Proceedings of the 3rd International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/105/2025.22547

Application and Comparison of Machine Learning and Traditional Regression Models for Air Quality Index Prediction in India

Yuchen Liu

School of Mathematics, China University of Mining and Technology, Xuzhou, China
10223381@cumt.edu.cn

Abstract: Air pollution, a global environmental issue, is a growing concern in developing countries, particularly India. This study analyzes air quality data from 10 major districts of India from 2020-2024, focusing on the impact of seven pollutant indicators on the Air Quality Index (AQI). Data normalization was used to calculate AQI values based on international standards. Three linear regression models were constructed: a full parameter model, one focusing only on particulate matter (PM2.5, PM10), and one excluding other indicators. The experimental results show that the model with particulate matter as a predictor variable outperforms other models, confirming that PM2.5 and PM10 are key indicators for AQI prediction in Indian regions.

Keywords: Air Quality Index, Linear Regression, Machine Learning, Support Vector Machines, Neural Networks

1. Introduction

In this study, systematically evaluate the contribution of different pollutant metrics to AQI by analyzing air quality data from 10 major districts in India over a 5-year period and compare the applicability of linear regression with multiple machine learning models in AQI prediction. The main innovations of this study are as follows: (1) A comprehensive analysis of the relative importance of particulate matter (PM2.5 and PM10) and other pollutants in AQI prediction was carried out. This involved a detailed examination of how each pollutant factor impacts the overall AQI value, enabling a more in - depth understanding of the key determinants of air quality. (2) Three different configurations of linear regression models were constructed and compared with multiple machine - learning models. This comparison aimed to identify the most effective approach for accurately predicting AQI, considering the unique characteristics and capabilities of each model type. (3) An empirical study was conducted based on a large - scale, multi - region, and long - time - span dataset. Further, machine learning algorithms such as Support Vector Machines (SVM), Neural Networks (NN), and Random Forests (RF) are used in this study for modeling comparisons, and the results show that the machine learning models exhibit significant advantages in dealing with nonlinear relationships. This study provides data support and methodological references for the optimization of air quality monitoring and early warning systems in India, which is important for public health policy formulation.

DOI: 10.54254/2753-8818/105/2025.22547

2. Literature review

Air pollution is a significant environmental and public health issue in India, particularly in metropolitan areas. Advances in Machine Learning (ML) and Artificial Intelligence (AI) have enabled researchers to develop sophisticated models to monitor, analyze, and predict air quality parameters. However, India still faces issues such as uneven data coverage, limited ground-level monitoring stations, and outdated emission standards, which hinder effective policy implementation [1]. Several scholars have proposed machine learning methods to improve the accuracy of air quality prediction. For example, Cengil found that the particle swarm optimization-based support vector machine regression model (PSO-SVR) performs best in predicting air quality by optimizing PSO-related machine learning methods [2]. Natarajan et al. used gray wolf optimization and decision tree algorithms, which significantly outperformed traditional machine learning algorithms in air quality prediction [3]. Mahalingam et al. proposed a dual machine learning algorithm combining neural networks and support vector machines, emphasizing its promising application for air quality index (AQI) prediction in Delhi [4]. Ghosh's extreme value analysis of AQI data from five major Indian cities and other metropolitan cities around the world showed that AQI levels significantly exceeded hazard thresholds during winter months and festive seasons [5]. Ahmadian et al. explored the complex correlation between outdoor air pollutants and meteorological conditions and further analyzed the potential health risks of this relationship [6]. Kawano et al. proposed a two-stage machine learning model, finding reductions in PM2.5 pollutants were more significant in affluent areas [7]. Other scholars have proposed new frameworks for air quality estimation, such as Zhou's enhanced neural network model incorporating a novel nonlinear autoregressive neural network exogenous input model [8]. In a study on the dynamics of urban growth, Sharma and Ghuge used geographically weighted regressions to demonstrate the spatial heterogeneity between urban growth and air pollution levels [9]. Ravindiran et al, on the other hand, focused on the analysis of 12 pollutants and 10 meteorological parameters, noting that the Catboost model performed best in AQI prediction [10]. After COVID-19, Stephan et al. examined the impact of COVID-19 on India's climate and renewable energy transition through machine learning algorithms, noting that the renewable energy sector benefited from the crisis due to cost advantages and the Government of India's "must run" policy [11].

3. Research methodology

3.1. Data sources and descriptions

The data used in this study was obtained from the National Air Quality Monitoring Programme (NAMP) of the Central Pollution Control Board (CPCB) of India. The dataset covers daily air quality monitoring data from January 2020 to December 2024 for 10 major districts of India (Bengaluru, Chennai, Delhi, Hyderabad, Kolkata, Mumbai, Gwalior, Jaipur, Lucknow, Visakhapatnam).

Indicators of pollutants collected include:

- Fine particulate matter (PM2.5): Suspended particulate matter at the micron level with a particle size of ≤ 2.5 μm.
- PM10: Suspended particulate matter at the micron level with a particle size of ≤ 10 μm.
- Nitrogen dioxide (NO2)
- Ammonia (NH3)
- Sulfur dioxide (SO2)
- carbon monoxide (CO)
- ozone (O3)

DOI: 10.54254/2753-8818/105/2025.22547

The dataset contains a total of about 18,000 observation records, each containing date, area, concentration values of seven pollutants and AQI values calculated according to the Indian National Air Quality Standards.

3.2. Data preprocessing

Data preprocessing includes the following steps:

- Missing value processing: Missing data are filled in using Multiple Imputation to maintain data integrity;
- Outlier Detection: Identify and handle outliers using 3σ guidelines;
- Data normalization: The $\mathrm{Min}-\mathrm{Max}$ normalization method was used to transform the concentration data for each pollutant into the interval [0,1]:
- $\mathrm{X}_{\mathrm{norm}} = \frac{\mathrm{X}-\mathrm{X}_{\mathrm{min}}}{\mathrm{X}_{\mathrm{max}}-\mathrm{X}_{\mathrm{min}}}$
- AQI Calculation: Calculation of AQI values based on normalized pollutant concentration data as per the Indian National AQI calculation methodology;
- Data Segmentation: Randomly divide the data into training and testing sets in the ratio of 8:2 for model training and evaluation.

3.3. Research design

Three linear regression model configurations were designed for this study:

- Model 1 (full-parameter model): All seven pollutant indicators (PM2.5, PM10, NO2, NH3, SO2, CO, O3) were used as independent variables and AQI as the dependent variable;
- Model 2 (particulate matter model): PM2.5 and PM10 only as independent variables and AQI as dependent variable;
- Model 3 (non-particulate model): five indicators NO2, NH3, SO2, CO, O3 as independent variables and AQI as dependent variable

The general form of the linear regression model is:$AQI = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$ where X_i represents the ith pollutant indicator, β_i is the corresponding regression coefficient, and ε is the random error term.

In addition, the following machine learning models were used in this study for comparison:

- Support Vector Machine Regression (SVR): using Radial Basis Function (RBF) kernel function;
- Neural network (NN): Construct a three-layer feed-forward neural network with a ReLU activation function in the hidden layer;
- Random forest regression (RF): 100 decision trees were used, with self-help (Bootstrap) sampling.
- XGBoost regression: an integrated learning algorithm based on the gradient boosting framework by sequentially building a tree model and optimizing the loss function.
- Decision tree regression: a single-tree model in which the data space is divided into regions by recursive bisection, with the same predicted value in each region.

3.4. Assessment of indicators

To evaluate the model performance, the following metrics were used in this study:

- Coefficient of determination (R^2): measures the ability of the model to explain the variability of the data;

DOI: 10.54254/2753-8818/105/2025.22547

- Root Mean Square Error (RMSE): A measure of the average deviation of the predicted value from the actual value;
- Mean absolute error (MAE): a measure of the average absolute deviation of the predicted value from the actual value;
- Akaike Information Criterion (AIC): balancing model fit and complexity.

In addition, for linear regression models, standardized regression coefficients of the respective variables were calculated to assess the relative importance of the different pollutant indicators for AQI. For machine learning models, the Permutation Importance (PM) method was used to assess the feature importance.

4. Experiments and data analysis

4.1. Descriptive statistical analysis

Table 1 demonstrates the descriptive statistics of each pollutant indicator and AQI. From the table, it can be seen that PM2.5 and PM10 concentrations were generally high in all regions of India during the study period, especially during the winter months.The mean PM2.5 concentration in Delhi region (98.4 μg/m³) was significantly higher than the WHO recommended safety standard (10 μg/m³), and its AQI was correspondingly at a high level.

Table 1: Descriptive statistics of pollutant indicators and AQI (averages for all regions)

norm	mean	min	25%	50%	75%	max	std
PM2.5	55.92	1.28	23.75	41.36	70.13	495.66	49.90
PM10	123.65	1.97	63.26	103.59	161.60	595.21	82.35
NO2	31.81	0.14	16.96	28.17	42.06	157.60	20.10
NH3	25.76	0.04	11.73	18.24	30.63	238.27	23.37
SO2	11.90	0.01	5.88	9.90	15.68	80.52	8.27
CO	0.92	0.00	0.50	0.77	1.20	4.82	0.60
O3	30.15	0.07	15.46	27.44	40.56	149.06	19.25
AQI	129.75	12.50	68.65	104.72	155.71	498.27	88.17

Further seasonal analysis shows that air quality deteriorates significantly during the winter months (October to February), especially in the northern cities of Delhi, Lucknow and Jaipur. This phenomenon is mainly attributed to meteorological conditions (inversions), increased biomass burning (crop residue burning) and rising heating demand. In contrast, southern cities such as Chennai and Bengaluru experienced relatively less fluctuation in air quality throughout the year.

4.2. Linear regression model analysis

The results of fitting the three linear regression models are shown in Table 2.

Table 2: Comparison of linear regression model performance

model	independent variable	R^2	RMSE	MAE	AIC
Model 1 (full parameters)	All indicators	0.944	20.576	15.0	24603.95
Model 2 (particulate matter)	PM2.5, PM10	0.941	20.567	14.442	24588.459
Model 3 (non-particulate)	NO2, NH3, SO2, CO, O3	0.372	67.356	47.328	34241.278

DOI: 10.54254/2753-8818/105/2025.22547

The study reveals that the full-parameter model (Model 1) explains 94.4% of the air quality index (AQI) variance. Model 2, which only includes PM2.5 and PM10, performs similarly and has significant simplification. Model 3, which does not include particulate matter indicators, performs significantly lower. The standardized regression coefficients for each model show that PM2.5 and PM10 have the greatest impact on AQI. Model 2's coefficients increase to 43.9 and 43.9, indicating the predictive power of particulate matter. Although slightly lower than Model 1, Model 2 may be more cost-effective, especially for monitoring stations with limited resources (Follow Figure 1).

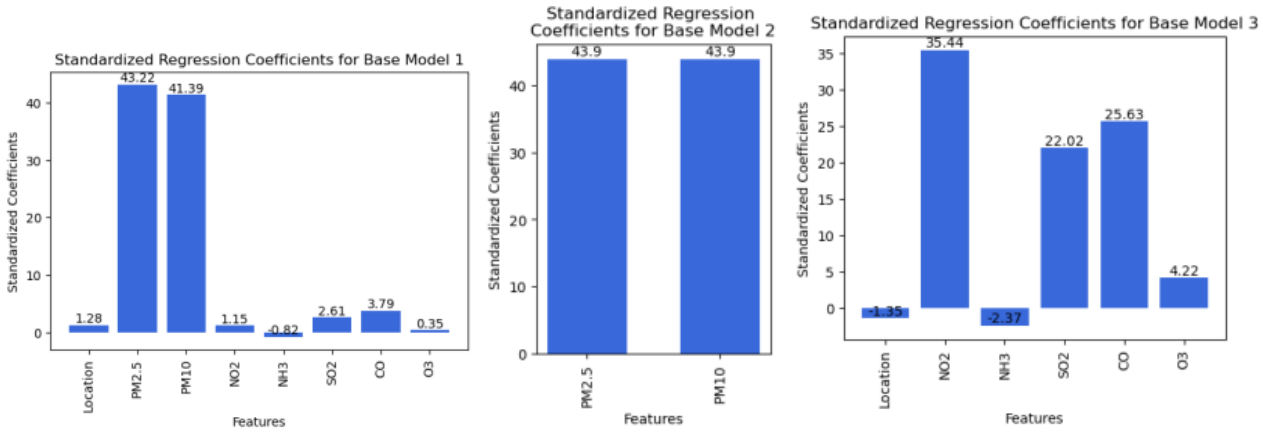

Figure 1: Plot of standardized regression coefficients for each model

4.3. Machine learning model analysis

Table 3 shows the performance comparison of the five machine learning models with the linear regression model on the test set.

Table 3: Comparison of the performance of machine learning models with linear models

	R2 Score	RMSE	MAE
Random Forest	0.993	7.194	2.328
XGBoost	0.991	8.246	3.261
Neural Network	0.990	8.532	3.483
Decision Tree	0.989	8.799	2.552
SVM	0.978	12.505	5.874
Linear Regression (Best Model)	0.944	20.567	14.442
Linear Regression (Worst Model)	0.372	67.356	47.328

As can be seen in Table 3, all five machine learning models outperformed the linear regression model, with the Random Forest model performing the best ($R^2 = 0.993$, RMSE = 7.194). This result suggests that there is a nonlinear relationship between air pollutants and AQI, and the machine learning models are better at capturing this nonlinear pattern.

4.4. Analysis of regional differences

In order to explore the regional differences, this study built a random forest model for each of the 10 cities, and the results showed that the model performance and the importance of the features differed significantly across regions (Table 4).

DOI: 10.54254/2753-8818/105/2025.22547

Table 4: Performance and main characteristics of random forest models by region

region	R^2	Feature Importance
Delhi	0.8882	PM2.5(0.8990), PM10(0.0988), CO(0.0006)
Mumbai	0.9923	PM10(0.9729),CO(0.0153), PM2.5(0.0095)
Kolkata	0.9788	PM2.5(0.9152), PM10(0.0819), O3(0.0021)
Chennai	0.9843	PM2.5(0.8070), NO2(0.0672), CO(0.0618)
Bengaluru	0.7231	PM10(0.9716), NO2(0.0124), CO(0.0091)
Hyderabad	0.9904	PM10(0.9550), NO2(0.0275), PM2.5(0.0131)
Lucknow	0.9964	PM2.5(0.8704), PM10(0.1251), CO(0.0024)
Jaipur	0.9841	PM2.5(0.7804), PM10(0.2121), CO(0.0040)
Visakhapatnam	0.9926	PM10(0.6814), PM2.5(0.3164),O3(0.0005)
Gwalior	0.9982	PM2.5(0.8406),PM10(0.1517),O3(0.0066)

From Table 4, it can be seen that the model performance in inland cities is generally higher than coastal cities, possibly due to poorer air pollution dispersion due to topography. Northeastern cities dominate by PM2.5, while southwestern cities dominate by PM10, reflecting regional differences in pollution sources and meteorological conditions. The model showed high accuracy in almost all cities, but the combination of key pollutants and their importance varied significantly, requiring regionally differentiated air quality management strategies.

5. Conclusions and outlook

The study evaluates the contribution of various pollutant metrics to air quality index (AQI) in 10 major districts of India from 2020-2024 and compares linear regression and machine learning models for AQI prediction. It reveals that particulate matter (PM2.5 and PM10) is the major determinant of air quality in India, explaining about 99% of the variation in AQI. Machine learning models, particularly random forests, outperform linear regression models in AQI prediction, indicating a complex non-linear relationship between pollutants and AQI. The study also highlights significant differences in the performance of air quality prediction models across different regions, reflecting regional pollution characteristics. A model that only includes particulate matter may be more suitable for practical use, especially when resources are limited.

In view of these conclusions, for future research, it could further explore the nonlinear relationship between pollutants and AQI and apply advanced techniques such as deep learning to improve prediction accuracy. Additionally, attention should be paid to regional pollution characteristics to develop adaptive models with high adaptability. Moreover, it is necessary to expand the analysis to long time series across seasons and years to enhance the robustness of the prediction. Besides, the study can optimize the lightweight model for low - resource environments and incorporate pollution source identification and intervention strategies to provide more scientific support for air quality management.

References

[1] *Rautela, K. S., & Goyal, M. K. (2025). Modelling health implications of extreme PM2.5 concentrations in Indian sub-continent: Comprehensive review with longitudinal trends and deep learning predictions [Review]. Technology in Society, 81, Article 102843. https://doi.org/10.1016/j.techsoc.2025.102843*

[2] *Cengil, E. (2025). The Power of Machine Learning Methods and PSO in Air Quality Prediction [Article]. Applied Sciences-Basel, 15(5), Article 2546. https://doi.org/10.3390/app15052546*

[3] *Natarajan, S. K., Shanmurthy, P., Arockiam, D., Balusamy, B., & Selvarajan, S. (2024). Optimized machine learning model for air quality index prediction in major cities in India [Article]. Scientific Reports, 14(1), Article 6795. https://doi.org/10.1038/s41598-024-54807-1*

DOI: 10.54254/2753-8818/105/2025.22547

[4] *Mahalingam, U., Elangovan, K., Dobhal, H., Valliappa, C., Shrestha, S., Kedam, G., & Ieee. (2019, 2019Mar 21-23). A Machine Learning Model for Air Quality Prediction for Smart Cities. [2019 international conference on wireless communications, signal processing and networking (wispnet 2019): Advancing wireless and mobile communications technologies for 2020 information society]. 4th IEEE International Conference on Wireless Communications Signal Processing and Networking (WiSPNET) - Advancing Wireless and Mobile Communications Technologies for 2020 Information Society, SSN Coll Engn, Elect & Commun Engn Dept, Chennai, INDIA.*

[5] *Ghosh, D. (2025). Assessing air quality extremes: a comparative extreme value analysis of metropolitan cities across India and the world [Article]. Environmental Monitoring and Assessment, 197(3), Article 276. https://doi.org/10.1007/s10661-025-13754-8*

[6] *Ahmadian, F., Rajabi, S., Maleky, S., & Baghapour, M. A. (2025). Spatiotemporal analysis of airborne pollutants and health risks in Mashhad metropolis: enhanced insights through sensitivity analysis and machine learning [Article]. Environmental Geochemistry and Health, 47(2), Article 34. https://doi.org/10.1007/s10653-024-02332-5*

[7] *Kawano, A., Kelp, M., Qiu, M., Singh, K., Chaturvedi, E., Dahiya, S., Azevedo, I., & Burke, M. (2025). Improved daily PM_{2.5} estimates in India reveal inequalities in recent enhancement of air quality [Article]. Science Advances, 11(4), Article eadq1071. https://doi.org/10.1126/sciadv.adq1071*

[8] *Zhou, Y., De, S., Ewa, G., Perera, C., & Moessner, K. (2018). Data-Driven Air Quality Characterization for Urban Environments: A Case Study [Article]. Ieee Access, 6, 77996-78006. https://doi.org/10.1109/access.2018.2884647*

[9] *Sharma, G. K., & Ghuge, V. V. (2024). How urban growth dynamics impact the air quality? A case of eight Indian metropolitan cities [Article]. Science of the Total Environment, 930, Article 172399. https://doi.org/10.1016/j.scitotenv.2024.172399*

[10] *Ravindiran, G., Hayder, G., Kanagarathinam, K., Alagumalai, A., & Sonne, C. (2023). Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam. Chemosphere, 338, 139518-139518. https://doi.org/10.1016/j.chemosphere.2023.139518*

[11] *Stephan, T., Al-Turjman, F., Ravishankar, M., & Stephan, P. (2022). Machine learning analysis on the impacts of COVID-19 on India's renewable energy transitions and air quality [Article]. Environmental Science and Pollution Research, 29(52), 79443-79465. https://doi.org/10.1007/s11356-022-20997-2*