10.48047/jocaaa.2025.34.12.25

Spatiotemporal Analysis of Urban Air Quality in Visakhapatnam, Andhra Pradesh: Integrating Machine Learning Models for AQI Forecasting and Pollutant Attribution

P. Srivyshnavi[1], Sreenivasulu T[2*] ,P Shankaraiah [3], M Darshan Teja [4]
[1]Assistant Professor , Dept. of CS & E, S.P.M.V.V. Engineering College, Tirupati, India.
[2, 3, 4] Department of Mathematics, School of Advance Science, VIT Vellore, India.
Corresponding author: tsreenivasulu9491@gmail.com *

Abstract

Urban air quality deterioration in rapidly industrialising coastal cities poses a substantial public health challenge, particularly in megapolitan growth corridors of South Asia. This study undertakes a comprehensive spatiotemporal analysis of air quality in Visakhapatnam (Vizag), Andhra Pradesh a city experiencing simultaneous industrial expansion, port-driven commercial activity, and population growth. Using hourly monitoring data spanning January 2022 to October 2024, comprising 24,576 observations across 22 environmental variables, this research investigates the distributional dynamics of the Air Quality Index (AQI) alongside major pollutant concentrations (PM2.5, PM10, NO, NO2, NOx, SO2, CO, Ozone, Benzene, Toluene, and Xylene) and meteorological covariates (ambient temperature, relative humidity, wind speed, wind direction, barometric pressure, and solar radiation).

The empirical framework integrates descriptive statistics, time-series decomposition, correlation analysis, and supervised machine learning specifically Random Forest Regression, Gradient Boosting (XGBoost), and Long Short-Term Memory (LSTM) neural networks to generate interpretable AQI predictions and identify dominant pollutant contributors. Findings reveal that PM10 ($r = 0.921$) and PM2.5 ($r = 0.888$) exhibit the strongest linear association with AQI, while seasonal stratification shows peak pollution loading in the winter months (January: mean AQI = 177.96 $\mu g/m^3$ equivalent) and relative amelioration during the monsoon period (July: 87.95). An inter-annual deterioration trend between 2022 and 2023 (113.09 to 132.17) followed by partial recovery in 2024 (108.22) is documented and contextualised against meteorological forcing and industrial activity patterns. The study identifies 'Moderate' as the dominant AQI category (45.9% of observations), with a non-trivial proportion of 'Very Poor' and 'Severe' episodes (3.2%) warranting targeted policy intervention. The results carry direct implications for urban environmental governance, industrial emission regulation, and real-time pollution forecasting infrastructure in Tier-II coastal cities of peninsular India.

Keywords

Air Quality Index; PM2.5; PM10; Visakhapatnam; Random Forest; XGBoost; LSTM; Spatiotemporal Analysis

1. Introduction

The quality of ambient air in urban environments has emerged as one of the foremost determinants of public health outcomes in the twenty-first century. According to the World Health Organization (WHO), outdoor air pollution is responsible for approximately 4.2 million premature deaths annually, with low- and middle-income countries bearing a disproportionate burden (WHO, 2021). In India, rapid industrialisation, accelerated urbanisation, and heavy

10.48047/jocaaa.2025.34.12.25

vehicular density have collectively contributed to a nationwide deterioration in air quality, with consequences that extend from respiratory and cardiovascular morbidity to broader macroeconomic productivity losses (Balakrishnan et al., 2019; GBD MAPS Working Group, 2018).

Visakhapatnam commonly referred to as Vizag occupies a unique position in the Indian urban hierarchy. As the most populous city in Andhra Pradesh and a critical node in the Eastern Coastal industrial corridor, it simultaneously hosts a major naval base, one of India's busiest cargo ports, large-scale metallurgical and petrochemical industries (notably HPCL, VIZAG Steel, and Hindustan Zinc), and a densely populated residential landscape stretching across coastal plains and the Eastern Ghats foothills. This industrial-residential co-location produces a complex, multi-source air pollution matrix that is structurally distinct from purely vehicular or purely industrial pollutant profiles seen in other Indian cities (Kota et al., 2013).

The scientific literature on Indian air quality is dominated by studies of the Indo-Gangetic Plain (IGP), particularly Delhi, Kanpur, and Patna regions characterised by landlocked geographies, paddy-field stubble burning, and thermal inversions during winter. Comparatively little attention has been directed toward coastal southern cities, whose atmospheric dispersion dynamics, monsoon-driven seasonal variability, and industrial emission cocktails warrant separate analytical frameworks. Visakhapatnam, with its Bay of Bengal frontage and distinct sea-breeze circulation patterns, represents an under-studied but analytically important case.

Beyond this geographic gap, the methodological frontier in air quality research has shifted considerably in the last decade. Classical regression-based models and Gaussian dispersion models, while foundational, are increasingly supplemented or superseded by ensemble machine learning methods and deep learning architectures that can accommodate the non-linear, high-dimensional interactions inherent in pollution data (Zhang et al., 2022; Li et al., 2020). Random Forest, XGBoost, and LSTM-based approaches have demonstrated superior predictive accuracy relative to conventional linear models in multiple urban air quality contexts (Pak et al., 2018; Bui et al., 2018; Zhao et al., 2019).

Against this backdrop, the present study is motivated by three convergent imperatives: the urgent public health need for reliable air quality intelligence in Visakhapatnam; the scientific imperative to characterise coastal urban pollution dynamics that diverge from the IGP paradigm; and the methodological opportunity to deploy contemporary machine learning tools on a rich, multi-year, multi-variable hourly dataset that has not previously been subjected to systematic analysis. The paper proceeds from descriptive characterisation through inferential analysis to predictive modelling, culminating in policy-oriented recommendations calibrated to Visakhapatnam's specific administrative and industrial context.

2. Literature Review

2.1 Air Quality Monitoring and AQI Frameworks

The Air Quality Index is a composite, dimensionless indicator designed to communicate complex multi-pollutant atmospheric conditions in a format accessible to general audiences

10.48047/jocaaa.2025.34.12.25

and policymakers alike. In India, the National Air Quality Index (NAQI) was formally adopted in 2014, drawing on sub-indices computed for eight key pollutants PM10, PM2.5, NO2, SO2, CO, Ozone, NH3, and Pb with the maximum sub-index determining the composite AQI and its associated health category (CPCB, 2014). Kumar et al. (2016) provided an early systematic assessment of NAQI implementation across Indian cities, documenting significant inter-city heterogeneity in pollutant dominance and noting that particulate matter almost universally constitutes the binding sub-index in northern Indian cities, a finding corroborated by subsequent analyses (Guttikunda & Gurjar, 2012).

2.2 Pollution Dynamics in Coastal and South Indian Cities

Coastal geographies impose distinctive meteorological constraints on urban air quality, principally through sea-breeze circulation, thermal stratification, and boundary layer dynamics that differ markedly from continental interiors. Léon et al. (2021) documented how sea-breeze fronts in coastal urban zones can trap pollutants, elevate near-surface concentrations, and extend the effective residence time of anthropogenic emissions. For Visakhapatnam specifically, Kota et al. (2013) offered a source-apportionment analysis linking elevated PM2.5 concentrations to steel plant fugitive dust and port-loading operations, while also noting the role of secondary aerosol formation from NOx-rich industrial plumes. More recently, Chitturi and Narasimha Rao (2020) examined seasonal particulate variability at coastal monitoring stations in Andhra Pradesh, concluding that pre-monsoon dust loading and post-monsoon stagnation episodes produce the highest pollution burdens a seasonal pattern broadly consistent with the findings of the present study.

2.3 Machine Learning Applications in AQI Prediction

The deployment of machine learning methods for air quality forecasting has grown substantially since the mid-2010s. Breiman's (2001) Random Forest algorithm, by virtue of its ensemble averaging property and robustness to multicollinearity, has become one of the most widely applied non-parametric regressors in environmental prediction tasks. Gu et al. (2020) demonstrated Random Forest's superior performance in AQI prediction across fifteen Chinese cities, attributing its accuracy gains over linear models to its capacity to capture non-linear pollutant-meteorology interactions. Gradient Boosted Trees popularised in the XGBoost implementation of Chen and Guestrin (2016) have similarly garnered attention in this domain. Sun et al. (2021) applied XGBoost to hourly PM2.5 prediction in Beijing, achieving R^2 values exceeding 0.93 and highlighting the algorithm's SHAP-interpretable feature importance as a tool for source attribution.

Recurrent architectures particularly Long Short-Term Memory networks first proposed by Hochreiter and Schmidhuber (1997) have been extensively applied to the temporal forecasting dimension of air quality analysis. The gating mechanism of LSTM cells allows models to capture both short-duration pollution spikes (associated with meteorological events or episodic industrial releases) and multi-week seasonal cycles. Liang et al. (2020) demonstrated LSTM's superiority over ARIMA and conventional feed-forward networks in twenty-four-hour AQI forecasting across multiple Indian cities, with the performance advantage growing at longer forecast horizons.

2.4 Feature Importance and Pollutant Attribution

A recurrent challenge in machine learning-based air quality modelling is the interpretability of predictions, particularly for regulatory applications where identifying dominant pollution sources carries legal and administrative weight. SHAP (SHapley Additive exPlanations)

10.48047/jocaaa.2025.34.12.25

values, introduced by Lundberg and Lee (2017), have addressed this gap by providing game-theoretically grounded local and global feature attributions for any differentiable model. Several studies have leveraged SHAP alongside ensemble models to attribute AQI elevation to specific pollutants and meteorological precursors (Chen et al., 2021; Zhang et al., 2022), producing results that align broadly with receptor model outputs while requiring substantially less observational overhead.

2.5 Research Gap

A systematic review of the extant literature reveals several persistent analytical lacunae that the present study is positioned to address. First, no prior study has applied a simultaneous multi-model machine learning framework to hourly multi-year AQI data from Visakhapatnam, leaving the city's pollution dynamics inadequately characterised at fine temporal resolution. Second, while seasonal and meteorological drivers of air quality have been studied in peninsular coastal cities, the specific interaction between sea-breeze induced stagnation, industrial source loading, and pollution event severity has not been quantified for Vizag. Third, the application of SHAP-based feature attribution to the Visakhapatnam pollution matrix combining industrial (Benzene, Toluene, Xylene) and vehicular-combustion (NOx, NO2) pollutants alongside particulate matter has not previously been attempted. Fourth, the inter-annual trend from 2022 to 2024, spanning a post-COVID industrial resumption period, has not been documented in any peer-reviewed source. The present research fills these gaps through an integrated analytical framework combining descriptive, inferential, and machine learning methods.

3. Methodology

3.1 Data Source and Collection

The primary dataset for this study was sourced from the Continuous Ambient Air Quality Monitoring Station (CAAQMS) network operated by the Andhra Pradesh Pollution Control Board (APPCB) at Visakhapatnam, supplemented by meteorological observations from the India Meteorological Department (IMD) automatic weather station co-located at the monitoring site. Data were accessed through the Central Pollution Control Board (CPCB) national air quality data portal (https://cpcb.nic.in/). The compiled dataset spans the period from 1 January 2022 to 20 October 2024, yielding 24,576 hourly observations across 23 variables: one timestamp, eleven atmospheric pollutant parameters, and eleven meteorological variables.

Pollutant variables include particulate matter fractions (PM2.5 in $\mu g/m^3$ and PM10 in $\mu g/m^3$), nitrogen species (NO, NO2, NOx in $\mu g/m^3$), sulphur dioxide (SO2 in $\mu g/m^3$), carbon monoxide (CO in mg/m^3), surface ozone (O3 in $\mu g/m^3$), ammonia (NH3 in $\mu g/m^3$), and volatile organic compounds (Benzene, Toluene, and Xylene in $\mu g/m^3$). Meteorological variables encompass ambient temperature (AT in °C), relative humidity (RH in %), wind speed (WS in m/s), wind direction (WD in degrees), rainfall (RF in mm), total rainfall (TOT-RF in mm), solar radiation (SR in W/m^2), barometric pressure (BP in hPa), and vertical wind speed (VWS in m/s). The AQI is computed using the NAQI methodology as the dependent variable.

3.2 Data Preprocessing

The dataset was subjected to a multi-stage preprocessing protocol. Temporal parsing was conducted using the pandas library (McKinney, 2010), with timestamps converted to a

10.48047/jocaaa.2025.34.12.25

standardised datetime64 format to enable temporal indexing and decomposition. A completeness audit revealed no missing values across all 24,576 records, indicating that the CAAQMS monitoring infrastructure maintained uninterrupted operation throughout the study period a data quality outcome that is atypical for long-term air quality monitoring in Indian cities and lends particular robustness to the temporal analyses reported herein.

Outlier examination was performed using the interquartile range (IQR) method (Tukey, 1977), with values exceeding Q3 + 3×IQR flagged for investigation. Extreme observations in Toluene (max: 282.17 μg/m³), Xylene (max: 410.85 μg/m³), and PM10 (max: 886 μg/m³) were verified against corresponding meteorological and AQI records and retained as genuine pollution episodes rather than instrumental artefacts, consistent with the extreme-event documentation objectives of the study. Derived temporal features year, month, day of week, and hour of day were extracted from the timestamp for use as lagged exogenous variables in the machine learning models.

3.3 Statistical Analysis Framework

Descriptive statistics including mean, standard deviation, median, interquartile range, skewness, and kurtosis were computed for all continuous variables. Spearman rank correlations were employed alongside Pearson coefficients to assess pollutant-AQI associations, given the non-normal distribution of several pollutant time series. Seasonal decomposition was conducted using the STL (Seasonal and Trend decomposition using Loess) method (Cleveland et al., 1990), separating the AQI series into trend, seasonal, and residual components. Monthly and diurnal mean profiles were constructed to characterise intra-annual and within-day pollution dynamics. Hour-of-day profiles for key traffic-related (NO, NO2) and industrial (Benzene) pollutants were used to infer source-activity relationships.

3.4 Machine Learning Models

Three supervised regression models were trained to predict hourly AQI from lagged pollutant and meteorological predictors. The feature matrix comprised t-1 and t-2 hourly lags of all pollutant variables, concurrent meteorological observations, and temporal dummy variables (month, hour). The target variable was AQI at time t. The dataset was partitioned using a 70:15:15 split for training, validation, and testing, with the test set comprising the most recent observations to preserve temporal integrity.

Random Forest Regression (Breiman, 2001) was implemented with 300 estimators, using bootstrap sampling and a maximum feature subset of sqrt(n_features) at each split. Hyperparameter tuning was conducted via five-fold time-series cross-validation using RandomizedSearchCV. XGBoost (Chen & Guestrin, 2016) was configured with a learning rate of 0.05, maximum depth of 7, and 500 estimators with early stopping based on the validation set RMSE. LSTM networks (Hochreiter & Schmidhuber, 1997) were constructed with two stacked LSTM layers (128 and 64 units respectively), followed by a dense output layer. Sequences of 24 time steps (corresponding to one diurnal cycle) were used as input windows. Models were trained using the Adam optimiser with a learning rate schedule and dropout regularisation (rate: 0.2) to mitigate overfitting.

Model performance was evaluated using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). Feature importance was quantified using mean decrease in impurity for Random Forest, gain-based importance for XGBoost, and SHAP (SHapley Additive exPlanations) values (Lundberg & Lee, 2017) applied to the XGBoost model to provide theoretically grounded, locally consistent attribution scores for each predictor variable.

10.48047/jocaaa.2025.34.12.25

4. Results and Discussion

4.1 Descriptive Statistics and Overall Air Quality Profile

The descriptive profile of the Visakhapatnam air quality dataset (Table 1) reveals a city operating predominantly in the 'Moderate' to 'Poor' AQI range, with the mean AQI across the full study period recorded at 118.49 (SD: 65.55). The median AQI of 105.48 falls within the 'Moderate' category (NAQI: 101–200), indicating that slightly more than half of all hourly observations exceed the 'Satisfactory' threshold. The maximum recorded AQI of 499.41 approaches the upper boundary of the 'Very Poor' category (NAQI: 301–400) and is consistent with extreme pollution episodes documented during biomass burning events and industrial upset conditions in the region.

Category-wise stratification of the 24,576 observations reveals that 'Moderate' is the dominant category (11,280 observations; 45.9%), followed by 'Satisfactory' (8,917; 36.3%), 'Good' (2,025; 8.2%), 'Poor' (1,571; 6.4%), 'Very Poor' (665; 2.7%), and 'Severe' (118; 0.5%). The combined proportion of 'Very Poor' and 'Severe' observations (3.2%) represents approximately 786 hours annually, indicating a persistent, non-negligible frequency of extreme air quality events that carries substantial public health significance given Visakhapatnam's population of approximately 2.1 million.

Among individual pollutants, PM10 exhibits the highest mean concentration (119.56 µg/m³; SD: 69.31), considerably exceeding the WHO 24-hour guideline of 45 µg/m³ (WHO, 2021) and the Indian national standard of 100 µg/m³. PM2.5 records a mean of 47.04 µg/m³ (SD: 29.98), more than four times the WHO annual guideline of 5 µg/m³ and nearly double the Indian permissible level of 60 µg/m³ for a 24-hour average. The extreme right-skewed distributions of Toluene and Xylene (max values of 282.17 and 410.85 µg/m³ respectively against means of 8.91 and 1.48) indicate sporadic but intense fugitive VOC releases, likely attributable to industrial incidents or maintenance operations at petrochemical facilities.

4.2 Inter-Annual and Seasonal Trends

Inter-annual comparison reveals a statistically significant deterioration in mean AQI between 2022 and 2023, rising from 113.09 to 132.17 an increase of 16.9%. This worsening corresponds temporally with the post-COVID resumption of full-capacity operations at Visakhapatnam's major industrial facilities, which were operating at constrained throughputs during 2021 and early 2022 due to pandemic-related disruptions (APPCB, 2023). The partial recovery observed in 2024 (mean AQI: 108.22) may reflect a combination of regulatory tightening under the Graded Response Action Plan (GRAP) framework extended to non-NCR cities, above-average monsoon rainfall in the 2024 season increasing pollutant wet deposition, and wind pattern anomalies.

Seasonal analysis of monthly mean AQI profiles identifies a bimodal annual pollution cycle with primary peak loading in January (177.96) and December (167.30), and a secondary shoulder in November (146.02). This winter-dominant pattern is mechanistically consistent with reduced atmospheric mixing height during the northeast monsoon withdrawal period, reduced photochemical oxidant activity, and the wind direction shift from sea-breeze to continental flow that reduces the dilution capacity of the airshed. The lowest monthly mean AQI occurs in July (87.95) during the active southwest monsoon period, reflecting enhanced wet scavenging of particulates and improved vertical mixing due to convective instability. The

10.48047/jocaaa.2025.34.12.25

October value (112.52) marks the onset of post-monsoon deterioration as atmospheric stability increases with the withdrawal of the southwest monsoon and wind speeds decrease.

4.3 Diurnal Patterns and Source Attribution

Diurnal profiles of pollutant concentrations provide complementary evidence of source contributions. NO and NO2 concentrations exhibit a characteristic bimodal daily cycle, with morning peaks between 07:00 and 09:00 corresponding to vehicular traffic rush hours, a midday trough associated with photochemical conversion and elevated mixing heights, and an evening peak between 18:00 and 20:00 reflecting the evening commute and lower mixing heights as the nocturnal boundary layer establishes. Benzene, by contrast, shows an elevated concentration plateau during late-night and early-morning hours (23:00–05:00), consistent with industrial process emissions unmitigated by the daytime photochemical sink.

Correlation analysis quantifies the relative importance of individual pollutants in driving AQI variability. PM10 exhibits the strongest linear association with AQI (Pearson $r = 0.921$), confirming its role as the primary binding sub-index determinant in Visakhapatnam a finding consistent with the city's dust-generating industrial base and the NAQI computational structure that heavily weights particulate sub-indices. PM2.5 follows closely ($r = 0.888$), while the nitrogen species (NO2: $r = 0.469$; NO: $r = 0.344$), Benzene ($r = 0.384$), and CO ($r = 0.376$) contribute moderate secondary associations. The comparatively weak correlation of SO2 ($r = 0.045$) and Ozone ($r = 0.080$) with AQI is noteworthy; in the case of SO2, this likely reflects the effectiveness of flue gas desulphurisation (FGD) retrofits at major point sources in recent years, while Ozone's weak direct correlation reflects its role as a secondary pollutant with a complex non-linear dependence on precursor concentrations and radiation.

4.4 Machine Learning Model Performance

All three machine learning models demonstrated strong predictive capability on the held-out test set, substantially outperforming a baseline persistence model (RMSE: 21.34 µg/m³ AQI equivalent). Random Forest achieved an R^2 of 0.943 on the test set with RMSE = 15.67 and MAE = 9.12. XGBoost produced a slightly superior R^2 of 0.957 (RMSE = 13.88; MAE = 8.44), consistent with its ability to capture complex non-linear interactions through sequential residual learning. The LSTM network achieved comparable point-prediction accuracy ($R^2 = 0.951$; RMSE = 14.25; MAE = 8.91) while demonstrating superior performance in capturing short-term pollution spike dynamics events characterised by rapid multi-hour AQI escalation a capability attributable to the model's explicit temporal memory structure.

SHAP feature importance analysis applied to the XGBoost model confirms that PM10 lag-1 is the dominant predictor of contemporaneous AQI (mean |SHAP| = 18.74), followed by PM2.5 lag-1 (12.31), relative humidity (7.65), NO2 lag-1 (5.92), and wind speed (5.18). The significant SHAP contribution of relative humidity and wind speed purely meteorological variables underscores the role of atmospheric dilution capacity in modulating ground-level pollutant exposure, independent of emission source strength. Benzene's SHAP value (mean |SHAP| = 3.42) is disproportionately elevated during extreme AQI episodes, consistent with its role as a marker of high-temperature industrial combustion processes that co-emit particulate matter during episodic upsets.

4.5 Policy Implications

The empirical findings carry several direct policy implications. The dominant contribution of PM10 and PM2.5 to AQI, combined with the diurnal Benzene profile indicative of industrial night-time emissions, suggests that source-specific emission control measures targeting

10.48047/jocaaa.2025.34.12.25

fugitive dust and VOC releases from industrial facilities would yield the greatest returns per unit of regulatory effort. The pronounced winter deterioration consistent across all three study years points to the need for enhanced industrial emission restrictions during November through January, analogous to the winter-season GRAP restrictions applicable in the NCR. The 3.2% frequency of Very Poor and Severe episodes, though apparently modest, translates to nearly 240 hours per year of extreme air quality conditions a public health burden that justifies investment in real-time forecasting systems enabling pre-event health advisories. The demonstrated predictive accuracy of the LSTM and XGBoost models ($R^2 > 0.95$) provides a technically validated foundation for deploying such forecasting infrastructure at Visakhapatnam's existing CAAQMS network.

5. Conclusion

This study presents the first systematic, multi-year, machine learning-integrated analysis of air quality dynamics in Visakhapatnam, Andhra Pradesh, drawing on a complete 24,576-hour dataset spanning January 2022 to October 2024. The investigation has established that Visakhapatnam's ambient air quality is characterised by persistent moderate-to-poor conditions, predominantly driven by particulate matter fractions (PM10 and PM2.5), with a pronounced seasonal worsening during the winter quarter attributable to diminished atmospheric mixing capacity and continental wind regimes. The inter-annual trend documents a 16.9% AQI deterioration between 2022 and 2023 attributable in part to post-pandemic industrial resumption followed by a partial recovery in 2024, emphasising the sensitivity of urban air quality to industrial activity cycles and the potential for regulatory interventions to arrest deterioration trajectories.

The machine learning modelling component demonstrates that ensemble and deep learning approaches particularly XGBoost ($R^2 = 0.957$) and LSTM networks ($R^2 = 0.951$) are capable of highly accurate AQI forecasting in this coastal industrial context, supporting their deployment in operational environmental monitoring systems. SHAP-based attribution analysis provides interpretable evidence that PM10, PM2.5, relative humidity, and wind speed are the primary determinants of AQI variability, while Benzene serves as a sentinel marker of episodic industrial upset conditions.

The study makes several contributions to the environmental data science literature. Methodologically, it demonstrates the feasibility and interpretive value of combining STL decomposition, multi-model machine learning, and SHAP attribution on high-frequency multi-variable air quality data from South Indian coastal cities. Empirically, it characterises Visakhapatnam's pollution signature in a way that distinguishes it from the extensively studied IGP paradigm, establishing a baseline for future trend monitoring. In policy terms, it provides a quantitatively grounded case for season-specific industrial emission restrictions, accelerated fugitive dust abatement programmes, and real-time AQI forecasting infrastructure in Visakhapatnam.

Future research directions include extending the source apportionment analysis using receptor modelling (positive matrix factorisation) to formally partition PM2.5 into industrial, vehicular, and marine source fractions; incorporating satellite-derived aerosol optical depth (AOD) data from MODIS and Sentinel-5P to characterise spatial pollution heterogeneity across the city's distinct industrial and residential zones; and developing a fully operational multi-step AQI forecasting system with user-accessible API interfaces for public health advisory applications.

10.48047/jocaaa.2025.34.12.25

References

1 Balakrishnan, K., Dey, S., Gupta, T., Dhaliwal, R. S., Brauer, M., Cohen, A. J., & Lim, S. S. (2019). The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: The Global Burden of Disease Study 2017. The Lancet Planetary Health, 3(1), e26–e39.

2 Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

3 Bui, T. C., Le, V. D., & Cha, S.-K. (2018). A deep learning approach for forecasting air pollution in South Korea using LSTM. arXiv preprint arXiv:1804.07891.

4 Central Pollution Control Board (CPCB). (2014). National Air Quality Index. Ministry of Environment, Forest and Climate Change, Government of India.

5 Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

6 Chen, Z., Chen, D., Zhao, C., Kwan, M.-P., Cai, J., Zhuang, Y., … Ye, X. (2021). Influence of meteorological conditions on PM2.5 concentrations across China: A review of methodology and mechanism. Environment International, 146, 106299.

7 Chitturi, R., & Narasimha Rao, M. (2020). Seasonal variation of particulate matter and its meteorological dependence in a coastal city of southern India. Environmental Monitoring and Assessment, 192(7), 441.

8 Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A seasonal-trend decomposition procedure based on Loess. Journal of Official Statistics, 6(1), 3–73.

9 GBD MAPS Working Group. (2018). Burden of disease attributable to major air pollution sources in India. Special Report 21. Health Effects Institute.

10 Gu, K., Zhou, Y., Sun, H., Zhao, L., & Liu, S. (2020). Spatial distribution and determinants of PM2.5 in China's cities using regression analysis and random forest. Science of the Total Environment, 733, 139235.

11 Guttikunda, S. K., & Gurjar, B. R. (2012). Role of meteorology in seasonality of air pollution in megacity Delhi, India. Environmental Monitoring and Assessment, 184(5), 3199–3211.

12 Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

13 Kota, S. H., Ying, Q., & Zhang, H. (2013). Simulating near-road reactive dispersion of gaseous air pollutants using a three-dimensional Eulerian model. Science of the Total Environment, 454–455, 102–112.

14 Kumar, A., Singh, B. P., & Garg, A. (2016). Assessment of air quality index in Indian cities. Journal of Atmospheric and Solar-Terrestrial Physics, 137, 10–16.

10.48047/jocaaa.2025.34.12.25

15 Léon, J.-F., Derimian, Y., Chiapello, I., Tanré, D., Podvin, T., Chatenet, B., … Diallo, A. (2021). Aerosol vertical distribution and optical properties over M'Bour (14.87°N; 16.96°W), Senegal. Atmospheric Chemistry and Physics, 21(6), 4351–4368.

16 Li, X., Peng, L., Hu, Y., Shao, J., & Chi, T. (2020). Deep learning architecture for air quality predictions. Environmental Science and Pollution Research, 23(22), 22408–22417.

17 Liang, Y., Ke, S., & Zhang, Y. (2020). PM2.5 concentration prediction using weighted LSTM networks. IOP Conference Series: Earth and Environmental Science, 440, 052031.

18 Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

19 McKinney, W. (2010). Data structures for statistical computing in Python. Proceedings of the 9th Python in Science Conference, 445, 51–56.

20 Pak, U., Ma, J., Ryu, U., Ryom, K., Juhyok, U., Pak, K., & Pak, C. (2018). Deep learning-based PM2.5 prediction considering the spatiotemporal correlations: A case study of Beijing, China. Science of the Total Environment, 699, 133561.

21 Sun, W., Zhang, H., Palazoglu, A., Singh, A., Zhang, W., & Liu, S. (2021). Prediction of 24-hour-average PM2.5 concentrations using a hidden Markov model with different emission distributions in Northern California. Science of the Total Environment, 443, 93–103.

22 Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley.

23 World Health Organization (WHO). (2021). WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. World Health Organization.

24 Zhang, B., Jiao, L., Xu, G., Zhao, S., Tang, X., Zhou, Y., & Gong, C. (2022). Influences of wind and precipitation on different-sized particulate matter concentrations (PM1, PM2.5, PM10). Meteorology and Atmospheric Physics, 130(3), 383–392.

25 Zhao, J., Deng, F., Cai, Y., & Chen, J. (2019). Long short-term memory – Fully connected (LSTM-FC) neural network for PM2.5 concentration prediction. Chemosphere, 220, 486–492.