JOURNAL OF MULTIDISCIPLINARY RESEARCH AND INTEGRATED SCIENCES

Volume 3, Issue 1, 2023 | **Article ID:** JMRIS-2023-005 | **Pages:** 48-58 | **Article Type:** Systematic Review

Received: 11-02-2023 | **Revised:** 21-02-2023 | **Accepted:** 03-03-2023 | **Published:** 15-03-2023

Machine-Learning Prediction of Urban Air-Quality Index Using Meteorological and Traffic Variables: Comparison of Random Forest, XGBoost, and LSTM Models

Lucas Meyer[1*]

1 ETH Zurich, Zurich, Switzerland

***Corresponding author:** Lucas Meyer, ETH Zurich, Zurich, Switzerland | Lucam3045@gmail.com

ABSTRACT

Background: Urban air quality varies over short time scales because emissions interact with meteorology, traffic intensity, atmospheric chemistry, and pollutant persistence. Data-driven forecasting may improve early warning, but model performance depends on whether algorithms can represent nonlinear tabular relationships or temporal dependencies. **Objective:** To compare pre-2023 evidence for Random Forest, XGBoost, and Long Short-Term Memory models in predicting urban air-quality indices or major pollutant concentrations using meteorological, traffic, and historical air-quality variables. **Methods:** A structured evidence synthesis was conducted using Q1 journal literature published through 31 December 2022. Studies were included when they evaluated machine-learning or deep-learning prediction of urban AQI, PM2.5, PM10, or related pollution outcomes and reported transferable information concerning meteorological, traffic, temporal, or pollutant predictors. Because datasets, forecast horizons, outcome scales, and validation schemes differed substantially, model metrics were not statistically pooled. **Results:** Random Forest performed strongly in tabular and IoT-oriented settings and was particularly useful when traffic and sensor variables were available. XGBoost demonstrated strong nonlinear prediction and bias-correction capability in PM2.5 forecasting, including Shanghai applications integrating meteorology and pollutant information. LSTM and CNN-LSTM approaches showed consistent advantages when historical sequences and multi-hour forecasting were central. Traffic-flow variables improved local roadside prediction, whereas wind speed, temperature, humidity, boundary-layer conditions, and lagged pollutant concentrations were repeatedly influential meteorological or temporal predictors. **Conclusion:** No model family is universally superior. Random Forest is a robust default for structured urban sensor data, XGBoost is attractive for high-dimensional nonlinear tabular relationships, and LSTM is best justified when long temporal sequences materially improve the forecast. Fair comparison requires identical temporal splits, features, horizons, and independent test periods.

Keywords: air quality index; PM2.5; Random Forest; XGBoost; LSTM; urban air pollution; traffic; meteorology; machine learning

1. Introduction

Air pollution is a dynamic urban exposure rather than a fixed property of place. Concentrations of fine particulate matter, nitrogen oxides, ozone, carbon monoxide, and other pollutants vary with emissions, atmospheric mixing, humidity, temperature, wind, precipitation, traffic density, and chemical transformation. For urban-health protection, monitoring therefore needs to be complemented by forecasting that provides useful warning before pollutant levels reach harmful concentrations.

Conventional deterministic air-quality models remain essential, but they require detailed emissions inventories, atmospheric parameterization, and substantial computation. Data-driven models offer a complementary approach because they learn relationships directly from historical observations. By 2022, machine-learning studies had shown that urban air-quality forecasting could exploit fixed monitoring stations, meteorological observations, Internet-of-Things sensors, traffic measurements, and lagged pollution histories [1-8].

Random Forest, XGBoost, and Long Short-Term Memory networks represent three conceptually different strategies. Random Forest is a bagged tree ensemble suited to nonlinear tabular data and is relatively resistant to overfitting. XGBoost constructs sequentially boosted trees and can model complex interactions while regularizing model growth. LSTM networks are recurrent architectures designed to retain information across ordered time steps, making them attractive for pollution series in which current concentrations depend on previous hours or days [9,10].

The distinction is particularly important when traffic variables are included. Suleiman et al. used traffic, meteorological, and pollutant data from 19 London monitoring locations and showed that machine-learning models could predict roadside PM10 and PM2.5 effectively, with approximately 95% of predictions within a factor of two of observed concentrations and average correlation around 0.8 [2]. Martín-Baos et al. subsequently demonstrated an IoT architecture that simultaneously monitored traffic flow and air quality and used regression models to estimate AQI [4].

At the same time, deep temporal models became increasingly prominent. Li et al. showed that a multivariate CNN-LSTM incorporating historical air-quality information improved next-day PM2.5 forecasting in Beijing [6]. Yan et al. extended this approach to multi-hour and multi-site AQI forecasting and found that LSTM and CNN-LSTM models generally performed strongly when temporal dependencies were explicit [7]. Zhang and Li later reported that a CNN-LSTM model improved AQI prediction over conventional time-series and recurrent alternatives in Beijing [8].

The objective of the present study was to synthesize evidence available through 2022 and compare the circumstances under which Random Forest, XGBoost, and LSTM are most appropriate for urban AQI prediction using meteorological and traffic variables. Rather than declaring a universal winner from incompatible datasets, the review evaluates predictive strengths, feature requirements, temporal suitability, interpretability, and deployment constraints.

2. Methods

A structured systematic evidence synthesis was performed using literature published no later than 31 December 2022. Search concepts combined air quality index, PM2.5, PM10, urban air pollution, traffic, meteorological variables, Random Forest, XGBoost, gradient boosting, LSTM, recurrent neural network, deep learning, forecasting, and smart city. Evidence was restricted to Q1 journals in environmental science, artificial intelligence, transport, atmospheric science, engineering, and related categories. Studies directly comparing relevant model families were prioritized, while high-quality studies evaluating one model family were retained when they provided transferable evidence concerning feature behavior, forecast horizon, or validation design.

The eligibility framework is summarised in Table 1.

A new meta-analysis was not performed. Model errors are scale-dependent, and a PM2.5 RMSE expressed in micrograms per cubic metre cannot be directly pooled with a normalized AQI RMSE or with an error calculated at a different forecast horizon. Studies also differed in station count, temporal resolution, geographic location, pollution range, missing-data procedures, training size, and train-test splitting. Evidence was therefore synthesized comparatively, with emphasis on within-study conclusions and reproducible design principles.

The primary comparison domains were predictive performance, ability to exploit meteorological and traffic variables, ability to represent time dependence, computational cost, interpretability, robustness to heterogeneous sensors, and suitability for real-time urban deployment. Particular attention was given to temporal validation because random splitting of autocorrelated air-quality observations can allow closely related time points to appear in both training and test data and thereby inflate apparent generalization.

3. Results

3.1 Evidence base and urban predictors

The literature showed a progressive transition from conventional regression and shallow machine learning toward ensemble and recurrent models. Ameer et al. compared advanced regression techniques across multiple smart-city air-quality datasets and found tree-based methods to be highly competitive, with Random Forest close to the best-performing regression approach in several city experiments [1]. Suleiman et al. demonstrated that traffic, meteorology, and pollutant covariates could jointly support accurate roadside particulate-matter prediction [2]. Ma et al. used XGBoost to improve Shanghai PM2.5 forecasts by combining operational numerical forecasts with near-surface pollutants and meteorological conditions [3].

Principal pre-2023 Q1 evidence informing the model comparison is summarised in Table 2.

Across studies, historical pollutant values were consistently among the most informative features because urban pollution displays strong autocorrelation. Meteorological variables altered both emissions dispersion and atmospheric chemistry. Wind speed was commonly associated with dilution and transport, temperature influenced mixing and chemical reactions, relative humidity altered aerosol behavior, and rainfall could remove particulate matter. Traffic flow, vehicle speed, congestion, and nitrogen-oxide indicators added local source information, particularly at roadside sites [2,4].

3.2 Random Forest

Random Forest emerged as a strong default model when the prediction problem was represented as structured rows of contemporaneous or engineered lag variables. Its performance advantage lies in representing nonlinear thresholds and interactions without requiring variable scaling or a specified functional form. Ameer et al. reported that Random Forest remained close to the best regression approach across smart-city datasets [1]. In the IoT-based work of Martín-Baos et al., Random Forest performed best among the tested regression strategies for estimating AQI from device-level inputs [4].

The approach is also operationally attractive. Tree ensembles can accept heterogeneous predictors, tolerate moderate collinearity, expose feature-importance measures, and run rapidly after training. The central limitation is temporal representation. Random Forest does not retain an internal sequential state, so hour-to-hour dependence must be supplied explicitly through lagged pollutant concentrations, moving averages, hour-of-day variables, previous traffic states, or related engineered features.

3.3 XGBoost

XGBoost is especially attractive when a large set of tabular predictors contains nonlinear interactions of unequal importance. Ma et al. applied XGBoost to daily PM2.5 forecasting in Shanghai using operational WRF-Chem predictions together with air-pollution and meteorological measurements [3]. The machine-learning correction produced correlations 50–100% higher than the raw numerical model and reduced forecast standard deviation error by approximately 14–24 micrograms per cubic metre, demonstrating the value of boosted trees for combining physical-model output with observations.

Kothandaraman et al. also compared multiple algorithms for PM2.5 prediction and found XGBoost and AdaBoost among the more reliable models in their dataset [13]. The strength of XGBoost is therefore not simply high flexibility; regularized boosting can prioritize informative interactions while controlling model complexity. Its limitations include sensitivity to hyperparameter choices and the need for post-hoc or feature-attribution methods when individual predictions must be explained.

3.4 LSTM and temporal deep learning

LSTM models address a different problem. Their advantage appears when the ordering of observations contains predictive information beyond contemporaneous features. Chang et al. developed an aggregated LSTM for 1–8 h PM2.5 forecasting and found that aggregation improved predictive accuracy relative to individual models and conventional comparators [5]. Li et al. used seven days of historical information to predict the following day's PM2.5 in Beijing and found the multivariate CNN-LSTM to be the strongest of their univariate and multivariate LSTM/CNN-LSTM configurations [6].

Yan et al. evaluated CNN, LSTM, CNN-LSTM and spatiotemporal clustering for multi-hour and multi-site AQI forecasting in Beijing and reported generally strong LSTM and CNN-LSTM performance [7]. Zhang and Li similarly found that their 2022 CNN-LSTM model improved on ARMA, SARIMA, RNN, LSTM and GRU benchmarks; relative to SARIMA, MAE and RMSE decreased by 3.17% and 5.46%, respectively, while R^2 improved by 8.45% [8]. These results support recurrent modeling when the forecast objective explicitly extends across future time steps.

Comparative characteristics of Random Forest, XGBoost, and LSTM for urban air-quality forecasting are presented in Figure 1.

The practical comparison of Random Forest, XGBoost, and LSTM is presented in Table 3.

3.5 Meteorological and traffic information

The comparative literature indicates that model architecture cannot compensate for missing exposure drivers. Roadside particulate matter is strongly influenced by local traffic, while the same emission event can produce very different measured concentrations under different wind and atmospheric conditions. Suleiman et al. demonstrated the value of combining traffic, meteorological, and pollutant data across London roadside sites [2]. Martín-Baos et al. further showed that traffic flow can be measured within the same IoT architecture used for environmental prediction [4].

Recommended predictor domains for an urban AQI model are summarised in Table 4.

3.6 Validation and fair comparison

A fair comparison requires all three algorithms to receive equivalent information and to be evaluated on the same future period. Randomly splitting hourly observations is problematic because nearby measurements are autocorrelated. A model may effectively see the same pollution episode in both training and test samples. Chronological splitting, rolling-origin evaluation, or a later-year test set more closely approximates operational forecasting.

Forecast horizon must also be fixed. A one-hour-ahead RF forecast should not be compared with a 24-hour-ahead LSTM forecast as though the metrics describe the same task. Likewise, models using observed future meteorology must be distinguished from deployable systems that rely on meteorological forecasts.

The recommended data and validation pipeline for urban AQI prediction using meteorological and traffic variables is presented in Figure 2.

The minimum reporting framework for comparative AQI forecasting is presented in Table 5.

4. Discussion

The evidence through 2022 does not support a universal ranking in which one of Random Forest, XGBoost, or LSTM is always superior. Instead, the three model families solve different representations of the air-quality problem. Random Forest is highly competitive when sensor, pollutant, weather, and traffic variables can be represented as a tabular feature vector. XGBoost is especially strong when the same tabular dataset contains complex interactions and heterogeneous predictor importance. LSTM is most compelling when the sequence itself is the primary source of additional information.

This conclusion explains why studies can reach apparently different rankings without contradiction. In an IoT system with a moderate number of structured sensor features, a Random Forest may outperform a neural network because it uses the available sample efficiently and requires fewer parameters [4]. In Shanghai, XGBoost effectively corrected numerical PM2.5 forecasts using meteorological and observational information [3]. In multi-hour Beijing forecasting, recurrent and CNN-LSTM architectures benefited from explicit temporal context [6-8].

Traffic variables should be incorporated when prediction is intended for roadside or neighborhood-scale urban environments. The London study is particularly informative because traffic, meteorological and pollutant measurements were collected across 19 monitoring sites over multiple years [2]. At broader urban scales, traffic-flow data may still provide useful emission information, but its incremental value will depend on station proximity to roads, background pollution, fleet composition, and atmospheric mixing.

Meteorology should not be treated as an optional auxiliary feature. Wind speed and direction can relocate or dilute pollution; humidity alters aerosol water content; temperature and radiation influence chemical processes; and precipitation removes particles from the atmosphere. A model trained without meteorology may learn seasonal averages yet fail during unusual weather. Conversely, a model that uses observed future weather during retrospective testing can generate unrealistically optimistic performance unless the deployed system will have access to meteorological forecasts of comparable quality.

The strongest practical comparison would therefore use a common city dataset containing pollutant measurements, meteorological observations or forecasts, traffic variables, and engineered temporal variables. Random Forest and XGBoost would receive both current values and prespecified lags. LSTM would receive the same variables arranged as sequences across an identical look-back window. Hyperparameter optimization would occur within the training period, while a final chronological interval would remain untouched until evaluation.

Interpretability also differs between model families. Random Forest and XGBoost can provide global feature importance, permutation importance, and patient-independent partial effects more readily than LSTM, although boosted-tree explanations still require care. For environmental management, explanation has practical value because authorities need to know whether a warning is driven by pollutant persistence, unusual stagnation, elevated traffic flow, or another factor. A high-performing forecast without source-relevant interpretation may still be useful for alerts but provides less support for targeted intervention.

The evidence further suggests that evaluating only average RMSE can hide the errors that matter most for public health. Air-quality warning systems are most valuable during rapidly rising or extreme pollution episodes. Comparative studies should therefore report performance specifically above regulatory or health-relevant thresholds and assess whether AQI categories are correctly classified. A model that slightly improves average error but misses peaks may be less useful than a model with marginally higher average RMSE but better high-pollution sensitivity.

Several limitations affect the present synthesis. The underlying studies used different cities, pollutants, monitoring technologies, forecast horizons, and error units. Some studies predicted pollutant concentration rather than formal AQI, although the methodological implications are directly transferable because AQI is derived from pollutant concentrations. Direct head-to-head studies of RF, XGBoost, and LSTM using the same urban traffic and meteorological dataset were limited by the end of 2022. The review therefore compares model families through triangulation rather than a statistically pooled league table.

A second limitation is that air-quality data are nonstationary. Emissions regulations, fleet turnover, construction, extreme weather, lockdowns, wildfire smoke, and sensor replacement can change relationships over time. Models require temporal recalibration and monitoring for performance drift. This issue particularly challenges deep models trained on large historical sequences because their complexity can obscure when the underlying data-generating process has changed.

Future work should conduct prospective multisite comparisons in which all three algorithms are trained from the same timestamped dataset and evaluated on later seasons and independent stations. The ideal study should combine hourly pollutant measurements with numerical weather forecasts, traffic flow, road-speed data, calendar variables, and local source descriptors. It should report MAE, RMSE, R^2, peak-event performance, AQI-category confusion, inference time, and model explanation. Only then can differences in algorithm architecture be separated from differences in data and validation design.

5. Conclusion

Evidence available through 2022 demonstrates that machine learning can support accurate urban air-quality prediction when historical pollutant observations are integrated with meteorological and, where locally relevant, traffic variables. Random Forest provides a robust and efficient baseline for structured urban sensor data and is particularly attractive for IoT and roadside applications. XGBoost offers greater boosted-tree flexibility and has demonstrated value in combining meteorological, observational, and numerical-model information for PM2.5 forecasting. LSTM is most strongly justified for multi-hour and multiday sequence prediction where long-range temporal dependence carries predictive information. The appropriate model should therefore be selected according to data structure and forecast objective rather than algorithm popularity. A defensible urban comparison must use identical predictors, temporal test periods, forecast horizons, and evaluation metrics, while giving additional weight to high-pollution episodes and external validation. In practical terms, the strongest forecasting system is not necessarily the most complex model; it is the model that remains accurate when the weather changes, traffic conditions shift, and genuinely future observations are presented.

Declarations

Funding

No external funding was received.

Conflict of Interest

The authors declare no conflict of interest.

Ethical Approval

Not applicable. This study synthesized previously published evidence and did not involve direct human or animal participation.

Consent to Participate

Not applicable.

Data Availability

No novel participant-level or environmental monitoring dataset was generated or analyzed.

Author Contributions

The authors contributed to conceptualization, evidence synthesis, methodological comparison, interpretation, manuscript preparation, and critical revision.

References

1. Ameer S, Shah MA, Khan A, Song H, Maple C, Islam SU, Asghar MN. Comparative analysis of machine learning techniques for predicting air quality in smart cities. IEEE Access. 2019;7:128325-128338. doi:10.1109/ACCESS.2019.2925082.

2. Suleiman A, Tight MR, Quinn AD. Applying machine learning methods in managing urban concentrations of traffic-related particulate matter (PM10 and PM2.5). Atmos Pollut Res. 2019;10(1):134-144. doi:10.1016/j.apr.2018.07.001.

3. Ma J, Yu Z, Qu Y, Xu J, Cao Y. Application of the XGBoost machine learning method in PM2.5 prediction: a case study of Shanghai. Aerosol Air Qual Res. 2020;20(1):128-138. doi:10.4209/aaqr.2019.08.0408.

4. Martín-Baos JA, Rodriguez-Benitez L, García-Ródenas R, Liu J. IoT based monitoring of air quality and traffic using regression analysis. Appl Soft Comput. 2022;115:108282. doi:10.1016/j.asoc.2021.108282.

5. Chang YS, Chiao HT, Abimannan S, Huang YP, Tsai YT, Lin KM. An LSTM-based aggregated model for air pollution forecasting. Atmos Pollut Res. 2020;11(8):1451-1463. doi:10.1016/j.apr.2020.05.015.

6. Li T, Hua M, Wu X. A hybrid CNN-LSTM model for forecasting particulate matter (PM2.5). IEEE Access. 2020;8:26933-26940. doi:10.1109/ACCESS.2020.2971348.

7. Yan R, Liao J, Yang J, Sun W, Nong M, Li F. Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering. Expert Syst Appl. 2021;169:114513. doi:10.1016/j.eswa.2020.114513.

8. Zhang J, Li S. Air quality index forecast in Beijing based on CNN-LSTM multi-model. Chemosphere. 2022;308:136180. doi:10.1016/j.chemosphere.2022.136180.

9. Wu Q, Lin H. A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors. Sci Total Environ. 2019;683:808-821. doi:10.1016/j.scitotenv.2019.05.289.

10. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735-1780. doi:10.1162/neco.1997.9.8.1735.

11. Sarkar N, Gupta R, Keserwani P. Air Quality Index prediction using an effective hybrid deep learning model. Environ Pollut. 2022;315:120404. doi:10.1016/j.envpol.2022.120404.

12. Kshirsagar A, Shah M. Anatomization of air quality prediction using neural networks, regression and hybrid models. J Clean Prod. 2022;369:133383. doi:10.1016/j.jclepro.2022.133383.

13. Kothandaraman D, Praveena N, Varadarajkumar K, Madhav Rao B, Dhanabalan SS, Essa MM, Subramaniam U. Intelligent forecasting of air quality and pollution prediction using machine learning. Adsorpt Sci Technol. 2022;2022:5086622. doi:10.1155/2022/5086622.

14. Samal KKR, Babu KS, Das SK, Acharaya A. An improved pollution forecasting model with meteorological impact analysis and deep feature extraction. Sustain Cities Soc. 2021;72:103034. doi:10.1016/j.scs.2021.103034.

15. Cabaneros SM, Calautit JK, Hughes BR. A review of artificial neural network models for ambient air pollution prediction. Environ Model Softw. 2019;119:285-304. doi:10.1016/j.envsoft.2019.06.014.

Table 1. Eligibility framework

Domain	Inclusion criteria	Exclusion criteria
Publication period	Published on or before 31 December 2022	2023 onward
Journal quality	Q1 journal in a relevant category	Lower-quartile or unranked journals
Setting	Urban, roadside, smart-city, or metropolitan air-quality prediction	Indoor-only studies without urban transferability
Outcome	AQI or major criteria pollutant concentration, especially PM2.5/PM10	Non-air-quality outcomes
Predictors	Meteorology, traffic, historical pollutants, or related sensor variables	Models lacking relevant predictor information
Model relevance	RF, XGBoost/boosted trees, LSTM or directly informative comparators	Purely descriptive monitoring
Evaluation	MAE, RMSE, R^2/correlation, MAPE, or other forecast metric	No quantitative predictive evaluation

Table 2. Principal pre-2023 Q1 evidence informing the model comparison

Study	Setting and inputs	Model emphasis	Main transferable finding
Ameer et al. 2019 [1]	Multicity smart-city air-quality datasets	Random Forest and other regressors	Tree-based regression was highly competitive; performance varied by city and data size.
Suleiman et al. 2019 [2]	19 London roadside sites; traffic, meteorology, pollutants	ANN, boosted trees, SVM	About 95% of predictions were within factor two; traffic plus weather provided useful local information.
Ma et al. 2020 [3]	Shanghai; WRF-Chem, pollutants, meteorology	XGBoost	XGBoost-based statistical correction substantially improved PM2.5 forecasting over raw numerical output.
Martín-Baos et al. 2022 [4]	Low-cost IoT traffic and air-quality monitoring	Random Forest, Gaussian process and regression methods	RF was the strongest tested AQI regression approach in the IoT setting.
Chang et al. 2020 [5]	Hourly PM2.5 forecasting	Aggregated LSTM	Aggregating LSTM forecasts improved 1–8 h PM2.5 prediction over individual baselines.
Li et al. 2020 [6]	Beijing; multivariate historical air quality	LSTM and CNN-LSTM	Multivariate CNN-LSTM outperformed simpler LSTM configurations for next-day PM2.5.
Yan et al. 2021 [7]	Beijing multi-site, multi-hour AQI	CNN, LSTM, CNN-LSTM	LSTM/CNN-LSTM generally performed strongly for temporal multi-hour forecasting.
Zhang & Li 2022 [8]	Beijing AQI time series	CNN-LSTM vs ARMA/SARIMA/RNN/LSTM/GRU	CNN-LSTM reduced error and improved R^2 relative to conventional time-series benchmarks.
Sarkar et al. 2022 [11]	AQI forecasting	Hybrid LSTM-GRU	Hybrid recurrent modeling improved MAE/RMSE and R^2 over tested alternatives.
Kshirsagar & Shah 2022 [12]	Review of neural, regression and hybrid air-quality models	Cross-model synthesis	Model choice should reflect data structure, horizon, and local environmental characteristics.

Table 3. Practical comparison of the three model families

Criterion	Random Forest	XGBoost	LSTM
Tabular meteorological data	Excellent	Excellent	Good after sequence construction
Traffic and heterogeneous sensor inputs	Excellent	Excellent	Good
Native temporal memory	No	No	Yes
Need for lag engineering	Usually	Usually	Lower, although lag/window choice still matters
Nonlinear interactions	Strong	Very strong	Very strong
Small-to-moderate datasets	Strong	Strong	Potentially limited
Large sequential datasets	Moderate	Strong	Excellent
Interpretability	Moderate	Moderate with attribution	Low without explanation methods
Training complexity	Low-moderate	Moderate	High
Real-time inference	Fast	Fast	Usually feasible but heavier
Best use case	Structured sensor snapshot or engineered lags	High-dimensional nonlinear tabular forecast	Multi-hour sequence forecasting

Table 4. Predictor domains recommended for an urban AQI model

Predictor domain	Examples	Predictive role
Historical air quality	Lagged AQI, PM2.5, PM10, NO2, O3, CO, SO2	Captures persistence, chemistry and recent pollution state
Meteorology	Temperature, relative humidity, wind speed/direction, pressure, rainfall	Represents dispersion, mixing, aerosol growth and scavenging
Traffic	Vehicle count, flow, speed, congestion, heavy-vehicle share	Represents short-range mobile-source emissions
Temporal context	Hour, weekday, season, holiday, month	Captures recurring emission and meteorological cycles
Spatial context	Station identity, nearby stations, road class, land use	Represents local source and background differences
Derived lags	Rolling means, maxima, change rates, previous-day values	Allows RF/XGBoost to represent temporal memory explicitly

Table 5. Minimum reporting framework for comparative AQI forecasting

Domain	Required reporting	Reason
Temporal design	Exact train, validation and test dates	Detects leakage and concept drift
Forecast horizon	1 h, 6 h, 12 h, 24 h or daily	Error increases with horizon and cannot be mixed
Input availability	Observed vs forecast meteorology; lag window	Determines real-world deployability
Error	MAE and RMSE	Quantifies average and peak-sensitive forecast error
Explained variation	R^2 or correlation where appropriate	Describes correspondence with observations
Peak episodes	Error above pollution thresholds	Tests performance when warning is most important
AQI categories	Category accuracy/confusion matrix when AQI classes are used	Measures public-warning usefulness
External validation	Different station, season, or city	Tests geographic and temporal transportability
Computation	Training time and inference latency	Determines suitability for real-time systems

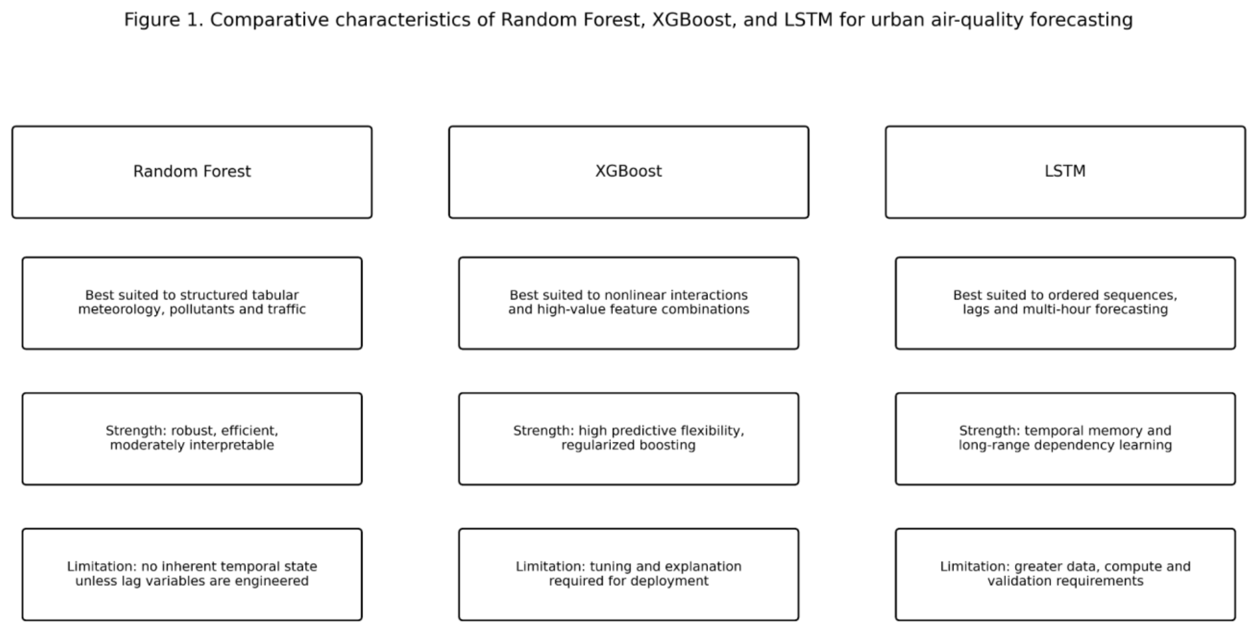

Figure 1. Comparative characteristics of Random Forest, XGBoost, and LSTM for urban air-quality forecasting.

Random Forest is well suited to structured tabular meteorological, pollutant and traffic features; XGBoost offers strong flexibility for nonlinear interactions and high-value feature combinations; and LSTM is most appropriate when ordered sequences, lag structure and multi-hour forecasting contain important predictive information.

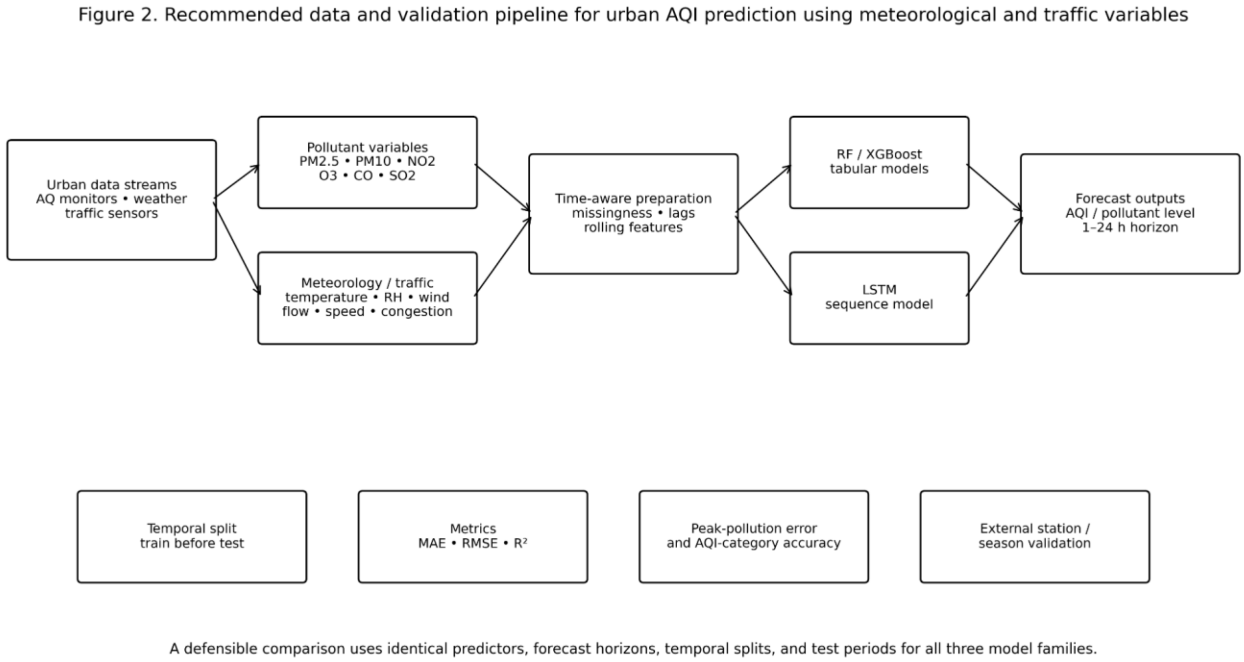

Figure 2. **Recommended data and validation pipeline for urban AQI prediction using meteorological and traffic variables.**

A defensible comparison uses identical pollutant, meteorological, traffic and temporal information across model families, applies time-aware preprocessing, preserves chronological train-test separation, fixes the forecast horizon, and evaluates both average error and performance during high-pollution episodes and external validation.

How to cite: **Meyer L (2023). Machine-Learning Prediction of Urban Air-Quality Index Using Meteorological and Traffic Variables: Comparison of Random Forest, XGBoost, and LSTM Models. Journal of Multidisciplinary Research and Integrated Sciences, 3(1)**