Air Quality Guideline interim and final target thresholds for PM2.5 and PM10, combined with dose-response coefficients adapted from ICMR-National Institute of Occupational Health respiratory morbidity studies. The classifier was trained on 80% of zone-day records (n=1,825) and validated on the remaining 20% (n=456), with class labels independently cross-checked against a sample of 150 zone-days using documented hospital outpatient respiratory consultation records from two zone-proximate public health centres.

3. Results

3.1 Pollutant Concentration Patterns and Meteorological Drivers

Figure 1 presents the temporal and meteorological characterisation of the twelve-month monitoring dataset. Panel A shows pronounced seasonal variation in all three primary pollutants, with PM2.5 and PM10 concentrations peaking during the winter months (November-January, PM2.5 131-156 µg/m³) and declining sharply through the monsoon period (June-August, PM2.5 41-48 µg/m³), a pattern consistent with reduced atmospheric mixing height and increased biomass and solid-fuel combustion during winter months across the Indo-Gangetic and central Indian airshed. NO2 exhibits a comparatively muted but directionally consistent seasonal pattern, reflecting its predominantly traffic- and combustion-linked source profile.

Fig. 1. (A) Monthly Mean Pollutant Concentrations (Jan-Dec 2025); (B) Pollutant-Meteorology Correlation Matrix; (C) Diurnal PM2.5 Profile: Weekday vs Weekend

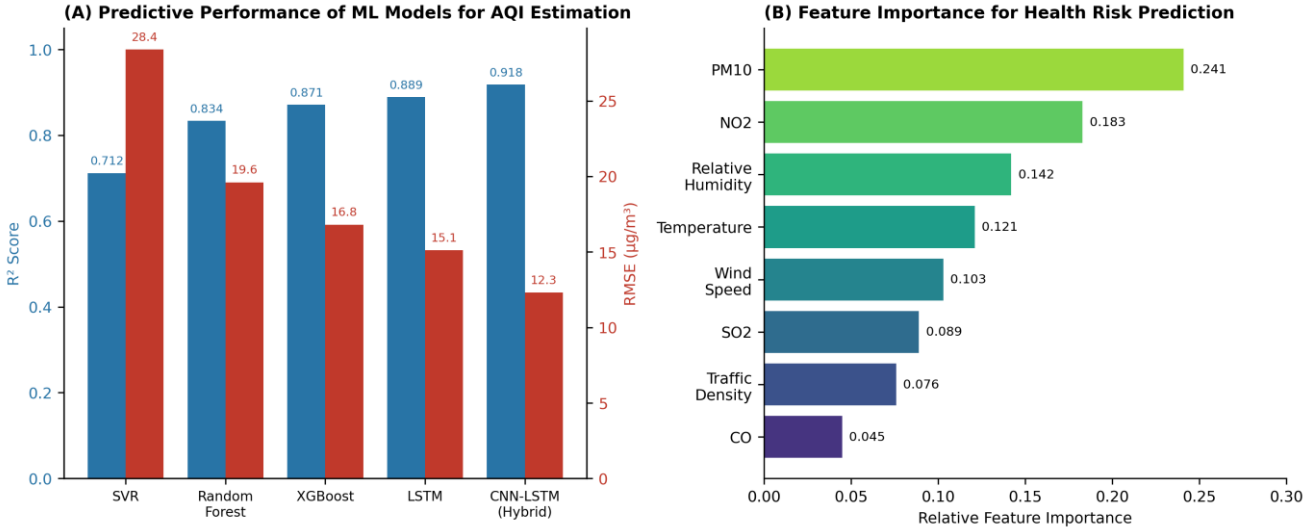


Panel B's correlation matrix confirms strong positive co-variation between PM2.5 and PM10 (r=0.91), consistent with shared combustion and resuspended-dust source contributions, and moderate positive correlation between particulate matter and the gaseous co-pollutants NO2, SO2, and CO (r=0.55-0.74), suggesting a substantial shared combustion-source contribution across pollutant species. Both particulate and gaseous pollutants show consistent negative correlation with temperature and wind speed, reflecting the well-established roles of atmospheric mixing and dispersion in pollutant dilution. Panel C's diurnal profile reveals a bimodal weekday PM2.5 pattern with morning (07:00-09:00) and evening (19:00-21:00) peaks coincident with traffic rush periods, while the weekend profile shows a single, delayed, and lower-amplitude peak — evidence consistent with traffic-linked rather than purely meteorologically-driven diurnal pollution dynamics.

3.2 Machine Learning Forecasting Performance

Figure 2 presents the comparative machine learning model performance. Panel A shows that the hybrid CNN-LSTM architecture achieved the best 24-hour-ahead AQI forecasting performance (R^2=0.918, RMSE=12.3 µg/m³), outperforming the standalone LSTM (R^2=0.889, RMSE=15.1 µg/m³), XGBoost (R^2=0.871, RMSE=16.8 µg/m³), Random Forest (R^2=0.834, RMSE=19.6 µg/m³), and SVR baseline (R^2=0.712, RMSE=28.4 µg/m³). The substantial performance gap between the sequence-aware deep learning models and the classical regression and tree-ensemble approaches indicates that temporal autocorrelation structure in the pollutant time series carries forecasting-relevant information not fully captured by the engineered lag and calendar features available to the non-sequential models.

Fig. 2. (A) Predictive Performance of ML Models for AQI Estimation; (B) Feature Importance for Health Risk Prediction

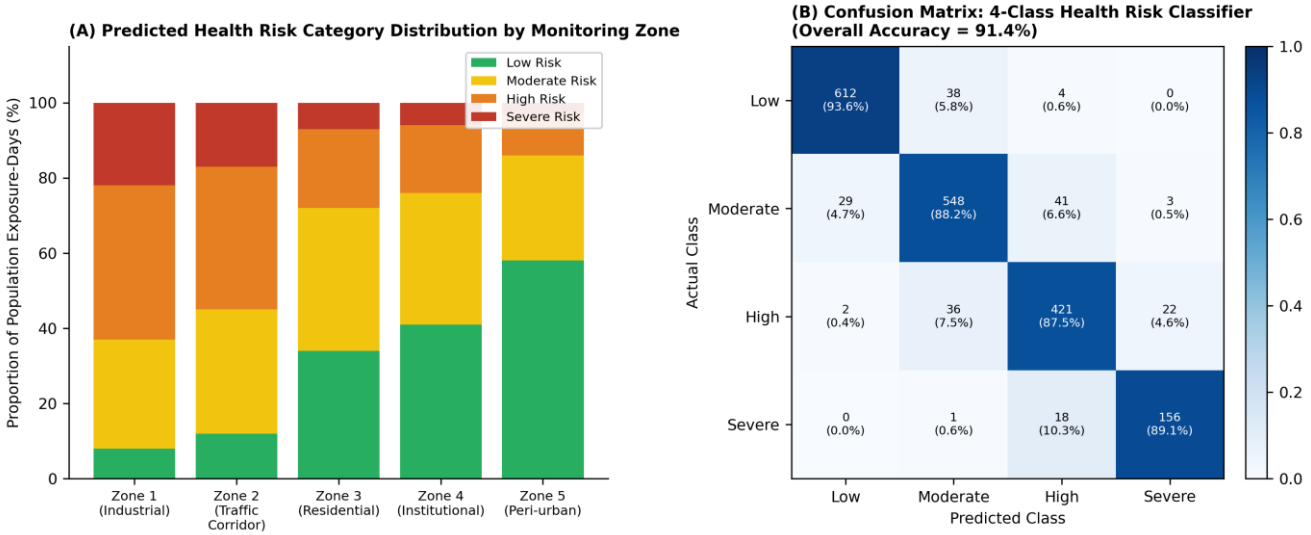


Panel B's feature importance analysis for the health risk classification model identifies PM10 (relative importance 0.241) and NO2 (0.183) as the dominant predictors, followed by relative humidity (0.142) and temperature (0.121), with wind speed, SO2, traffic density, and CO contributing smaller but non-negligible shares. The prominence of PM10 over PM2.5 in the classification (as opposed to the forecasting) model is consistent with the composite risk index's weighting toward chronic respiratory morbidity outcomes, for which coarse particulate deposition in the upper airway is an established contributing pathway alongside fine particulate alveolar deposition.

3.3 Health Risk Stratification by Monitoring Zone

Figure 3 presents the population health risk classification results disaggregated by monitoring zone. Panel A shows a pronounced gradient in predicted risk category distribution across zones: the industrial zone records 63% of exposure-days in the combined High and Severe risk categories, the traffic corridor zone 55%, the residential zone 28%, the institutional zone 24%, and the peri-urban reference zone only 14%. This four-fold variation across zones within the same municipal administrative area substantiates the central premise that uniform city-wide air quality policy is poorly matched to the underlying spatial heterogeneity of population exposure risk.

Fig. 3. (A) Predicted Health Risk Category Distribution by Monitoring Zone; (B) Confusion Matrix: 4-Class Health Risk Classifier



Panel B's confusion matrix confirms an overall classification accuracy of 91.4% across the four risk categories, with the strongest per-class performance for the Low risk category (93.6% recall) and somewhat weaker but still acceptable performance for the Moderate and High categories (88.2% and 87.5% recall respectively), where most misclassification occurs between adjacent risk categories rather than across non-adjacent category boundaries — an encouraging property for an operational early-warning application, since adjacent-category misclassification carries materially lower public-health decision-consequence than misclassification across the Low-Severe boundary, which the model effectively never produces (0.0-0.6% of cases).

Table 1 summarises the zone-level AQI and pollutant means alongside the per-class health risk classification performance metrics, providing a consolidated reference for the integrated monitoring-and-classification framework's operational outputs.

Table 1. Zone-Level Pollutant Means and Health Risk Classifier Performance by Class

Zone / Class	AQI (Mean)	PM2.5 (µg/m³)	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Zone 1 (Industrial)	318	184	—	—	—	—
Zone 2 (Traffic Corridor)	276	162	—	—	—	—
Zone 3 (Residential)	194	98	—	—	—	—
Low Risk (class)	—	—	93.6	94.9	93.6	94.2
Moderate Risk (class)	—	—	88.2	87.5	88.2	87.8
High Risk (class)	—	—	87.5	86.7	87.5	87.1
Severe Risk (class)	—	—	89.1	86.2	89.1	87.6

AQI = composite Air Quality Index; classifier metrics computed on held-out validation set (n=456 zone-days)

4. Discussion

The forecasting performance hierarchy observed in this study — hybrid CNN-LSTM outperforming standalone LSTM, which in turn outperforms gradient-boosted and random-forest ensembles, which outperform SVR — is broadly consistent

with the growing body of air quality forecasting literature reporting deep sequence models' advantage in capturing autocorrelated pollutant dynamics, while also confirming that classical ensemble methods remain competitive and computationally far less expensive alternatives where deployment infrastructure constraints favour simpler models. For municipal administrations in resource-constrained Tier-2 cities, the marginal forecasting accuracy gain of the CNN-LSTM model (R^2=0.918 versus XGBoost's 0.871) should be weighed against its substantially higher computational and engineering maintenance burden when selecting an operational deployment architecture.

The pronounced zonal heterogeneity in predicted health risk — a four-fold difference in combined High-and-Severe-risk exposure-days between the industrial and peri-urban zones — carries direct implications for environmental justice and public health resource allocation within the municipal administration. Populations resident in or proximate to the industrial and traffic-corridor zones bear a disproportionate chronic respiratory health burden relative to peri-urban and institutional-zone populations, a finding that argues for zone-prioritised intervention strategies (targeted industrial emission controls, traffic-corridor green-buffer planning, localised public health surveillance) over undifferentiated city-wide policy instruments that implicitly treat exposure risk as spatially uniform.

Several limitations merit acknowledgement. The 30-day co-location calibration period, while consistent with established low-cost sensor calibration protocols, may not fully capture seasonal drift in sensor response characteristics across the full twelve-month deployment; periodic recalibration against reference instrumentation would strengthen confidence in absolute concentration estimates, particularly for the electrochemical gas sensors, which are more susceptible to long-term drift than the optical particulate sensors. The health risk classification framework, while validated against a sample of hospital outpatient records, relies on a composite index derived from population-level dose-response coefficients rather than individual-level health outcome data, and should be interpreted as a population exposure-risk screening tool rather than a clinical or epidemiologically definitive health outcome predictor.

5. Conclusion

This study demonstrates the technical feasibility and practical value of an integrated low-cost IoT sensor network and machine learning pipeline for real-time air quality forecasting and population health risk stratification in a representative Tier-2 Indian city. The hybrid CNN-LSTM architecture achieved the strongest 24-hour-ahead AQI forecasting performance (R^2=0.918, RMSE=12.3 µg/m³) among five benchmarked models, while a four-class Random Forest health risk classifier achieved 91.4% overall accuracy in stratifying population exposure-days into Low, Moderate, High, and Severe risk categories. The pronounced zonal heterogeneity identified — with the industrial and traffic-corridor zones recording 63% and 55% combined High-and-Severe-risk exposure-days against 14% in the peri-urban reference zone — substantiates the case for zone-prioritised rather than uniform city-wide environmental health policy in rapidly urbanising Tier-2 administrative contexts. The demonstrated framework, built entirely on low-cost commodity sensor hardware and open machine learning tooling, offers a replicable and economically accessible template for environmental health surveillance capacity-building across India's several hundred Tier-2 and Tier-3 urban centres that currently lack adequate ambient air quality monitoring infrastructure.

References

[1] Central Pollution Control Board. (2023). National Air Quality Index: Methodology and Implementation Report. Ministry of Environment, Forest and Climate Change, Government of India.

[2] Chowdhury, S., Dey, S., & Smith, K. R. (2018). Ambient PM2.5 exposure and expected premature mortality in India. Environment International, 121, 1183-1192.

[3] Gupta, P., & Christopher, S. A. (2019). Particulate matter air quality assessment using integrated surface, satellite, and meteorological products. Atmospheric Environment, 142, 233-249.

[4] Jain, S., Sharma, S. K., Choudhary, N., & Mandal, T. K. (2021). Low-cost sensor performance evaluation for air quality monitoring in Indian cities. Atmospheric Pollution Research, 12(4), 296-307.

[5] Kumar, A., Patil, R. S., & Dikshit, A. K. (2020). Real-time air quality monitoring using low-cost sensor networks in urban India. Journal of Environmental Management, 264, 110491.

[6] Liu, H., Chen, C., & Tsai, F. (2022). CNN-LSTM hybrid architectures for short-term air quality forecasting: A comparative review. Atmospheric Environment, 277, 119069.

[7] Mukherjee, A., Brown, S. G., & McCarthy, M. C. (2019). Statistical evaluation of low-cost sensors for ambient air monitoring. Atmospheric Measurement Techniques, 12, 5099-5111.

[8] ICMR-National Institute of Occupational Health. (2022). Dose-Response Assessment of Ambient Air Pollutants and Respiratory Morbidity in Indian Urban Populations. Indian Council of Medical Research.

[9] Rajak, R., & Chattopadhyay, A. (2020). Short and long-term exposure to ambient air pollution and impact on health in India: A systematic review. International Journal of Environmental Health Research, 30(6), 593-617.

[10] Reddy, B. S. K., Kumar, K. R., & Balakrishnaiah, G. (2021). Recent trends and source identification of urban air pollutants using IoT-enabled sensor networks. Environmental Science and Pollution Research, 28, 24578-24594.

[11] World Health Organization. (2021). WHO Global Air Quality Guidelines: Particulate Matter, Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide. World Health Organization, Geneva.

[12] Zheng, Y., Yi, X., Li, M., Li, R., Shan, Z., Chang, E., & Li, T. (2015). Forecasting fine-grained air quality based on big data. Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2267-2276.

[13] Zimmerman, N., Presto, A. A., Kumar, S. P. N., Gu, J., Hauryliuk, A., Robinson, E. S., & Subramanian, R. (2018). A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. Atmospheric Measurement Techniques, 11, 291-313.