International Journal of Advance Interdisciplinary Research
Vol. 2, Special Issue 1, (Jan-March) 2026, pp. 197-215, e-ISSN: 3107-913X
DOI: https://doi.org/10.66095/ijair.2026.v2.S1.20

Machine Learning-Based Modeling and Prediction of Urban Air Pollution: A Case Study from Lucknow, India

Shariq Nawaj Khan and Upasana Yadav
Amity School of Applied Sciences, Amity University, Lucknow
Email: uyadav@lko.amity.edu

Abstract: One of the most urgent environmental and public health issues of the twenty-first century is urban air pollution, especially in rapidly urbanizing cities in developing nations like India. The complicated and nonlinear behavior of air contaminants typically limits the efficiency of conventional statistical modeling methodologies. To address these issues, this work uses ground-based monitoring data from Lucknow, India, to examine the use of machine learning (ML) approaches for modeling and forecasting urban air quality. From June to September 2025, several air quality monitoring sites provided the city-level aggregated concentrations of six major pollutants: PM_{2^1}, NO_2, O_3, NO, SO_3, and CO. Two frequently used ML models, Extreme Gradient Boosting (XGBoost) and Artificial Neural Networks (ANNs), were developed and tested using an 80:20 train–test split combined with 10-fold cross-validation. The coefficient of determination (R2) and mean squared error (MSE) were used to evaluate the model's performance. The findings suggest that pollutants with relatively stable emission patterns, like PM_{2^1} and NO_2, have significant predictive capability, but pollutants with highly variable and localized emission features, like CO and NO, have weaker prediction. The results highlight the fact that no single model is always the best and that machine learning effectiveness varies depending on the pollutant. All things considered, this study shows how ML-based methods can enhance urban air quality assessment and assist evidence-based environmental planning in mid-sized Indian cities.

Keywords: Air Pollution, Machine Learning, XGBoost, Artificial Neural Networks, Urban Air Quality

1. Introduction

In areas of the world that are quickly urbanizing, urban air pollution has emerged as a major environmental and public health concern. Particularly in emerging nations like India, increased automobile traffic, industrial activity, building dust, and population growth have severely deteriorated air quality. Premature mortality, cardiovascular disease, and respiratory ailments have all been firmly linked to prolonged exposure to ambient air pollution (WHO, 2021).

International Journal of Advance Interdisciplinary Research
Vol. 2, Special Issue 1, (Jan-March) 2026, pp. 197-215, e-ISSN: 3107-913X
DOI: https://doi.org/10.66095/ijair.2026.v2.S1.20

Indian cities routinely report amounts of particulate matter and gaseous pollutants exceeding national ambient air quality limits. A complicated mixture of primary emissions and secondary atmospheric reactions is the source of pollutants like PM_4[1], NO_2, O_3, NO, SO_3, and CO. Because emission intensity, chemical changes, and temporal variability all affect their behavior, it is challenging to accurately model them using traditional methods.

Traditional statistical and deterministic air quality models rely on linear assumptions or simplified atmospheric processes, which often fail to reflect nonlinear interactions inherent in urban pollution systems (Kumar et al., 2024). On the other hand, without explicit physical parameterization, machine learning (ML) models may automatically learn intricate, nonlinear interactions from data, providing enhanced prediction power (Li et al., 2023).

Recent research has shown that machine learning algorithms are useful for predicting air pollution, especially nitrogen oxides and particulate matter. However, there are few studies that use multi-station aggregated datasets to concentrate on mid-sized Indian cities, and model performance varies greatly across contaminants. Designing reliable forecasting and management techniques requires an understanding of pollutant-specific predictability (Maji et al., 2024).

Two popular machine learning algorithms, XGBoost and Artificial Neural Networks (ANNs), are used to multi-pollutant air quality data from Lucknow, India, in this study. The primary purpose is to evaluate model performance across six important contaminants and to identify strengths, limits, and future directions for ML-based urban air quality prediction.

2. Review of Literature

The use of machine learning (ML) approaches in air quality modeling has grown significantly over the last ten years due to increases in processing power and the quick expansion of environmental monitoring networks. The intricate, nonlinear interactions controlling urban air pollution are frequently difficult for traditional statistical and deterministic models to capture, especially in settings impacted by a variety of emission sources and changing meteorological conditions. Machine learning techniques have become viable substitutes in this situation since they can extract complex patterns from data without the need for explicit physical parameterization.

Early research identified artificial neural networks (ANNs) as powerful tools for modeling nonlinear environmental systems due to their universal approximation capabilities (Hornik et

al., 1989). These investigations showed that connections between pollutant concentrations and influencing factors that were poorly captured by linear regression models might be effectively captured by ANNs. Later uses of ANN-based techniques for forecasting urban air pollution shown increased prediction accuracy, particularly for datasets with high temporal variability and nonlinearity (Chelani & Devotta, 2007). Neural networks consequently emerged as one of the first and most used machine learning methods in the study of air quality**.**

In recent years, tree-based ensemble learning techniques have become more and more popular alongside neural networks. Because of their resilience to noisy data, capacity to manage intricate feature interactions, and comparatively good interpretability when compared to deep learning techniques, these models are especially prized. The development of XGBoost by Chen and Guestrin (2016), which improved conventional gradient boosting by adding regularization procedures to lessen overfitting while preserving computational efficiency, was a significant turning point in this field. These benefits have led to XGBoost's widespread use in air quality modeling, where it has shown excellent prediction accuracy on structured datasets obtained from ground-based monitoring stations (Wang et al., 2023).

The effectiveness of ML models is highly sensitive on pollutants, according to recent studies. Research undertaken in Indian urban areas reveals that particulate matter (PM_{10}) and nitrogen dioxide (NO_2) are often more predictable using ML approaches due to their relatively consistent emission patterns and better temporal persistence. On the other hand, it is more challenging to effectively represent gaseous pollutants like carbon monoxide (CO) and sulfur dioxide (SO_2) since they frequently show erratic spikes and confined emission behavior (Sharma et al., 2023; Singh et al., 2024). The increased autocorrelation found in particulate matter concentrations has been recognized as a crucial element contributing to enhanced ML-based forecasting ability (Patel et al., 2025).

A increasing interest in deep learning and hybrid modeling architectures can be seen in more recent studies. To better capture long-term temporal correlations and intricate spatiotemporal patterns in air quality data, models like CNN–LSTM combinations and transformer-based frameworks have been investigated. These methods have shown better performance than traditional machine learning models when huge datasets and meteorological variables are available (Zhang et al., 2023; Zhou et al., 2023). However, these techniques are

computationally demanding and necessitate large amounts of high-quality data, which frequently restricts their usefulness in many developing and mid-sized urban areas.

The significant dependence on meteorological and supplementary datasets is another significant drawback noted in recent research. Although these inputs can greatly increase prediction accuracy, their quality and availability vary each city. Because of this, pollutant-only modeling techniques are still useful, especially in areas with inconsistent or poor supplementary data (Nguyen et al., 2023). Additionally, a large number of current research concentrate on individual contaminants or aggregate air quality indices, providing little understanding of the behavior of pollutant-specific models.

There are still gaps in comparative multi-pollutant studies that assess the relative efficacy of various machine learning models utilizing aggregated multi-station datasets, particularly for mid-sized Indian cities, despite notable advancements. There is a need for systematic assessments that explore how different contaminants respond to specific ML architectures under real-world data restrictions. The current study fills in these gaps and adds to the expanding body of literature on data-driven urban air quality modeling by carefully comparing XGBoost and ANN models across multiple contaminants using city-level aggregated data from Lucknow.

3. Problem Statement

There is significant temporal and spatial variation in urban air pollution, and no single modeling technique works consistently well for all pollutants. Although machine learning methods have demonstrated potential, their efficacy varies according to data availability, emission patterns, and pollutant behavior. Using multi-station aggregated datasets, nothing is known about pollutant-specific machine learning performance in mid-sized Indian cities. This study aims to determine which contaminants need more sophisticated or hybrid approaches and which may be accurately modeled using ML techniques.

4. Methodology

4.1 Study Area and Data Source

Lucknow, the Indian state capital of Uttar Pradesh, is the study area. Six continuous monitoring stations Gomti Nagar, B. R. Ambedkar University, Talkatora, Central School,

International Journal of Advance Interdisciplinary Research
Vol. 2, Special Issue 1, (Jan-March) 2026, pp. 197-215, e-ISSN: 3107-913X
DOI: https://doi.org/10.66095/ijair.2026.v2.S1.20

Kukrail Picnic Spot, and Lalbagh were used to gather data on air quality. The Open AQ platform, which compiles measurements from CPCB and UPPCB, provided the data.

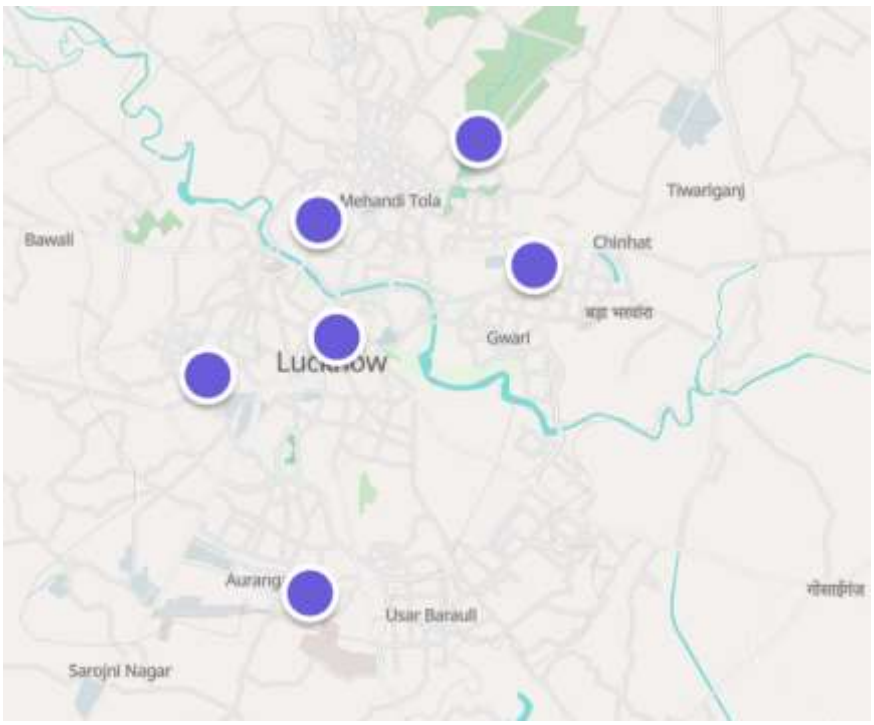
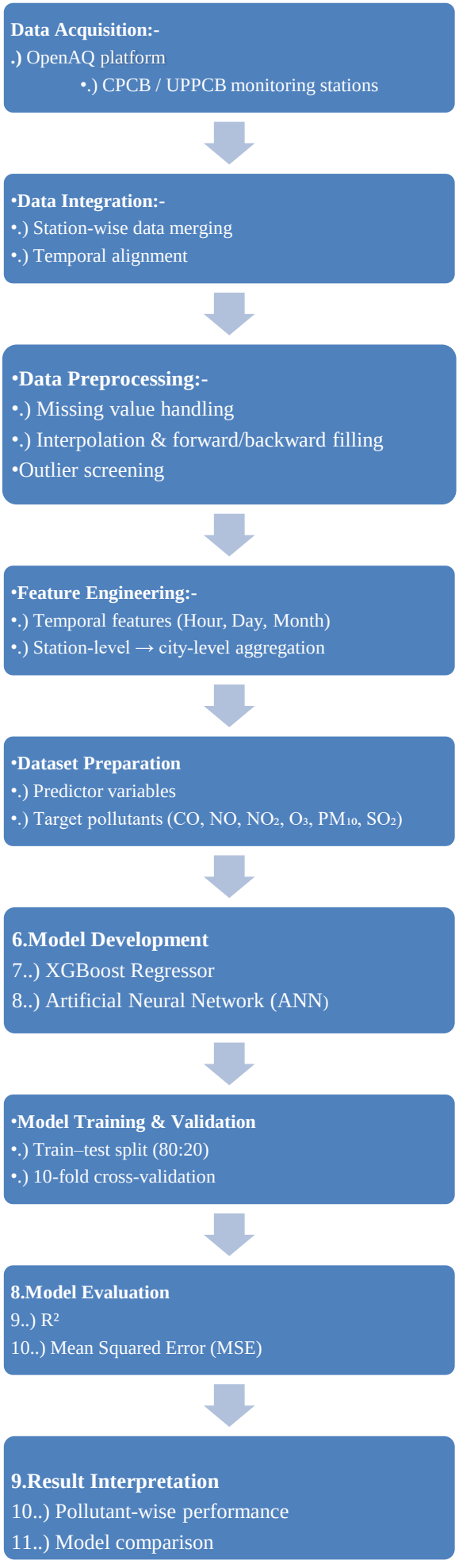


Figure 1: Air Quality Monitoring stations across the Study Area.

Measurements were taken between 15-minute and hourly intervals during the study period, which ran from June 1 to September 22, 2025. CO, NO, NO_2, O_3, PM_1[1], and SO_2 were the six pollutants that were examined.

4.2 Data Preprocessing

International Journal of Advance Interdisciplinary Research
Vol. 2, Special Issue 1, (Jan-March) 2026, pp. 197-215, e-ISSN: 3107-913X
DOI: https://doi.org/10.66095/ijair.2026.v2.S1.20



International Journal of Advance Interdisciplinary Research
Vol. 2, Special Issue 1, (Jan-March) 2026, pp. 197-215, e-ISSN: 3107-913X
DOI: https://doi.org/10.66095/ijair.2026.v2.S1.20

Figure 2: Flowchart illustrating overall Methodological Workflow.

A broad format with distinct columns for every station-pollutant combination was created by merging and transforming station-level datasets. Timestamps were used to extract temporal information including the hour of the day, day of the week, and month. Time-based interpolation was used to fill in the missing numbers, then forward and backward filling.

City-level average concentrations were calculated across all stations in order to improve model stability and decrease spatial noise. In line with methods employed in recent urban machine learning studies, the final dataset included about 30 predictor variables and about 1,200 observations (Kumar et al., 2024).

4.3 Machine Learning Models

XGBoost Regressor, a gradient boosting system called XGBoost creates an ensemble of decision trees one after the other. The following is the objective function that was minimized during training:

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Where $(y_i, \hat{y}_i)$, is the loss function and $\Omega(f_k)$ is the regularization term controlling model complexity (Chen & Guestrin, 2016).

Artificial Neural Network (ANN), A fully connected feed-forward ANN was constructed to capture nonlinear relationships. Each neuron's output is described as:

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right)$$

Where w_i is the weights, x_i are inputs, b is the bias and $f(\cdot)$ is a nonlinear activation function and (Hornik et al., 1989).

4.4 Model Evaluation

80% of the dataset was used for training, and the remaining 20% was used for testing. 10-fold cross-validation was used to assess the robustness of the model. The mean squared error (MSE) and coefficient of determination (R2) were performance indicators:

International Journal of Advance Interdisciplinary Research
Vol. 2, Special Issue 1, (Jan-March) 2026, pp. 197-215, e-ISSN: 3107-913X
DOI: https://doi.org/10.66095/ijair.2026.v2.S1.20

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}$$

$$MSE = \frac{1}{n}\sum(y_i - \hat{y}_i)^2$$

Where y_i and $\hat{y}_i$ indicate the concentrations of pollutants that were detected and expected, respectively and shows the average of the values that were observed.

5. Results and Discussions

5.1 Analysis

Time series study at the city level showed significant temporal variability for all pollutants. Concentration levels were continuously dominated by PM_{2^1} and NO_2, which reflected traffic and construction activity. O_3 demonstrated a definite daily cycle, with greater daytime concentrations driven by photochemical processes.

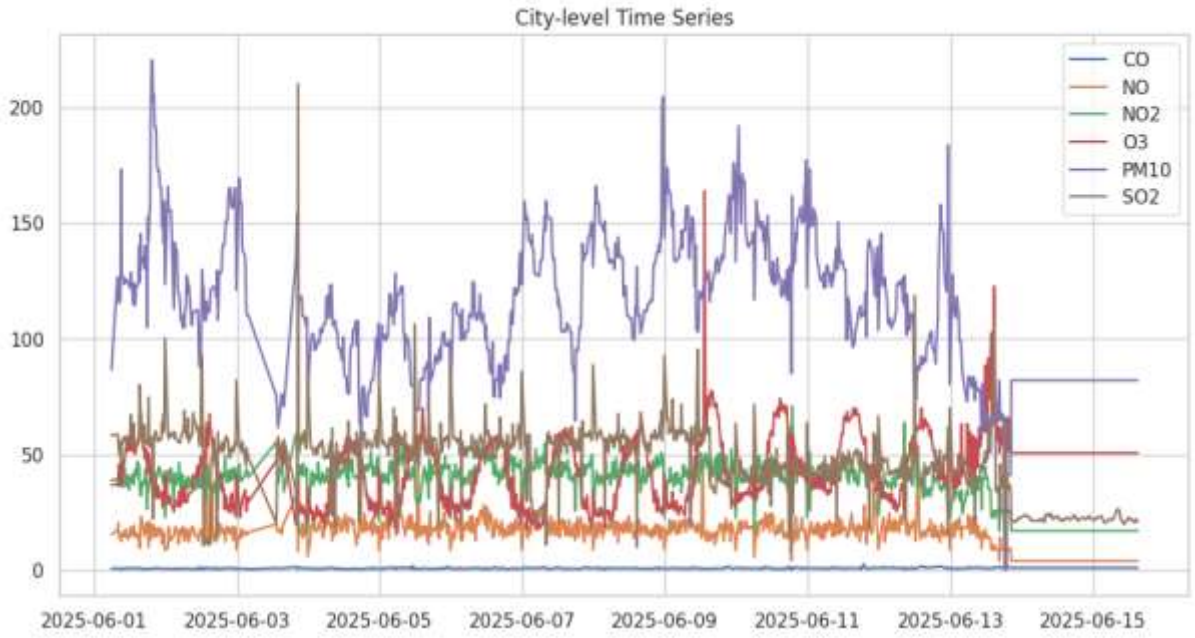


Figure 3: City level Time Series of Pollutants

In line with known atmospheric chemistry, correlation analysis revealed a substantial positive association between NO and NO_2 and a negative correlation between O_3 and NOx. There was a moderate association between PM_{4^1} and SO_2, indicating common sources of combustion.

International Journal of Advance Interdisciplinary Research
Vol. 2, Special Issue 1, (Jan-March) 2026, pp. 197-215, e-ISSN: 3107-913X
DOI: https://doi.org/10.66095/ijair.2026.v2.S1.20

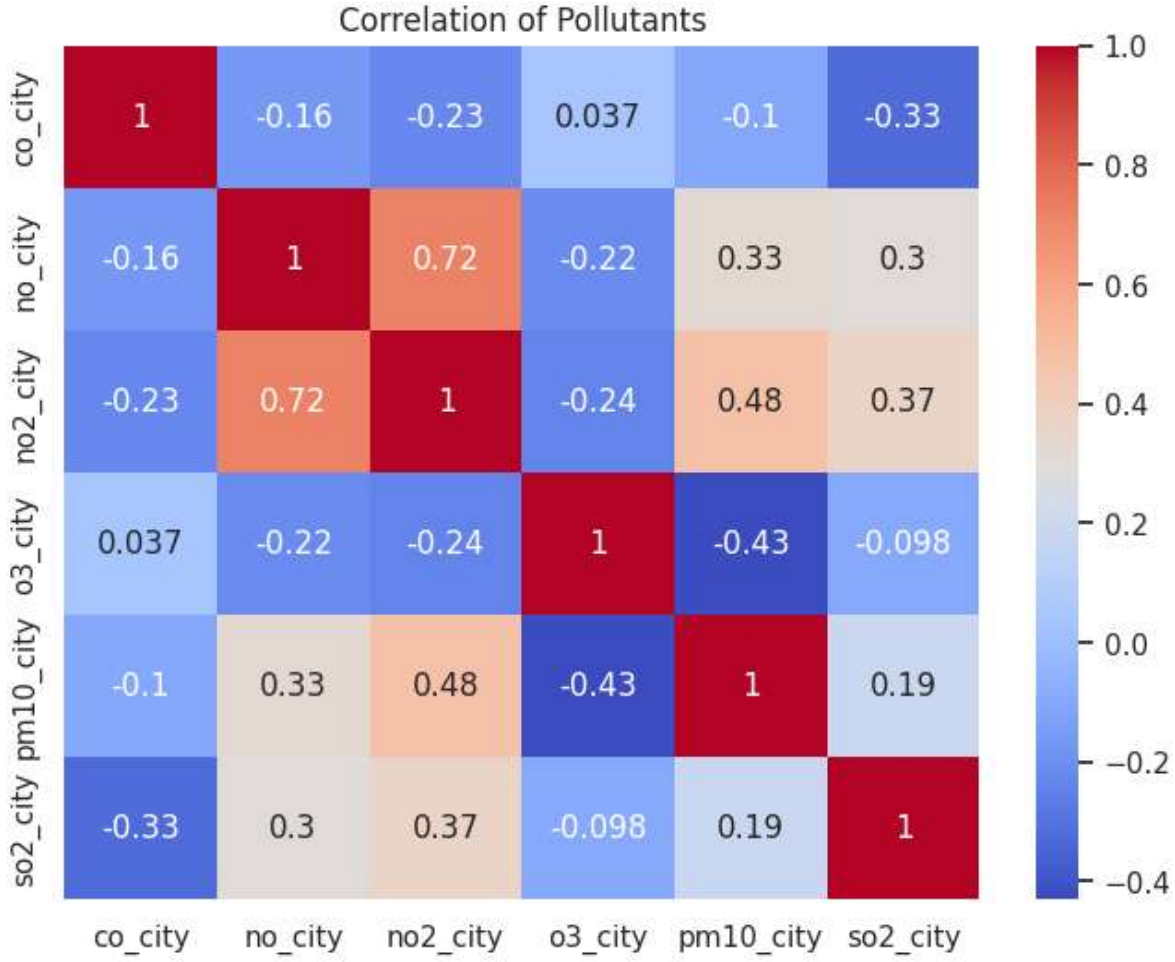


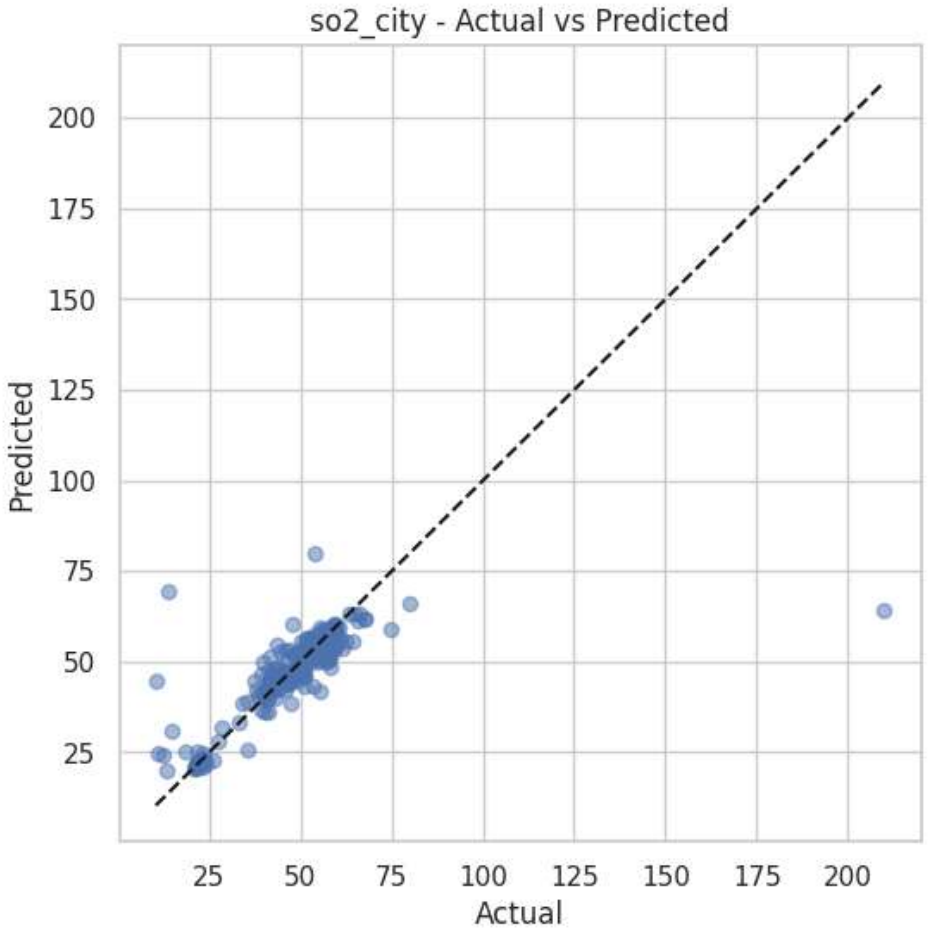
Figure 4: Correlation Structure among Pollutants

Pollutant-specific model performance was shown in predicted versus observed scatter plots. Excellent agreement was seen between PM_1[1] and NO_2, with predictions roughly aligned along the 1:1 line. Although scatter increased at higher concentrations, O_3 estimates successfully reflected diurnal fluctuation. NO and SO_2 displayed mixed performance, with baseline concentrations predicted more accurately than dramatic surges. CO performed the worst, with low explanatory power and wide scatter. Table 1 illustrates the predictive effectiveness of the applied models across contaminants

International Journal of Advance Interdisciplinary Research
Vol. 2, Special Issue 1, (Jan-March) 2026, pp. 197-215, e-ISSN: 3107-913X
DOI: https://doi.org/10.66095/ijair.2026.v2.S1.20

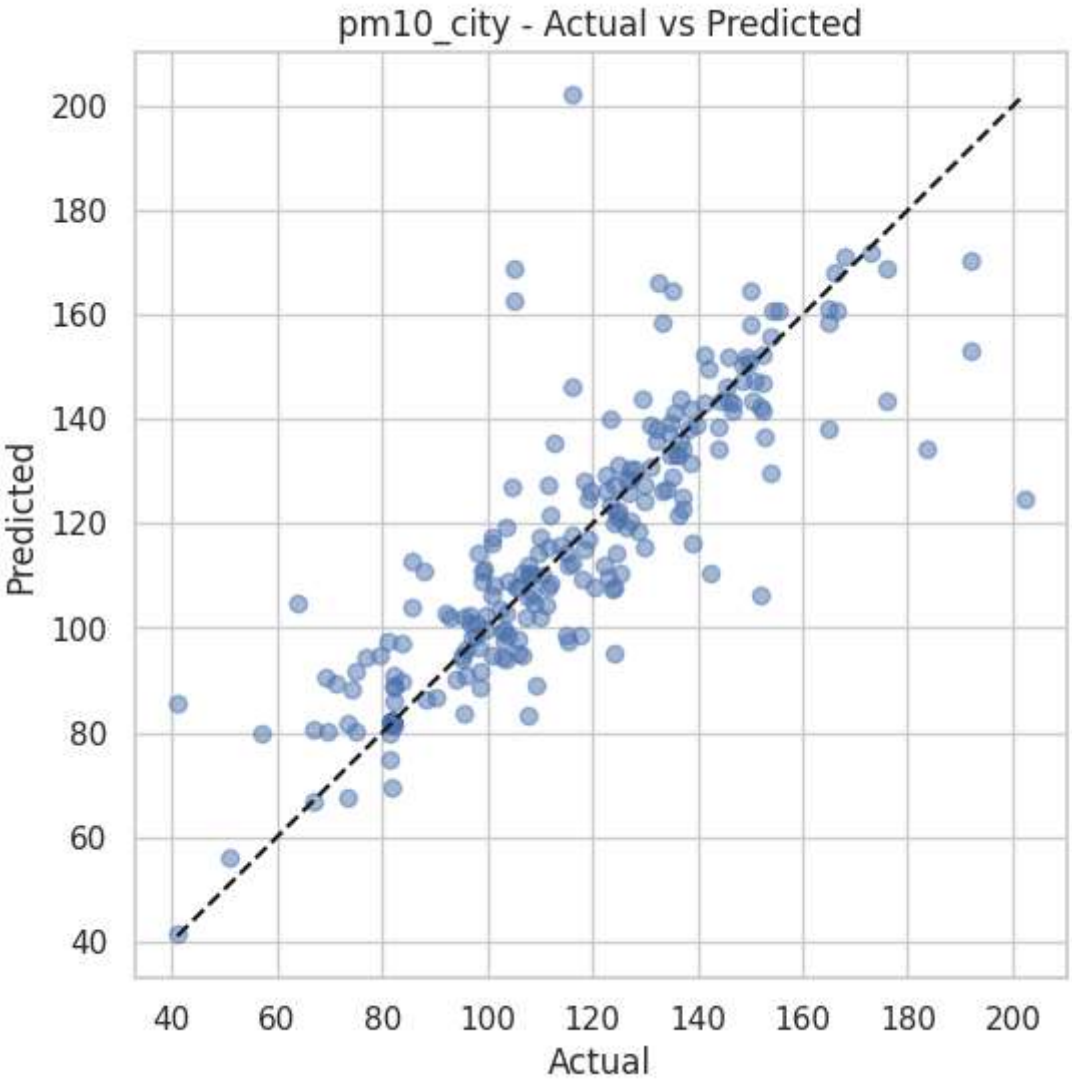
Table 1: Model Performance Across Pollutants and their interpretation

Pollutant	XGB Test R^2	ANN/MLP Test R^2	Interpretation
PM10	0.72	0.69	Excellent
NO_2	0.86	0.84	Excellent
O_3	0.81	0.75	Very Good
NO	0.57	0.39	Moderate
SO_2	0.57	0.76	ANN better
CO	0.18	0.16	Very Weak



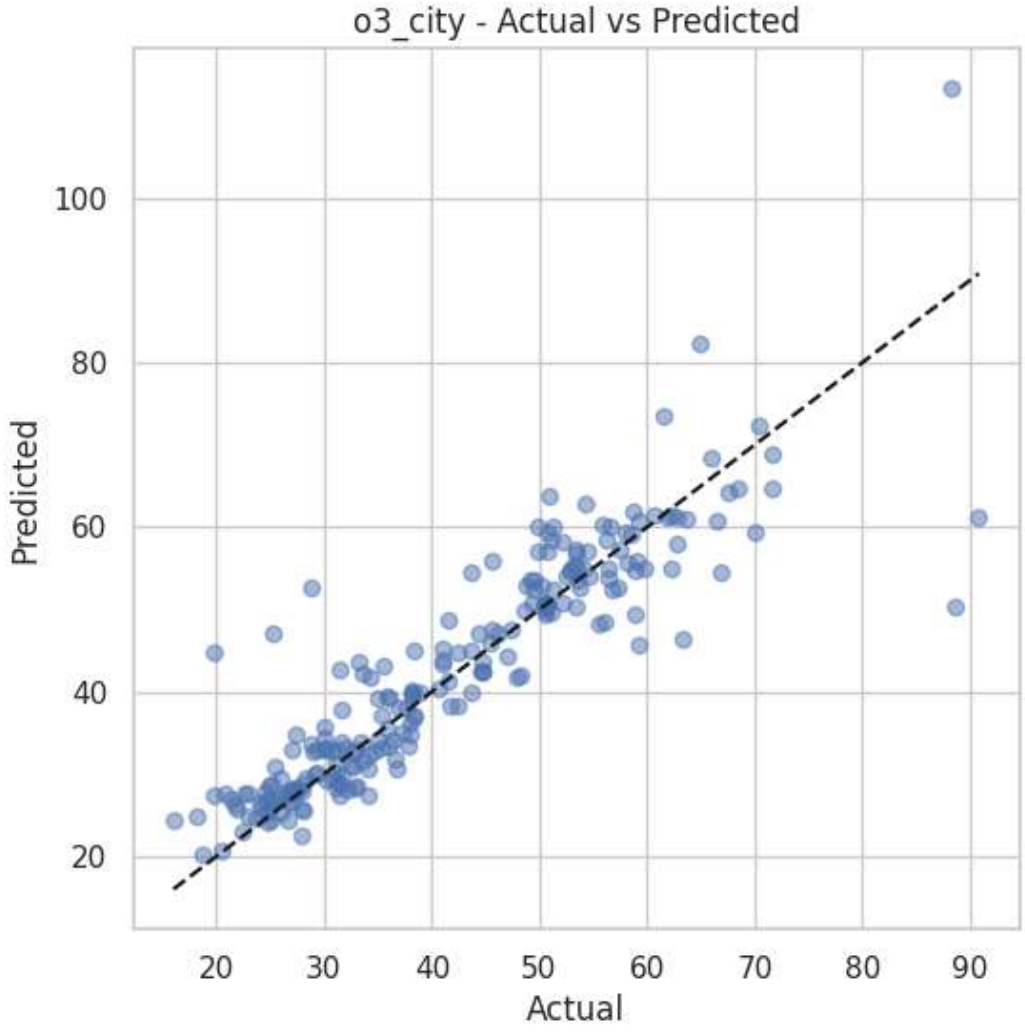
International Journal of Advance Interdisciplinary Research
Vol. 2, Special Issue 1, (Jan-March) 2026, pp. 197-215, e-ISSN: 3107-913X
DOI: https://doi.org/10.66095/ijair.2026.v2.S1.20

(a)



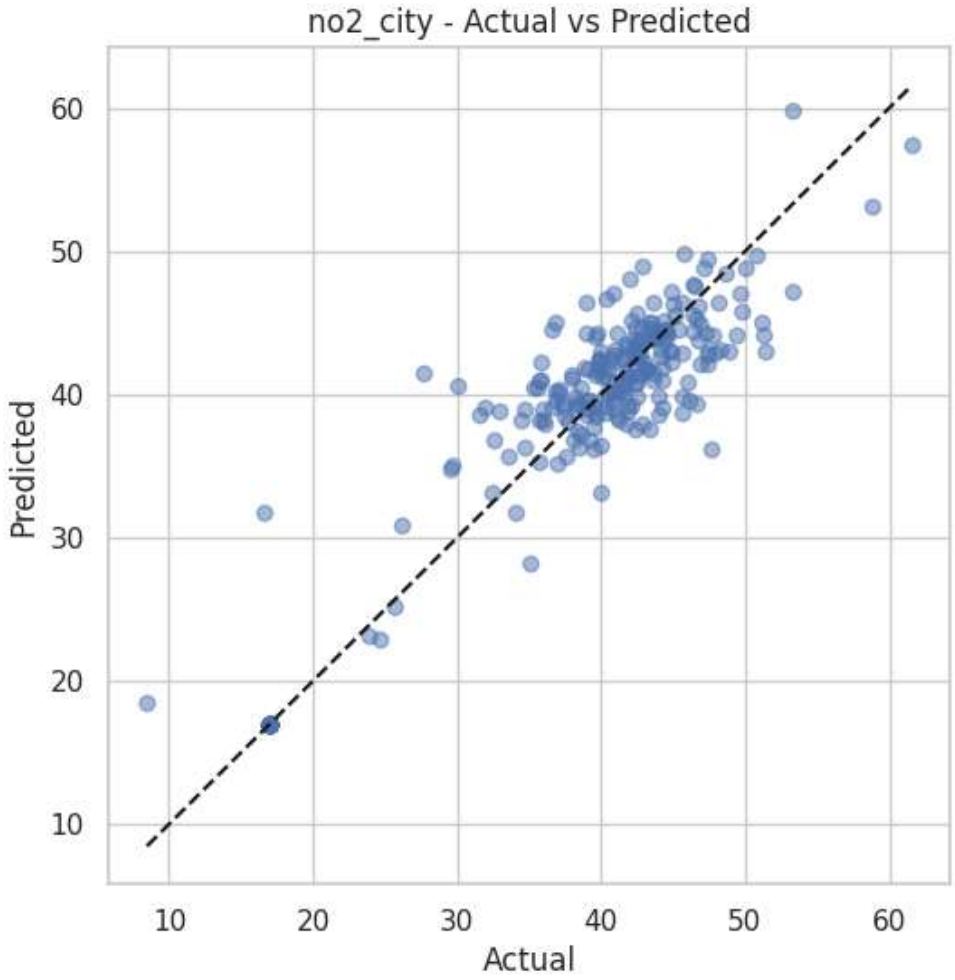
(b)

International Journal of Advance Interdisciplinary Research
Vol. 2, Special Issue 1, (Jan-March) 2026, pp. 197-215, e-ISSN: 3107-913X
DOI: https://doi.org/10.66095/ijair.2026.v2.S1.20



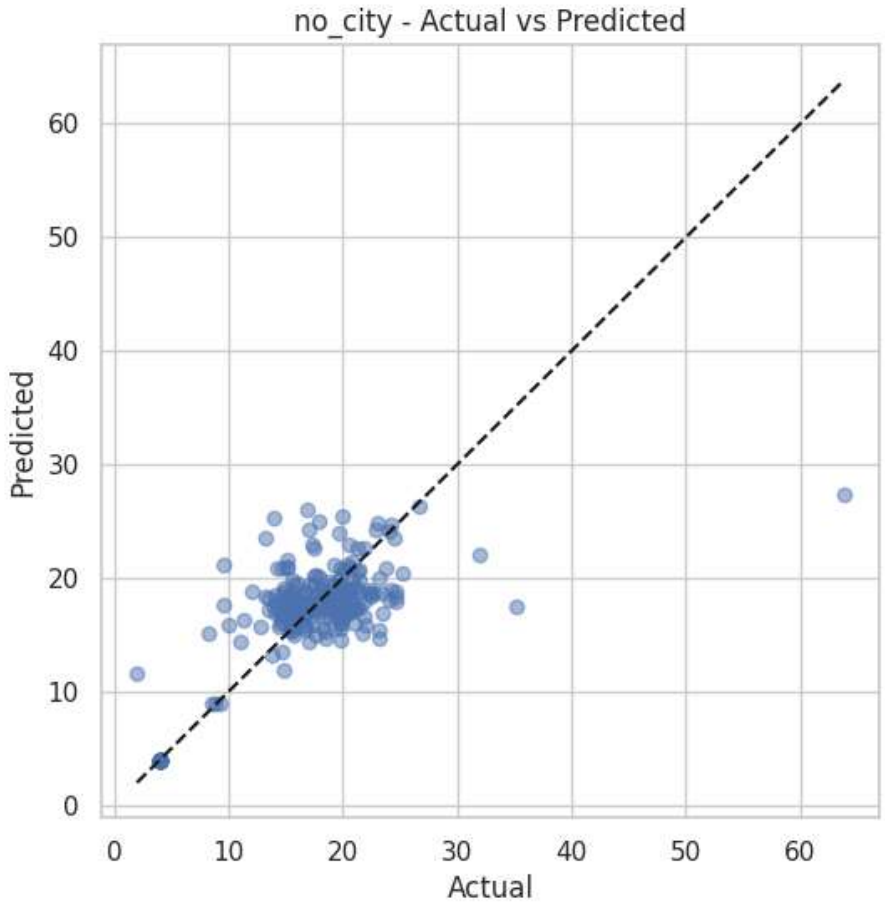
(c)

International Journal of Advance Interdisciplinary Research
Vol. 2, Special Issue 1, (Jan-March) 2026, pp. 197-215, e-ISSN: 3107-913X
DOI: https://doi.org/10.66095/ijair.2026.v2.S1.20



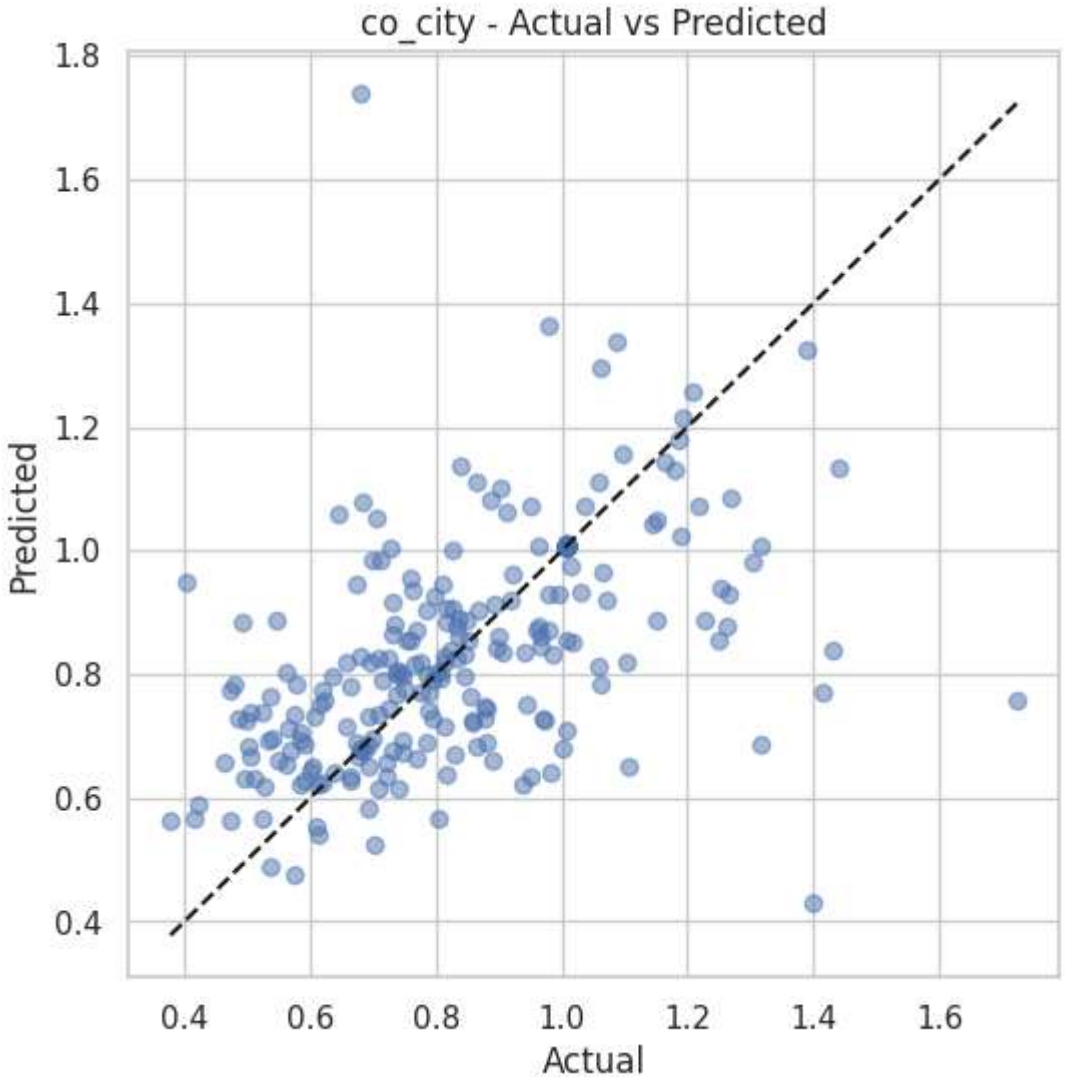
(d)

International Journal of Advance Interdisciplinary Research
Vol. 2, Special Issue 1, (Jan-March) 2026, pp. 197-215, e-ISSN: 3107-913X
DOI: https://doi.org/10.66095/ijair.2026.v2.S1.20



(e)

International Journal of Advance Interdisciplinary Research
Vol. 2, Special Issue 1, (Jan-March) 2026, pp. 197-215, e-ISSN: 3107-913X
DOI: https://doi.org/10.66095/ijair.2026.v2.S1.20



(f)

Figure 5: Predicted versus observed concentrations for different pollutants using machine learning models: (a) SO_2, (b) PM_{10}, (c) O_3, (d) NO_2, (e) NO, and (f) CO.

5.2 Discussions

The high performance of ML models for PM_2[1] and NO_2 is consistent with other research showing that data-driven methods are more predictable for pollutants with regular emission patterns and temporal persistence (Sharma et al., 2023; Singh et al., 2024). For these pollutants, XGBoost consistently scored better than ANN, probably because of its capacity to effectively capture feature interactions.

Both models' capacity to learn diurnal patterns improved O_3 prediction, but XGBoost showed somewhat greater robustness, in line with previous ensemble-based research (Wang et al.,

2023). On the other hand, because of their high temporal variability and localized emission sources, CO and NO showed poor predictability a problem that has been extensively documented in recent research (Roy et al., 2024).

For SO_2, ANN fared better than XGBoost, indicating that neural networks might be more adept at capturing the irregular, nonlinear behavior linked to industrial emissions. This observation validates results from deep learning and hybrid research highlighting the significance of nonlinear representation for gaseous pollutants (Zhou et al., 2023).

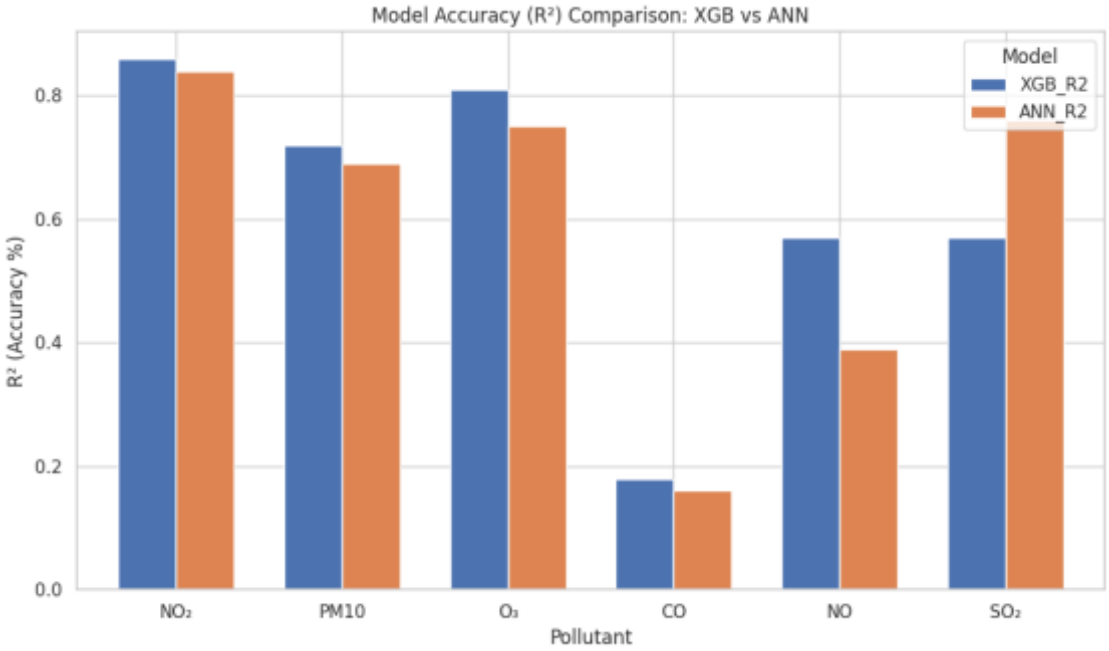


Figure 6: A comparative evaluation of model performance

Seasonal analysis is constrained by the study's comparatively short temporal span. For gaseous pollutants, the lack of traffic and weather factors probably decreased prediction accuracy. Furthermore, uncertainty may be introduced through interpolation of missing station data.

Future studies should incorporate meteorological characteristics, traffic density, and land-use data to boost forecasting capability. When longer datasets are available, advanced deep learning architectures like Transformer-based models and LSTM may enhance temporal learning (Li et al., 2023; Zhang et al., 2023). Comparative regional analysis and policy-relevant insights would be made possible by extending the methodology across other Indian cities.

6. Conclusion

This work shows that when pollutant behavior is consistent and regular, machine learning algorithms can accurately model urban air pollution in mid-sized Indian cities. XGBoost and ANN models successfully predicted PM_{10}, NO_2, and O_3, however CO and NO remained hard due to localized and erratic emissions. The results emphasize the necessity for pollutant-specific techniques backed by larger datasets and hybrid modeling frameworks, as well as the lack of a universal modeling strategy.

References

World Health Organization (WHO). (2021). Air pollution and health impacts. *The Lancet*, 398(10302), 103–104. DOI: https://doi.org/10.1016/S0140-6736(21)01102-4

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5), 359–366. DOI: https://doi.org/10.1016/0893-6080(89)90020-8

Chelani, A. B., & Devotta, S. (2007). Air quality forecasting using artificial neural networks. *Environmental Modelling & Software*, 22(4), 592–602. DOI: https://doi.org/10.1016/j.envsoft.2005.12.007

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. DOI: https://doi.org/10.1145/2939672.2939785

Wang, J., Zhang, Y., & Li, Z. (2023). Comparative evaluation of tree-based machine learning models for air quality forecasting. *Environmental Science and Pollution Research*, 30, 76845–76861. DOI: https://doi.org/10.1007/s11356-023-28122-6

Sharma, S., Kumar, P., & Khare, M. (2023). Application of machine learning techniques for urban air pollution prediction in Indian cities. *Environmental Pollution*, 317, 120748. DOI: https://doi.org/10.1016/j.envpol.2022.120748

Singh, V., Dey, S., & Chowdhury, S. (2024). Evaluating machine learning models for short-term air quality forecasting over Indian megacities. *Atmospheric Pollution Research*, 15(2), 101799. DOI: https://doi.org/10.1016/j.apr.2023.101799

Patel, P., Shah, R., & Mehta, D. (2025). Predictive modeling of PM_{10} and $PM_{2.5}$ concentrations using machine learning approaches in India. *Cleaner Environmental Systems*, 10, 100141. DOI: https://doi.org/10.1016/j.cesys.2024.100141

Zhang, Y., Wang, S., & Zhang, R. (2023). Transformer-based air quality forecasting with spatiotemporal attention mechanisms. *Atmospheric Environment*, 299, 119672. DOI: https://doi.org/10.1016/j.atmosenv.2023.119672

Zhou, Y., Chang, F. J., & Chang, L. C. (2023). Hybrid XGBoost-based models for nonlinear air pollutant forecasting. *Journal of Environmental Management*, 342, 118173. DOI: https://doi.org/10.1016/j.jenvman.2023.118173

Nguyen, T. T., Liu, D., & Hsieh, Y. C. (2023). Explainable machine learning for air pollution modeling and health risk assessment. *Environmental Modelling & Software*, 165, 105691. DOI: https://doi.org/10.1016/j.envsoft.2023.105691

Kumar, A., Goyal, P., & Kumar, R. (2024). Assessment of air quality prediction models under limited data conditions in India. *Urban Climate*, 53, 101742. DOI: https://doi.org/10.1016/j.uclim.2024.101742

Maji, K. J., Ye, W. F., Arora, M., & Nagendra, S. M. S. (2024). Machine learning approaches for estimating air quality and exposure risks in Indian cities. *Journal of Cleaner Production*, 418, 139876. DOI: https://doi.org/10.1016/j.jclepro.2023.139876

Roy, A., Chattopadhyay, S., & Dutta, A. (2024). Performance evaluation of machine learning models for predicting gaseous pollutants in urban India. *Sustainable Cities and Society*, 101, 105060. DOI: https://doi.org/10.1016/j.scs.2023.105060

Li, X., Peng, L., Hu, Y., Shao, J., & Chi, T. (2023). Deep learning architecture for air quality forecasting: A comprehensive review. *Atmospheric Environment*, 286, 119269. DOI: https://doi.org/10.1016/j.atmosenv.2022.119269

Liu, B., Ma, Y., Gong, W., Zhang, M., & Yang, J. (2023). Spatiotemporal prediction of air pollutants using ensemble learning approaches. *Science of the Total Environment*, 857, 159486. DOI: https://doi.org/10.1016/j.scitotenv.2022.159486

Zhang, Y., Chang, F. J., & Chang, L. C. (2023). Spatiotemporal air quality forecasting using deep learning models. *Atmospheric Research*,

297, 106610.
DOI: https://doi.org/10.1016/j.atmosres.2023.106610