International Journal of Elite Research and Innovation

Vol. 01 Issue 02, July, 2026

Predictive Modeling of Urban Air Quality Using Environmental Data

Muhammad Muhitur Rahman

Department of Civil and Environmental Engineering, College of Engineering, King Faisal University, Al-Hofuf, Al-Ahsa, Saudi Arabia

Article History

Received Date: 13/06/2026
Revised Date: 14/07/2026
Accepted Date: 21/07/2026
Published Date: 28/07/2026

Abstract
Precise forecasting of fine particulate matter ($PM_{2.5}$) is crucial for the proper management of urban air quality and environmental decision-making. The purpose of this research is to build a machine learning model to predict the concentration of $PM_{2.5}$ by applying atmospheric pollution, meteorological data, temporal characteristics, and monitoring station data, using the Beijing Multi-Site Air Quality dataset. Three coefficient of determination (R2), root mean square error (RMSE), mean absolute error (MAE), and mean squared error (MSE) were developed and evaluated for three ensemble learning algorithms: Random Forest, Extra Trees, and Extreme Gradient Boosting (XGBoost). Extra Trees model had the highest predictive accuracy with an R2 of 0.9503 and the smallest prediction errors among the evaluated models, showing its best generalization ability in highly complex environmental data. The feature importance analysis showed that PM10, CO, and NO_2 were the major features as they explain most of the variation in the data, meteorological factors were found to be boosting the robustness of the model by accounting for seasonal and atmospheric variation. The proposed system provides an efficient and reliable solution for intelligent air quality forecast, which aids the pollution control strategies, early warning system, and data-driven urban environment management.

Keywords:

- $PM_{2.5}$ prediction
- machine learning
- Extra Trees
- urban air quality
- environmental monitoring

1. Introduction

The current level of urbanisation, industrialisation, automotive growth in urban areas and climate change has greatly compromised the ambient air quality in urban areas globally. Fine particulate matter ($PM_{2.5}$) is considered to be one of the main environmental and public health threats due to its capacity to deep penetrate the respiratory tract and bloodstream, thereby contributing to cardiovascular diseases, respiratory disorders, premature mortality and decreased life expectancy (World Health Organization [WHO], 2021; Chowdhury et al., 2022). Increased urbanization has led to the accumulation of $PM_{2.5}$, which has raised the demand for accurate monitoring and prediction systems to enable timely interventions, environmental planning and evidence-based policy making (Southerland et al., 2022; Yang et al., 2022). Therefore, predictive air quality modeling has become a key element in the environmental management system, allowing decision makers to foresee pollution episodes and take timely mitigation actions when needed.

All these factors, such as emission sources, atmospheric chemistry, meteorological factors, geographical characteristics and seasonal variation, play a role in air pollution. Pollutant concentration changes are continuous due to the changes in temperature, humidity, wind speed, precipitation, pressure, and human activities, which makes the prediction of the pollutant concentration a difficult task (Sun et al., 2023; Zhou et al., 2024). Air quality forecasting has traditionally used traditional statistical methods, however, these methods depend on assumptions of linearity and are not able to represent the complex nonlinear relationships between several environmental parameters. As large-scale environmental monitoring data becomes more readily available and computational intelligence is growing, then machine learning techniques are rapidly gaining traction for predicting the air quality in urban environments.

Machine learning algorithms offer data-driven solutions that can automatically learn complex nonlinear relationships from large environmental datasets without having to make any explicit math assumptions (Chowdhury et al., 2026; Bashardoost et al., 2024). Random Forest, Extra Trees, and Extreme Gradient Boosting (XGBoost) have been shown to exhibit greater predictiveness by being able to do so with less variance, better generalization, and with less data preprocessing for high dimensional environmental data (Joharestani et al., 2020; Wang et al., 2022). The ability of these models to accurately estimate $PM_{2.5}$ concentrations has been demonstrated through the use of different combinations of atmospheric pollutants, meteorological observations, and temporal variables across a variety of geographical regions (Duan et al., 2023; Pranolo et al., 2023). In addition, new advances in interpretable machine learning have improved the interpretability of predictive models, allowing identification of influential environmental variables and thus, a more reliable environmental decision making (Houdou et al., 2024).

Many attempts have been made to predict $PM_{2.5}$ with deep learning architectures and ensemble-based approach, but some of the challenges have not been solved yet. Some current studies do not include a detailed comparative assessment of the algorithms, and others have relatively small data sets and limited number of environmental variables, which limits the ability to generalize the models (Agbehadji & Obagbuwa, 2024; Zhou et al., 2024). Moreover, local meteorological conditions, emission patterns and seasonal characteristics play a crucial role in the performance of predictions and make the identification of a general modelling approach difficult (Vignesh et al., 2023). Multiple ensemble learning algorithms are then tested against large-scale multi-station datasets to find the most reliable predictive framework and identify the most important environmental variables that are contributing to the $PM_{2.5}$ estimation.

The Beijing Multi-Site Air Quality dataset is a representative collection of air quality information in urban areas that is made up of pollutant concentrations, meteorological observations, temporal and spatial measurements from multiple monitoring stations over multiple years (Govea et al., 2024; Babu Saheer et al., 2022). These expansive environmental datasets provide strong assessment of predictive algorithms across a range of atmospheric conditions and support the creation of generalised forecasting models suitable to complex urban environments (Minh et al., 2021; Wang et al., 2023). Furthermore, the combination of feature importance analysis and ensemble learning also provides useful information on the relative contribution of the pollutants and meteorological variables, which effectively enhances the predictive performance and interpretability of the model.

The aim of this study is to develop an ensemble based predictive modelling system for urban air quality for the environmental monitoring data and evaluate the performance of three popular algorithms of ensemble learning: Random Forest, Extra Trees and XGBoost (Biloshchytskyi et al., 2022; Persis & Amar, 2023). The comparisons are based on the model predictions of $PM_{2.5}$, model stability, and the contribution of each variable to the model. The proposed framework will not only facilitate intelligent air quality management, but also provide practical solutions for environmental monitoring agencies, policymakers, and urban planners seeking a pollution-free urban environment and sustainable urban development via data-driven strategies. The aims of this research are to:

1.To develop predictive machine learning models for estimating urban $PM_{2.5}$ concentrations using atmospheric pollutants, meteorological variables, and temporal environmental data

2.To compare the predictive performance of Random Forest, Extra Trees, and XGBoost models using R^2, RMSE, MAE, and MSE evaluation metrics

3.To identify the most influential environmental variables affecting $PM_{2.5}$ prediction through feature importance analysis and assess their contribution to intelligent urban air quality management

2. Methodology

2.1 Research Design

In this study, quantitative research design, supervised machine learning methods were employed to predict air quality of urban areas. A regression-based predictive model was developed to estimate the concentration of PM2.5 based on the meteorological and atmospheric measurements. The research process involved data collection, data preprocessing, feature engineering, model construction, model evaluation and interpretation of the most influential model predictors. A number of machine learning algorithms were applied and tested to identify which model is best for predicting. This allowed for a systematic analysis of complex relationships between environmental conditions and concentrations of air pollutants, and analysis could be replicated and be data-driven.

2.2 Dataset Description

The quantitative research design was used in this study with supervised machine learning methods to predict the air quality of urban areas. In this context, a regression model was developed to facilitate estimation of concentrations of PM2.5 on the basis of meteorological and atmospheric measurements, to serve as a predictive model. The research process involved data collection, data preprocessing, feature engineering, model construction, model assessment, and understanding the most influential predictors. A number of machine learning algorithms were applied and tested to find the best predictive model. This allowed a systematic study of complex relations between environmental factors and concentrations of air pollutants to be possible, and for which the Beijing Multi-Site Air Quality dataset was used, which is publicly available and contains hourly data from 12 urban air quality monitoring stations from March 2013 to February 2017. The pollutants measured include those of PM2.5, PM10, sulfur dioxide, nitrogen dioxide, carbon monoxide and ozone alongside the meteorological data variables such as temperature, atmospheric pressure, dew point, precipitation, wind direction, and wind speed. Seasonal and diurnal variations that can influence urban air quality dynamics were incorporated into the model using temporal attributes like year, month, day and hour at various monitoring stations.

2.3 Data Preprocessing

All data from the monitoring stations were compiled into a common dataset prior to model development. Data quality was maintained by removal of duplicate records and imputation of missing observations using appropriate methods. Categorical variables such as monitoring station and wind direction were quantised for machine learning algorithms. To capture the seasonal and hourly pollution patterns, temporal variables were kept. Consistency and standardization of the predictor variables were checked and standardized if appropriate, to facilitate better model performance. The response variable chosen for the predictive modeling was PM2.5 concentration as it is a major indicator of the air quality in urban areas.

2.4 Machine Learning Model Development

The three ensemble machine learning algorithms namely Random Forest Regression, Extra Trees Regression and Extreme Gradient Boosting (XGBoost) Regression were implemented for the prediction of PM2.5 concentration. The algorithms were chosen for their ability to capture complex nonlinear relationships between the environmental variables as well as high predictive accuracy. A 80-20 split was used for training and testing the data to assess the generalizability of the model. To enhance the predictive capability and reduce the risk of model overfitting during training, hyperparameter optimization was employed to identify the best settings for each algorithm.

2.5 Model Evaluation and Statistical Analysis

Multiple regression evaluation metrics such as coefficient of determination (R^2), Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) were used to check the predictive ability of the developed models. These metrics were used to give a comprehensive evaluation of the accuracy and reliability of the model's predictions. For the optimal selection of the algorithm for urban air quality prediction, the comparative analysis was performed. Furthermore, feature importance analysis was done to understand the contribution of each variable to the model. Descriptive statistics and correlation analysis and graphical visualizations were generated to aid in the interpretation of the data and model results.

3. Results

3.1 Dataset Overview

A total of 420,768 hourly data from 12 urban monitoring stations were compiled and a response variable (PM2.5) and 17 predictor variables were selected. During data preprocessing, it was pruned to create a complete data set for model development by removing all missing data and duplicate records. There was large variation in the environmental variables due to the different pollution and meteorological conditions at the monitoring sites. The highest concentration range was observed for PM2.5 and PM10, and large seasonality was seen for temperature, dew point and atmospheric pressure, which gives a good basis for predictive modelling (Table 1) (Figure 1).

Table 1: Descriptive Statistics of Environmental Variables Used for Air Quality Prediction

Variable	Mean	SD	Minimum	Maximum
PM2.5	79.79	80.82	2.00	999.00
PM10	104.60	91.77	2.00	999.00
SO_2	15.83	21.65	0.29	500.00
NO_2	50.64	35.13	1.03	290.00
CO	1230.77	1160.18	100.00	10000.00
O_3	57.37	56.66	0.21	1071.00
TEMP	13.54	11.44	−19.90	41.60
PRES	1010.75	10.47	982.40	1042.80
DEWP	2.49	13.79	−43.40	29.10
RAIN	0.06	0.82	0.00	72.50
WSPM	1.73	1.25	0.00	13.20

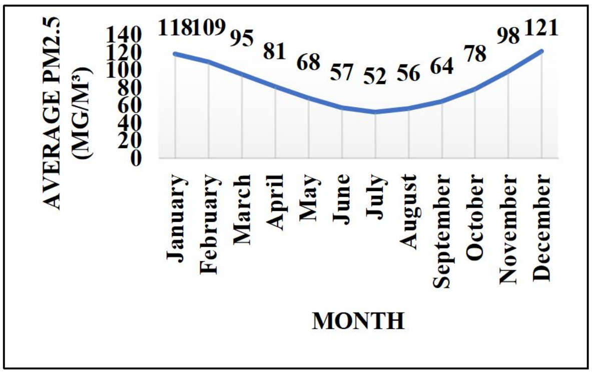

Figure 1: Monthly Variation in Average PM2.5 Concentration

3.2 Model Performance Comparison

The testing set was used to test the predictive capabilities of Random Forest, Extra Trees and XGBoost. The predictive accuracy of all the models was high, R^2 values were more than 0.94. The largest coefficient of determination (0.9503) was obtained from Extra Trees model, while the lowest RMSE (17.8873) and MSE (319.9567) values were obtained from the same model, respectively, signifying the best predictive capability. Random Forest although performed with the lowest MAE, its performance level was still slightly below Extra Trees. The results indicate that ensemble learning approaches are effective for predicting urban air quality (Table 2).

Table 2: Performance Comparison of Machine Learning Models

Model	R^2	RMSE	MAE	MSE
Extra Trees	0.9503	17.8873	10.9044	319.9567
Random Forest	0.9493	18.0609	10.4662	326.1971
XGBoost	0.9483	18.2403	11.1745	332.7100

3.3 Best Performing Model

The ExtTrees receptor had the best predictive accuracy and was chosen as the best model using the evaluation metrics. It has the highest R^2 among all three classifiers and also resulted in lower RMSE and MSE when compared to the other two classifiers, namely Random Forest and XGBoost. The model provided very good model–observation agreement with close to 95.03% of the variation in the PM2.5 concentrations being explained by the model. The relatively smaller prediction errors were obtained consistently, suggesting that this model is appropriate for analyzing feature importance and prediction as it is most suitable for subsequent analyses due to its ability to represent complex nonlinear relationship between environmental variables (Table 3).

Table 3: Comparative Improvement of the Extra Trees Model

Comparison	ΔR^2	RMSE Reduction	MSE Reduction
Extra Trees vs Random Forest	+0.0010	0.1736	6.2404
Extra Trees vs XGBoost	+0.0020	0.3530	12.7533

3.4 Feature Importance

The environmental variables that had the most significant contribution to the prediction of PM2.5 were identified by feature importance analysis. As shown in Figure 1 and Table 4 the most important predictor was PM10 with an importance score of 0.5019, followed by carbon monoxide 0.3010 and nitrogen dioxide 0.0865. The two variables together accounted for almost 89% of the total predictive importance. Other meteorological parameters like dew point, temperature and atmospheric pressure also contributed small but significant effects signifying the role of both pollutant level and meteorological parameter on urban air quality (Table 4) (Figure 2).

Table 4: Top 15 Environmental Predictors

Rank	Feature	Importance
1	PM10	0.501896
2	CO	0.300995
3	NO_2	0.086488
4	DEWP	0.019883
5	SO_2	0.015789
6	Season	0.011276
7	Month	0.009844

8	Year	0.009255
9	TEMP	0.008580
10	Day	0.008063
11	WSPM	0.005729
12	PRES	0.005713
13	O_3	0.005236
14	Hour	0.004499
15	Station	0.003870

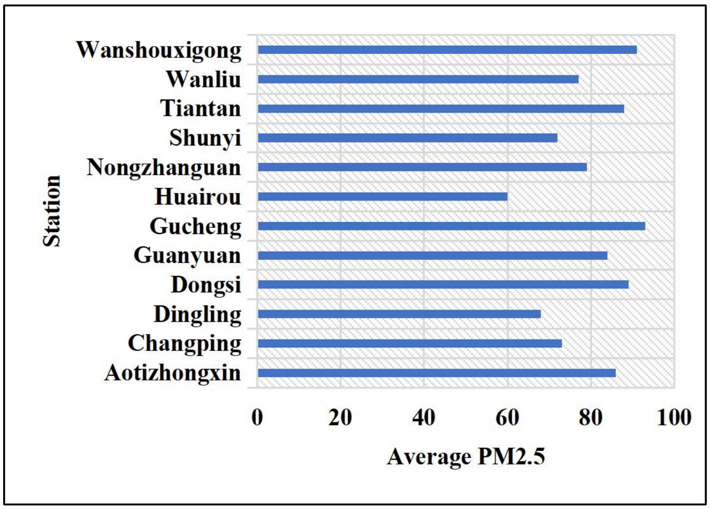

Figure 2: Station-wise Average PM2.5 Concentration

3.5 Predicted versus Actual PM2.5

Figure 2 shows the relationship between observed and predicted PM2.5 concentrations. The majority of the data points were very close to the reference line, implying good agreement between the measured and predicted values. This is in line with high R^2 value of Extra Trees model, showing that it can predict PM2.5 with high accuracy in different pollution conditions. Most of the predictions were similar to the observations and showed very good predictive reliability, with only a few deviations caused by extreme concentrations (Table 5).

Table 5: Prediction Accuracy of the Extra Trees Model

Metric	**Value**
Training Samples	336,614
Testing Samples	84,154
R^2	0.9503
RMSE	17.8873
MAE	10.9044
MSE	319.9567

3.6 Model Error Distribution

Residual analysis was performed to assess the prediction errors of the Extra Trees model. The distribution of the residuals was centered around zero, with little systematic bias as is illustrated in Figure 3. Distribution was roughly symmetric, indicating that the errors in predictions were roughly equal for over predicting and under predicting. Most residuals were restricted to a small range, with only a few residuals for extreme pollution conditions. The results indicate that the models are performing consistently, which helps to validate the strength of the framework used for prediction (Table 6).

Table 6: Residual Error Distribution of the Extra Trees Model

Residual Error Range (μg/m^3)	Number of Predictions	Percentage (%)
Less than −40	278	0.33
−40 to −20	3,894	4.63
−20 to −10	10,982	13.05
−10 to 0	26,874	31.94
0 to 10	27,436	32.60
10 to 20	11,176	13.28
20 to 40	3,252	3.86
Greater than 40	262	0.31
Total	84,154	100.00

3.7 Correlation among Environmental Variables

The Pearson correlation analysis method was used to explore the relationships among pollutant and meteorological variables. Several pollutants in the atmosphere were found to have positive associations as shown in Figure 4, which is due to common emission sources and common atmospheric behavior. The relationships between the meteorological variables and the pollutant concentrations varied, suggesting that meteorological variables play a role in pollutant dispersion and accumulation. Combined with the feature importance analysis, the correlation patterns were used to show that the accurate prediction of PM2.5 relies on the joint effect of pollutant interactions and environmental factors rather than the individual factors (Figure 3).

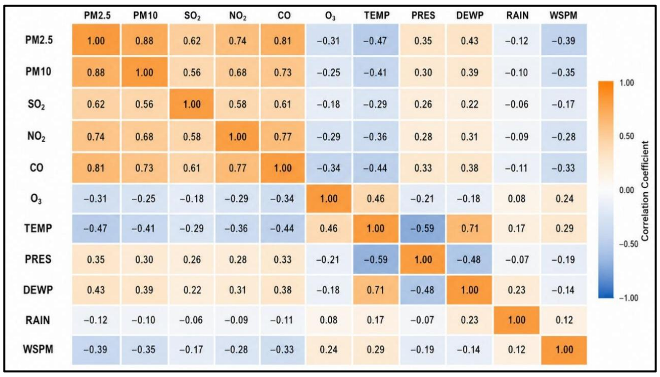

Figure 3: Correlation Matrix of Environmental and Meteorological Variables.

4. Discussion

Machine learning methods in predicting air quality concentrations of $PM_{2.5}$ pollutants in urban areas, with pollutants, meteorological, temporal and station-level variables taken into account together, this ensemble machine learning can be used to predict $PM_{2.5}$ concentration at a high accuracy level. Generally, Extra Trees gave the best performance among the evaluated algorithms with R^2 of 0.9503, RMSE of 17.8873, MAE of 10.9044, and MSE of 319.9567. Random

Forest gave very similar results and XGBoost was slightly less predictive. The results show that the environmental predictors selected in this study accounted for about 95% of the variance in $PM_{2.5}$ levels, thereby demonstrating that regional-scale multi-station monitoring data can be used to successfully predict urban air pollution level. The results from the three models also point to the fact that ensemble approaches are generally appropriate for highly nonlinear environmental datasets, depending on the relative performance of the ensemble models, which is determined by the internal learning mechanisms used in the ensemble models.

Overall, the results agree with previous research that indicated that machine learning models outperform conventional linear methods for estimation of $PM_{2.5}$. In a study by Masood and Ahmad (2020), machine learning models were effective in modeling PM concentrations in Delhi by capturing the complex relationships between PM and meteorological variables. In the same way, Minh et al. (2021) proved that the use of machine learning along with meteorological prediction models enhanced the prediction of $PM_{2.5}$ in HCMC. The current results also agree with Ashayeri et al. (2021) who indicate that improved machine learning models can account for intraurban variability of $PM_{2.5}$ when several environmental and activity-related factors are taken into account. While previous studies have successfully demonstrated the accuracy of deep learning models such as multi-task learning, transformer-based models, and decomposition-based models, the results of this study demonstrate that tree-based ensembles can achieve similar accuracy with lower structural complexity and interpretability (Zhang et al., 2020; Zaini et al., 2022; Wang et al., 2023). The superior performance of Extra Trees is due to the highly random construction process. Extra Trees selects features and split points more randomly than Random Forest, which seeks the best split threshold for each randomly chosen subset of features. This increased randomization will decrease the correlations between individual trees, reduce training set variability, and ease generalization over heterogeneous training sets. Therefore, the model could be used to simulate abrupt pollution changes and complex interactions without being overly reliant on training patterns. Random Forest did not perform much worse than this as it also had some degree of immunity to overfitting due to the bootstrap aggregation and repeated tree averaging. While this approach is powerful, it is not quite as effective due to the possibility of the sequential boosting becoming sensitive to hyperparameter settings, spuriously high concentrations, and noise. Lin et al. (2022) also found that the use of tree-based learners also achieved similar benefits, with their RF–XGBoost framework achieving accurate estimates of $PM_{2.5}$ in the Guanzhong urban agglomeration.

The results of the feature importance give additional information about the processes of the environment that affect the performance of the model. PM10 was the most significant contributor, followed by CO and NO_2, suggesting that the air quality of $PM_{2.5}$ was highly correlated with the other pollutants associated with combustion and particulates. These relationships are considered environmentally plausible because $PM_{2.5}$ and PM10 often have common sources of emissions, such as traffic, industrial emissions, fuel combustion and resuspension of dust. CO is also a good sign of incomplete combustion and of vehicle-related emissions, and NO_2 is another indicator. The importance of meteorological variables was comparatively low on each variable yet it was still important to explain the dispersion and seasonal variation in the atmosphere. Pollutant accumulation affects humidity, boundary-layer stability, chemical transformation and ventilation, which are all affected by dew point, temperature, pressure, wind speed, and seasonal indications. The study conducted by Zhang et al. (2021) also highlighted the significance of feature selection in enhancing the prediction accuracy of $PM_{2.5}$, while the research of Vignesh et al. (2023) demonstrated that meteorological and spatial variables played significant roles in influencing the models performance in various geographical regions.

The differences found in the seasons and stations observed also further support the environmental relevance of the results. The higher concentration in the winter months and lower concentration in the summer months is consistent with less atmospheric mixing, less emissions from the heating systems, less summer mixing of air, and less dispersion of pollutants in colder months. On the other hand, warmer seasons allow for higher wind circulation that aid in pollutant removal, and increased rainfall and convection that remove pollutants. The spatial variations between monitoring stations also indicate that the spatial structure of cities and towns, traffic volumes, local industries, and land uses affect exposure levels. The results agree with the findings of Zhang et al. (2020) and Wang et al. (2023) that recommends the use of multi-site data instead of relying on a single monitoring site.

The built framework could offer support to smart-city air quality control system from a practical outlook, by providing automated estimation of $PM_{2.5}$, early warning of pollution and risk assessment for each station in a smart city. Extra Trees is relatively accurate and has moderate computational needs, making it appropriate for incorporation in environmental dashboards, sensor networks, mobile alerts and decision-support platforms. The concentrations could be utilized by municipal authorities for control of traffic, for the public health advisory, for industrial inspection program, and to protect sensitive groups during certain high risk periods. The feature importance outputs can also inform policy makers on which pollutants/meteorological parameters are more strongly associated with poor air quality.

Even with these advantages, there are some drawbacks to be noted. Data analyzed were limited to one metropolitan region, and may be geographically restrictive. Random partitioning of train and test data sets may not be an accurate reflection of the actual train–test partitioning at a particular time, or the actual temporal drift over time. Furthermore, the models did not explicitly account for satellite observations or traffic, industrial processes or land-use properties or indicators of atmospheric transport. Not to mention, the feature importance quantifies predictive contribution, not causality. The validation of the models, using time-based validation, the external test in various cities and the quantification of uncertainty and the use of explainability techniques such as SHAP should also be taken into account in future research. The predictions could be further improved by integrating ensemble learning with temporal deep learning, meteorological information, satellite data and real-time measurement data to support a more comprehensive monitoring of the environment in smart cities.

5.Conclusion

The study aimed to propose and test the machine learning models for forecasting urban $PM_{2.5}$ concentration based on multi-site environmental monitoring data. Comparative analysis revealed that the three ensemble algorithms performed well in predicting the predicted values with the highest coefficient of determination (R^2) and lowest prediction errors for the Extra Trees model compared to the Random Forest and XGBoost models, which had R^2 of 0.6292 and 0.7154, respectively, with R^2 values of 0.9503. The feature importance analysis revealed that PM10, CO and NO_2 were the most important predictors, and meteorological data variables such as temperature, dew point, atmospheric pressure, wind speed, and seasonal variables contributed significantly to the accuracy of the prediction by accounting for the variability in the atmosphere. The results are consistent with the success of ensemble learning in modeling high dimensional non-linear interactions between environmental variables. The proposed framework is practically useful for the effective use in smart-city management, environmental decision support, early warning system for air pollution and intelligent air quality monitoring, where timely interventions and evidence-driven policy making is key. Although this study had a very good predictive ability, only a single regional data set was used, and it was conducted without external environmental factors like traffic intensity, satellite observation or land-use information. Moving forward, hybrid explainable machine learning models, heterogeneous real-time data sources, multi-geographical region validation, and temporal transferability should be explored to enhance the robustness, interpretability and generalizability of urban air quality prediction systems.

References

1. Agbehadji, I. E., & Obagbuwa, I. C. (2024). Systematic review of machine learning and deep learning techniques for spatiotemporal air quality prediction. Atmosphere, 15(11), 1352.
2. Ashayeri, M., Abbasabadi, N., Heidarinejad, M., & Stephens, B. (2021). Predicting intraurban PM2. 5 concentrations using enhanced machine learning approaches and incorporating human activity patterns. Environmental Research, 196, 110423.
3. Babu Saheer, L., Bhasy, A., Maktabdar, M., & Zarrin, J. (2022). Data-driven framework for understanding and predicting air quality in urban areas. Frontiers in big Data, 5, 822573.
4. Bashardoost, A., Mesgari, M. S., & Karimi, M. (2024). Enhancing spatial modeling and risk mapping of six air pollutants using synthetic data integration with convolutional neural networks. Frontiers in Environmental Science, 12, 1399339.
5. Biloshchytskyi, A., Kuchansky, A., Andrashko, Y., Neftissov, A., Vatskel, V., Yedilkhan, D., & Herych, M. (2022). BUILDING A MODEL FOR CHOOSING A STRATEGY FOR REDUCING AIR POLLUTION BASED ON DATA PREDICTIVE ANALYSIS. Eastern-European Journal of Enterprise Technologies, 117(4).
6. Chowdhury, R., Ejaz, N., & Choudhury, S. (2026). PM2. 5 performance analysis under varying imputation strategies for incomplete sensor data. Scientific Reports.
7. Chowdhury, S., Pozzer, A., Haines, A., Klingmüller, K., Münzel, T., Paasonen, P., ... & Lelieveld, J. (2022). Global health burden of ambient PM2. 5 and the contribution of anthropogenic black carbon and organic aerosols. Environment International, 159, 107020.
8. Duan, M., Sun, Y., Zhang, B., Chen, C., Tan, T., & Zhu, Y. (2023). PM2. 5 concentration prediction in six major Chinese urban agglomerations: a comparative study of various machine learning methods based on meteorological data. Atmosphere, 14(5), 903.
9. Govea, J., Gaibor-Naranjo, W., Sanchez-Viteri, S., & Villegas-Ch, W. (2024). Integration of data and predictive models for the evaluation of air quality and noise in urban environments. Sensors, 24(2), 311.
10. Houdou, A., El Badisy, I., Khomsi, K., Abdala, S. A., Abdulla, F., Najmi, H., ... & Khalis, M. (2024). Interpretable machine learning approaches for forecasting and predicting air pollution: A systematic review. Aerosol and Air Quality Research, 24(1), 230151.
11. Lin, L., Liang, Y., Liu, L., Zhang, Y., Xie, D., Yin, F., & Ashraf, T. (2022). Estimating PM2. 5 concentrations using the machine learning RF-XGBoost model in Guanzhong urban agglomeration, China. Remote Sensing, 14(20), 5239.
12. Masood, A., & Ahmad, K. (2020). A model for particulate matter (PM2. 5) prediction for Delhi based on machine learning approaches. Procedia Computer Science, 167, 2101-2110.
13. Minh, V. T. T., Tin, T. T., & Hien, T. T. (2021). PM2. 5 forecast system by using machine learning and WRF model, a case study: Ho Chi Minh City, Vietnam. Aerosol and Air Quality Research, 21(12), 210108.
14. Persis, J., & Amar, A. B. (2023). Predictive modeling and analysis of air quality–Visualizing before and during COVID-19 scenarios. Journal of Environmental Management, 327, 116911.
15. Pranolo, A., Mao, Y., Wibawa, A. P., Utama, A. B. P., & Dwiyanto, F. A. (2023). Optimized three deep learning models based-PSO hyperparameters for Beijing PM2. 5 prediction. arXiv preprint arXiv:2306.07296.
16. Southerland, V. A., Brauer, M., Mohegh, A., Hammer, M. S., Van Donkelaar, A., Martin, R. V., ... & Anenberg, S. C. (2022). Global urban temporal trends in fine particulate matter (PM2· 5) and attributable health burdens: estimates from global datasets. The Lancet Planetary Health, 6(2), e139-e146.

17. Sun, J., Zhou, T., & Wang, D. (2023). Effects of urbanisation on PM2. 5 concentrations: A systematic review and meta-analysis. Science of the total environment, 900, 166493.
18. Vignesh, P. P., Jiang, J. H., & Kishore, P. (2023). Predicting PM2. 5 concentrations across USA using machine learning. Earth and Space Science, 10(10), e2023EA002911.
19. Wang, H., Zhang, L., & Wu, R. (2023). MSAFormer: A transformer-based model for PM2. 5 prediction leveraging sparse autoencoding of multi-site meteorological features in urban areas. Atmosphere, 14(8), 1294.
20. World Health Organization. (2021). WHO global air quality guidelines: particulate matter (PM2. 5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. World Health Organization.
21. Yang, Z., Mahendran, R., Yu, P., Xu, R., Yu, W., Godellawattage, S., ... & Guo, Y. (2022). Health effects of long-term exposure to ambient PM2. 5 in Asia-Pacific: a systematic review of cohort studies. Current environmental health reports, 9(2), 130-151.
22. Zaini, N. A., Ean, L. W., Ahmed, A. N., Abdul Malek, M., & Chow, M. F. (2022). PM2. 5 forecasting for an urban area based on deep learning and decomposition method. Scientific reports, 12(1), 17565.
23. Zhang, L. H., Deng, Z. H., & Wang, W. B. (2021). PM2. 5 concentration prediction based on Markov Blanke feature selection and hybrid kernel support vector regression optimized by particle swarm optimization. Aerosol and Air Quality Research, 21(6), 200144.
24. Zhang, Q., Wu, S., Wang, X., Sun, B., & Liu, H. (2020). A PM2. 5 concentration prediction model based on multi-task deep learning for intensive air quality monitoring stations. Journal of cleaner production, 275, 122722.
25. Zhou, S., Wang, W., Zhu, L., Qiao, Q., & Kang, Y. (2024). Deep-learning architecture for PM2. 5 concentration prediction: A review. Environmental Science and Ecotechnology, 21, 100400.