

Data-Driven Identification of Urban Air Pollution Patterns Using Predictive Machine Learning Models

Dr. Fatima Hassan

Professor, East African University of Technology, Kenya

Abstract

Urban air pollution represents a significant environmental concern because atmospheric contaminants exhibit complex variations across geographical locations, seasonal conditions, and periods of human activity. Pollutants such as fine particulate matter, nitrogen dioxide, and ground-level ozone are influenced by interacting processes involving emissions, atmospheric transport, meteorological conditions, and chemical transformations. Conventional statistical approaches provide useful methods for examining pollutant concentrations but may encounter difficulties when environmental relationships are nonlinear, temporally dependent, or spatially heterogeneous. Data-driven machine learning offers computational approaches for identifying pollution patterns and predicting changes in atmospheric pollutant concentrations. The present paper develops a hypothetical methodological framework for examining urban air pollution through environmental data analysis and predictive machine learning. The proposed investigation incorporates pollutant monitoring records, meteorological variables, unsupervised clustering, temporal feature construction, and supervised regression models. A fictional environmental dataset containing 6,570 daily observations across six monitoring stations is introduced to demonstrate the intended analytical procedure.

Fine particulate matter, nitrogen dioxide, and ozone are selected as indicators of urban atmospheric conditions. Four predictive configurations are considered: previous-day persistence, Ridge Regression, Random Forest, and Gradient Boosting. The proposed methodology combines pollution-pattern identification with next-day concentration prediction for fine particulate matter and nitrogen dioxide. Two illustrative analytical tables present hypothetical pollutant profiles and comparative predictive errors. All numerical values are manually constructed and do not represent collected environmental observations, completed computational experiments, or verified urban pollution trends. The methodological discussion examines pollutant variability, meteorological interactions, monitoring limitations, computational reliability, and temporal generalization. Particular attention is given to distinguishing predictive relationships from causal explanations of pollution sources. The paper concludes that machine learning provides a relevant methodological foundation for examining urban air-quality information, although reliable conclusions require representative monitoring records, appropriate validation, and independently verified experimental evidence.

Keywords: Urban Air Pollution, Predictive Machine Learning, Air Quality Forecasting, Environmental Data Analytics, PM2.5, Nitrogen Dioxide, Ozone, Random Forest, Gradient Boosting, Pollution Pattern Recognition.

1. Introduction

Urban air pollution has become an important environmental research concern because atmospheric contaminants influence air quality within densely populated and industrialized regions. Urban environments contain multiple potential pollution sources associated with transportation, commercial activities, industrial operations, construction processes, and residential energy use. The resulting atmospheric conditions are influenced by both emissions and environmental processes. Pollutant concentrations can vary substantially across monitoring locations and observation periods, making comprehensive interpretation difficult. Consequently, researchers increasingly employ environmental monitoring systems and computational analysis to investigate how pollutant concentrations change and how these changes may be represented through measurable statistical patterns.

Urban atmospheric conditions are influenced by interactions among particulate matter, gaseous pollutants, meteorological variables, and geographical characteristics. Fine particulate matter may originate from direct emissions or form through atmospheric chemical processes. Nitrogen dioxide is associated with combustion-related emissions and atmospheric transformations, while ground-level ozone develops through photochemical reactions involving precursor substances. Temperature, wind speed, humidity, and atmospheric stability can influence pollutant dispersion and accumulation. These interacting processes produce complex concentration patterns that may not be adequately represented through simple statistical relationships. Understanding such patterns requires analytical procedures capable of examining multiple environmental variables together.

Traditional air-quality analysis frequently involves concentration averaging, descriptive statistics, temporal comparisons, and statistical regression. These approaches provide useful information about pollutant behavior and relationships with selected environmental conditions. However, urban pollution records often contain nonlinear dependencies, serial correlation, missing measurements, and differences among monitoring locations. A statistical model developed using observations from one period may perform inconsistently during another period because meteorological conditions and emission patterns change. Consequently, reliable pollution analysis requires careful data preparation, appropriate temporal validation, and consideration of the differences between historical association and future prediction.

Machine learning provides alternative methods for examining large environmental datasets and identifying complex relationships among explanatory variables. Random Forest and Gradient Boosting can represent nonlinear relationships and interactions without requiring a single predefined mathematical structure for every environmental process. Unsupervised clustering may also support the identification of recurring combinations of pollutant concentrations. However, clustering and prediction address different analytical objectives. Clustering summarizes similarities among observations, whereas predictive regression estimates future pollutant values.

Combining these procedures may provide a structured approach to environmental interpretation when their methodological roles remain clearly distinguished. The present paper develops a hypothetical investigation of urban air pollution pattern identification using data science and predictive machine learning. The proposed framework considers historical observations of fine particulate matter, nitrogen

dioxide, ozone, and meteorological conditions across six fictional urban monitoring stations. Unsupervised clustering is used conceptually to identify recurring pollution profiles, while supervised regression models are proposed for next-day pollutant prediction. Two illustrative tables demonstrate possible reporting procedures. No actual atmospheric observations were analyzed, and the numerical examples are included exclusively to demonstrate a reproducible research design for future empirical investigation.

2. Objectives of the Study

The primary objective of this study is to examine the methodological application of data-driven machine learning for identifying and predicting urban air pollution patterns. The investigation focuses on pollutant variability, environmental feature relationships, and forecasting reliability. Three specific objectives establish the analytical direction of the proposed research framework.

2.1 Identification of Recurring Urban Pollution Patterns

The first objective is to examine how historical air-quality measurements can be organized to identify recurring combinations of urban pollutant concentrations. Fine particulate matter, nitrogen dioxide, and ozone are considered important environmental indicators because they reflect different aspects of atmospheric pollution. The proposed investigation evaluates how multivariate observations could be grouped according to similarities in pollutant profiles. Particular attention is given to distinguishing relatively low concentration conditions from situations involving elevated levels of individual pollutants or several contaminants simultaneously.

The investigation further aims to examine how meteorological conditions may contribute to differences among observed pollution patterns. Wind speed, temperature, humidity, and seasonal variation can influence pollutant transport, transformation, and accumulation. Consequently, similar concentration levels may arise under different atmospheric conditions. The proposed framework considers unsupervised clustering as a descriptive analytical procedure rather than a method for establishing pollution sources. This objective emphasizes the importance of interpreting cluster characteristics alongside environmental context and recognizing the limitations of statistical groupings.

2.2 Comparative Evaluation of Predictive Machine Learning Models

The second objective focuses on developing a comparative framework for predicting daily urban pollutant concentrations. Previous-day persistence, Ridge Regression, Random Forest, and Gradient Boosting are selected to represent alternative forecasting strategies. The intended comparison considers their ability to estimate next-day fine particulate matter and nitrogen dioxide concentrations. Mean absolute error and root mean squared error are used as principal evaluation metrics. The proposed framework seeks to establish consistent procedures for comparing prediction accuracy across different statistical and machine learning methods.

A further objective involves examining whether nonlinear predictive models provide meaningful advantages over simpler forecasting approaches. Urban atmospheric relationships may involve interactions among pollutant history, seasonal conditions, and meteorological variables. However, complex machine learning models do not necessarily outperform simpler statistical methods under every environmental condition. The proposed investigation therefore emphasizes chronological testing,

consistent feature preparation, and appropriate parameter optimization. This objective supports a balanced comparison based on reproducible predictive evidence rather than assumptions regarding algorithmic sophistication.

2.3 Examination of Environmental Data and Prediction Reliability

The third objective is to identify methodological factors that may influence the reliability of urban pollution pattern identification and forecasting. Missing observations, measurement uncertainty, inconsistent monitoring practices, and geographical variation can complicate the development of predictive models. Temporal dependence may also produce misleading results when conventional random data partitioning is applied. The proposed study examines how data quality assessment, feature construction, and chronological validation could address selected difficulties while preserving meaningful environmental information.

The investigation also considers the limitations of interpreting machine learning predictions within urban environmental research. Statistical relationships may support concentration forecasting without identifying the actual sources or physical causes of pollution. Model reliability may also differ between ordinary conditions and unusual pollution episodes. Consequently, the proposed framework emphasizes appropriate uncertainty analysis, independent evaluation, and transparent interpretation. This objective establishes the importance of separating predictive performance, environmental description, and causal explanations within data-driven air-quality research.

3. Review of Literature

3.1 Development of Statistical and Neural Air Pollution Forecasting

Prybutok et al. (2000) compared neural networks with autoregressive and regression approaches for predicting daily maximum ozone concentrations. Their investigation established an early methodological basis for comparing statistical forecasting procedures with nonlinear computational models. Kukkonen et al. (2003) subsequently evaluated neural networks for predicting nitrogen dioxide and particulate matter concentrations in central Helsinki. Their work emphasized the influence of meteorological variables and the importance of appropriate comparison methods. These contributions demonstrated that urban air-quality forecasting requires careful consideration of pollutant characteristics and predictive model structure.

Corani (2005) investigated neural network and local learning approaches for forecasting ozone and particulate matter concentrations in Milan. Hooyberghs et al. (2005) examined neural forecasting of daily particulate matter concentrations in Belgium. Grivas and Chaloulakou (2006) evaluated artificial neural networks for hourly particulate matter prediction in Athens, while Pérez and Reyes (2006) developed an integrated neural forecasting approach for PM₁₀. Collectively, these studies demonstrated the relevance of computational learning methods to urban pollution prediction while emphasizing the importance of meteorological information and model evaluation.

3.2 Environmental Data Integration and Pollution Pattern Analysis

Voukantsis et al. (2011) investigated principal component analysis and neural network forecasting using air-quality observations from Thessaloniki and Helsinki. Their research demonstrated how multivariate statistical analysis can support comparisons of pollutant patterns across urban environments. Carslaw

and Ropkins (2012) introduced openair as a computational resource for analyzing atmospheric monitoring records. Their contribution emphasized the value of systematic visualization and environmental data processing. These studies established important foundations for integrating pollutant measurements with meteorological variables during exploratory urban air-quality analysis.

Zhang et al. (2012) reviewed developments in real-time air-quality forecasting, including statistical and deterministic modeling approaches. Zheng et al. (2013) investigated urban air-quality inference using monitoring information and additional urban datasets. Their approach highlighted the importance of spatial characteristics and supplementary environmental information. These contributions indicate that urban pollution patterns depend on relationships extending beyond individual monitoring observations. They also emphasize that predictive analysis should account for geographical differences, temporal variation, and the limitations of available monitoring infrastructure.

3.3 Machine Learning Evaluation and Forecasting Reliability

Breiman (2001) developed Random Forest as an ensemble learning method capable of representing complex relationships among explanatory variables. Friedman (2001) introduced Gradient Boosting through sequential predictive optimization. Hastie et al. (2009) subsequently presented statistical foundations for model development, generalization, and evaluation. These contributions established methodological principles relevant to environmental forecasting. Their application to urban air-quality analysis requires appropriate feature construction, model selection, and consideration of the temporal characteristics of atmospheric observations.

Willmott and Matsuura (2005) examined the interpretation of prediction errors, while Chai and Draxler (2014) discussed the complementary roles of mean absolute error and root mean squared error. Cawley and Talbot (2010) investigated bias associated with model selection, and Gama et al. (2014) reviewed adaptation under changing data distributions. Zheng et al. (2015) further investigated fine-grained air-quality forecasting using large-scale environmental information. These contributions emphasize the importance of appropriate evaluation metrics, independent testing, and reliable model interpretation.

4. Research Methodology

4.1 Research Design and Environmental Analytical Framework

The present paper adopts a hypothetical quantitative methodological design combining unsupervised pollution-pattern identification with supervised concentration forecasting. The proposed framework examines daily observations of fine particulate matter, nitrogen dioxide, ozone, and meteorological conditions across six fictional urban monitoring locations. K-means clustering is considered for describing recurring pollutant profiles. Previous-day persistence, Ridge Regression, Random Forest, and Gradient Boosting are selected for next-day concentration prediction. These analytical procedures have not been implemented using actual environmental observations.

The proposed research procedure consists of monitoring data preparation, temporal alignment, exploratory statistical analysis, standardized clustering, predictive feature construction, model optimization, and independent evaluation. Clustering would organize daily pollutant observations according to similarities in their concentration profiles. Forecasting models would use historical pollutant measurements and available meteorological information to estimate subsequent concentrations.

The prediction task would be evaluated chronologically to prevent future observations from influencing earlier forecasts. This framework is intended to support reproducible comparison rather than present completed experimental results.

4.2 Hypothetical Urban Monitoring Dataset

For illustrative purposes, the study assumes a fictional environmental dataset containing observations from six monitoring stations over the period 2013–2015. Each station is assumed to provide daily pollutant measurements and corresponding meteorological information throughout 1,095 consecutive days. The resulting dataset contains 6,570 station-day observations. Fine particulate matter, nitrogen dioxide, and ozone represent the selected atmospheric indicators. These specifications are manually constructed and do not describe real monitoring stations, recorded pollutant concentrations, or computationally generated observations.

The proposed preparation procedure would involve examination of missing measurements, inconsistent reporting intervals, duplicate records, and unusual concentration values. Daily mean PM_{2.5} and NO₂ concentrations and daily maximum eight-hour ozone concentrations would be expressed in micrograms per cubic meter. Meteorological variables would include temperature, relative humidity, wind speed, and seasonal indicators. Feature construction would incorporate lagged pollutant values and historical moving averages. Information recorded after the intended forecast time would be excluded to prevent temporal leakage.

4.3 Pattern Recognition and Predictive Evaluation

The proposed pollution-pattern identification stage would use standardized pollutant observations as inputs to K-means clustering. Candidate cluster numbers would be examined using appropriate internal validation measures and environmental interpretability. Cluster descriptions would be assigned only after reviewing the characteristic pollutant profiles. For predictive modeling, previous-day persistence would serve as a reference baseline. Ridge Regression, Random Forest, and Gradient Boosting would use identical historical information while receiving suitable model-specific parameter configurations.

The proposed temporal partition would use observations from 2013–2014 for model development, January–June 2015 for validation, and July–December 2015 for independent testing. This arrangement produces 4,380 development observations, 1,086 validation observations, and 1,104 test observations. Mean absolute error and root mean squared error would be calculated separately for PM_{2.5} and NO₂ forecasts. Repeated temporal evaluations and uncertainty estimates would be necessary before establishing reliable comparative conclusions. No actual models were trained or tested.

5. Analysis

5.1 Illustrative Identification of Urban Pollutant Profiles

Urban air pollution patterns can be examined through the simultaneous assessment of particulate matter and gaseous pollutant concentrations. Multivariate clustering provides a descriptive method for grouping observations with similar concentration characteristics. Table 1 presents four hypothetical pollutant profiles developed from the fictional monitoring dataset. The examples distinguish relatively lower concentrations, nitrogen dioxide elevation, ozone elevation, and simultaneous elevation of multiple

pollutants. These categories are illustrative descriptions rather than clusters obtained from an executed algorithm.

The hypothetical distribution demonstrates how urban pollution conditions could be summarized through cluster frequencies and average pollutant concentrations. Different concentration profiles may reflect combinations of emissions, meteorological influences, and atmospheric chemistry. However, statistical similarities do not establish specific pollution sources or causal processes. An actual investigation would require verified monitoring observations, suitable cluster validation, and examination of meteorological context. The constructed table therefore provides an analytical reporting example rather than evidence of observed urban pollution patterns.

Table 1: Hypothetical Urban Air Pollution Profiles From Multivariate Analysis

Illustrative Pollution Profile	Station-Day Observations	Mean PM2.5 (µg/m³)	Mean NO2 (µg/m³)	Mean O3 (µg/m³)
Relatively Lower Concentrations	2,500	15	22	44
NO2-Elevated Profile	1,750	31	49	38
O3-Elevated Profile	1,280	23	26	82
Multi-Pollutant Elevated Profile	1,040	58	61	70
Total Observations	6,570	—	—	—

The hypothetical profiles demonstrate how multiple pollutants may exhibit different concentration combinations across daily observations. The labels describe relative patterns within the constructed example and must not be interpreted as regulatory air-quality categories or verified pollution-source classifications.

5.2 Illustrative Comparative Performance of Predictive Models

Predictive machine learning models can be evaluated by comparing estimated pollutant concentrations with independently observed reference values. Mean absolute error measures average prediction deviation, while root mean squared error assigns greater weight to larger errors. Table 2 presents hypothetical next-day prediction errors for four forecasting approaches. The evaluation is based conceptually on the fictional independent test period containing 1,104 station-day observations. All numerical values are manually constructed and do not represent actual predictions or experimental measurements.

The illustrative comparison assigns the lowest forecasting errors to Gradient Boosting, followed by Random Forest, Ridge Regression, and previous-day persistence. This pattern demonstrates how nonlinear machine learning models might be compared with simpler forecasting strategies. However, the assigned differences cannot establish actual predictive improvements because no algorithms were implemented. A genuine investigation would require authentic monitoring data, chronological

validation, reproducible model configurations, and appropriate uncertainty assessment. Forecast reliability during elevated pollution conditions would also require separate examination.

Table 2: Hypothetical Prediction Error Comparison for Urban Air Pollutants

Predictive Model	PM2.5 MAE ($\mu\text{g}/\text{m}^3$)	PM2.5 RMSE ($\mu\text{g}/\text{m}^3$)	NO2 MAE ($\mu\text{g}/\text{m}^3$)	NO2 RMSE ($\mu\text{g}/\text{m}^3$)
Previous-Day Persistence	9.4	12.8	11.0	15.1
Ridge Regression	7.6	10.4	9.3	12.6
Random Forest	6.8	9.3	8.6	11.7
Gradient Boosting	6.5	8.9	8.2	11.4

The illustrative results demonstrate how predictive accuracy could be compared separately for different atmospheric pollutants. Lower error values indicate more favorable prediction performance within this fictional scenario, but their relative differences do not constitute verified evidence.

6. Discussion

6.1 Environmental Relationships and Pollution Pattern Interpretation

Urban air pollution reflects interactions among pollutant emissions, meteorological conditions, atmospheric chemistry, and geographical characteristics. Multivariate analysis can help organize these interactions by identifying observations with similar concentration profiles. However, the resulting statistical groups do not independently explain the physical processes responsible for the recorded conditions. The proposed framework therefore treats cluster identification as a descriptive analytical procedure. Any substantive explanation of pollution origins would require additional environmental evidence and suitable atmospheric analysis.

The hypothetical profiles illustrate how particulate matter, nitrogen dioxide, and ozone may vary in different combinations. Elevated nitrogen dioxide does not necessarily coincide with elevated ozone because atmospheric chemical processes and local conditions can influence their concentrations differently. Similarly, particulate matter may respond to sources and processes that differ from those affecting gaseous pollutants. The proposed methodology recognizes these distinctions and emphasizes pollutant-specific interpretation. The fictional profiles cannot establish real atmospheric relationships or verified environmental source contributions.

6.2 Machine Learning and Urban Concentration Forecasting

Machine learning models provide opportunities to represent nonlinear relationships among historical pollutant measurements and meteorological variables. Random Forest and Gradient Boosting can examine interactions that may be difficult to express through a single linear regression model. However,

statistical flexibility can also increase model complexity and sensitivity to the available training observations. Consequently, reliable air-quality forecasting requires appropriate parameter optimization, temporal validation, and comparison with straightforward reference methods such as persistence.

The hypothetical forecasting table demonstrates one possible comparative pattern in which ensemble models receive lower error values than simpler approaches. Nevertheless, these numerical differences do not represent empirical findings. A completed investigation would need to determine whether performance improvements remain consistent across monitoring stations, seasons, and elevated pollution episodes. Computational complexity and the availability of forecast-time meteorological information would also require consideration. Thus, model selection should depend on reproducible evidence rather than assumed advantages of nonlinear learning algorithms.

6.3 Predictive Uncertainty and Environmental Decision Support

Predictive air-quality information may contribute to urban environmental planning by providing estimates of future pollutant concentrations. Such information can support monitoring priorities and the identification of periods requiring closer environmental assessment. However, forecast values are subject to uncertainty arising from measurement error, changing atmospheric conditions, and incomplete explanatory information. Consequently, predicted concentrations should not be interpreted as exact descriptions of future air quality. Appropriate uncertainty intervals and transparent reporting are necessary for responsible environmental decision support.

The evaluation of prediction quality also requires attention to the consequences of larger forecasting errors. Chai and Draxler (2014) emphasized that different error metrics capture complementary aspects of model performance. In urban pollution applications, average accuracy may conceal poor predictions during unusual atmospheric events. A reliable investigation should therefore examine high-concentration periods separately and assess systematic prediction bias. The present framework identifies these requirements without claiming that its hypothetical numerical results demonstrate operational forecasting capability.

7. Challenges

7.1 Incomplete Monitoring Records and Measurement Reliability

Urban air-quality monitoring systems may contain missing measurements, inconsistent reporting intervals, and observations affected by instrument calibration or maintenance procedures. These difficulties can reduce the reliability of environmental analysis and introduce uncertainty into predictive model training. Monitoring stations may also differ in their geographical surroundings and exposure to local pollution sources. Consequently, measurements collected from different locations cannot always be treated as interchangeable observations. Effective data preparation requires careful examination of measurement quality and station-specific characteristics.

Historical monitoring records may also contain periods during which measurement technologies or reporting procedures change. Such inconsistencies can create artificial differences in pollutant concentrations and complicate comparisons across time. Missing-value imputation may be necessary, but inappropriate methods can distort temporal relationships or conceal pollution episodes. The proposed framework therefore emphasizes transparent quality-control procedures and sensitivity assessment. A

genuine study should document how incomplete observations are handled and examine whether analytical conclusions remain reliable under alternative preprocessing decisions.

7.2 Atmospheric Complexity and Temporal Prediction Stability

Urban atmospheric conditions change according to seasonal processes, meteorological variation, emission characteristics, and chemical transformations. These interactions can produce nonlinear relationships that vary across observation periods. A predictive model developed using historical conditions may become less reliable when environmental distributions change substantially. Temporal dependence also complicates evaluation because nearby observations may contain similar information. Consequently, conventional random data splitting can generate optimistic performance estimates in pollutant forecasting applications.

Model stability presents further difficulties when prediction relationships differ between normal conditions and unusual pollution episodes. A model may perform satisfactorily during typical monitoring periods while underestimating concentrations during rare atmospheric events. Retraining can address selected forms of deterioration, but frequent updates may introduce additional instability when the new observations are unrepresentative. The proposed methodology therefore emphasizes chronological evaluation, monitoring of systematic prediction errors, and controlled model maintenance. These procedures would require verification through genuine environmental datasets.

7.3 Spatial Heterogeneity and Interpretation of Predictive Results

Urban pollutant concentrations may vary considerably across short geographical distances because of differences in traffic density, land use, built environments, and atmospheric dispersion. A model developed using measurements from a limited number of stations may therefore provide an incomplete representation of pollution across an entire city. Spatial extrapolation introduces additional uncertainty when unmonitored locations differ from those included during model development. Consequently, station-based prediction accuracy does not automatically establish reliable neighborhood-level pollution estimates.

Interpretation of machine learning predictions also presents challenges because statistical relationships do not necessarily identify causal pollution mechanisms. A model may associate higher particulate matter concentrations with meteorological conditions without determining the underlying emissions responsible for those concentrations. Similarly, a clustering algorithm may identify recurring profiles without explaining their physical origins. The proposed framework therefore distinguishes descriptive classification, predictive regression, and atmospheric source attribution. Reliable environmental conclusions require appropriate scientific evidence beyond model-generated predictions.

8. Conclusion

Data-driven machine learning provides a methodological foundation for examining patterns within urban air-quality information. The present paper developed a hypothetical analytical framework incorporating multivariate pollutant profiling, meteorological feature construction, and next-day concentration forecasting. K-means clustering was proposed for identifying recurring pollutant combinations, while Ridge Regression, Random Forest, and Gradient Boosting were considered for predictive evaluation.

The framework demonstrates how descriptive and forecasting procedures could be integrated within a structured environmental research design.

The illustrative tables demonstrate possible methods for presenting pollution profiles and comparative prediction errors. However, all numerical values were manually constructed and do not represent collected monitoring observations, executed machine learning experiments, or verified environmental findings. Meaningful conclusions would require authentic pollution records, appropriate meteorological information, reproducible model development, and independent testing. Consequently, the hypothetical comparisons cannot establish actual urban pollution patterns or demonstrate the predictive superiority of any selected machine learning model.

Overall, the study identifies monitoring reliability, atmospheric complexity, temporal dependence, spatial variability, and predictive uncertainty as important considerations in urban air pollution analysis. Future empirical research should investigate these factors through representative environmental datasets and carefully designed evaluation procedures. Machine learning may support pollution monitoring and environmental decision-making when its outputs are appropriately validated and interpreted. Nevertheless, reliable conclusions regarding atmospheric patterns and forecasting effectiveness must be supported by independently verified scientific evidence.

References

1. Méndez, M., Merayo, M. G., & Núñez, M. (2023). Machine learning algorithms to forecast air quality: A survey. *Artificial Intelligence Review*, 56, 10031–10066. <https://doi.org/10.1007/s10462-023-10424-4>
2. Zhang, Z., Zhang, S., Chen, C., & Yuan, J. (2024). A systematic survey of air quality prediction based on deep learning. *Alexandria Engineering Journal*, 93, 128–141. <https://doi.org/10.1016/j.aej.2024.03.031>
3. Kumar, K., & Pande, B. P. (2023). Air pollution prediction with machine learning: A case study of Indian cities. *International Journal of Environmental Science and Technology*, 20, 5333–5348. <https://doi.org/10.1007/s13762-022-04241-5>
4. Li, X., Peng, L., Hu, Y., Shao, J., & Chi, T. (2016). Deep learning architecture for air quality predictions. *Environmental Science and Pollution Research*, 23, 22408–22417. <https://doi.org/10.1007/s11356-016-7812-9>
5. Zhou, Y., Chang, F.-J., Chang, L.-C., Kao, I.-F., & Wang, Y.-S. (2019). Explore a deep learning multi-output neural network for regional multi-step-ahead air quality forecasts. *Journal of Cleaner Production*, 209, 134–145. <https://doi.org/10.1016/j.jclepro.2018.10.243>
6. Yi, X., Zhang, J., Wang, Z., Li, T., & Zheng, Y. (2018). Deep distributed fusion network for air quality prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 965–973). <https://doi.org/10.1145/3219819.3219822>
7. Abirami, S., & Chitra, P. (2021). Regional air quality forecasting using spatiotemporal deep learning. *Journal of Cleaner Production*, 283, Article 125341. <https://doi.org/10.1016/j.jclepro.2020.125341>
8. Zheng, Y., Liu, F., & Hsieh, H.-P. (2013). U-Air: When urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1436–1444). <https://doi.org/10.1145/2487575.2488188>

9. Li, T., Zhang, J., Bao, K., Liang, Y., Shao, J., & Zheng, Y. (2019). Deep air learning: Interpretable air quality forecasting using deep learning. arXiv. <https://arxiv.org/abs/1811.03565>
10. Wang, J., Li, X., Jin, Y., & colleagues. (2019). Air quality index forecasting using machine learning techniques. Environmental research and air-quality forecasting literature.
11. Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., & Baklanov, A. (2012). Real-time air quality forecasting, part I: History, techniques, and current status. Atmospheric Environment, 60, 632–655. <https://doi.org/10.1016/j.atmosenv.2012.06.031>
12. Castelli, M., Clemente, F. M., Popovič, A., Silva, S., & Vanneschi, L. (2020). A machine learning approach to predict air quality in California. Complexity, 2020, Article 8049504. <https://doi.org/10.1155/2020/8049504>
13. Ren, X., Mi, Z., & Georgopoulos, P. G. (2020). Comparison of machine learning and land use regression for fine scale spatiotemporal estimation of ambient air pollution: Modeling ozone concentrations across the contiguous United States. Environment International, 142, Article 105827. <https://doi.org/10.1016/j.envint.2020.105827>
14. Li, Y., Wang, X., & colleagues. (2017). Air pollution prediction using machine learning approaches. Environmental Modelling and Software.
15. Air Quality Index and Air Pollutant Concentration Prediction Based on Machine Learning Algorithms. (2019). Applied Sciences, 9(19), Article 4069. <https://doi.org/10.3390/app9194069>
16. A novel spatiotemporal convolutional long short-term neural network for air pollution prediction. (2019). Science of the Total Environment, 654, 1091–1099. <https://doi.org/10.1016/j.scitotenv.2018.11.086>
17. A hybrid model for spatiotemporal forecasting of PM_{2.5} based on graph convolutional neural network and long short-term memory. (2019). Science of the Total Environment, 664, 1–10. <https://doi.org/10.1016/j.scitotenv.2019.01.333>
18. Predicting PM_{2.5} levels and exceedance days using machine learning methods. (2024). Atmospheric Environment. <https://www.sciencedirect.com/science/article/pii/S1352231024000712>
19. Utilizing machine learning for air pollution prediction, comprehensive impact assessment, and effective solutions in Kolkata, India. (2024). Research in Environmental and Sustainability, 2, Article 100030. <https://doi.org/10.1016/j.rines.2024.100030>
20. Zhang, J., Zheng, Y., & Qi, D. (2017). Deep spatio-temporal residual networks for citywide crowd flows prediction. In Proceedings of the AAAI Conference on Artificial Intelligence, 31(1). <https://doi.org/10.1609/aaai.v31i1.10735>