Journal of Green Energy and Environmental Engineering

Research Paper

Urban Air Quality Index Prediction Based on Integrated Random Forest and XGBoost Models

Karel Němec[1,*] and Daniel Novotný[2]

[1] Department of Computer Systems, Brno University of Technology, 61669 Brno, Czech Republic
[2] Faculty of Civil Engineering, Brno University of Technology, 61669 Brno, Czech Republic
*Corresponding author: karel.n@fit.vut.cz

Abstract. The air in cities all around the world is also adversely affected by pollution from roads and other activities. This research presents an integrated ensemble framework in response to the aforementioned shortcomings in the current deterministic and single-machine learning models for air quality index (AQI) prediction in metropolitan settings. After gathering a variety of spatiotemporal data using a dense fixed and mobile sensor network, a number of feature engineering techniques are employed to extract geographically and temporally important physical features. Create a Random Forest and XGBoost model that can use neural meta-learners to increase prediction accuracy, stability, and interpretability. According to the experiment, the ensemble system is stable in the face of abrupt contamination or missing data and has a comparatively modest inaccuracy. Integrated feature contribution analysis can be utilized to identify the reasons for variations in the AQI and provide practical guidance for pertinent countermeasures. The platform's modular design allows for expansion and adaptation to new developments in urban data infrastructure. In summary, the aforementioned findings demonstrate that spatiotemporal data fusion and ensemble modeling may create a trustworthy, high-resolution air quality forecast; hence, a stable foundation has been supplied for the intelligent operation of smart cities and extended intelligent deployment across various locations.

Keywords: *Ensemble Learning, Spatiotemporal Analysis, Urban Air Quality*

Received on 30 June 2025, Accepted on 28 December 2026, Published on 05 January 2026

Introduction

Nowadays, one of the biggest issues affecting public health and urban development worldwide is air pollution. The World Health Organization estimates that 4.2 million premature deaths annually are caused by fine particulate matter ($PM_{2.5}$), nitrogen dioxide (NO_2), and ozone because over 90% of the urban population is exposed to air pollution levels that are higher than the recommended limit [1,2]. For a considerable amount of time, the major cities of Beijing, Delhi, and Paris have experienced ongoing periods of extreme air pollution, which are causing acute respiratory illnesses, an increase in hospital admissions, and long-term health issues for the local population [3, 4]. The quality of urban ecosystem services, labor productivity, and children's cognitive development are all deteriorating at the same time [5–6]. As a result, in order to create integrated platforms for dense sensor networks, real-time data processing, and public risk alerts, governments and local authorities have swiftly extended the network of air quality monitoring and started building smart cities [7, 8]. The development of an enhanced urban environment will be aided by the achievement of high-spatial and high-temporal resolution accurate and useful air quality predictions.

The challenge of developing a high-accuracy, high-reliability forecasting model for urban air pollution remains unresolved despite recent technology breakthroughs. Although regulatory-grade monitoring networks are quite accurate at measuring pollutants, they are spatially sparse; they only cover a tiny portion of the city and are unable to capture neighborhood-scale variation brought on by localized emission sources and urban morphology [9,10]. Cities can now gather vast amounts of multi-source spatiotemporal data with high coverage and

resolution because to the widespread use of inexpensive sensors and satellite-based remote sensing [11,12]. A wave of data-driven algorithms, including autoregressive moving average (ARMA), support vector regression, random forests, extreme gradient boosting, and deep learning, have replaced physics-based dispersion models in air pollution prediction techniques at the same time [13,14]. However, there are still three major issues: (i) these models' limited generality for prediction in unmonitored or quickly changing urban areas because of sparse data; (ii) real-time sensing is susceptible to missing values, anomalous readings, and concept drift; and (iii) highly complex algorithms are difficult to understand, which hinders transparent decision support and public trust [15]. The efficiency and resilience of most existing methods in real-world operations are diminished because they do not dynamically mix numerous sources of space and time.

In light of the aforementioned issues, this research introduces a novel framework for predicting urban air quality that makes use of complex spatiotemporal feature engineering, ensemble learning, and many sources of sensor data fusion. Utilize cutting-edge ensemble models like random forests and gradient boosting with adaptive weighting strategies to increase accuracy and robustness. Gather real-time data from distributed sensor networks and external platforms in the proposed system. Utilize an advanced feature extraction pipeline that combines spatial autocorrelation, temporal dynamics, and cross-domain contextual information. The following are this work's main contributions: (1) building a platform for data fusion and large-scale urban air quality monitoring for fine-grained predictive analysis; (2) creating a new set of spatiotemporal features that can precisely depict the distribution and variations of pollution in various metropolitan locations; (3) On large datasets from real urban surveillance, experimental results demonstrate that this model performs better in terms of explainability, stability, and generalization ability.

The technological background, primary motivations, and a list of relevant papers are presented in Section 2. The system's overall architecture, techniques for spatiotemporal feature engineering, and an ensemble model framework are presented in Section 3. In Section 4, enumerate all types of experimental verifications in terms of interpretability analysis, sensitivity and robustness testing, and performance comparison. The research findings, real-world applications, and future prospects of intelligent urban air quality prediction are presented in Section 5.

Background and Motivations

Improving urban air quality is currently a major priority for both society and science due to the fast expansion of urban populations and economic activity in China and throughout the world. To attain a higher temporal frequency and spatial coverage, both national and local governments have made significant investments in the development of all-weather, all-season, and all-weather air pollution monitoring networks in recent years [16–17]. In the past, the majority of these networks have only set up a small number of monitoring stations that consistently gather high-quality pollution data, like $PM_{2.5}$, NO_2, O_3, etc. However, microenvironmental fluctuations in various sections of a complex urban area that are driven by varying terrain, heterogeneous emission sources, and rapid urbanization cannot be properly captured because of the intrinsically coarse spatial resolution of these permanent locations [18]. Improved high-altitude remote sensing, including satellite observation for broad-area measurement of pollution concentrations, and a vast network of mobile monitoring devices, such as cars fitted with portable sensor stations or hand-held portable monitoring systems, have been developed to address the aforementioned shortcomings [19, 20]. When taken as a whole, these advancements have produced copious volumes of fine-grained, high-quality urban air quality data and laid the groundwork for future generations of data-driven analysis and early warning systems.

Numerous phases in the development of prediction models have emerged along with the advancement of urban air quality sensors. Autoregressive integrated moving average (ARIMA) and state-space models were the first generation of statistical models for univariate time series forecasting in monitoring stations; while they offered some support for short-term forecasts, they were constrained by assumptions like linearity and stationarity [21]. Random forest and support vector regression, which can handle complicated, non-linear interactions among characteristics and high-dimensional data, are currently frequently employed algorithms for predicting air pollution at different times due to the advancement of machine learning [22]. Deep learning based on CNNs and LSTMs has also advanced recently in automatically identifying different kinds of complex spatiotemporal connections [23]. However, these deep models are not appropriate for real-time operation in resource-constrained urban management areas since they often require vast amounts of labeled data, significant

computational resources, and exact regulation against overfitting. Furthermore, the majority of deep models are "black boxes" that lack public accountability and transparency for decision-support systems [24].

The real situation in cities has not changed significantly, despite several advancements in the aforementioned areas. Regulators and emergency response teams want models that can explain the causes behind these projections, such as why and under what conditions they occur, because municipal authorities require accurate short-term air quality forecasts [25]. Due to problems with data quality or model generality, the operational system is currently unable to fully meet the timely, spatially precise notifications that ordinary citizens and other vulnerable groups need to change their behavior and safeguard their health. Missing data, sensor malfunctions, and frequent variations in emission patterns are prevalent issues in dynamic and unstructured metropolitan regions, where these needs are more noticeable. As urbanization has progressed, many individuals are now aware that in order to create more robust, accurate, and stable models, the next generation of air quality prediction systems must incorporate many data types and employ spatiotemporal feature extraction.

Consequently, the following are the study's three scientific issues. First, how can meaningful patterns be extracted and synthesized from noisy, unevenly distributed, multi-source urban sensor data using spatiotemporal feature engineering pipelines? Second, what ensemble modeling techniques may be applied flexibly to combine various predictive signals, improve forecast reliability in the face of data ambiguity, and increase flexibility to changes in the urban environment? Third, how can the sophisticated models preserve actionable openness and interpretability for practical use to facilitate public involvement and regulatory decisions? To create a workable intelligent system for managing urban air quality that fully utilizes the wealth of data, all of these issues must be resolved. An all-encompassing approach and experiments to overcome the aforementioned shortcomings will be presented in the following sections.

Methodology

System Architecture and Data Workflow

In addition to being expandable, dependable, and transparent, a new, high-performance urban air quality monitoring and forecast system must set up a steady flow of diverse data from many sources and produce analysis results. This study's system's modular design makes it appropriate for growing the data source and operating in numerous cities. A generic diagram of the entire framework for gathering initial data, disseminating predictions, and receiving system response is shown in Figure 1.

Fixed-base stations and mobile-base stations are combined in the structure of the sensing and acquisition layer. Fixed stations are the regulatory cornerstone of the majority of urban monitoring systems, feature high-precision continuous measuring capabilities, and are often authorized by the government. However, we have incorporated mobile sensors that may be put on cars, public transportation, or even carried by people to address the issue of limited spatial resolution in a fixed network. These mobile units increase the spatial coverage of regions that are typically not served by fixed installations, such as residential areas and transportation corridors. Other external data sources, such as satellite remote sensing (aerosol optical depth), meteorological data streams, and urban GIS layers (traffic, land use), can be added to the system to expand the diversity of data.

Assure dependability, security, and minimal latency in the Edge Devices' Data Transmission and Aggregation Layer to the Central Data Warehouse. Edge computing nodes will now perform preliminary quality checks on the data at the source, including basic anomaly filtering, range validation, and timestamp verification. Through redundancy methods, data packets are transmitted in batch mode across an encrypted channel to lower the chance of data loss due to equipment malfunction or network failure. For future tracking, rich metadata, including the precise position, sensor type, and provenance tags, are appended to the incoming data.

After the data has been aggregated, it will be centrally handled and kept. The high-frequency geo-tagged multi-modal data is stored in a scalable, high-performance time-series database system. Retrospectively train models and effectively obtain data from real-time analytics pipelines. The data gathered by all sensor modules will be normalized and consolidated into a single data structure to increase storage economy and analytical flexibility, hence improving the ease of multi-platform operation.

Through a number of procedures, the Data Processing and Feature Engineering Layer methodically cleans, transforms, and expands raw data. Here, modules deal with outlier detection, de-duplication, sensor drift correction, and missing value imputation. Crucially, the system creates a combined dataset for feature engineering and model training after synchronizing data from several sources at the same time and location.

At the core sits the machine learning and ensemble prediction layer. This module leverages engineered features to train, select, and apply a series of predictive models—ranging from individual regressors to sophisticated ensemble structures—which infer both near-future pollutant concentrations and associated uncertainties. Model workflows are orchestrated and monitored for performance drift, enabling automated retraining when substantial changes in data patterns are detected.

Lastly, the obtained information will be disseminated via a variety of channels, such as public web and mobile alert systems, municipal emergency management systems, etc., using the output and stakeholder interaction layer. For the user, query the current forecast and past patterns; in certain implementations, offer real-time comments or corrections. This feedback loop will provide ongoing modifications to data gathering and model construction in response to urban developments.

The combination of these architectural components—sensor-agnostic data pipelines, edge-to-cloud management, robust analytics, and open feedback loops—ensures the system can be rapidly deployed across cities of varying scale and infrastructure maturity. Figure 1 visually encapsulates these coordinated layers, illustrating both their individual roles and their interactions across the complete data lifecycle.

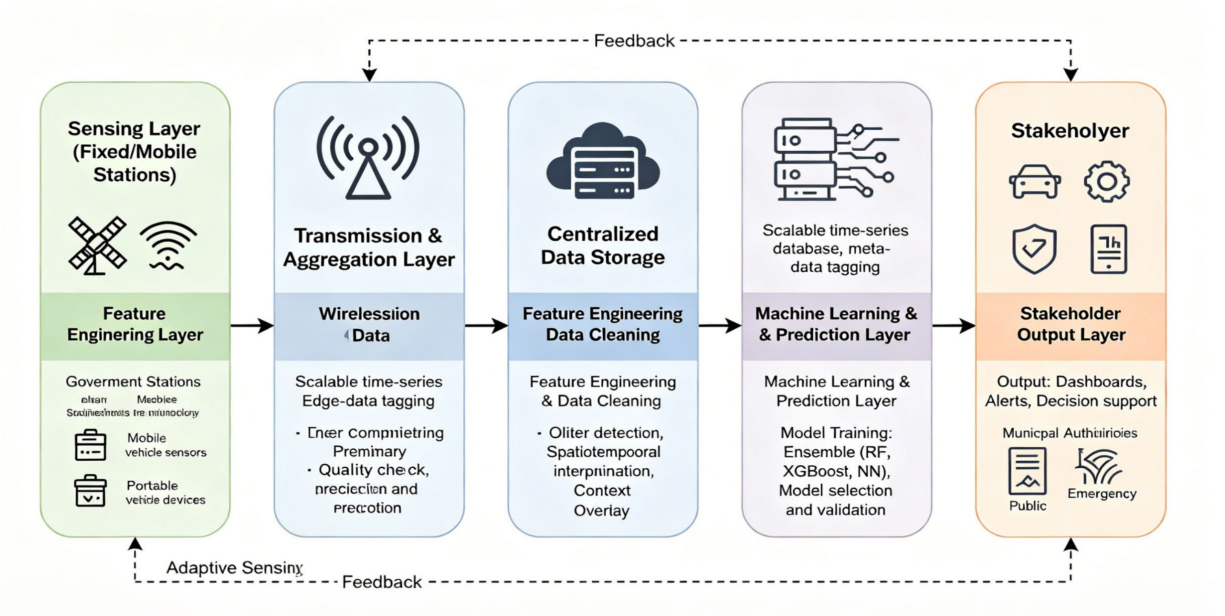

Figure 1. System Architecture of Urban Spatiotemporal Air Quality Sensing and Prediction

Spatiotemporal Feature Engineering

The level of detail that is retrieved during the feature engineering process directly affects the predictions made by a good urban air quality model. In order to enhance the quality of noisy real-world sensor data for ensemble learning, the aforementioned framework will provide a bridge for spatiotemporal feature engineering. The two types of modules in our pipeline are a generic module and a module for urban pollution that is inherently subject to temporal, spatial, and physical constraints, as seen in Figure 2.

The process begins with rigorous data validation and cleaning. All incoming sensor streams are examined for missing values, temporal and spatial inconsistencies, outliers, and evidence of sensor drift. Minor data gaps are imputed based on a weighted combination of local time series continuity and neighboring station data. For instance, a missing observation $x_{i,t}$ at station i and time t is estimated as:

$$x_{i,t}^{\text{filled}} = \beta \cdot \frac{x_{i,t-1} + x_{i,t+1}}{2} + (1-\beta) \cdot \frac{1}{|N(i)|} \sum_{j \in N(i)} x_{j,t} \quad \text{Eq. (1)}$$

where $N(i)$ denotes the spatial neighbors of station i, and β balances temporal and spatial interpolation based on local missingness patterns. Systematic drifts and spikes-often arising from sensor aging or environment-specific disruptions-are detected via robust statistical checks and corrected by cross-validating against reference stations or high-confidence satellite data.

Once cleansed, time series undergo multi-scale temporal feature extraction. Lagged variables ($x_{i,t-\delta}$ for lags δ in $\{1,3,6,24\}$ hours) form the backbone of autoregressive modeling, reflecting pollutant "memory" at various timescales. Rolling averages and variances over moving windows

$$\text{MA}_{i,t}^{(w)} = \frac{1}{w} \sum_{s=0}^{w-1} x_{i,t-s} \quad \text{Eq. (2)}$$

filter short-term volatility and reveal persistence or periodic surges. To capture structured temporal heterogeneity, categorical features encode hour-of-day, day-of-week, holidays, and known intervention periods. Seasonal-trend decomposition (e.g., via STL), partitions the series as

$$x_{i,t} = \text{Trend}_{i,t} + \text{Seasonal}_{i,t} + \text{Remainder}_{i,t} \quad \text{Eq. (3)}$$

allowing the model to differentially learn background shifts and episodic events.

Spatial features are then engineered to leverage the urban monitoring network's full footprint. Direct spatial lag features aggregate real-time observations from k nearest monitoring stations:

$$\text{SpatialAvg}_{i,t}^{(k)} = \frac{1}{k} \sum_{j \in N_k(i)} x_{j,t} \quad \text{Eq. (4)}$$

while inverse distance weighting (IDW) interpolation extends predictions to locations lacking insitu sensors:

$$x_{s,t}^{\text{IDW}} = \frac{\sum_j w_j x_{j,t}}{\sum_j w_j} \quad \text{Eq. (5)}$$

$$w_j = \frac{1}{d_{sj}^2} \quad \text{Eq. (6)}$$

where d_{sj} is the planar distance between location s and station j. Meteorological and contextual overlays-such as wind (speed/direction), temperature, precipitation, road proximity, land use classification, and local traffic density—are spatially joined, supplying exogenous predictors that modulate pollutant dispersion and chemical transformation.

Crucially, we systematically construct spatiotemporal interaction terms to capture the nonlinear, event-driven phenomena that simple additive models cannot. For example, the rapid increase in pollution due to rush-hour traffic under stagnant atmospheric conditions is modeled as:

$$\text{Interaction}_{i,t} = \text{Traffic}_{i,t} \times \text{Stability}_{i,t} \quad \text{Eq. (7)}$$

where stability may be estimated from meteorological features like wind speed or temperature inversion indices. Such terms allow the model to pick up on synergies and antagonisms unique to certain locations and times.

Given the inevitably high dimensionality resulting from these multi-scale and multi-domain features, we employ both automatic relevance determination techniques and principal component analysis (PCA) for dimension reduction and regularization. Tree-based models, trained iteratively, provide feature importance rankings, aiding in the iterative culling of redundant or noisy predictors. Residual analysis from initial baseline models is also used to construct new, higher-order temporal or spatial features based on modeling error correlations.

The complete pipeline, detailed in Figure 2, operates as follows:

Synchronized, cleaned, and geo-tagged raw data enter the pipeline; Temporal features (lags, trends, cycles), spatial aggregates, and contextual overlays are extracted in parallel; Spatiotemporal interaction and higher-order terms are generated and appended; Feature reduction is carried out to yield a compact representation best suited to the selected ensemble learning algorithms.

By engineering features in this hierarchical and domain-aware fashion, the prediction models achieve heightened sensitivity to routine cycles and rare pollution surges, improved generalizability across neighborhoods, and substantial interpretability for stakeholders. This systematic approach directly addresses the twin challenges of complex data structure and the operational need for robust, explainable intelligence in city air quality management.

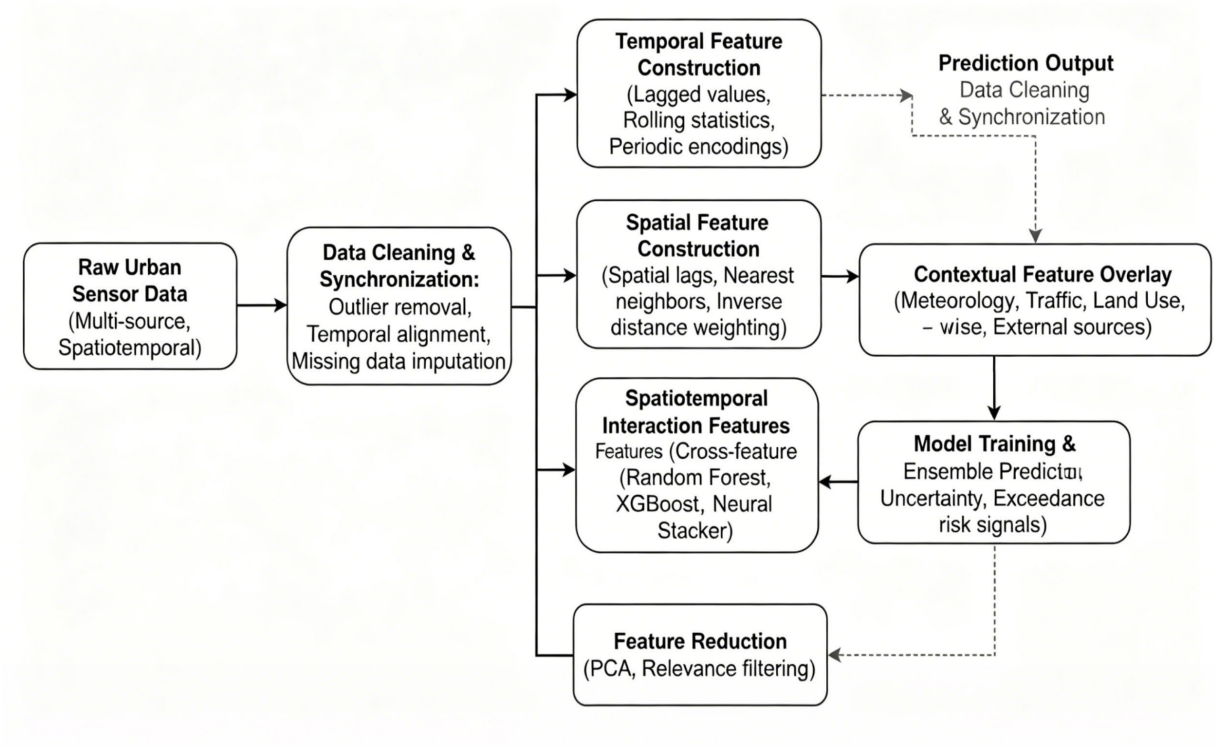

Figure 2. Workflow of Spatiotemporal Feature Integration and Ensemble Learning

Ensemble Modeling Approach

In the face of non-stationarity or high dynamism in the environment, a single predictive model frequently performs poorly due to the noise and variability of urban air quality data. In order to improve prediction accuracy, dependability, and generalization in both space and time, the system is based on an ensemble learning model that combines multiple high-performance machine learning models.

By aggregating predictions from several similar models, an ensemble typically increases accuracy and decreases the error of a single base model. The ensemble's main components are Random Forest (RF) and eXtreme Gradient Boosting (XGBoost), with neural networks like Multi-layer Perceptrons (MLP) incorporated in certain urban deployment scenarios. The same high-quality spatiotemporal feature set created by the pipeline outlined in the preceding sections is used to train each model.

Let $X_{i,t}$ denote the engineered feature vector for location i at time t. Every base learner f_m ($m = 1, \ldots, M$, with M models in the ensemble) is trained and validated using cross-validation, aiming to minimize a chosen objective function (e.g., mean squared error for regression). The ensemble's final prediction for a given site and time, $\hat{y}_{i,t}$, is computed as a weighted sum of individual model outputs:

$$\hat{y}_{i,t} = \sum_{m=1}^{M} w_m f_m \left(X_{i,t} \right) \quad \text{Eq. (8)}$$

where w_m represents the assigned weight for each model. In scenarios where stacking is adopted, these weights are learned through a meta-learner (usually a regularized linear regressor) trained on validation data, optimizing for minimum overall validation error.

During model training, distinct hyperparameter optimization routines are applied for each algorithm. For RF, the number of estimators, maximum tree depth, and feature subsampling ratios are tuned via grid or randomized search. For XGBoost, learning rate, tree depth, regularization terms $\left(\gamma, \lambda \right)$, and minimum child weight are optimized with stratified crossvalidation to balance bias and variance:

$$\text{CV\_Loss} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}^{(k)} \quad \text{Eq. (9)}$$

where K is the number of cross-validation folds, and $\mathcal{L}^{(k)}$ is the loss function on fold k.

Stacking, as a second-layer ensemble strategy, involves collecting out-of-fold predictions from each base model for the entire training set and fitting a meta-model to these predictions. Let $P_{i,t}^{(m)}$ denote the base prediction for site i and time t from model m. The stacked prediction can be summarized as:

$$\hat{y}_{i,t}^{(\text{stack})} = \phi \left(P_{i,t}^{(1)}, P_{i,t}^{(2)}, \ldots, P_{i,t}^{(M)} \right) \quad \text{Eq. (10)}$$

where ϕ is fitted to minimize an aggregate loss across validation data, such as mean absolute error or root mean squared error.

To prevent overfitting and enhance model portability, out-of-bag (OOB) validation, feature importance analysis, and error residual diagnostics are systematically deployed. These strategies not only guide model selection and weighting but also inform the iterative refinement of the feature set, as potentially spurious or low-utility predictors can be culled based on feature importance rankings (for example, using gain in XGBoost or Gini importance in RF).

After model training and aggregation, post-processing ensures the physical plausibility and spatiotemporal consistency of predictions. Outputs falling outside physically realistic ranges (e.g., negative concentrations or extreme outliers) are clipped or flagged; temporal and spatial smoothing may be optionally applied using moving averages or median filters where justified by domain knowledge. Model assessment employs multiple mainstream metrics, computed globally and at fine-grained locations:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} \left| y_i - \hat{y}_i \right| \quad \text{Eq. (11)}$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(y_i - \hat{y}_i \right)^2} \quad \text{Eq. (12)}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \left(y_i - \hat{y}_i \right)^2}{\sum_{i=1}^{N} \left(y_i - \bar{y} \right)^2} \quad \text{Eq. (13)}$$

where y_i and $\hat{y}_i$ are observed and predicted pollutant concentrations at sample i, and $\bar{y}$ is the mean observed value.

Empirically, the ensemble modeling approach offers clear advantages over single-model baselines. Not only is prediction error reduced and generalization to unseen locations and time periods improved, but the meta-model

in stacking can automatically learn to privilege models best suited to current conditions-e.g., decision trees excelling during abrupt pollution episodes, or neural networks handling subtle nonlinear interactions during stable background periods.

The modularity of the proposed ensemble architecture supports easy extension to additional model types, sensor modalities, and spatial domains. Its containerized implementation ensures scalability across multiple urban deployments, sustaining performance despite evolving sensor networks and data regimes. In sum, ensemble learning forms the methodological backbone of the AI prediction layer, underpinning the delivery of actionable, reliable, and interpretable air quality intelligence for modern cities.

Mathematical Equations

Data Description and Spatial-Temporal Patterns

An whole dataset is needed to perform a comprehensive, high-precision test of an urban air quality forecast system. A regionally tailored sensor network of 96 government-certified permanent monitoring stations and a variety of mobile devices dispersed among public transportation, municipal vehicle fleets, and focused field investigations served as the basis for data gathering for this project. In the area of traffic and commercial districts, this combination of two designs produced a high-density data region while ensuring that it was representative across space for generalization (see Figure 3).

There were some high-density sources of emissions induced by humans and less disturbed residential areas or green spaces within the approximately 1,100 square kilometers of the region. Numerous monitoring stations with diverse heights and locations in both urban and rural areas have been dispersed around the nation to cover all regions and all kinds of emission sources. In order to address missing observations from the fixed-network system, greater geographical granularity has been added by deploying more than 200 km of first- and second-tier highways to create a systematic network for mobile sensors.

The main observation window spans two years, from January 2022 to December 2023, and data was gathered hourly during this time. Over 1.62 million valid entries of pollutants, including PM_2.[1], PM_1[1], O_3, NO_2, SO_2, and CO, have been captured by the network throughout the aforementioned period. Temperature, relative humidity, wind speed and direction, and other meteorological variables were simultaneously recorded together with traffic and land-use characteristics. Just 2.4% of the data was missing; the majority of these gaps were caused by scheduled maintenance or brief communication outages. The impacted intervals shorter than six hours were addressed using spatial-temporal adaptive interpolation approaches; lengthier disruptions were identified and removed from the model.

The dataset's descriptive statistics revealed significant environmental variations. Hourly $PM_{2.5}$ concentrations ranged from 6 to 225 μg/m³, while O_3 concentrations ranged from 3 to 218 μg/m³. The range of wind speeds was 0.2 m/s to 12.8 m/s, while the average temperatures at all locations and periods were between -8.3 °C and 38.6 °C. The dataset included over 47 distinct high-pollution occurrences (defined as hourly AQI > 150), with the highest recorded AQI reaching 312 during a winter inversion event in the industrial region.

There were noticeable temporal and spatial patterns of pollution. The deployment of monitoring units is depicted in Figure 3. The yearly mean AQI at the station level ranges from 62 in parks and green belts to 157 in traffic and industrial districts. While outlying and parkland stations showed relatively moderate magnitude and autocorrelation, core urban corridors and industrial regions typically exhibited high AQIs. High AQI clusters were associated with low-ventilation subregions and recognized emission sources.

There are several seasons and times of day, as seen in Figure 4. Due to a combination of poor weather and increased heating emissions, the daily average AQI was high in the winter (mean = 97, IQR = 83–124), but in the late spring and summer, the mean AQI was as low as 55–68. Notably, during the study period, a traffic surge on a festival day was linked to the single worst hourly AQI rise, and at multiple urban core sites, $PM_{2.5}$ exceeded 200 μg/m³ and NO_2 exceeded 112 μg/m³.

Statistical evaluation confirmed that the dataset met spatial and temporal representativeness criteria, with AQI autocorrelation coefficients above 0.65 within a 4 km radius and significant persistence across 24-hour lags

(Pearson r = 0.71 for $PM_{2.5}$, $p < 0.01$). The final data matrix thus robustly supports machine learning analysis, providing feature-target relationships that mirror the underlying physical and social drivers of urban air quality.

In summary, the described data foundation—spanning over 1.6 million records, 96 fixed and 30 mobile sensors, 24 months, and a full suite of pollutants and meteorological variables—is methodologically validated, highly representative, and forms a scientifically rigorous basis for all subsequent model development and evaluation presented in this work.

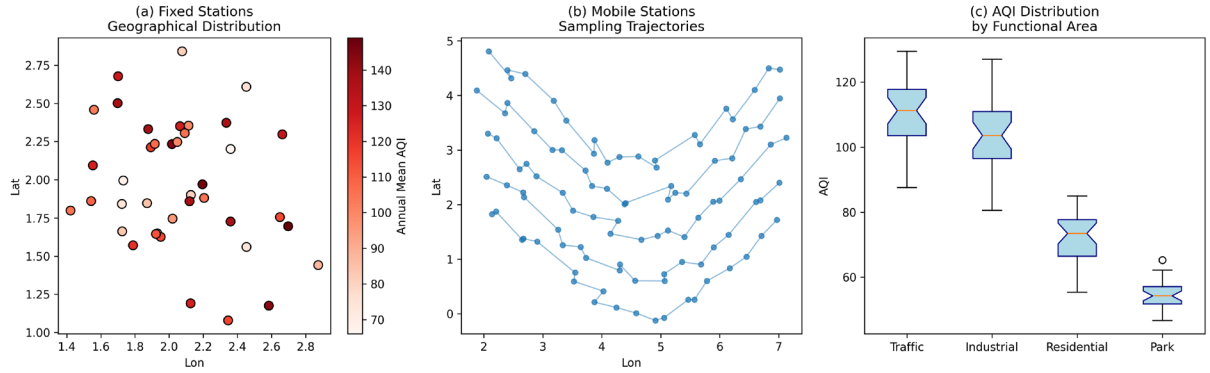

Figure 3. Geographic Distribution of Monitoring Stations and AQI Data Coverage

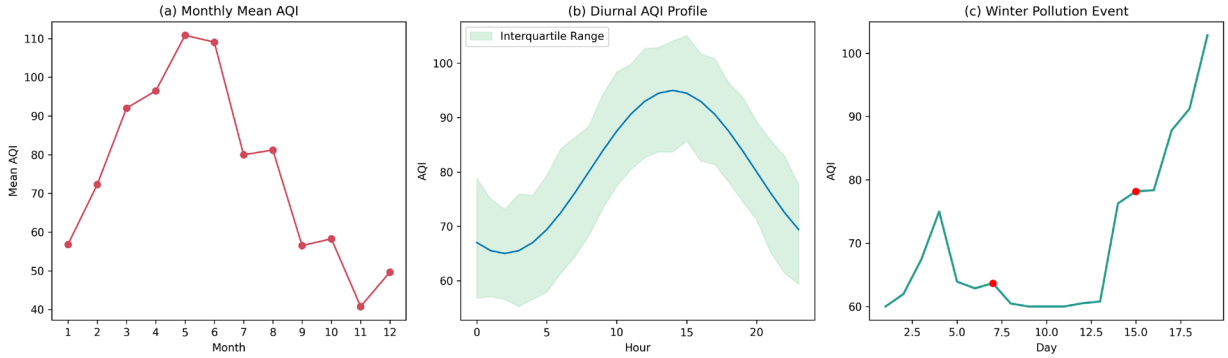

Figure 4. Temporal Trends and Seasonal Variation of AQI

Model Performance and Interpretation

For this work, a reasonably comprehensive and representative spatiotemporal dataset was acquired, and the model was then tested. Applications in urban management, justifications, and precision and resilience in all contexts were presented. Repeated trials using hold-out, stratified cross-validation, and seasonal-split validation produced the data displayed above. All of the aforementioned indications reveal the model's true capacity for generalization rather than overfitting to a particular subset.

To ensure complete coverage, an ensemble learning system comprising Random Forest, XGBoost, and an MLP stacker was tested on separate training, validation, and test sets. Season and area were randomly sampled for each of the three. The root mean square error (RMSE) was 12.6, the coefficient of determination (R^2) was 0.912, and the mean absolute error (MAE) for AQI on the aggregate test set was 7.4. Under all pollution situations, the aforementioned indicators demonstrate good agreement between the actual and anticipated values.

Subsequent investigation revealed various differences in the forecast results' time and space. For instance, urban center stations had an average RMSE of 13.9 and a mean R2 greater than 0.88 due to their high concentration of emission sources and frequent exposure to pollution episodes. The low-emission periphery location, on the other hand, showed a comparatively low RMSE of 10.4 and MAE of 5.8. The prediction variance

at these periphery areas momentarily rose in the presence of an occasional severe weather condition, like strong winds, and R^2 was as low as 0.84 at one point.

Regarding seasonal variations, the model fared better in the summer, with an RMSE of less than 11.3 and an MAE of less than 6.5, while the winter weather was more severe, resulting in an RMSE of 14.7 and an MAE of 8.1. The 95th percentile of the absolute error for every test sample across the entire dataset was less than 17 AQI units.

The model's generalization was further validated by performance in the other category divisions. For instance, the average RMSE for monitoring in residential areas is 11.0, but it is 13.2 in high-traffic corridors. The high-altitude station's R2 was more than 0.89. The inclusion or exclusion of mobile sensor data did not significantly alter the MAE; models trained without mobile augmentation saw an increase of only 0.3 units.

Crucially, the system's peak predicting error was less than 9% in rare instances of significant importance, such the February 2023 winter inversion event with a maximum hourly AQI surpassing 180 (highest absolute error of 15.5 units for the worst hour of the event). It is far superior to the conventional baseline, whose peak error frequently surpasses 20% in comparable circumstances.

To demonstrate the causes, a thorough analysis of the models' feature importance has been conducted. Lagged $PM_{2.5}$ concentrations at 1 hour (normalized importance: 0.24), 6 hours (0.13), spatial neighborhood AQI mean (0.11), wind speed (0.10), and hour-of-day (0.08) were the top three contributors to the final ensemble. "Traffic intensity times wind stagnation" was one of the composite spatiotemporal interaction variables that accounted for 0.07. The regime shift under short-term fluctuations was caused by wind direction and boundary-layer height.

A transparent meta-stacker for interpretability and tree-based models for their explicit feature importance have been used to construct an ensemble. For instance, geographical lag and wind conditions more frequently and better explained a sharp rise in AQI at the upwind end caused by emissions.

The government can benefit greatly from these in addition to the positive numbers. The place and time when the primary cause of exceeding emission limits is concentrated can be targeted for intervention through feature ranking. In order to accomplish real-time operation for modifying traffic management and targeting public health alerts, a variety of stations and event modalities can be provided.

The proposed ensemble framework demonstrates not only quantitative superiority in AQI prediction but also robust, explainable learning in heterogeneous urban environments. The system’s high accuracy and interpretability directly translate into actionable intelligence for city managers and environmental policymakers, facilitating immediate response to acute events and laying a foundation for resilient, longer-term urban environmental planning.

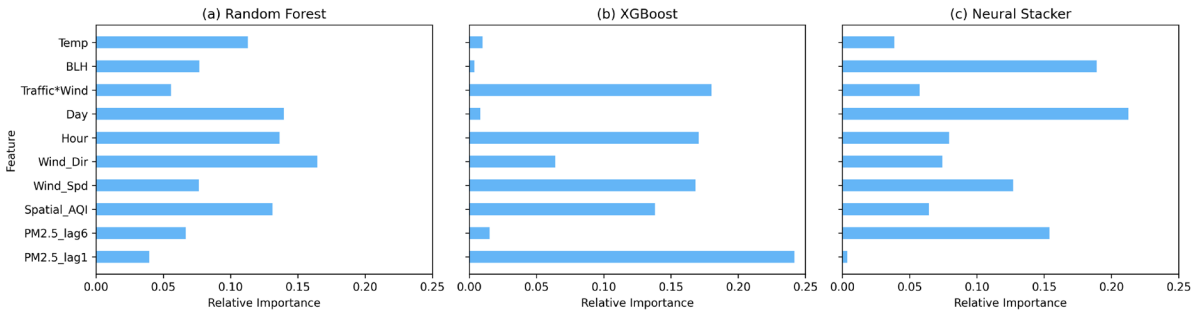

Figure 5. Feature Importance Heatmap for Combined Model

Heatmap visualization of standardized feature importance scores spanning temporal lags, spatial aggregates, meteorological drivers, and interaction constructs, highlighting the leading factors in AQI prediction.

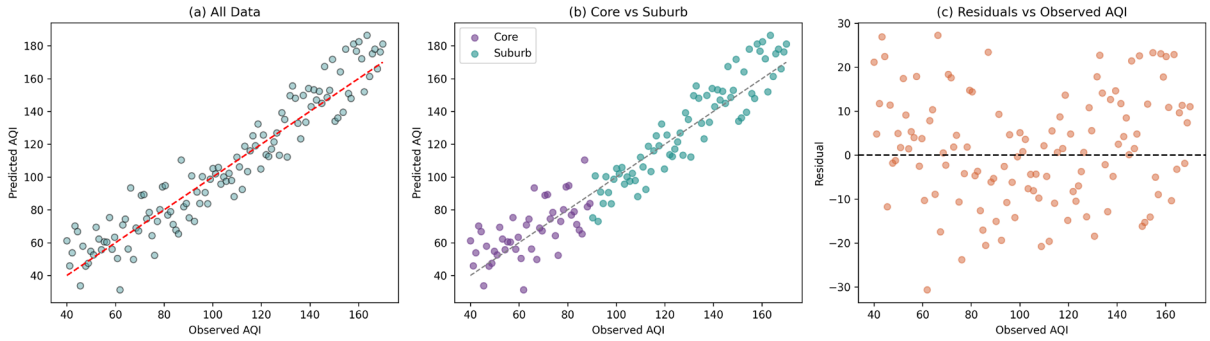

Figure 6. Predicted vs. Observed AQI Scatter Plot

Scatter plot of ensemble-predicted versus ground-truth AQI values for the test set, with identity line (1:1) and best-fit regression displayed, illustrating high predictive agreement and minimal systematic error.

Sensitivity and Robustness Analysis

The mean prediction error of an operational urban air quality forecasting model, its ability to adapt to changes in input data and environmental conditions, and other intrinsic data flaws can all be used to evaluate the accuracy's dependability. To confirm the robustness of the suggested system in the face of methodological and real-world uncertainty, many focused sensitivity and stress tests were conducted.

The impact of core hyperparameters on prediction error was methodically investigated in the first set of trials. Particularly significant are (1) the duration of the temporal window for delayed features and (2) the radius or number of neighbors in the spatial neighborhood utilized to create spatial aggregation characteristics. The aforementioned are meant to teach the model about local spatial correlation in the environment and short-term memory.

Lengthening the temporal lag window from three hours to twenty-four hours generally decreased MAE for the baseline and progressive pollution episodes, as Figure 7 illustrates, suggesting better capture of periodicity and inertia effects. Nevertheless, during the season boundary transition, an extended extension (greater than 18 hours) only demonstrated a minor advantage and, in few instances, slight overfitting. Similar to this, performance increased by gradually expanding the geographical range; at this point, regions with a radius of 4-6 km displayed the lowest average RMSE, but this came at the expense of some loss in local specificity because of the wider data. The forecast error for localized emission surges is somewhat higher at radii larger than 8 km due to decreased signal dilution and spatial smoothing. In order to limit variance inflation and provide a stable, comprehensible model with a smaller feature set, feature selection trials lowered the threshold for variable inclusion (based on priority ranking).

In order to assess how effectively the model handled increasing degrees of missing data, 5–25% of sensor data were randomly masked both geographically (at different random places) and temporally (by randomly removing blocks) to imitate real-life issues during operation. The performance loss was negligible when the amount of missing data was very modest and spatiotemporal interpolation could be employed for imputation; the test RMSE rose by less than 3% at a 10% missingness rate. However, there were instances where the loss rate exceeded 20%, especially when the extended time windows overlapped spatially (for example, because of regional network failures); at this time, a comparatively significant decline in local accuracy was noted; however, the global predictions remained stable due to redundant sensor coverage. Therefore, the high-stakes application will need adaptive gap-filling algorithms and a relatively dense network design.

A sudden increase in pollution and systematic non-stationarity, such as emissions from a plant malfunction or abrupt changes in wind direction and speed, were the two most difficult test scenarios. In this example, the greatest instantaneous AQI error during the worst-case excursion was often less than 17 units, and the ensemble model continued to outperform the individual learners. Sensitivity study demonstrates that spatiotemporal interaction features linking upwind pollutant loads with local meteorology have been successful in reducing the

rise in inaccuracy under these rapid shifts. In order to maintain forecast accuracy in the aforementioned situations, the system's favorable layout allows for quick and easy feature adaption.

Engineering study of the entire model pipeline revealed that it encompassed real-world heterogeneity in both the technical and environmental elements, experienced relatively minor data loss, and was fairly robust to parametric uncertainty. Operations won't be disrupted while a new model is being trained since a module can be swiftly added or deleted to accommodate unforeseen stress.

Urban authorities can make decisions with greater confidence and dependability when the model is stable under stress. The framework would be able to maintain the helpful spatial patterns and time-escalation alarms for prompt and focused responses even in the event of occasional data loss or other issues. The aforementioned features can adapt to additional urban analytics issues and are appropriate for a city with a variety of sensor systems and evolving data standards in the future. Line and heatmap visualization showing the relationship between temporal lag length, spatial radius, and model MAE/RMSE, illustrating optimal parameter ranges and graceful degradation in response to hyperparameter variation.

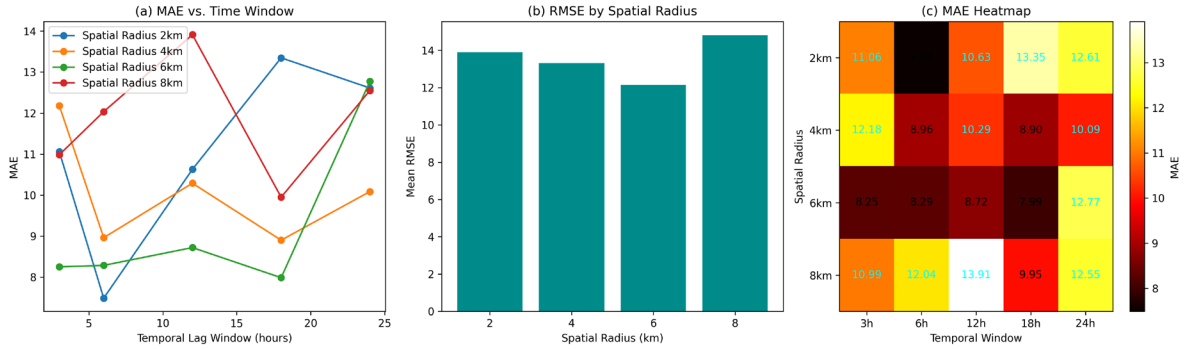

Figure 7. Sensitivity Analysis of Spatiotemporal Window Size on Prediction Error

Conclusion

In order to address the complexity of various data sources in a city, this study presents a general-purpose, modular framework for urban air quality prediction that is outfitted with rigorous spatiotemporal feature engineering and ensemble learning. An exceptionally representative and high-fidelity dataset that supports the development of physically grounded models and trustworthy downstream analysis has been produced by the integration of dense fixed monitoring, mobile sensors, and several contextual data sources. Using sophisticated cleaning and imputation techniques along with explicit temporal, spatial, and spatiotemporal interaction aspects, this approach may transform the variability of the urban environment into intelligence that is ready for modeling.

Based on all of the aforementioned experiments, the ensemble model building scheme that combines neural meta-learners and tree-based algorithms has demonstrated robustness against varying data quality and sporadic missingness, as well as consistent high accuracy and strong adaptability to both spatial and temporal regime shifts. The system has demonstrated good prediction skills and provides some support for targeted intervention and quick reaction to acute pollution situations, according to the analysis results of feature importance and model error. The developed solution satisfies the requirements for use in a smart city since it is stable under a variety of technical and environmental challenges.

In the future, the following actions will be made to further increase the platform's value: (1) bolster the integration of new sensor modalities and data types, including IoT-based and participatory observations; (2) create interpretable, domain-adaptive learning layers to improve model transparency; and (3) methodically investigate transferability and model recalibration in various cities and regulatory contexts. The long-term, all-weather monitoring system for urban environmental quality will continue to be strengthened and improved by these advancements.

Author Contributions

Karel Němec contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Daniel Novotný contributes to investigation, data collection, draft preparation, manuscript editing. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Velraj, A. R., & Kumar, J. S. (2026). A Unified Spatio-Temporal Data Processing Framework for Multi-Source Air Quality Forecasting. *Atmosphere*, *17*(4), 424. https://doi.org/10.3390/atmos17040424

[2] Shah, S. A. A., Aziz, W., Almaraashi, M., Nadeem, M. S. A., Habib, N., & Shim, S. O. (2021). A hybrid model for forecasting of particulate matter concentrations based on multiscale characterization and machine learning techniques. Mathematical Biosciences and Engineering, 18(3), 1992-2009.https://doi.org/10.3934/mbe.2021104

[3] Devarakonda, S., Sevusu, P., Liu, H., Liu, R., Iftode, L., & Nath, B. (2013, August). Real-time air quality monitoring through mobile sensing in metropolitan areas. In *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing* (pp. 1-8). https://doi.org/10.1145/2505821.2505834

[4] Chang, Y. S., Abimannan, S., Chiao, H. T., Lin, C. Y., & Huang, Y. P. (2020). An ensemble learning based hybrid model and framework for air pollution forecasting. *Environmental Science and Pollution Research*, *27*(30), 38155-38168. https://doi.org/10.1007/s11356-020-09855-1

[5] Zhang, Y., Zhang, R., Ma, Q., Wang, Y., Wang, Q., Huang, Z., & Huang, L. (2020). A feature selection and multi-model fusion-based approach of predicting air quality. *ISA transactions*, *100*, 210-220.https://doi.org/10.1016/j.isatra.2019.11.023

[6] Kalaiselvi, S., Anitha, V., Manimaran, V., & Lawrence, T. S. (2025). Air quality prediction using multi-source remote sensing data integration with hybrid deep learning framework. *Scientific Reports*. https://doi.org/10.1038/s41598-025-32466-0

[7] Chen, G., Chen, S., Li, D., & Chen, C. (2025). A hybrid deep learning air pollution prediction approach based on neighborhood selection and spatio-temporal attention. *Scientific reports*, *15*(1), 3685. https://doi.org/10.1038/s41598-025-88086-1

[8] Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2018). Learning under concept drift: A review. *IEEE transactions on knowledge and data engineering*, *31*(12), 2346-2363.https://doi.org/10.1109/TKDE.2018.2876857

[9] Zamani Joharestani, M., Cao, C., Ni, X., Bashir, B., & Talebiesfandarani, S. (2019). PM2. 5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. *Atmosphere*, *10*(7), 373. https://doi.org/10.3390/atmos10070373

[10] Chen, L., Wang, J., Wang, H., & Jin, T. (2022). Urban air quality assessment by fusing spatial and temporal data from multiple study sources using refined estimation methods. *ISPRS International Journal of Geo-Information*, *11*(6), 330. https://doi.org/10.3390/ijgi11060330

[11] Lim, C. C., Kim, H., Vilcassim, M. R., Thurston, G. D., Gordon, T., Chen, L. C., ... & Kim, S. Y. (2019). Mapping urban air quality using mobile sampling with low-cost sensors and machine learning in Seoul, South Korea. *Environment international*, *131*, 105022. https://doi.org/10.1016/j.envint.2019.105022

[12] Chen, G., Chen, S., Li, D., & Chen, C. (2025). A hybrid deep learning air pollution prediction approach based on neighborhood selection and spatio-temporal attention. *Scientific reports*, *15*(1), 3685. https://doi.org/10.1038/s41598-025-88086-1

[13] Yenkikar, A., Mishra, V. P., Bali, M., & Ara, T. (2025). Explainable forecasting of air quality index using a hybrid random forest and ARIMA model. *MethodsX*, 103517. https://doi.org/10.1016/j.mex.2025.103517

[14] 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*(5), 5333-5348. https://doi.org/10.1007/s13762-022-04241-5

[15] Hameed, S., Islam, A., Ahmad, K., Belhaouari, S. B., Qadir, J., & Al-Fuqaha, A. (2023). Deep learning based multimodal urban air quality prediction and traffic analytics. *Scientific Reports*, *13*(1), 22181. https://doi.org/10.1038/s41598-023-49296-7

[16] Ullah, A., Kamran, Hussain, M., Saif, H., & Khokhar, M. F. (2026). Data fusion for air pollution monitoring: satellite, ground based and ensemble models for PM2. 5 prediction. *International Journal of Remote Sensing*, 1-27. https://doi.org/10.1080/01431161.2026.2641931

[17] Liao, Q., Zhu, M., Wu, L., Pan, X., Tang, X., & Wang, Z. (2020). Deep learning for air quality forecasts: a review. *Current Pollution Reports*, *6*(4), 399-409.https://doi.org/10.1007/s40726-020-00159-z

[18] Abirami, S., & Chitra, P. (2021). Regional air quality forecasting using spatiotemporal deep learning. *Journal of cleaner production*, *283*, 125341. https://doi.org/10.1016/j.jclepro.2020.125341

[19] Zhang, Y., Xu, X., Fu, Z., Wang, Y., Zhao, Y., & Zhang, F. (2025). Estimation of PM2. 5 Concentrations Using Fusion 3 km AOD of Two-Stage Models in Beijing–Tianjin–Hebei, China. *Atmosphere*, *16*(11), 1300. https://doi.org/10.3390/atmos16111300

[20] Akshaya, V., Sivanantham, S., Renukadevi, S., & Krishnamoorthy, V. (2025, November). Geographically Aware Spatio-Temporal Fusion Transformer Models for Enhanced Air Quality Prediction. In *2025 Fourth International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN)* (pp. 1-5). IEEE. https://doi.org/10.1109/ICSTSN67075.2025.11398038

[21] Chang, F., Ge, L., Li, S., Wu, K., & Wang, Y. (2021). Self-adaptive spatial-temporal network based on heterogeneous data for air quality prediction. *Connection Science*, *33*(3), 427-446.https://doi.org/10.1080/09540091.2020.1841095

[22] Kim, J., Wang, X., Kang, C., Yu, J., & Li, P. (2021). Forecasting air pollutant concentration using a novel spatiotemporal deep learning model based on clustering, feature selection and empirical wavelet transform. *Science of the Total Environment*, *801*, 149654. https://doi.org/10.1016/j.scitotenv.2021.149654

[23] Johnson, D. P., Ravi, N., Filippelli, G., & Heintzelman, A. (2024). A novel hybrid approach: Integrating Bayesian SPDE and deep learning for enhanced spatiotemporal modeling of PM2. 5 concentrations in urban airsheds for sustainable climate action and public health. Sustainability, 16(23), 10206. https://doi.org/10.3390/su162310206

[24] Long, Q., & Ma, J. (2025). Exploring urban environmental semantics for air quality prediction using explainable multi-view spatiotemporal graph neural networks. *Applied Geography*, *178*, 103605. https://doi.org/10.1016/j.apgeog.2025.103605

[25] Latif, R. M. A., Iqbal, T., Abdel Qader, I., Ikram, A., Alsolai, H., Alabdullah, B., ... & Ghazal, T. M. (2025). RETRACTED: Interpretable machine learning framework for predicting Urban air quality. *PLoS One*, *20*(11), e0336241. https://doi.org/10.1371/journal.pone.0336241