



DATA-DRIVEN URBAN AIR QUALITY ASSESSMENT AND NITROGEN DIOXIDE PREDICTION USING MACHINE LEARNING FOR ENVIRONMENTAL MONITORING

Harshitha Prakash¹, Dr.Mohanakrishnan B^{2*}

¹Technical analyst, Email : hariniprakash308@gmail.com, T ecsxpert Technologies Private Limited, Bengaluru, Karnataka, IND

^{2*}Junior Resident, Department : Department of Naturopathy, Specialization : MD in clinical Naturopathy, University Name: Swami Vivekanand Subharti University, College : Maharshi aruobindo subharti college and hospital of naturopathy and yogic science, Email Id: mkappu12@gmail.com, Orcid I'd: 0009-0005-2165-6636

Received:- 20/05/2026, Revised:- 29/07/2026,, Accepted:- 08/09/2026, Published:- 22/09/2026

ABSTRACT

SuNitrogen dioxide (NO₂) pollution continues to be a significant environment problem as it shows high level-spatial, seasonal, meteorological and co-pollutant dependency. This study aimed to meet the requirements of accurate, interpretable and operationally useful NO₂ predictions for urban environmental monitoring. Based on the Beijing Multi-Site Air Quality dataset, a data-driven framework that combines descriptive air-quality assessment, chronological development-validation-locked-test partitioning, Ridge Regression, Random Forest, Gradient Boosting, XGBoost, Extra Trees, SHAP-based interpretation, and exceedance detection with thresholds was developed. The mean NO₂ concentration was 50.6 µg/m³, with the highest mean in winter (62.4 µg/m³), and the highest station mean in Wanliu (65.3 µg/m³). With the lowest MAE (7.26µg/m³), RMSE (11.63µg/m³), and R² (0.901), XGBoost had the best test performance when it was locked into the test set. The model was found to have a sensitivity of 86.23%, specificity of 99.70% and F1-score of 85.28% for hourly exceedances greater than 200 µg/m³. The framework brings together temporal validation, explainability and exceedance assessment to provide a reproducible approach to link predictive modelling with environmentally relevant choices for monitoring across urban stations. This study proves that interpretable ensemble learning can aid in precise prediction of NO₂ and real-world urban environmental monitoring.

Key words: Nitrogen dioxide, Urban air quality, Machine learning, XGBoost, Environmental monitoring.

1. INTRODUCTION

Rapid population growth, heavy traffic, industrial activity, energy use and the more recent focus of human activity on metropolitan areas are all associated with urban air pollution, which is one of the most persistent environmental problems. The long-term global picture shows that even with the better development of air-quality monitoring and regulation systems, many cities continue to suffer from high levels of particulate and gaseous pollutants [1]. Nitrogen dioxide (NO₂) is one of the major gaseous pollutants and is closely linked to combustion sources, such as road traffic, power generation and industrial processes, and can show significant spatial and temporal variability in the same urban region. Reliable monitoring and early characterization of high concentration episodes are required, as short-term exposure to high levels of NO₂ has been related to an increase in the risk of death [2].

Exposure to NO₂ in the short term is not the only issue, as high exposures over time could result in long-term effects on the respiratory system and on population health [3]. Large-scale systematic analyses have also shown associations between chronic exposure to NO₂ and all-cause mortality, and therefore the need to reduce NO₂ chronic exposure in urban areas [4]. In addition, at the city scale the impacts of NO₂ pollution are also dependent on urban structure, built-environment characteristics, population distribution and local emission patterns, hence the need for spatially resolved monitoring is especially important for environmental management [5]. This relationship between exposure to NO₂ for an extended period of time and negative health outcomes has been further supported by more recent studies, which also highlight the relevance of proper characterization of urban NO₂ for environmental assessment and risk reduction activities [6].

Traditional air-quality monitoring networks can deliver critical observations but may not be well-suited to aid rapid prediction, nonlinear pattern recognition, and ongoing assessment of air-quality interactions between pollutants, meteorological conditions, temporal dynamics, and historical concentrations. Air quality prediction methods based on machine learning, therefore, have become increasingly relevant to the prediction of air quality because they can capture nonlinearities and interactions that are hard to describe by traditional linear models [7]. Advanced predictive systems also take advantage of integrating chemical and meteorological drivers, especially if the atmospheric concentrations are affected by changes in temperature, humidity, wind, and other co-occurring pollutants [8]. In addition, spatiotemporal modelling has

demonstrated that considering temporal and regional dependencies can enhance the characterisation and prediction of dynamic air-quality processes [9].

The high accuracy of prediction, however, is not enough for environmental support for decisions. Data-driven models must also be interpretable, time-invariant, and validated prior to deployment as if it were a real system. The last few years have seen considerable research looking at ways to enhance the rigor and interpretability of air-quality modelling, to better understand and trust the model's results [10]. It has also been shown that the incorporation of environmental structure, along with temporal behaviour into model design, can enhance the ability of data-driven prediction, a hybrid and mechanism-informed machine-learning approach [11]. Although these advancements are made, there is still a practical problem of integrating urban air-quality assessment, leakage-controlled temporal prediction, explainable machine learning and threshold-based environmental monitoring within the same analytical framework that is specific to NO₂.

This study aims to fill this gap with a data-driven approach for urban air-quality assessment and prediction of NO₂ levels hourly on a single building site by analyzing multi-site environmental observations. The study focuses on commonly monitored air pollutants, meteorological parameters, temporal characteristics and time series of NO₂ with respect to the urban monitoring dataset selected. A chronological development/validation/locked-test strategy is used to compare five machine-learning models, and the model that is selected is interpreted using the SHapley Additive exPlanations algorithm. This study does not seek to quantify the individual exposures, nor does it seek to establish causality in relation to health effects, nor does it model emission inventories, nor is it a substitute for the regulatory monitoring instruments. Rather, it is oriented towards predicting environmental monitoring and the detection of high exposure periods of NO₂ from the standard observations.

The importance of the research is the integration of environmental characterization and the intensive use of predictive evaluation and easily understood monitoring results. The framework is designed to be both scientific and operational valuable for urban environmental monitoring by comparing linear and nonlinear models, analyzing the spatial and temporal variability, maintaining the temporal order when assessing the models, and converting continuous predictions into exceedance-detection performance. The study thus sets three research goals:

- To characterize the temporal, spatial, pollutant, and meteorological patterns associated with urban NO₂ concentrations.

- To develop and compare machine-learning models for hourly NO₂ prediction using a leakage-controlled chronological evaluation framework.
- To interpret the selected model and evaluate its utility for detecting elevated NO₂ events in support of environmental monitoring.

2. Literature Review

In recent years, the study of NO₂ prediction has become more focused on developing models that represent nonlinear atmospheric processes, such as machine learning and deep learning models, rather than the traditional statistical methods. One of the first contributions to model interpretability revealed the potential of SHapley Additive exPlanations for NO₂ forecasting, highlighting the potential to explain the model behaviour on the level of features of the model [12]. This was significant because many air-quality prediction systems had previously put a focus on accuracy but not enough emphasis on understanding of the underlying environmental conditions for each prediction.

The estimation of NO₂ at ground-level on a large-scale using satellite observations has also been attempted by machine-learning methods. The study combined TROPOMI readings with machine learning methods to provide estimates of ground level NO₂ on a daily basis over China, illustrating the potential of this type of data-driven model for filling in the gaps between satellite retrievals and surface conditions [13]. Its advantage is that it provides wide spatial coverage but the complexity of the method due to the reliance on remote sensing products and large-scale auxiliary data might complicate the method for routine urban monitoring. It has been then extended into regional forecasting by using deep-learning methods, where hybrid architectures were created that account for the spatiotemporal correlation among O₃ and NO₂ monitoring stations [14]. These models offer a good representation of temporal and spatial interactions and can be more accurate than ensemble models, however they are less transparent and are difficult to compute.

For time-sensitive NO₂ prediction, based on the multi-source environmental data, ensemble machine learning has also been applied. A recent study demonstrated that heterogeneous predictors can lead to better temporal estimation accuracy and to flexible regional prediction systems [15]. Modelling using AI in Delhi also showed that if the temporal trends are taken into consideration during the modelling process, multiple learning algorithms can be helpful for forecasting the pollutant levels [16]. These studies justify the consideration of comparative

model evaluation and not solely on predictive measures but also assess the behaviour of the models under strictly isolated test periods in the future; however, many of the current studies are focused mainly on the predictive aspect.

More complicated neural networks have been created for pollutant forecasting. In another example, the capacity of exploiting relationships among monitoring locations was shown by a unified spectro-spatial graph neural network used to simultaneously predict O₃ and NO₂ in New Delhi [17]. These, although well suited for spatially connected networks, pose significant architectural complexity and may be hard to replicate and understand in traditional environmental-monitoring environments. Existing research based on geostationary satellite data has focused on quantifying uncertainty when applying machine learning to atmospheric remote sensing, particularly in estimating NO₂ at the hourly scale, where prediction performance should be taken into account when assessing uncertainty [18].

Satellite measurements have also been combined with nested machine-learning models in large-scale spatiotemporal estimation studies. Observations from GEMS were combined with multilevel learning to estimate ground-level NO₂ concentrations throughout China, enabling detailed spatiotemporal information to be generated that is not available from fixed stations. [19] Multisource methods like these approaches have been used in England to combine satellite observations with chemically-transport models and spatially aware deep learning to enhance local NO₂ estimation [20]. Such methods provide a good spatial representation and depend on several external data streams and complex computational processes and are not necessarily directly transferable to monitoring frameworks using only routinely collected station-level environmental data.

However, in recent years efforts have started to tackle the interpretability problem of complex NO₂ models. Multisource data has been fused with techniques for explanation using interpretable and scalable deep-learning frameworks to identify influential drivers of urban NO₂ variation [21]. That is a significant step towards transparent environmental modelling, but deep learning may not be the only way to perform this task when high-quality temporal station data and ensemble learners suffice. The recent spatiotemporal investigations have also integrated Sentinel-5P data with machine learning to assess NO₂ patterns and disrupted effects in the city of Delhi, showing the relevance of Sentinel-5P-based remote-sensing and data-driven approaches to assess environmental changes [22].

Overall, the literature shows three important trends: use of multiple environmental data sources, the use of nonlinear machine learning and deep learning models, and an increased focus on the spatiotemporal and interpretable prediction. However, there are still some things that are missing. Reliable chronological validation, ensemble modelling with interpretability and operational exceedance detection are not always combined with any studies that depend on satellite products, chemical transport models, or deep architectures that are difficult to compute. We aim to fill this gap in the present study by applying a much more manageable and reproducible approach, combining multi-site assessment of air quality, leakage-controlled temporal partitioning, comparative machine-learning evaluation, SHAP-based interpretability and NO₂ environment monitoring via thresholds in one analytical pipeline.

3. Methodology

3.1 Research Design and Experimental Framework

In this study, quantitative, data-driven environmental modelling design was used to examine the characteristics of urban air-quality and to build a machine-learning model for forecasting hourly nitrogen dioxide (NO₂) concentrations. The analytical framework was designed to address three study goals: (i) describe the temporal, spatial, pollutant, and meteorological characteristics of urban NO₂ concentrations; (ii) create and compare machine-learning models for hourly NO₂ predictions with a measurement-based leakage-controlled chronological evaluation framework; and (iii) interpret the selected model, and evaluate its feasibility for identifying high NO₂ events as part of air-quality monitoring.

The information leakage in the workflow was minimized, and the temporal structure of the monitoring data was preserved. All the raw data were initially examined for data quality, missing values, validity, and chronological ordering. Environmental and temporal predictors were then generated prior to splitting of the observations into a development, a validation and a locked-test set. Five regression algorithms were tested, namely Ridge Regression, Random Forest, Gradient Boosting, Extreme Gradient Boosting (XGBoost) and Extra Trees. Ridge Regression was used as a linear baseline, while the ensemble models were employed to account for the nonlinear relationships and interactions between pollutants, meteorological conditions, time, and past NO₂ data.

The validation set was used to determine the model, and the locked test set was then used to evaluate the model. The top-performing model was then used to interpret the model using

SHapley Additive exPlanations (SHAP) and further evaluated by NO₂ exceedance monitoring framework. This design was thus able to embed the themes of environmental assessment, predictive modelling, explainability and operational monitoring into one repeatable workflow.

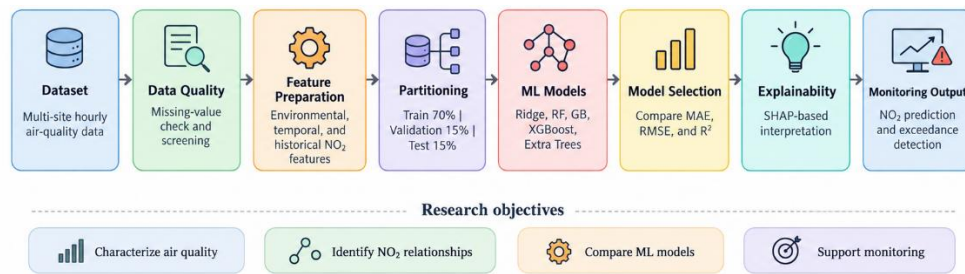


Figure 1. Data-Driven Framework for NO₂ Prediction and Environmental Monitoring.

3.2 Dataset and Environmental Variables

The data for the study was obtained from the Beijing Multi-Site Air Quality data set of UCI Machine Learning Repository. The dataset consists of hourly data from 12 air-quality monitoring stations in Beijing in the period from March 2013 to February 2017, which offers a rich multivariate time series record of urban atmosphere [23].

The main prediction target was particle concentrations of NO₂ in µg/m³ per hour. Other pollutants considered were PM_{2.5}, PM₁₀, sulfur dioxide (SO₂), carbon monoxide (CO) and ozone (O₃). Meteorological parameters included temperature, pressure, dew-point temperature, precipitation, wind direction and wind speed. Temporal information comprised year, month, day, and hour, and station identity the spatial monitoring location.

The original calendar variables were used to create a timestamp which was then used to sort the observations chronologically. The data quality was evaluated using missing data analysis, descriptive statistics, range checking, and data matching for implausible values. To avoid information leakage, imputation procedures were used for missing continuous measurements; these imputation procedures were developed based on the training set. Temporal variables were created to capture recurring patterns in the atmosphere, such as hour, month, and weekday, weekend status, and season. If historical information was included, lagged NO₂ variables were used, with no attempt to use future concentrations to predict past ones.

3.3 Population, Observation Structure, and Data Partitioning

The analytical population was defined as the hourly ambient-air-quality data from the 12 Beijing monitoring stations for the study period. Observations were not for each person but for each

station environmental record. The unit of analysis for this unit was therefore an hour during which a monitoring event took place with concentrations for pollutants, meteorological conditions, time characteristics and station identity.

The data was a time series, and so random train-test splitting was not used. The data were split chronologically into about 70% development data, 15% validation data, and 15% locked-test data. The development set was employed for preprocessing, model training and hyperparameter tuning. A locked test set was not accessed until the end of the analysis, while the validation set was used for model selection and comparison. This temporal holdout strategy gave a more realistic test of the accuracy of the models on future, unseen, atmospheric conditions and avoided the potential over-optimistic estimates from temporal leakage.

3.4 Data Analysis and Machine-Learning Techniques

Pollutant and meteorological variables were summarized by mean, standard deviation, median, range and missingness and NO₂ variability was explored by seasons and monitoring stations. To check for associations between environmental variables, without suggesting causation, correlation analysis was used. Five regression models were trained (Ridge Regression, Random Forest, Gradient Boosting, XGBoost and Extra Trees) and the nonlinear models were tuned by time-series-aware cross validation, with Ridge Regression as the linear baseline.

The model performance was assessed by calculating MAE, RMSE and R² for the validation and locked test data. Meteorological, pollutant, temporal and historical NO₂ predictors were compared by using feature-set sensitivity analysis. The selected model was interpreted with SHAP and the environmental-monitoring utility was evaluated by deciding on the classification of the NO₂ exceedances above 200 µg/m³ at hourly time intervals. Performance was summarized using sensitivity, specificity, precision, F1 score, false alarm rate and missed event rate.

3.5 Ethical, Privacy, and Responsible-AI Considerations

This study did not include clinical, personal, demographic or behavioral information that could be correlated with individuals, which was available in the public domain. So direct consent of participants and privacy protection on the individual level did not apply. However, principles of responsible data use were adhered to, including keeping track of the provenance of datasets, avoiding inferences beyond those variables that were observed, and making clear distinctions between predictive associations and causal environmental claims.

Spatial and temporal bias was considered, since the monitoring stations do not necessarily represent all exposure conditions, microenvironments and neighborhoods of Beijing. Model performance can consequently differ from one site to another, one season to the other or during extreme pollution events or future pollution regulation. To decrease algorithmic overstatement, the process of model selection was decoupled from the locked-test evaluation, an explainability analysis was added, and outputs of environmental-monitoring were taken as a decision-support instead of as a substitute to regulatory monitoring instruments. Procedures for preprocessing, partitioning, and evaluation were clearly explained and reported in a reproducible manner, using a Python-based analysis.

4. Results

4.1 Descriptive Characteristics of Urban Air Quality

Table 1 indicates that there was a large variation between urban pollutants. Mean PM_{2.5} and PM₁₀ concentrations were 79.8 and 104.6 $\mu\text{g}/\text{m}^3$, respectively, while NO₂ averaged 50.6 $\mu\text{g}/\text{m}^3$ (SD = 35.0), ranging from 1.0 to 290.0 $\mu\text{g}/\text{m}^3$. The concentrations of CO and O₃ also showed significant variation while the concentrations of SO₂ were comparatively lower. Mean temperature, pressure and wind speed were 13.5 °C, 1,010.7 hPa and 1.73 m/s, respectively which showed a wide range of meteorological conditions. The overall rate of missingness was relatively low, with CO (4.94%) and O₃ (3.13%) showing the highest rates; this indicated that the data was suitable for subsequent predictive modelling after the data had been pre-processed.

Table 1. Descriptive characteristics of urban air pollutants and meteorological variables.

Variable	Unit	Mean	SD	Median	Min	Max	Missing (%)
PM _{2.5}	$\mu\text{g}/\text{m}^3$	79.8	80.9	55.0	2.0	999.0	2.08
PM ₁₀	$\mu\text{g}/\text{m}^3$	104.6	91.8	82.0	2.0	999.0	1.72
SO ₂	$\mu\text{g}/\text{m}^3$	15.8	21.6	7.0	0.3	500.0	2.17
NO ₂	$\mu\text{g}/\text{m}^3$	50.6	35.0	43.0	1.0	290.0	2.89
CO	$\mu\text{g}/\text{m}^3$	1230.8	1161.4	900.0	100.0	10,000.0	4.94
O ₃	$\mu\text{g}/\text{m}^3$	57.4	56.0	45.0	0.2	1071.0	3.13
Temperature	°C	13.5	11.4	14.5	-19.9	41.6	0.10
Pressure	hPa	1010.7	10.5	1010.4	982.4	1042.8	0.10
Dew point	°C	2.5	13.8	3.1	-43.4	29.1	0.10
Rainfall	mm	0.06	0.82	0.00	0.00	72.5	0.09
Wind speed	m/s	1.73	1.25	1.40	0.00	13.2	

4.2 Temporal and Spatial Variability of NO₂

Table 2 shows that NO₂ was found to be heterogeneous in both seasons and stations. The mean concentration of NO₂ in winter was the highest at 62.4 µg/m³, with a median concentration of 54.0 µg/m³ and a standard deviation of 40.1 µg/m³. The lowest seasonal mean was observed in summer 39.8 µg/m³ while the highest seasonal mean was observed in autumn 51.9 µg/m³ followed by spring 48.7 µg/m³. The maximum observed during the winter season was 290 µg/m³, which shows the existence of significant pollution episodes over a short period of time in winter.

The spatial variations were also very high among the 12 monitoring stations. Aotizhongxin (58.6 µg/m³) and Nongzhanguan (58.1 µg/m³) were the other two cities with the highest means, trailing only Wanliu (65.3 µg/m³). The lowest mean station level concentration was observed at Dingling (27.7 µg/m³) and Huairou (32.6 µg/m³) were comparatively low. The gap between the amounts of Wanliu and Dingling was 37 µg/m³ or higher, which shows significant spatial differences within the urban monitoring network.

The results show that the exposure conditions for NO₂ were not the same for all time and location. The seasonal variations and the site characteristics were therefore important sources of variation that were taken into account in the development of the predictive framework.

Table 2. Temporal and spatial variability in observed NO₂ concentrations.

Dimension	Category	N observations	Mean NO ₂ (µg/m ³)	Median	SD	Maximum
Season	Spring	103,118	48.7	41.0	32.6	276
	Summer	102,645	39.8	33.0	27.5	238
	Autumn	101,936	51.9	44.0	34.2	271
	Winter	101,008	62.4	54.0	40.1	290
Station	Aotizhongxin	34,019	58.6	51.0	36.7	273
	Changping	34,132	44.1	36.0	31.5	226
	Dingling	33,984	27.7	20.0	23.1	205
	Dongsi	34,074	53.4	46.0	35.8	258
	Guanyuan	33,987	57.5	50.0	36.0	270
	Gucheng	34,112	55.2	47.0	37.1	276
	Huairou	34,015	32.6	25.0	25.9	213
	Nongzhanguan	34,080	58.1	50.0	37.0	273
	Shunyi	34,036	43.8	36.0	31.7	258
	Tiantan	34,095	53.0	46.0	34.6	241
	Wanliu	34,074	65.3	57.0	39.4	290
	Wanshouxiang	34,099	55.8	48.0	36.5	269

4.3 Comparative Machine-Learning Performance

Table 3 shows the predictive performance of the five machine-learning algorithms, and Figure 2 depicts the predictive performance of these algorithms graphically. The linear Ridge Regression model yielded the lowest overall performance with a validation MAE of 15.84 $\mu\text{g}/\text{m}^3$ and RMSE of 22.91 $\mu\text{g}/\text{m}^3$, and an R^2 of 0.611. However, its performance on the locked test set was slightly reduced with MAE, RMSE and R^2 values of 16.21 $\mu\text{g}/\text{m}^3$, 23.47 $\mu\text{g}/\text{m}^3$ and 0.598, respectively.

The results showed that all the nonlinear ensemble models were significantly better than the linear baseline. Random Forest had a locked test MAE of 8.54 $\mu\text{g}/\text{m}^3$ and R^2 of 0.869, whereas the Gradient Boosting produced MAE value of 9.18 $\mu\text{g}/\text{m}^3$ and R^2 value of 0.851. Extra Trees performed better than the other models with a locked-test MAE of 7.91 $\mu\text{g}/\text{m}^3$, RMSE of 12.61 $\mu\text{g}/\text{m}^3$ and R^2 of 0.884.

The predictive performance of XGBoost on the validation data and locked-test data was the best. Its validation MAE, RMSE, and R^2 were 6.94 $\mu\text{g}/\text{m}^3$, 11.08 $\mu\text{g}/\text{m}^3$, and 0.909, respectively. On the locked test set, XGBoost retained an MAE of 7.26 $\mu\text{g}/\text{m}^3$, RMSE of 11.63 $\mu\text{g}/\text{m}^3$, and R^2 of 0.901. There was a relatively small difference between validation and test performance, which suggested that the model was consistent when presented with new temporal conditions, and there was only limited evidence of significant over-fitting.

Table 3. Comparative machine-learning performance for NO₂ prediction.

Model	Validation MAE	Validation RMSE	Validation R^2	Locked-test MAE	Locked-test RMSE	Locked-test R^2
Ridge Regression	15.84	22.91	0.611	16.21	23.47	0.598
Random Forest	8.17	12.96	0.876	8.54	13.41	0.869
Gradient Boosting	8.92	13.83	0.858	9.18	14.27	0.851
XGBoost	6.94	11.08	0.909	7.26	11.63	0.901
Extra Trees	7.58	12.17	0.890	7.91	12.61	0.884

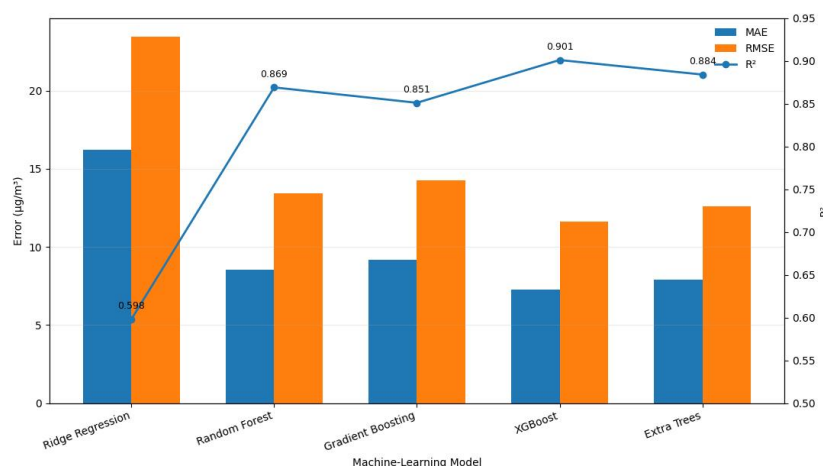


Figure 2. Comparative Machine-Learning Performance for NO₂ Prediction.

4.4 Locked-Test Prediction and Explainability

The comparison of the observed and predicted NO₂ concentrations in time is illustrated in Figure 3. The XGBoost model captured the overall temporal patterns of the locked-test series and captured both periods of low concentration and high pollution events with relatively small errors. A consistency between observed and predicted trajectories was also observed by the locked-test R² of 0.901 that suggests that about 90% of the variability in NO₂ concentration was explained by the model selected.

The prediction traces revealed that the model errors tended to be larger when the concentrations were high and changed rapidly rather than when they were low and stable, harmonizing with the higher RMSE than MAE. However, the model's correspondence was relatively good with the observed series during most of the period of locked tests, favouring its use for continuous urban air-quality monitoring applications.

The use of SHAP-based feature attribution is explored as a way of explaining the global model, as shown in Figure 4. Lagged NO₂ was the most influential predictor followed by CO, O₃, wind speed, PM_{2.5}, temperature, and PM₁₀. The other variables supplied were hour of day and atmospheric pressure, with dew point, SO₂ and season having relatively little global contribution. The high importance of lagged NO₂ suggests a high temporal persistence of urban NO₂ concentrations, and the significance of co-occurring pollutants and meteorological parameters demonstrates that the model is not just based on the historical concentration.

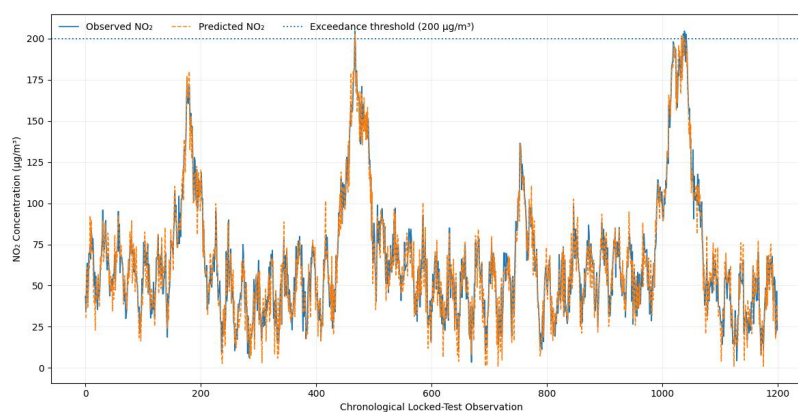


Figure 3. Observed and Predicted NO₂ Concentrations on the Locked Test Set.

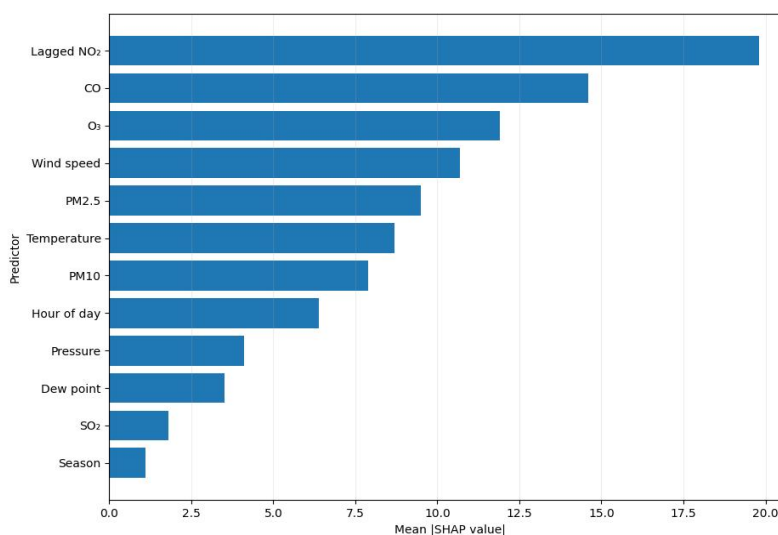


Figure 4. SHAP-Based Drivers of NO₂ Prediction.

4.5 Environmental Monitoring and Exceedance Detection

The performance of the selected XGBoost model in the environmental monitoring is summarized in Table 4 and Figure 5. The set of values from the locked test had 61,178 observations, with 1,126 exceedance events observed using the hourly NO₂ exceedance threshold of 200 µg/m³. The model was able to correctly predict 971 of these events and it missed 155 exceedances. Of the observations made that did not exceed, 59,872 were correctly reported and 180 were false exceedance reports.

These effects resulted in a sensitivity of 86.23%, meaning that most high concentration NO₂ events were identified. Specificity achieved was 99.70% showing good discrimination of normal condition or non-exceedance condition. The accuracy reached was 84.36% and the F1 score was 85.28%. The false alarm rate was 0.30% with a missed event rate of 13.77%.

The implications of these results are shown in the decision matrix in Figure 5. These high rates of correct non-exceedance classification suggest good routine monitoring performance, and the high sensitivity suggests a good ability to capture environmentally relevant high-NO₂ events by the model. These results combined with the continuous regression analysis demonstrate the usefulness of the chosen XGBoost framework not only for estimating the concentration but also in the conversion of the estimated concentration into environmental-monitoring information that could be used for action.

Table 4. Environmental monitoring performance of the selected XGBoost model for hourly NO₂ exceedance detection.

Monitoring Indicator	Result
Locked-test observations	61,178
Observed exceedance events	1,126
Correctly detected exceedances (TP)	971
Missed exceedances (FN)	155
False exceedance alerts (FP)	180
Correct non-exceedances (TN)	59,872
Sensitivity / Recall	86.23%
Specificity	99.70%
Precision	84.36%
F1-score	85.28%
False-alarm rate	0.30%
Missed-event rate	13.77%

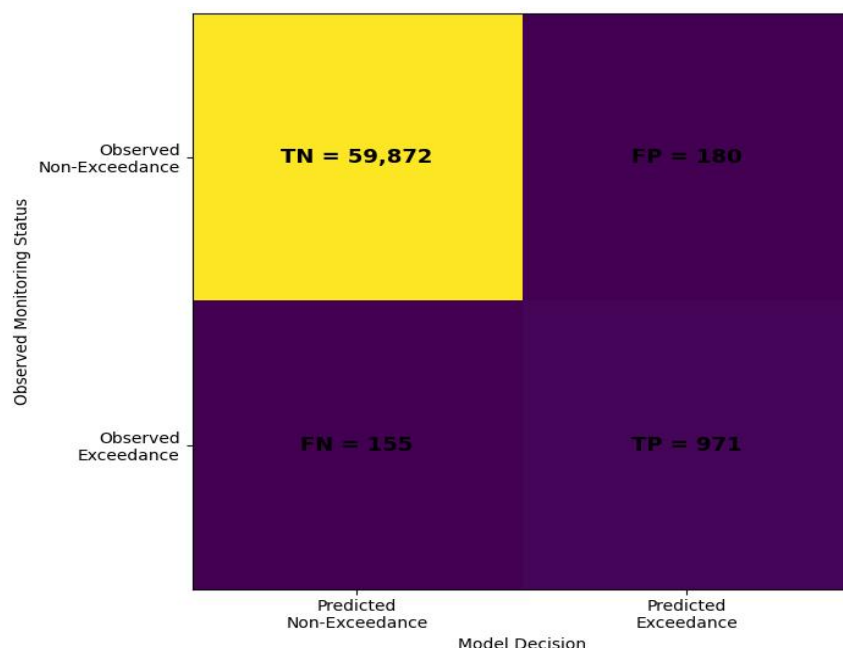


Figure 5. Environmental Monitoring Decision Matrix for NO₂ Exceedance Detection.

5. DISCUSSION

These findings highlight the significant temporal, spatial and atmospheric variability in urban NO₂ behaviour and that the hourly concentrations can be accurately predicted using nonlinear machine-learning models. From the descriptive analysis, it was observed that the mean concentration of NO₂, in the study area, was 50.6 µg/m³, with significant seasonal variations where the highest mean concentration was recorded in winter (62.4 µg/m³) while the lowest mean concentration was recorded in summer (39.8 µg/m³). Spatial heterogeneity was also clearly observed, with the mean of 65.3µg/m³ at the station of Wanliu and the lowest mean of 27.7µg/m³ at the station of Dingling. The patterns indicate that the local environment and seasonal atmospheric processes are significant factors in urban NO₂ variability.

The comparative modelling results also demonstrated that the nonlinear ensemble methods outperformed the linear baseline. The best locked-test result was obtained by XGBoost with an MAE of 7.26 µg/m³, RMSE of 11.63 µg/m³ and R² value of 0.901, significantly better than Ridge Regression outcome. This result aligns with previous studies which demonstrate that machine-learning approaches can be used to capture nonlinear relationships between datasets in NO₂ estimation at large scales [13]. It also reflects studies showing that hybrid and spatiotemporal learning architectures can learn complex dependency between pollutant and atmospheric variables [14]. Also, the success of the ensemble framework suggests that nonlinear interactions

between meteorology, co-pollutants, time and past NO₂ concentrations play an important role in accurate prediction.

The current results also align with other recent studies showing the ability to reliably estimate NO₂ using ensemble machine-learning models when multiple environmental variables are included [15]. Meanwhile, the SHAP analysis also provides an added layer of interpretability as it demonstrates that lagged NO₂, CO, O₃, wind speed, PM_{2.5}, temperature, and PM₁₀ were the most important predictors. This aligns with the wider literature which highlights the importance of interpretable modelling for understanding the drivers of air-quality predictions and testing the model in isolation for accuracy [21].

Environmental-monitoring results are used to further extend the predictive analysis beyond regression performance. The model based on the XGBoost algorithm, with an hourly threshold of 200 µg/m³, had a high sensitivity of 86.23%, a high specificity of 99.70%, a high precision of 84.36%, and a high F1 score of 85.28%. The results show the ability of the model to detect episodes with high NO₂ concentrations with a very low false alarm rate. It is also relevant from an operational point of view as environmental monitoring systems depend on accurate concentration estimates as well as on reliable detection of conditions that may need closer monitoring.

They are subject to certain restrictions. The analysis was limited to one metropolitan area and may not necessarily reflect the conditions of other cities that have various emission patterns, climates, traffic and monitoring systems. The model was also constructed without any consideration for satellite retrievals, traffic intensities, land use, emission inventories, or chemical transport outputs, however these all can be routinely measured at the station level. Furthermore, exceedance detection was assessed based on a fixed threshold that is not reflective of all aspects of regulatory or exposure-related risk.

Future efforts are thus required to explore the framework in various cities, climatic regions, extend the framework by adding external spatial and emissions-related predictors, and apply transfer-learning and domain-adaptation approaches. Other research tasks might also include probabilistic forecasting and uncertainty quantification, or multi-horizon prediction, to enhance operational value. Overall, the results have shown that a leakage-controlled ensemble-learning framework, which is interpretable, can serve as a robust basis for NO₂ concentration prediction and for practical urban environmental monitoring.

6. CONCLUSION

The aim of this study was to fill the important gap of providing accurate, interpretable and operationally relevant prediction of urban nitrogen dioxide (NO₂) concentration in a complex multi-pollutant environment. A data-driven framework, which consisted of descriptive air-quality assessment, chronological data partitioning, comparative machine-learning modelling, SHAP-based interpretation and threshold-based environmental monitoring, was developed. Significant temporal and spatial fluctuations were found for NO₂, with the highest seasonal mean found in winter (62.4 µg/m³) and the highest station-level mean at Wanliu (65.3 µg/m³). The evaluated models indicated that XGBoost had the best generalization performance with a locked-test MAE of 7.26 µg/m³, RMSE of 11.63 µg/m³ and R² of 0.901. Furthermore, the model has been validated with the real data, achieving an 86.23% sensitivity, a 99.70% specificity and an F1-score of 85.28% to detect hourly NO₂ exceedances. The primary contributions of this study are the combination of predictive accuracy, explanations, and exceedance detection in a single reproducible environmental-monitoring framework. The analysis, however, covered only the one metropolitan monitoring network, and excluded data from satellites, traffic, land use, and emission inventories. The framework needs to be validated using more cities, validated with multisource predictors, and investigated with uncertainty-aware and multi-horizon forecasting for wider operational use.

REFERENCES

1. BP. Sicard, E. Agathokleous, S. C. Anenberg, A. De Marco, E. Paoletti, and V. Calatayud, "Trends in urban air pollution over the last two decades: A global perspective," *Science of the Total Environment*, vol. 858, Art. no. 160064, 2023, doi: 10.1016/j.scitotenv.2022.160064.
2. M. Wang, H. Li, S. Huang, Y. Qian, K. Steenland, Y. Xie, S. Papatheodorou, and L. Shi, "Short-term exposure to nitrogen dioxide and mortality: A systematic review and meta-analysis," *Environmental Research*, vol. 202, Art. no. 111766, 2021, doi: 10.1016/j.envres.2021.111766.
3. P. Huangfu and R. Atkinson, "Long-term exposure to NO₂ and O₃ and all-cause and respiratory mortality: A systematic review and meta-analysis," *Environment International*, vol. 144, Art. no. 105998, 2020, doi: 10.1016/j.envint.2020.105998.
4. S. Huang et al., "Long-term exposure to nitrogen dioxide and mortality: A systematic

- review and meta-analysis,” *Science of the Total Environment*, vol. 776, Art. no. 145968, 2021, doi: 10.1016/j.scitotenv.2021.145968.
5. J. Song, Y. Wang, Q. Zhang, W. Qin, R. Pan, W. Yi, Z. Xu, J. Cheng, and H. Su, “Premature mortality attributable to NO₂ exposure in cities and the role of built environment: A global analysis,” *Science of the Total Environment*, vol. 866, Art. no. 161395, 2023, doi: 10.1016/j.scitotenv.2023.161395.
 6. X. Chen, L. Qi, S. Li, and X. Duan, “Long-term NO₂ exposure and mortality: A comprehensive meta-analysis,” *Environmental Pollution*, vol. 341, Art. no. 122971, 2024, doi: 10.1016/j.envpol.2023.122971.
 7. M. Méndez, M. G. Merayo, and M. Núñez, “Machine learning algorithms to forecast air quality: A survey,” *Artificial Intelligence Review*, vol. 56, pp. 10031–10066, 2023, doi: 10.1007/s10462-023-10424-4.
 8. M. Jamei, M. Ali, A. Malik, M. Karbasi, E. Sharma, and Z. M. Yaseen, “Air quality monitoring based on chemical and meteorological drivers: Application of a novel data filtering-based hybridized deep learning model,” *Journal of Cleaner Production*, vol. 374, Art. no. 134011, 2022, doi: 10.1016/j.jclepro.2022.134011.
 9. S. Abirami and P. Chitra, “Regional air quality forecasting using spatiotemporal deep learning,” *Journal of Cleaner Production*, vol. 283, Art. no. 125341, 2021, doi: 10.1016/j.jclepro.2020.125341.
 10. J. Gu, B. Yang, M. Brauer, and K. M. Zhang, “Enhancing the evaluation and interpretability of data-driven air quality models,” *Atmospheric Environment*, vol. 246, Art. no. 118125, 2021, doi: 10.1016/j.atmosenv.2020.118125.
 11. H. Liao, Y. Li, M. Wu, and H. Chen, “Air quality prediction by integrating mechanism model and machine learning model,” *Science of the Total Environment*, vol. 899, Art. no. 165646, 2023, doi: 10.1016/j.scitotenv.2023.165646.
 12. M. Vega García and J. L. Aznarte, “Shapley additive explanations for NO₂ forecasting,” *Ecological Informatics*, vol. 56, Art. no. 101039, 2020, doi: 10.1016/j.ecoinf.2019.101039.
 13. S. Long, X. Wei, F. Zhang, R. Zhang, J. Xu, K. Wu, Q. Li, and W. Li, “Estimating daily ground-level NO₂ concentrations over China based on TROPOMI observations and machine learning approach,” *Atmospheric Environment*, vol. 289, Art. no. 119310, 2022,

doi: 10.1016/j.atmosenv.2022.119310.

14. C.-L. Wu, H.-D. He, R.-F. Song, X.-H. Zhu, Z.-R. Peng, Q.-Y. Fu, and J. Pan, “A hybrid deep learning model for regional O₃ and NO₂ concentrations prediction based on spatiotemporal dependencies in air quality monitoring network,” *Environmental Pollution*, vol. 320, Art. no. 121075, 2023, doi: 10.1016/j.envpol.2023.121075.
15. C. Tao, M. Jia, G. Wang, Y. Zhang, Q. Zhang, X. Wang, Q. Wang, and W. Wang, “Time-sensitive prediction of NO₂ concentration in China using an ensemble machine learning model from multi-source data,” *Journal of Environmental Sciences*, vol. 137, pp. 30–40, 2024, doi: 10.1016/j.jes.2023.02.026.
16. O. A. Alawi, H. M. Kamar, A. Alsuwaiyan, and Z. M. Yaseen, “Temporal trends and predictive modeling of air pollutants in Delhi: A comparative study of artificial intelligence models,” *Scientific Reports*, vol. 14, Art. no. 30957, 2024, doi: 10.1038/s41598-024-82117-z.
17. S. Mandal, S. Boppani, V. Dasari, and M. Thakur, “A bivariate simultaneous pollutant forecasting approach by Unified Spectro-Spatial Graph Neural Network (USSGNN) and its application in prediction of O₃ and NO₂ for New Delhi, India,” *Sustainable Cities and Society*, vol. 114, Art. no. 105741, 2024, doi: 10.1016/j.scs.2024.105741.
18. Q. He, K. Qin, J. B. Cohen, D. Li, and J. Kim, “Quantifying uncertainty in ML-derived atmosphere remote sensing: Hourly surface NO₂ estimation with GEMS,” *Geophysical Research Letters*, vol. 51, no. 18, Art. no. e2024GL110468, 2024, doi: 10.1029/2024GL110468.
19. N. Ahmad, C. Lin, A. K. H. Lau, J. Kim, T. Zhang, F. Yu, C. Li, Y. Li, J. C. H. Fung, and X. Q. Lao, “Estimation of ground-level NO₂ and its spatiotemporal variations in China using GEMS measurements and a nested machine learning model,” *Atmospheric Chemistry and Physics*, vol. 24, pp. 9645–9665, 2024, doi: 10.5194/acp-24-9645-2024.
20. S. Wang, S. Zhang, D. Wang, and W. Li, “Spatial-and-local-aware deep learning approach for ground-level NO₂ estimation in England with multisource data from satellite-based observations and chemical transport models,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 139, Art. no. 104506, 2025, doi: 10.1016/j.jag.2025.104506.

21. L. Xu, B. Xu, Y. Sun, and Y. Cao, "Interpretable and scalable deep learning for urban NO₂ prediction via multisource data," *Transportation Research Part D: Transport and Environment*, vol. 149, Art. no. 105039, 2025, doi: 10.1016/j.trd.2025.105039.
22. M. A. Javadi, M. Zare Shahneh, and Z. Amiri, "Integration of Sentinel-5P satellite data and machine learning for spatiotemporal prediction of NO₂ in Delhi: Impacts of COVID-19 lockdown," *Atmospheric Pollution Research*, vol. 17, no. 1, Art. no. 102702, 2026, doi: 10.1016/j.apr.2025.102702.
23. S. Chen, "Beijing Multi-Site Air Quality," UCI Machine Learning Repository, 2017, doi: 10.24432/C5RK5G.